Universidad de Málaga

Escuela Técnica Superior de Ingeniería de Telecomunicación

Programa de Doctorado en Ingeniería de Telecomunicación

TESIS DOCTORAL POR COMPENDIO

Performance assessment of mobile networks in Industry 4.0

Autor:

David Segura Ramos

Directores:

Emil Jatib Khatib

Raquel Barco Moreno

2024

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D. **David Segura Ramos**

Estudiante del programa de doctorado en **Ingeniería de Telecomunicación** de la Universidad de Málaga, autor de la tesis presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: **Performance assessment of mobile networks in Industry 4.0**.

Realizada bajo la tutorización de **Raquel Barco Moreno** y dirección de **Emil Jatib Khatib** y **Raquel Barco Moreno**.

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a **8** de **Noviembre** de **2024**.

SEGURA RAMOS DAVID - 79057585F Firmado digitalmente por SEGURA RAMOS DAVID - 79057585F Fecha: 2024.11.08 10:55:12 +01'00'	BARCO MORENO RAQUEL - 25695879A Firmado digitalmente por BARCO MORENO RAQUEL - 25695879A Fecha: 2024.11.10 21:46:23 +01'00'
Fdo.: David Segura Ramos Doctorando	Fdo.: Raquel Barco Moreno Tutor de tesis

JATIB KHATIB EMIL - Firmado por: JATIB KHATIB EMIL -

BARCO MORENO RAQUEL - 25695879A
Firmado digitalmente por BARCO MORENO RAQUEL - 25695879A
Fecha: 2024.11.11 07:22:14 +01'00'

Fdo.: Emil Jatib Khatib y Raquel Barco Moreno
Directores de tesis

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

AUTORIZACIÓN PARA LA LECTURA DE LA TESIS

Por la presente, la Dra. Raquel Barco Moreno, y el Dr. Emil Jatib Khatib, profesores del Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga,

CERTIFICAN

Que D. David Segura Ramos, ha realizado en el Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga bajo su dirección, el trabajo de investigación correspondiente a su TESIS DOCTORAL titulada:

"**Performance assessment of mobile networks in Industry 4.0**"

En dicho trabajo, se han propuesto aportaciones originales para la evaluación y análisis del rendimiento de las redes de comunicaciones móviles e inalambricas en el escenario industrial. Los resultados de dicha tesis han dado lugar a las diversas publicaciones en revista, así como a aportaciones a congresos, superando el requisito de 1 punto ANECA del programa de doctorado regulado por el Real Decreto 99/2011.

Por todo ello, y dada la unidad temática de las distintas contribuciones y la metodología común seguida en todas ellas, los directores consideran que esta tesis es apta para su presentación al Tribunal que ha de evaluarla y AUTORIZA la presentación de la tesis por COMPENDIO DE PUBLICACIONES en la Universidad de Málaga. Igualmente, certifica que las publicaciones que avalan la tesis no han sido empleadas en trabajos anteriores a la misma.

Málaga, 8 de Noviembre de 2024

BARCO MORENO RAQUEL - 25695879A
Firmado digitalmente por BARCO MORENO RAQUEL - 25695879A
Fecha: 2024.11.10 21:46:50 +01'00'

JATI B KHA
Firmado por: JATIB KHATIB EMIL - 25726182S

Fdo.: Raquel Barco Moreno

Fdo.: Emil Jatib Khatib

Universidad de Málaga
Escuela Técnica Superior de Ingeniería de Telecomunicación
Programa de doctorado en ingeniería de telecomunicación

Reunido el tribunal examinador en el día de la fecha, constituido por:

Presidente: Dr. D./Dña. _____

Secretario: Dr. D./Dña. _____

Vocal: Dr. D./Dña. _____

para juzgar la Tesis Doctoral titulada *Performance assessment of mobile networks in Industry 4.0* realizada por D. David Segura Ramos, y dirigida por los doctores D. Emil Jatib Khatib y Dña. Raquel Barco Moreno, acordó por

_____ otorgar la calificación de

_____ y para que conste,

se extiende firmada por los componentes del tribunal la presente diligencia.

Málaga, a ____ de _____ de _____.

El Presidente:

Fdo.:_____

El Secretario: El Vocal:

Fdo.:_____ Fdo.:_____

Hold the vision, trust the process.

Acknowledgments

First of all, I would like to express my most sincere gratitude to my supervisors, Raquel and Emil. Thank you Raquel for giving me the opportunity to join the MobileNet research group where I had the opportunity to learn a lot. Thank you Emil for the time and effort you dedicated to this thesis, for your valuable advice and for being my guide during its development.

Moreover, I would like to express my gratitude to my colleagues in lab 1.3.1. They made it possible to work in a more enjoyable way, and made me able to learn a lot from them to be a better research. A special mention to Hao and Carlos Baena for their always willingness to help with questions that arose throughout the thesis. Also, to my colleagues Antonio, José Antonio and Sebastián, who joined me into the adventure of the research stay and we made a lot of travels together.

I would like to extend my gratitude to people at Aalborg University with whom I had the opportunity to collaborate during my research stay, Preben, Sebastian, Melisa, and Akif. Thank you all for making me feel integrated and welcome. I would like to make a special mention to Sebastian, for all his work, help and dedication involved during the research stay.

A special mention among these lines is for my family. Thanks to my parents, brother and sister for their unconditional support and for always believe in me. Finally, I would like to thank Ana for being my greatest support during these years, for your patience and for always making me laugh.

This thesis has been partially funded by the following projects:

- EDEL4.0: Seguridad y fiabilidad en las comunicaciones 5G/IoT para la Industria 4.0. Número de proyecto UMA18-FEDERJA-172, receiving funds from Junta de Andalucia and European Comission, within the framework of "Proyectos de I+D+i en el marco del Programa Operativo FEDER Andalucia 2014-2020".
- PENTA: Provisión de servicios PPDR a través de Nuevas Tecnologías de Acceso radio. Número de proyecto PY18-4647, receiving funds from Junta de Andalucía and European Comission, within the framework of "Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI 2020)".
- MAORI: Massive AI for the OpenRadIo b5G/6G network. Project number TSI-063000-2021-72, receiving funds from Ministerio de Asuntos Económicos y Transformación Digital and European Union - NextGenerationEU within the framework "Recuperación, Transformación, y Resiliencia".

Contents

Abstract

The world is currently undergoing a profound digital transformation across various socio-economic sectors, a shift often referred to as the *digital revolution*. In the context of the industrial sector, the advent of the fourth industrial revolution, commonly known as *Industry 4.0*, is transforming factories into smart factories. This revolution refers to the current trend of automation and integration of data exchange mechanisms in manufacturing processes, thereby making production and distribution processes more flexible, robust and efficient. To achieve this goal, Industry 4.0 relies on different enabler technologies such as the the integration of Cyber-Physical Systems (CPS), the use of the Internet of Things (IoT), cloud computing, Artificial Intelligence (AI), robotics, and Big Data analytics.

With the integration of these enabler technologies in the factory, new applications and use cases emerge to provide mobility and flexibility such as rearrangeable modules in production lines, Automated Guided Vehicles (AGVs), autonomous robots or connected worker solutions. These applications and use cases pose new challenges for their correct operation, such as requirements of low latency communications, high reliability, and high throughput. To adapt to the challenges launched by the new services and use cases contemplated in a smart factory, the Fifth Generation (5G) of mobile networks is emerging as an enabling technology for this transformation. This thesis addresses the integration of the 5G cellular network into the industrial scenario, by assessing and improving network performance for different Industry 4.0 use cases involved in a smart factory. In particular, this thesis has been divided into three parts where different techniques and optimizations of the 5G network are addressed.

The first part is focused on the assessment of the network performance for critical services. These services such as the communication of the AGVs impose requirements of low latency and high reliability in the network. In this part, the focus is set on the assessment of the latency. On the

one hand, a study of the numerologies introduced in 5G is performed in terms of latency. In addition, an empirical assessment and comparison of the scalability of the network with different technologies in an industrial environment is carried out.

The second part focuses on the development of optimization algorithms of the network. First, a dynamic algorithm for the activation of Packet Duplication (PD) when using Dual Connectivity (DC) is proposed. This algorithm is centered on the increase of the reliability for critical services while minimizing the waste of radio resources. Secondly, the Quality of Service (QoS) performance with different configurations of Network Slicing (NS) for the different traffic profiles involved within a distribution center is studied.

The third part of this thesis focuses on the evaluation of Cellular Internet of Things (CIoT) devices in the 5G network. In particular, the performance of the different optimizations proposed in the standard to reduce the signaling overhead in the data transmission of CIoT devices has been evaluated. In addition, an in-depth analysis of the security of the Early Data Transmission (EDT) optimization is provided, analysing its main vulnerabilities and providing a set of recommendations for manufacturers and researchers.

Resumen

El mundo está experimentando actualmente una profunda transformación digital en diversos sectores socioeconómicos, un cambio que a menudo se denomina revolución digital. En el contexto del sector industrial, la llegada de la cuarta revolución industrial, comúnmente conocida como Industria 4.0, está transformando las fábricas en fábricas inteligentes. Esta revolución se refiere a la tendencia actual de automatización e integración de mecanismos de intercambio de datos en los procesos de fabricación, haciendo así más flexibles, robustos y eficientes los procesos de producción y distribución. Para lograr este objetivo, la Industria 4.0 se apoya en diferentes tecnologías facilitadoras como la integración de sistemas ciberfísicos (*Cyber-Physical Systems*, CPS), el uso del internet de las cosas (*Internet of Things*, IoT), la computación en la nube, la Inteligencia Artificial (IA), la robótica y el análisis de Big Data.

Con la integración de estas tecnologías facilitadoras en la fábrica, surgen nuevas aplicaciones y casos de uso que aportan movilidad y flexibilidad, como módulos reorganizables en las líneas de producción, vehículos guiados automatizados (*Automated Guided Vehicles*, AGVs), robots autónomos o soluciones para trabajadores conectados. Estas aplicaciones y casos de uso plantean nuevos retos para su correcto funcionamiento, como los requisitos de comunicaciones de baja latencia, alta fiabilidad y alto rendimiento. Para adaptarse a los retos lanzados por los nuevos servicios y casos de uso contemplados en una fábrica inteligente, la quinta generación (5G) de redes móviles se perfila como una tecnología habilitadora de esta transformación.

Esta tesis aborda la integración de la red celular 5G en el escenario industrial, mediante la evaluación y mejora del rendimiento de la red para diferentes casos de uso de Industria 4.0 implicados en una fábrica inteligente. En concreto, esta tesis se ha dividido en tres partes donde se abordan diferentes técnicas y optimizaciones de la red 5G.

La primera parte se centra en la evaluación del rendimiento de la red para

servicios críticos. Estos servicios, como la comunicación de los AGV, requieren requisitos de baja latencia y alta fiabilidad en la red. En esta parte, la atención se centra en la evaluación de la latencia. Por un lado, se realiza un estudio de las numerologías introducidas en 5G en términos de latencia. Por otro lado, se realiza una evaluación empírica y una comparación de la escalabilidad de la red con diferentes tecnologías en un entorno industrial.

La segunda parte se centra en el desarrollo de algoritmos de optimización de la red. En primer lugar, se propone un algoritmo dinámico para la activación de la duplicación de paquetes (*Packet Duplication*, PD) cuando se utiliza la conectividad dual (*Dual Connectivity*, DC). Este algoritmo se centra en el aumento de la fiabilidad para los servicios críticos, minimizando al mismo tiempo el desperdicio de recursos radio. En segundo lugar, se estudia la calidad de servicio (*Quality of Service*, QoS) con diferentes configuraciones de *Network Slicing* (NS) para los diferentes perfiles de tráfico involucrados dentro de un centro de distribución.

La tercera parte de esta tesis se centra en la evaluación de los dispositivos IoT celulares (*Cellular IoT*, CIoT) en la red 5G. En particular, se ha evaluado el rendimiento de las diferentes optimizaciones propuestas en el estándar para reducir la sobrecarga de señalización en la transmisión de datos de los dispositivos CIoT. Además, se ofrece un análisis en profundidad de la seguridad de la optimización de la transmisión temprana de datos (*Early Data Transmission*, EDT), analizando sus principales vulnerabilidades y proporcionando un conjunto de recomendaciones para fabricantes e investigadores.

Acronyms

3GPP	Third Generation Partnership Project
5G	Fifth Generation of mobile networks
5GC	5G Core
5GS	5G System
AF	Application Function
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AKA	Authentication and Key Agreement
AMF	Access and Mobility Management Function
AMR	Autonomous Mobile Robot
AR	Augmented Reality
ARQ	Automatic Repeat reQuest
AS	Access Stratum
APN	Access Point Name
ARPF	Authentication credential Repository and Processing Function
AUSF	Authentication Server Function
BPSK	$\pi/2$-Binary Phase Shift Keying
BWP	Bandwidth Part

CA	Carrier Aggregation
CC	Component Carrier
CE	Coverage Enhancement
CIoT	Cellular Internet of Things
CN	Core Network
CP	Control Plane
CPS	Cyber-Physical Systems
CU	Central Unit
DC	Dual Connectivity
DFT-s-OFDM	Discrete Fourier Transform-spread-OFDM
DoS	Denial/Degradation of Service
DRB	Data Radio Bearer
DRX	Discontinuous Reception
DU	Distributed Unit
DY	Dolev-Yao
E2E	End-to-End
eDRX	Extended Discontinuous Reception
EDT	Early Data Transmission
eMBB	Enhanced Mobile Broadband
en-gNB	Evolved-Next Generation NodeB
eNB	Evolved Node B
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
EPC	Evolved Packet Core
EPS	Evolved Packet System

FL	Federated Learning
FR	Frequency Range
FTP	File Transfer Protocol
gNB	Next-generation NodeB
HARQ	Hybrid Automatic Repeat reQuest
HSS	Home Subscriber Server
InF	Indoor Factory
InF-DH	InF with Dense clutter and High base station height
InF-DL	InF with Dense clutter and Low base station height
InF-SH	InF with Sparse clutter and High base station height
InF-SL	InF with Sparse clutter and Low base station height
IoT	Internet of Things
IP	Internet Protocol
ITU	International Telecommunication Union
LOS	Line-of-Sight
LPWA	Low-Power Wide-Area
LTE	Long Term Evolution
LTE-M	LTE for Machine Type Communications
MAC	Medium Access Control
MitM	Man-in-the-middle
MCG	Master Cell Group
MCL	Maximum Coupling Loss
ME	Mobile Equipment
MEC	Mobile Edge Computing

MeNB	Master eNB
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MME	Mobility Management Entity
mMTC	Massive Machine Type Communications
MN	Master Node
multi-RAT	Multi-Radio Access Technology
MR-DC	Multi-Radio Dual Connectivity
MTC	Machine Type Communications
NAS	Non-Access Stratum
NB-IoT	NarrowBand Internet of Things
NF	Network Function
NFV	Network Function Virtualization
ng-eNB	Next-generation eNB
NG-RAN	Next Generation Radio Access Network
NLOS	Non-Line-of-Sight
NR	New Radio
NS	Network Slicing
NSA	Non-Stand Alone
NSSF	Network Slice Selection Function
NSSAAF	Network Slice-specific and Authentication and Authorization Function
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access

P-GW	Packet Data Network Gateway
PAPR	Peak-to-Average Power Ratio
PCC	Primary Component Carrier
PCell	Primary Cell
PCF	Policy Control Function
PD	Packet Duplication
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Packet Data Unit
PHY	Physical
PRB	Physical Resource Block
PSM	Power Saving Mode
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RA	Random Access
RAI	Release Assistance Indication
RAN	Radio Access Network
RAR	Random Access Response
RAT	Radio Access Tecnology
RE	Resource Element
RedCap	Reduced Capability
RF	Random Forest

RLC	Radio Link Control
RRC	Radio Resource Control
RSRP	Reference Signal Received Power
S-GW	Serving Gateway
SA	Stand Alone
SBA	Service-Based Architecture
SC-FDMA	Single Carrier Frequency Division Multiple Access
SCC	Secondary Component Carrier
SCell	Secondary Cell
SCG	Secondary Cell Group
SCS	Subcarrier Spacing
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Network
SDR	Software Defined Radio
SEAF	SEcurity Anchor Function
SeNB	Secondary eNB
SINR	Signal to Interference plus Noise Ratio
SMC	Security Mode Command
SMF	Session Management Function
SN	Secondary Node
SR	Service Request
SRB	Signaling Radio Bearer
TB	Transport Block
TBS	Transport Block Size

TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TSN	Time Sensitive Network
TTI	Transmission Time Interval
UDM	Unified Data Management
UDP	User Datagram Protocol
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
USIM	Universal Subscriber Identity Module
UWB	Ultra-Wide Band
VoIP	Voice over Internet Protocol
VoLTE	Voice over LTE
WISA	Wireless Interface to Sensors and Actuators

Part I

Background

Chapter 1

Introduction

This chapter provides an introduction to the work carried out during this thesis. First, Section 1.1 provides the motivation of this thesis, describing the fourth industrial revolution and indicating how cellular networks can be applied in this context. Next, the challenges identified and the objectives pursued in this thesis are outlined in Section 1.2. Finally, the structure of the document is described in Section 1.3.

1.1 Motivation

The advent of the fourth industrial revolution or Industry 4.0 [1] marks a transformative shift in manufacturing and the industrial sector. The term Industry 4.0 was used for the first time in 2011 in the assignment that the German government made to the Industry-Science Research Alliance for the consolidation of the leadership of the German Industry [2]. This initiative was subsequently extended to the rest of the European Union, and today, Industry 4.0 refers to the interconnection of machines and systems within production centers, as well as between them and the outside world. This digital revolution is transforming factories into smart factories, where digitization is key. In a connected factory, sensors, cloud storage and real-time data analysis are used to optimize production processes. Central to this revolution is the need for a robust, efficient, and improved flexibility of production and distribution processes. To achieve these needs, there are different enabler technologies that are in the core of Industry 4.0 (see Figure 1.1):

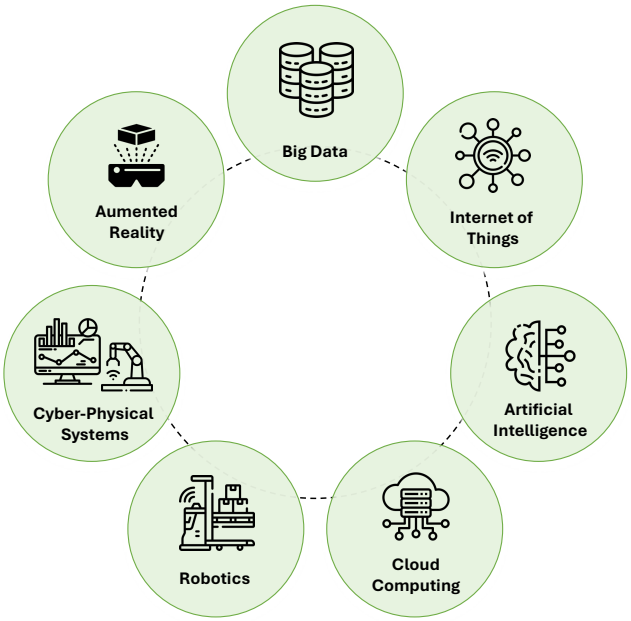


Figure 1.1: Industry 4.0 enabler technologies.

- Cyber-Physical Systems (CPS) [3, 4]. They integrate computing and network capacity into a physical process. CPS technologies enable the development of smart factories, where machinery and equipment are interconnected, allowing for real-time monitoring, control, and optimization.

- Internet of Things (IoT) [5]. IoT is a network of physical objects that have been embedded with sensors, software, and other technologies to enable them to connect and exchange data. In Industry 4.0, IoT facilitates the seamless flow of information across production lines, enhancing operational visibility and decision-making.

- Artificial Intelligence (AI) [6]. AI algorithms analyze vast amounts of data generated by CPS and IoT devices. This technology enables predictive maintenance, quality control, and adaptive manufacturing processes, reducing downtime and improving product quality.

- Cloud Computing [7]. Cloud computing plays a significant role in Industry 4.0 by providing the infrastructure and platform for storing, processing, and analyzing the large amounts of data generated by IoT devices and other sensors in the manufacturing process. In addition, cloud computing can provide the computing power needed to run AI algorithms.

- Augmented Reality (AR) [8]. The application of AR technology has the potential to enhance a range of processes, including training, maintenance, and design. By overlaying digital information onto the physical world, AR technology can provide workers with real-time data and instructions, thereby facilitating more efficient and effective workflows.

- Robotics [9, 10]. Robots and automation systems in Industry 4.0 are more intelligent, flexible, and collaborative. These systems can perform complex tasks alongside human workers, enhancing productivity and safety in manufacturing environments.

- Big Data analytics [11, 12]. The collection and analysis of large datasets allow for better forecasting, efficiency improvements, and the discovery of new insights. Data-driven decision-making is at the core of Industry 4.0, driving more responsive and agile manufacturing practices.

Although the Industry 4.0 concept is focused on manufacturing, the aforementioned technologies and principles are also applied across different industry sectors such as logistics, healthcare, agriculture or energy.

Traditional industrial networks are mainly based on wired connections and legacy wireless technologies. Some of the wired connections that have been used are ProfiNET [13], EtherCAT [14] and the set of Time Sensitive Networks (TSNs) protocols [15]. In the field of wireless technologies, the main technologies used are those based on the IEEE 802.11 family, commonly named Wi-Fi, but also customized solutions for factories based on IEEE 802.15.1 and 802.15.4, such as Wireless Interface to Sensors and Actuators (WISA) and WirelessHART [16]. Nevertheless, these networks often fall short in terms of scalability, flexibility, and real-time responsiveness required by modern industrial applications [17]. The dynamic nature of smart factories, autonomous systems, and complex supply chains needs a communication infrastructure that can seamlessly support a vast number of connected devices, facilitate real-time data exchange, and ensure high levels of security and reliability.

Cellular networks, with their widespread adoption, proven reliability, and continuous evolution, are uniquely positioned to address these needs, offering a foundational technology to propel Industry 4.0 forward. Cellular networks, particularly with the advent of the Fifth Generation of mobile networks (5G) and upcoming 6G technologies [18], offer unprecedented capabilities that align perfectly with the demands of

Industry 4.0. These include the support of use cases related to critical communications, that are known as Ultra-Reliable Low Latency Communications (URLLC), the massive use of machine-type devices, also known as Massive Machine Type Communications (mMTC), and Enhanced Mobile Broadband (eMBB). The ability to provide deterministic communication, support for a massive number of IoT devices, and high data throughput are critical enablers for applications such as predictive maintenance, remote monitoring, and autonomous robotics. Furthermore, the modular and scalable nature of cellular networks allows for tailored deployments in diverse industrial environments, from large-scale manufacturing plants to remote and isolated facilities. This flexibility supports the creation of private networks [19] dedicated to specific industrial needs, ensuring that the unique requirements of different sectors are met effectively.

The global push towards sustainability and efficiency in industrial operations [20] further underscores the importance of leveraging advanced communication networks. By enabling more efficient resource management, reducing downtime through predictive maintenance, and facilitating the seamless integration of renewable energy sources, cellular networks [21] contribute significantly to the sustainability goals of modern industries.

As the adoption and implementation of the cellular technology is progressively taking place in factories, especially the 5G technology [22], it is necessary to study its applicability, evaluating the network performance through the different services and use cases that are involved in a smart factory.

1.2 Challenges and objectives

The main objective of this thesis is to asses and improve cellular network performance in an indoor industrial environment. For this purpose, different techniques and optimizations to the network are addressed in this thesis. First, tasks related to the study of the latency performance of critical services and the scalability in the network are carried out. Secondly, different tools have been developed to assess network performance in an industrial environment and to improve the reliability of critical services through the use of the multi-connectivity solution. Thirdly, optimizations algorithms are developed and evaluated with the following purposes: to improve the reliability of critical services without resource wastage, and to enhance the Quality of Service (QoS) of the different traffic profiles involved in a factory. Finally, the performance of the different optimizations proposed by the Third Generation Partnership Project (3GPP)

for Cellular Internet of Things (CIoT) devices has been evaluated, also including a security analysis of the latest optimization.

In an indoor industrial scenario, there are many challenges present due to the characteristics of this particular scenario. First, in a factory, a harsh environment is present for radio propagation due to the presence of large metallic machines within crowded spaces [23]. This causes reflections and multi-path in the signal, making difficult the appropriate adjustment of the configuration in the network according to radio conditions. Another challenge regarding to the scenario that is also present is that the distribution can change from one day to another (i.e., moving and placing stock from one side of the factory to another), so network conditions and performance could vary. This is specially important for adjusting the appropriate configuration in the network to maximize the QoS of the different services and to fulfil service requirements.

As new use cases are introduced in the smart factory to enhance the flexibility, such as the mobility within the factory by using an Automated Guided Vehicle (AGV), it also introduces more demanding requirements in the network for the correct operation of these applications. In particular, the AGVs are vehicles that follow programmed paths to transport material and goods within the factory facility [24]. The AGVs communicate with a guidance control system from which they receive guidance commands. For the correct operation of the AGVs, the communication with the guidance control system needs to be in real-time with low latency and high reliability to avoid malfunctions or accidents in the factory. Therefore, these communications are considered as critical. In 5G, there are different approaches followed in the literature to reduce the latency of the communications in the Radio Access Network (RAN). The approaches encompass the use of mini-slots [25, 26], solutions in the scheduler [27–30], uplink grant free transmissions [31–34], and the use of a flexible numerology [35–38]. The numerology approach has been one of the mainly used techniques, with the use of a higher numerology to successfully reduce the latency for critical services. Despite the importance of this approach, the latency evaluation of the numerology in the literature only considers Line-of-Sight (LOS) conditions, and there are no studies that include the assessment in the industrial scenario. Given that Non-Line-of-Sight (NLOS) conditions are the most prevalent in a factory, it is necessary to conduct an assessment to study the numerology impact and identify the optimal configuration for critical services. Therefore, the first objective (Obj. 1) of this thesis consists in studying and assessing the impact of the different 5G numerologies on the latency experienced by an AGV in an indoor factory scenario under LOS and NLOS conditions.

Another aspect that must be taken into account in a smart factory is the network scalability. As many devices are connected in a smart factory and high traffic is coming from machinery, sensors, AGVs, etc., it may overload the network or decrease the performance, thereby not achieving the requirements of the different applications. In the case of critical services, this becomes even more important, as a low latency must be ensured despite an increase in the number of devices to maintain productivity and prevent accidents. Moreover, the technology selection in the factory is not clear for industrial manufacturers. Some will prefer low cost-technologies with lower performance, and others will prefer a very reliable network although this implies a higher cost. Different assessments of the network performance in an industrial scenario have been carried out in the literature with different technologies such as Long Term Evolution (LTE), 5G, Wi-Fi, and the use of Multi-Radio Access Technology (multi-RAT) connectivity [39–47]. However, the existing literature does not take into account the scalability of the network, with only one device attached to the network in the majority of the works. Therefore, following with this line, the second objective (Obj. 2) of this thesis is to assess and compare the scalability of the network with different technologies in an industrial environment. In this way, the study should provide a clear vision of which technology suits better the manufacturing sector.

Following the line started with Obj. 1, the critical services involved in the smart factory also have requirements of high reliability in the communication in addition to a low latency. In the 5G network, the use of multi-connectivity is proposed to enhance the reliability of critical communications. Multi-connectivity consists in establishing two or more links between a user and two or more radio access nodes, which are typically uncorrelated links. For instance, the two links can use different channels, different networks or even different network access technologies, namely multi-RAT [48]. Multi-connectivity is often adopted for improving communication aspects such as latency, reliability and throughput [49,50]. In the case of reliability, the most extended solution is the Packet Duplication (PD) approach [51,52]. This solution allows the transmission of the same data duplicated from different links. However, this solution comes at a cost in terms of network redundancy, as the duplication can lead to an inefficient use of network resources, resulting in a degradation of the overall network performance [53]. Thus, the third objective (Obj. 3) of this thesis consists in the enhancement of the reliability for critical communications, designing and developing a dynamic algorithm to control the activation of PD to avoid resource wastage in the network.

One of the industrial sectors where Industry 4.0 is adopted is smart logistics to add

flexibility and easily adapt to changes both at large and small volumes of moving stock. In smart logistics, distribution centers [54] play the role of nodes in the distribution network, where small batches of products (or even individual units) are received, stored for very short periods of time (days or hours), and redirected to the next distribution center. Within a distribution center, different traffic profiles with diverse requirements are present (i.e., workers with AR glasses, smart tags, and AGVs moving stock), which makes it challenging, as network resources need to be shared between these profiles and each traffic profile requires different network parameters such as the numerology. The application of the 5G network in smart logistics has been discussed in the literature in [55–59]. On the other hand, several works have been centered in optimizing the 5G network specifically for smart logistics, analyzing the specific particularities of the applications and the different environments where the processes take place [60–65]. One solution followed in the literature to tackle this challenge is the use of Network Slicing (NS), which allows the creation of subsets of the network for each traffic profile, with optimized configurations [66–68]. Despite that, the performance of the 5G network has not been assessed yet in a distribution center. Therefore, the fourth objective (Obj. 4) of this thesis is to evaluate the 5G network performance in a distribution center in terms of QoS for the different traffic profiles involved there.

CIoT devices are characterized by the transmission of small data to update information provided by sensors in the factory such as the temperature, the humidity, the state of a machine, an alarm, etc. These transmissions are characterized by a large transmission interval (i.e., one transmission per hour), and on each transmission a high signaling overhead is produced compared to the size of the data sent. Furthermore, the power constraints inherent to CIoT devices, which rely on batteries, underscore the importance of optimizing data transmission to save energy. To tackle with this challenge, many optimizations have been proposed in the standard by the 3GPP with two main objectives: to increase the battery life and to reduce the signaling overhead when transmitting data. These optimizations were first proposed in Release 13, namely Control Plane (CP) and User Plane (UP) CIoT optimizations; and a new optimization was introduced in Release 15 for infrequent and small data transmissions, namely Early Data Transmission (EDT). The enhancements to these optimizations in terms of battery life and the latency have been the subject of analysis in the literature, as evidenced by works such as [69–73]. Nevertheless, none of these employ the use of commercial equipment; rather, they employ analytical frameworks or simulators. Additionally, they operate under the assumption of different ideal scenarios, which do not

align with the actual implementation of the 3GPP standard. Thus, the fifth objective (Obj. 5) of this thesis consists in studying the latency impact of CIoT signaling optimizations in the network with commercial equipment.

Lastly, in relation with Obj. 5, much effort has been made in optimizing the transmissions of CIoT devices, however, it is of particular importance to also analyse the threats and vulnerabilities, and to ensure the security of these optimizations. As the security of the EDT optimization has not been studied in the state of the art, the sixth objective (Obj. 6) of this thesis consists in analyzing the security of the EDT optimization in the 5G network. In this way, the EDT optimization should be described in detail in its operation modes and the main vulnerabilities associated to this optimization must be analyzed.

In summary, the objectives that address the previous challenges are the following (see Figure 1.2):

Obj. 1. To study the impact of 5G numerologies on the latency for critical services.

The objective of this study is to analyse the behaviour of the different numerology configurations on the latency perceived by the users under different channel conditions and packet sizes. In this way, this study aims to lay the foundations for future optimizations to reduce the latency, as the appropriate numerology can be selected according to the radio conditions experienced.

Obj. 2. To assess and compare network scalability with different technologies in an industrial environment.

The aim of this objective is to empirically assess and compare the network performance in terms of latency and packet loss with different technologies in an indoor industrial scenario. In particular, the assessment should take into account different packet sizes and scenarios with varying number of devices transmitting data. As a result, this study should provide a clear vision of which technology suits better the manufacturing sector.

Obj. 3. To propose a mechanism to enhance reliability for critical services.

This objective refers to the design and development of an algorithm to fulfil reliability requirements for critical services. Thus, the proposed algorithm should be able to dynamically adapt and control the activation of PD to avoid resource wastage in the network.

Obj. 4. To evaluate the network performance in a distribution center.
The aim of this objective is to perform an assessment of the 5G network in a distribution center scenario, taking into account the different traffic profiles involved in this scenario. In particular, this work should compare the QoS of these traffic profiles under different logistics activities with different NS approaches.

Obj. 5. To study the impact of CIoT signaling optimizations in the network.
The objective of this study is to analyse the behaviour of the different CIoT signaling optimizations on the latency perceived by the user when transmitting infrequent small data into the network.

Obj. 6. To analyze the security of 5G EDT optimization for CIoT.
This objective is related to Obj. 5 and refers to an in-depth analysis of the security of the EDT optimization, describing in detail its operation modes and analyzing the main vulnerabilities associated to this optimization. As a result, a set of recommendations for vendors should be derived from the security analysis.

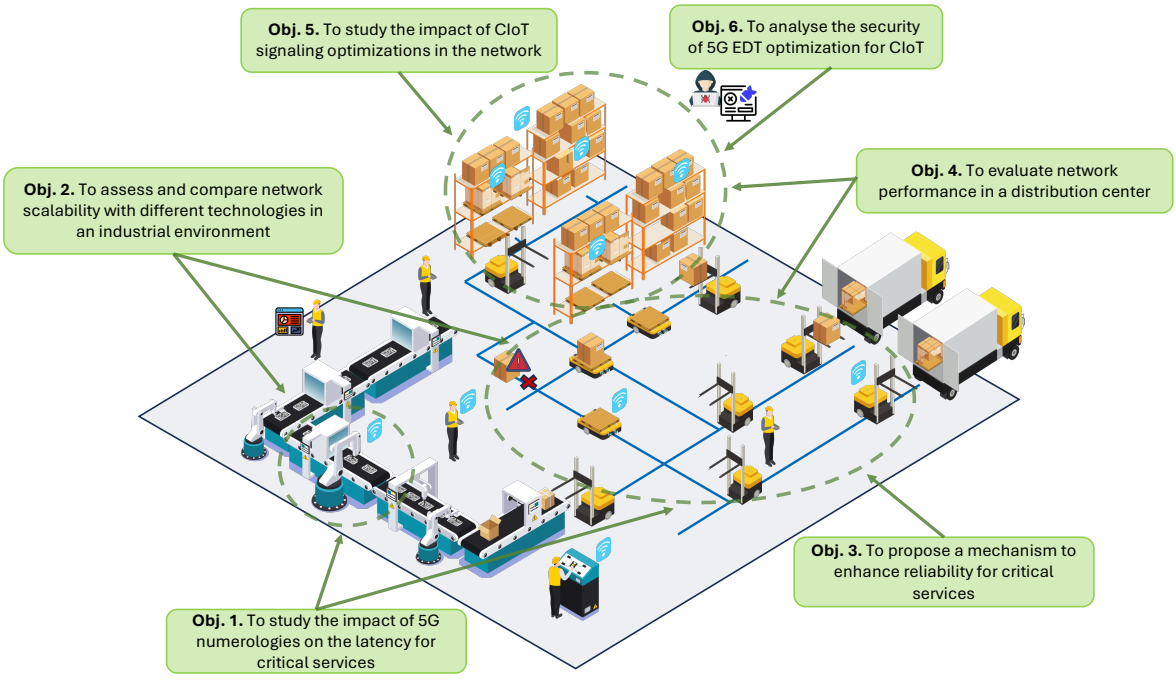


Figure 1.2: Conceptual scheme of the objectives addressed.

1.3 Document structure

This document has been structured in seven chapters grouped into three blocks for an easier understanding, as shown in Figure 1.3. The first block contains the background and knowledge required to understand the rest of the thesis and it is composed of two chapters. Chapter 1 consists in an introduction to the thesis, detailing the motivation that led to the research conducted and setting out the objectives to be addressed. Chapter 2 provides the technical background necessary to understand the content of the thesis. In this chapter, first the cellular technologies involved in this thesis are described: LTE, 5G and CIoT. In particular, their main characteristics and features are described for each of these technologies. Next, an overview of multi-connectivity in 5G is provided, presenting the different architectures and the benefits and challenges of this feature. Finally, this chapter ends with a description of the 5G network security, in particular, it is focused on the security architecture and procedures, along with the threat model and main attacks.

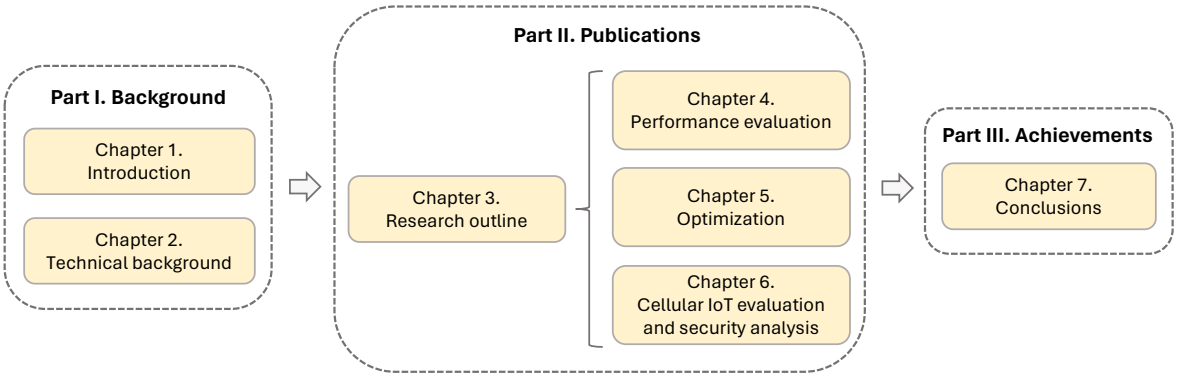


Figure 1.3: Document structure.

The second block corresponds to the publications that support this thesis. This block has been divided into different chapters according to their topic. This block includes a first chapter (Chapter 3) in which the research outline is presented. In particular, this chapter details the relationship between the publications and the challenges, objectives and chapters of this thesis. The research methodology followed in the development of this thesis is also presented in this chapter. The rest of the chapters included in this second block correspond to the publications that are directly related to the objectives established in Section 1.2.

Chapter 4 contains the results related to the performance of the cellular network and is associated with Obj. 1 and 2 of this thesis, including two publications. Specifically,

the first publication addresses the latency evaluation of the new numerologies introduced in 5G under different channel conditions. The second one provides an empirical comparison of the scalability performance of 5G, Wi-Fi 6, and multi-connectivity in terms of latency and packet loss. For that, measurement campaigns were performed in an indoor industrial scenario with commercial equipment.

Chapter 5 covers the work related to the development of algorithms that aims to improve the performance in the network. This chapter is related to Obj. 3 and 4 of this thesis and includes two publications. The first publication proposes a dynamic packet duplication algorithm in multi-connectivity scenarios for latency-constraint services in order to improve the reliability and reduce the resource consumption. In particular, the solution proposed is based on Machine Learning (ML), and a latency predictor is trained and evaluated. The second one introduces a novel open-source simulator with a realistic representation of a distribution center scenario. In particular, the floorplan, activities and applications have been developed and the 5G network performance is evaluated by comparing two NS strategies (static and dynamic). This comparison have been performed for all traffic profiles involved in a distribution center (eMBB, URLLC and mMTC).

Chapter 6 is related to the evaluation and the security analysis of CIoT in the context of 5G. This chapter is associated with Obj. 5 and 6 of this thesis and includes two publications. Specifically, the first publication evaluates and compares the different transmissions modes for CIoT, such as Release 13 optimization and EDT with commercial equipment. The evaluation takes into account different coverage levels and packet sizes and focuses on the latency performance. The second one describes in detail the EDT feature for CIoT devices and analyzes its main vulnerabilities with a set of recommendations.

Finally, the third block consists of Chapter 7, which provides an overview of the main results and conclusions of this thesis and the future lines of research are discussed.

This document also includes two appendices. Appendix A describes the evaluation tools and testbeds used for research. Finally, the summary of the thesis in Spanish is provided in Appendix B.

Chapter 2

Technical background

This chapter provides a review of the technical background necessary to follow the content of this thesis. The first section describes the cellular technologies involved in this thesis: LTE, 5G, and CIoT. More specifically, it describes the architecture and main components, along with the radio access technology. Section 2.2 describes the multi-connectivity feature for 5G networks, along with the benefits and challenges. Finally, Section 2.3 describes the security in cellular networks, in particular, it is focused on the security architecture and procedures for the 5G network. Moreover, the threat model along with the main attacks are also described.

2.1 Cellular technologies

This section aims to provide an overview of the cellular technologies that are the basis of this thesis. Section 2.1.1 describes the LTE network. Section 2.1.2 describes the 5G network, including its architecture and protocol stack. The new radio access technology in 5G is described in Section 2.1.3. Finally, Section 2.1.4 makes an overview of CIoT focused on Low-Power Wide-Area (LPWA) technologies.

2.1.1 LTE

The Evolved Packet System (EPS), also known as LTE, was first introduced in Release 8 as a mobile communication standard by the 3GPP [74]. The system differs from its predecessor technology, 3G, by using packet-switched networks for the delivery of all services, also including voice, which is commonly referred as Voice over Internet

Protocol (VoIP) and, in the LTE network, as Voice over LTE (VoLTE). LTE evolved in Release 10 with the introduction of LTE-Advanced, which provides an increased data rate and enhancements in multi-antenna techniques. At this point, the standard met the requirements of a 4G network [75].

Regarding the transmission in LTE, Orthogonal Frequency Division Multiple Access (OFDMA) is used in the downlink, while in the uplink Single Carrier Frequency Division Multiple Access (SC-FDMA) is used.

Network architecture

The architecture of LTE is divided into two parts [76]: the Evolved Universal Terrestrial Radio Access Network (E-UTRAN) and the Evolved Packet Core (EPC), as depicted in Figure 2.1. The EPC performs functions such as mobility management, network access control or the connection with external networks. The following elements compose the EPC:

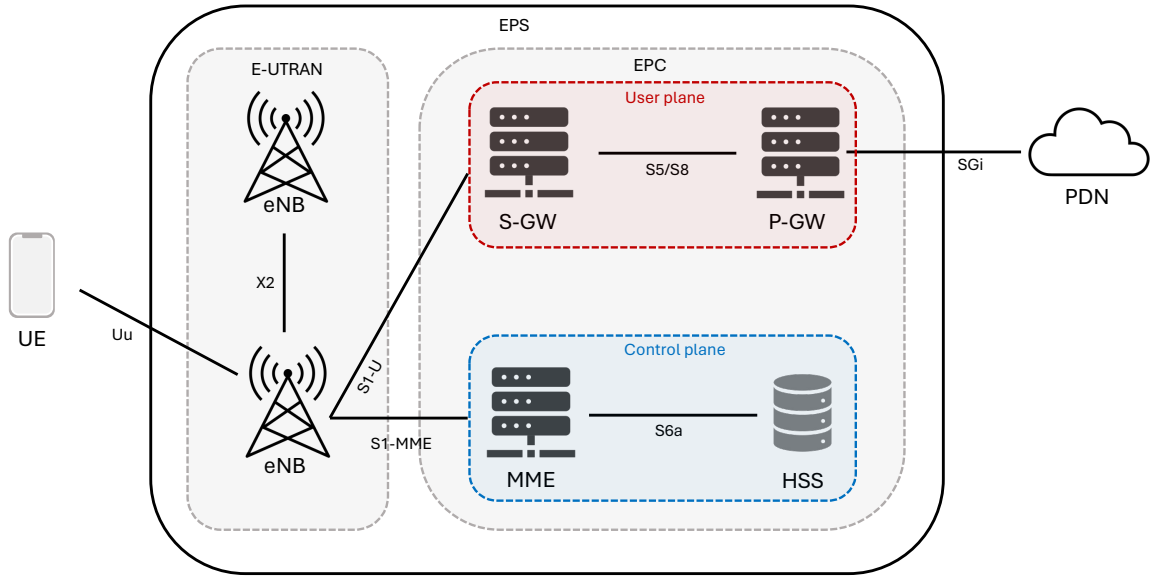


Figure 2.1: LTE architecture.

- Mobility Management Entity (MME): this element manages the CP between the User Equipment (UE) and the Core Network (CN). Its main functions are:
  - Non-Access Stratum (NAS) signaling and security.
  - Access Stratum (AS) security control.
  - EPS bearer control.

  - Roaming and authentication.
  - Idle state handling.

- Serving Gateway (S-GW): this element manages the UP between the UE and the CN. This element performs the following functions:
  - Local mobility anchor point for inter-eNB handover.
  - Mobility anchoring for inter-3GPP mobility.
  - E-UTRAN idle mode downlink packet buffering and initiation of network triggered service request procedure.
  - Packet routing and forwarding.
  - Transport level packet marking in the uplink and the downlink.

- Packet Data Network Gateway (P-GW): this element connects the EPC to external Internet Protocol (IP) networks. The main functions of the P-GW are the following:
  - Per-user based packet filtering.
  - UE IP address allocation.
  - Transport level packet marking in the uplink and downlink.

- Home Subscriber Server (HSS): database that stores users subscription data such as the QoS profile, roaming access restrictions or Access Point Name (APN) of Packet Data Network (PDN) to which the user can connect.

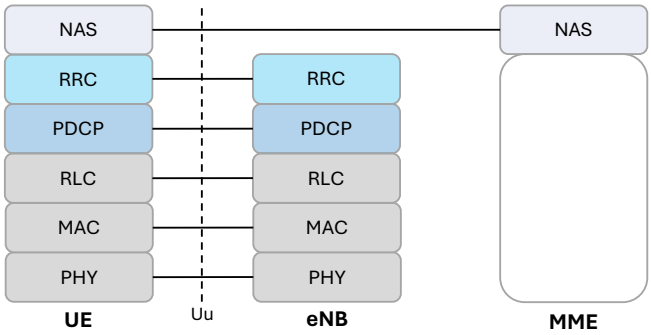
The E-UTRAN consists of base stations, providing E-UTRA UP and CP protocols terminations towards the UE. The E-UTRAN is composed with only one kind of element, the Evolved Node B (eNB), which establishes communication with the UEs via radio signal and with the CN elements. Its main functions are:

- Radio Resource Management functions (e.g., Radio Bearer Control, Admission Control, etc.).
- IP and Ethernet header compression.
- Selection of an MME at UE attachment.
- Routing of UP data towards S-GW.
- Scheduling and transmission of broadcast information and paging.

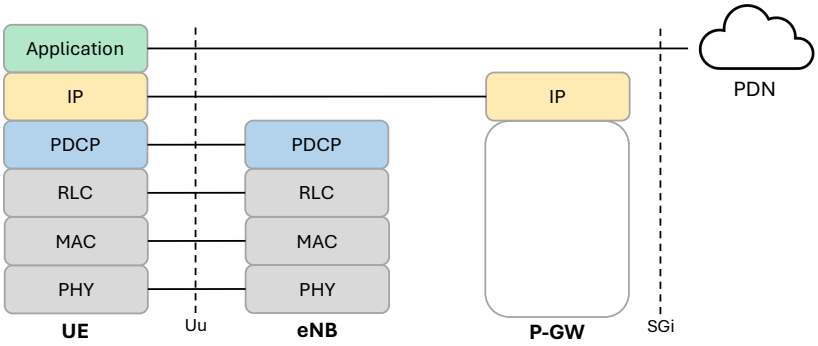
Protocol stack

The communication in LTE between the network elements is done by using a protocol stack. The protocol stack is different for the UP and the CP, along with the network core elements involved, as shown in Figure 2.2.

The UP is used for data transmission between the UE and the network. On the other hand, the CP is used between the network elements to establish connections and authentication. A brief description of the different layers is provided below.



(a) Control Plane.



(b) User Plane.

Figure 2.2: LTE protocol stack.

- Physical layer (PHY): the main function of this layer is the transmission of the signal over the radio channel. This layer includes the use of channel coding, modulation, resource element mapping and mapping to antennas.
- Medium Access Control (MAC): the main purpose of this layer is the allocation of radio resources. This layer also performs functions such as physical channel error corrections using the Hybrid Automatic Repeat reQuest (HARQ) technique or the control of the Random Access (RA) procedure to establish connection from the UE to the network.

- Radio Link Control (RLC): this layer provides a radio connection with an error detection and recovery with Automatic Repeat reQuest (ARQ). This layer also manages packet segmentation and reassembly.
- Packet Data Convergence Protocol (PDCP): this layer provides functions such as header compression and decompression, applies security functions, sequential delivery and data duplicated detection.
- Radio Resource Control (RRC): this layer manages the broadcast of system information (i.e., paging); establish, modify and release RRC connections; performs handover between cells; and encapsulates NAS messages in RRC messages.
- NAS: this layer manages direct signaling between the UE and the MME to establish and maintain communication sessions with the UE as it moves through the network.
- IP layer: this layer is a network protocol that provides bidirectional data transfer, ensuring that data packets are routed and delivered correctly across the network.

Radio Access technology

The physical layer of LTE is based on two multiple access technologies over the air interface: OFDMA and SC-FDMA for downlink and uplink transmissions, respectively. OFDMA allows multiple users to transmit data simultaneously on different subcarriers within the same frequency band. This is done by dividing the channel into a set of narrow subcarriers that are divided into groups according to the needs of each user. This parallel transmission enhances spectral efficiency and enables LTE to achieve higher data rates compared to previous technologies. However, the combination of a high number of subcarriers leads to a high Peak-to-Average Power Ratio (PAPR), which causes a high power consumption in the radio transceivers. In the uplink, this multiple access increases the battery consumption in the UEs. For this reason, OFDMA is not suitable for uplink and SC-FDMA is used instead. SC-FDMA utilizes a single-carrier transmitting signal, transmitting the data symbols in series over one wideband signal, with higher rate and more bandwidth.

Regarding the modulation scheme, in downlink each subcarrier is modulated using from Quadrature Phase Shift Keying (QPSK) to 1024-Quadrature Amplitude Modulation (QAM), whereas the modulation used in uplink ranges from $\pi/2$-Binary Phase Shift Keying (BPSK) to 256-QAM [77].

In terms of system bandwidth, the following values are available in LTE: 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz and 20 MHz. The physical radio resources can be thought of as a set of subcarriers in the frequency domain and a set of opportunities for modulated symbols in the time domain. The smallest resource unit in the physical layer of LTE is called Resource Element (RE), which consists of one symbol in the time domain and one subcarrier in the frequency domain. The REs are grouped together into logical structures that can be used for transmission and reception, called Physical Resource Block (PRB). A PRB is the smallest unit of resources that can be allocated to a user. The PRB consists of 180 kHz in the frequency domain and one slot (0.5 ms) in the time domain. In frequency, PRBs are either $12 \cdot 15$ kHz subcarriers or $24 \cdot 7.5$ kHz subcarriers wide, depending on the Subcarrier Spacing (SCS). The number of subcarriers used per PRB for most channels and signals is 12. On the other hand, the number of symbols per PRB varies depending on the length of the cyclic prefix used, with the common practice being the use of 7 symbols per time slot. In the time domain, LTE is composed of frames with a duration of 10 ms and there are ten subframes per frame with a duration of 1 ms and each subframe contains two slots. An overview of the LTE frame structure is depicted in Figure 2.3.

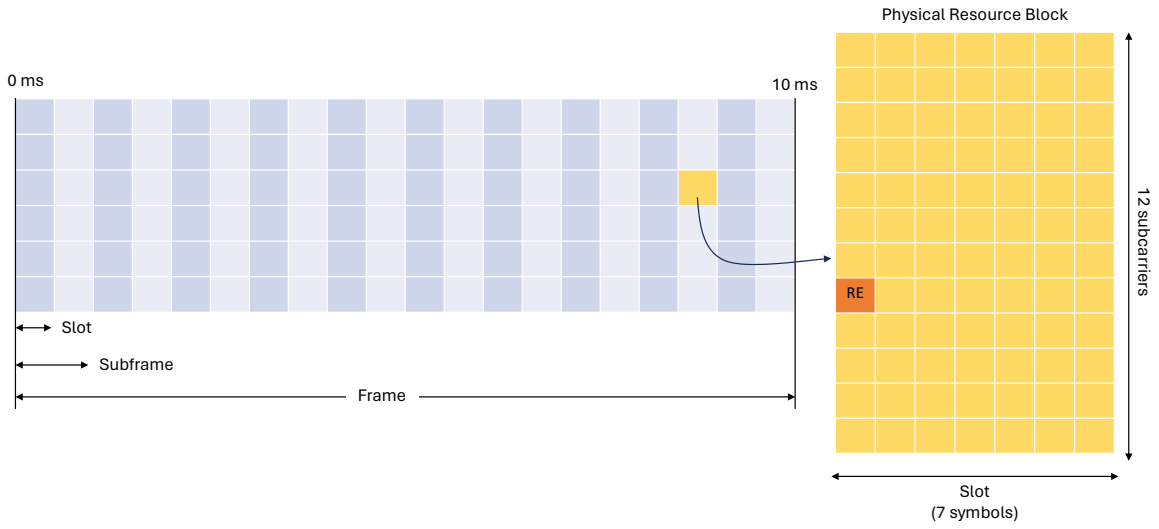


Figure 2.3: LTE frame structure.

2.1.2 5G

The 5G network was first introduced in Release 15 with the aim to provide more flexibility to support new services and applications. Unlike previous technologies that were focused on traditional mobile broadband, in 5G the services have been classified

into three categories according to their requirements [78]:

- eMBB: this service category is an evolution of traditional mobile broadband, with higher data rates and bandwidth. This category encompasses traditional human use cases such as web browsing or streaming multimedia content.
- URLLC: this service category aims to cover critical communications, with low latency and high reliability requirements. Some examples of this category include industrial automation, self-driving car or remote medical surgery.
- mMTC: this category covers massive connection of devices with low bandwidth requirement and non-critical delay. It is mainly focused on the IoT.

An overview of the requirements for the different services is depicted in Figure 2.4.

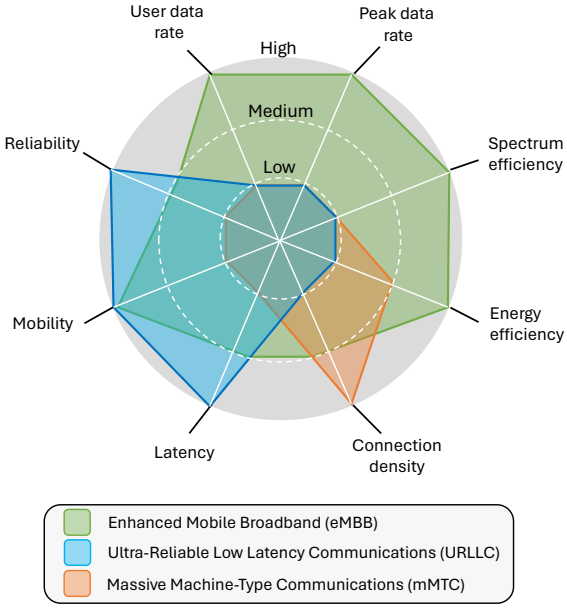


Figure 2.4: Requirements for the different 5G service categories.

To support these traffic profiles and add more flexibility to the network, the following key features have been introduced in 5G:

- Beamforming: this technique allows to transmit the signal directly in one or more specific directions by changing the phase and amplitude of the transmission when using multiple antennas.
- Multiple-Input Multiple-Output (MIMO): this technique allows to transmit different data streams multiplexed on the same spectral resources by using different

antennas, with the aim of improving the channel spectral efficiency. This technique is evolved from LTE MIMO and in 5G the number of antennas is increased.

- Numerologies: 5G provides flexibility in the frame configuration, that is, the SCS and the cyclic prefix. The main purpose of the numerology is to reduce the latency by reducing the slot duration in the frame structure. A more in depth detail of this feature is explain in Section 2.1.3.

- Network Slicing (NS): this technique allows to divide the physical network into different logical networks, commonly known as slices. Each slice can be configured with different parameters to be optimized for a specific application. The implementation of NS is simplified by using Software Defined Network (SDN) and Network Function Virtualization (NFV).

- Mobile Edge Computing (MEC): applications servers are moved to the network edge, thus, reducing the path that a packet must travel.

- Multi-connectivity: consists in establishing multiple radio links between the UE and the network. These links can be established using different components carriers in one node, using different nodes of a network or a combination of both. This feature can be used to improve the throughput if different information is transmitted or to improve the reliability if the same information is transmitted in all links. More details of this feature is explained in Section 2.2.

Network architecture

The 3GPP introduced in Release 15 the first specification of the 5G technology, where two deployment options are defined [79]: the Non-Stand Alone (NSA) and Stand Alone (SA). In the NSA architecture (see Figure 2.5), the 5G RAN is used in conjunction with the existing LTE and EPC infrastructure. This configuration provides the same services as LTE, but with improvements offered by the 5G New Radio (NR) such as lower latency. In this case, a new network element has been introduced in the RAN, the Evolved-Next Generation NodeB (en-gNB). The en-gNB is a network node that provides NR UP and CP towards the UE and it is connected to the EPC.

On the other hand, the SA architecture (see Figure 2.6) operates using the 5G System (5GS). The 5GS is divided into the Next Generation Radio Access Network (NG-RAN) and the 5G Core (5GC). Under this architecture, the different services

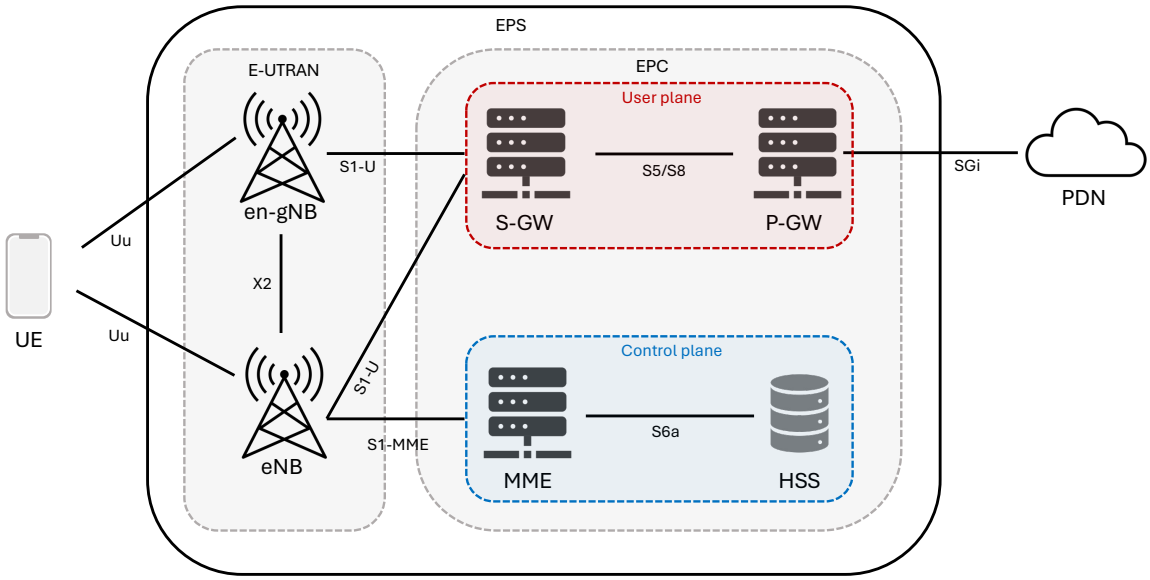


Figure 2.5: 5G NSA architecture.

introduced in 5G are offered. The new elements defined in 5G SA in the RAN part are described below [80, 81].

- Next-generation eNB (ng-eNB): network node that provides E-UTRA UP and CP towards the UE with capabilities to connect to the 5G core.
- Next-generation NodeB (gNB): network node that provides NR UP and CP towards the UE and it is connected to the 5GC. Its main functions are:
  - Radio resource management functions (e.g., radio bearer control, radio admission control or connection mobility control).
  - Selection of an Access and Mobility Management Function (AMF) at UE attachment.
  - Routing of UP data towards User Plane Function (UPF) and CP towards AMF.
  - Scheduling and transmission of paging messages and system broadcast information.
  - Connection setup and release.
  - Session management.
  - Support of NS.
  - Dual connectivity.

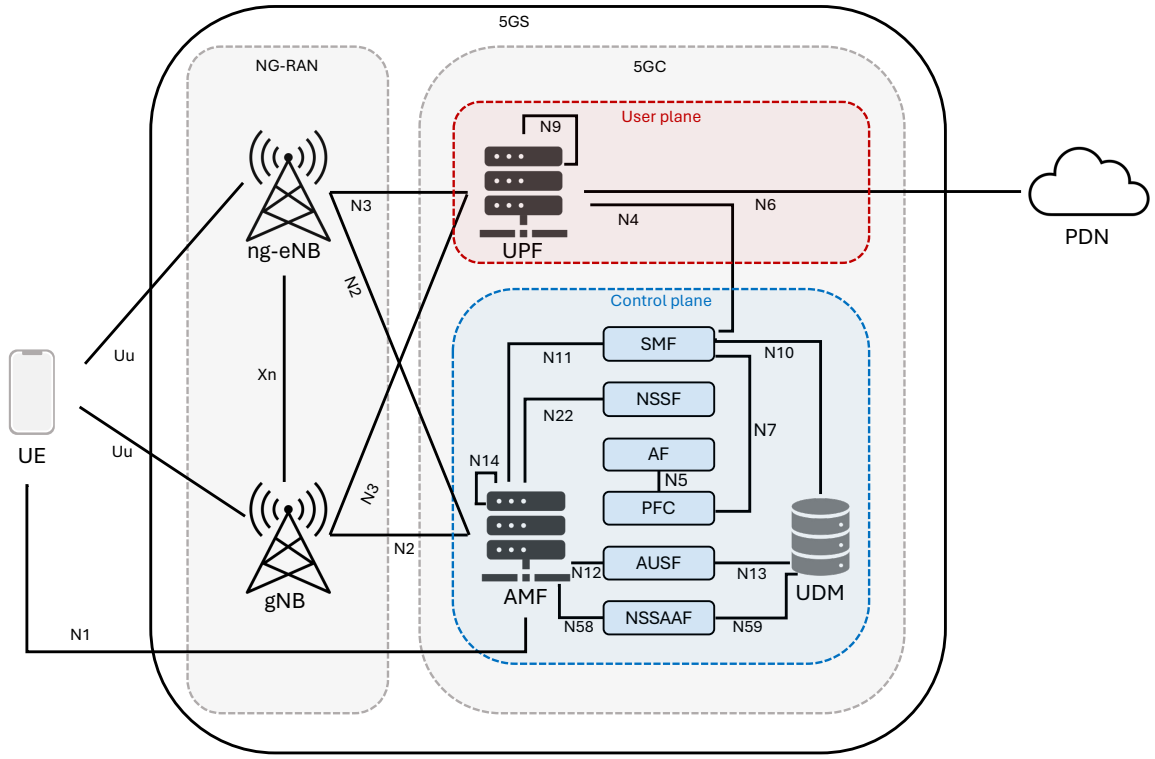


Figure 2.6: 5G SA architecture.

– QoS flow management and mapping to Data Radio Bearer (DRB).

The 5GC architecture relies on a Service-Based Architecture (SBA), where the different mobile CN functionalities (authentication, mobility management, etc.) are defined in terms of Network Functions (NFs) rather than by traditional network entities. This allows an open and modular service platform. A description of the different 5GC NFs is provided below [81].

- AMF: it performs the access, authentication and mobility in the network. The main functions of the AMF are the following:
  - NAS signaling termination and security.
  - AS security control.
  - Access authentication and authorization.
  - Mobility management control.
  - NS support.
  - Session Management Function (SMF) selection.

- SMF: it is in charge of the session management of the UEs. Its main functions are:
  - Session management.
  - UE IP address allocation and management.
  - Control part of policy enforcement and QoS.
  - Configuration of traffic steering at UPF.
  - Downlink data notification.
- UPF: it connects the UP between the NG-RAN and the 5GC; and connects the 5GC to external IP networks. The UPF performs the following functions:
  - Anchor point for intra-/inter-Radio Access Tecnology (RAT) mobility.
  - Packet routing and forwarding.
  - Packet inspection and UP part of policy rule enforcement.
  - QoS handling for the UP.
  - Downlink packet buffering and downlink data notification triggering.
- Authentication Server Function (AUSF): this function is in charge of the authentication in the network. In particular, it executes the following functions:
  - Authentication for 3GPP access and untrusted non-3GPP access.
  - Authentication of the UE.
- Policy Control Function (PCF): this function is in charge of the policy of the network. It performs the following functions:
  - Supports unified policy framework.
  - Provides policy rules to CP functions.
- Application Function (AF): interacts with the CN in order to provide services such as application influence on traffic routing, time synchronization service or interacting with the policy framework for policy control.
- Unified Data Management (UDM): the main functions of this element are the following:

  - Generation of 3GPP authentication credentials.
  - User identification handling.
  - Access authorization.
  - Support to service continuity.
  - Subscription management.

- Network Slice Selection Function (NSSF): this function selects the set of NS instances serving the UE.
- Network Slice-specific and Authentication and Authorization Function (NSSAAF): this function supports for NS specific authentication and authorization.

Protocol stack

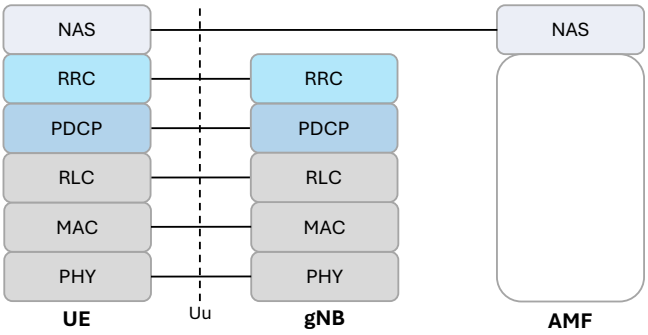
The protocol stack in 5G is based on previous LTE technology and it is depicted in Figure 2.7. However, an additional layer is included in the UP on top of the PDCP layer. This layer is called Service Data Adaptation Protocol (SDAP). The purpose of this layer is to map QoS flows to radio bearers and to provide QoS marking on data packets in the RAN for traffic prioritization purposes.

Moreover, some changes in many layers have been made in 5G:

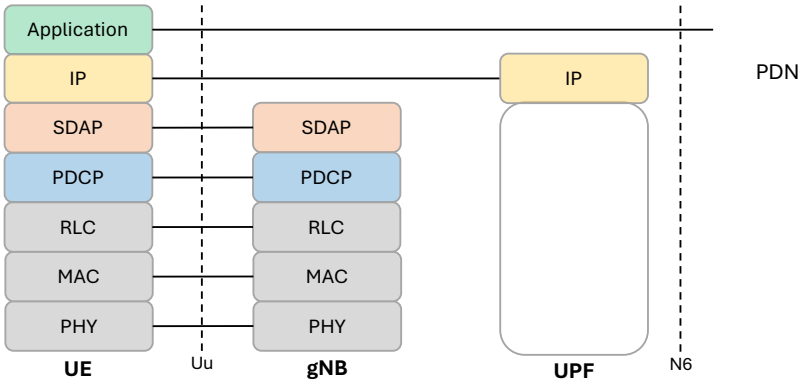
- RRC layer: support of a new RRC state with the aim of reducing the energy consumption and latency for devices that transmit small data with low frequency. This new state is namely RRC Inactive.
- PDCP layer: data integrity protection is added to the UP. Duplication is also added, mapping Packet Data Units (PDUs) to more than one logical channel and sending them over different component carriers.
- RLC/MAC layer: support for beam management procedures and transmission modes that use different numerologies and Transmission Time Intervals (TTIs).

2.1.3 New Radio access technology

As previously explained, the 5G technology was first standardized in Release 15 by the 3GPP. Its radio access technology has suffered changes compared to LTE with the aim



(a) Control Plane.



(b) User Plane.

Figure 2.7: 5G protocol stack.

to provide more flexibility and meet requirements such as lower latency and improved throughput in the communications. This radio access technology is known as NR.

5G NR is based on a flexible Orthogonal Frequency Division Multiplexing (OFDM) system, allowing it to operate in a wide range of bands, address different use cases and operate under multiple spectrum access. Regarding the waveform, OFDM with a cyclic prefix is used for the downlink while for the uplink, unlike LTE, it is possible to use OFDM with cyclic prefix or Discrete Fourier Transform-spread-OFDM (DFT-s-OFDM), the last one with the aim of minimizing the PAPR and to improve the uplink coverage [79].

With respect to the frequency operation in 5G, Release 15 allows frequencies up to 52.60 GHz. Moreover, two Frequency Ranges (FRs) are defined [82]:

- FR1: it covers frequencies from 410 MHz to 7.125 GHz.
- FR2: encompasses frequencies in the range of 24.25 GHz to 52.60 GHz. In this range, directional antennas are necessary due to propagation losses and interference.

Higher frequencies are considered in Release 17. In fact, a new FR has been defined, the FR2-2, which encompasses frequencies in the range of 52.60 GHz to 71 GHz.

Regarding the bandwidth for a component carrier, a maximum value of 100 MHz is supported for FR1, while for FR2 the maximum value is 400 MHz. Note that for the FR2-2 in Release 17, the maximum bandwidth value is increased up to 2 GHz. Nevertheless, the bandwidth configuration will differ depending on the SCS for data transmission configured in the network [82].

Numerology concept and frame structure

5G NR introduces a flexible SCS in its design. The SCS is formed as $15 \cdot 2^{\mu}$ where μ can adopt values of 0, 1, 2, 3 or 4, which results in SCS of 15, 30, 60, 120 and 240 KHz. This is commonly known as numerology (μ), defined by a SCS and the cyclic prefix, which can be normal or extended [79].

However, not all numerologies are suitable for each FR. The reason is that as the SCS increases, the symbol duration decreases, also reducing the cyclic prefix and this can led to inter-symbol interference due to OFDM signal characteristics. This will affect in FR1, since multi-path is present. However, as the frequency is increased, the multi-path problem is reduced, since propagation is predominantly LOS. Therefore, higher SCS are suitable for higher frequencies.

In the standard, the use of a numerology has been divided into the FR used and into data and synchronization channels [80]. In terms of data channels, $\mu = \{0, 1, 2\}$ is supported in FR1 and $\mu = \{2, 3\}$ in FR2. On the contrary, for synchronization channels, $\mu = \{0, 1\}$ is supported in FR1 and $\mu = \{3, 4\}$ in FR2. In Release 17, two numerologies have been introduced for FR2-2 (5 and 6), which corresponds to a SCS value of 480 and 960 kHz and they are supported by data and synchronization channels. The different 5G numerologies defined in the Release 17 are summarized in Table 2.1.

Numerology (μ)	**SCS (kHz)**	**Slots per subframe**	**Slot duration (ms)**	**Symbol duration (μs)**
0	15	1	1	71.42
1	30	2	0.5	35.71
2	60	4	0.25	17.85
3	120	8	0.125	8.92
4	240	16	0.0625	4.46
5	480	16	0.03125	2.26
6	960	16	0.015625	1.115

Table 2.1: 5G numerology configurations (Release 17).

With respect to the radio frame structure, in 5G NR the number of subcarriers is 12 for all numerologies. Moreover, with the aim of maintain compatibility with LTE, the frame duration remains fixed with a duration of 10 ms and the frame is divided into 10 subframes with a duration of 1 ms. Each subframe is composed of slots, which will vary depending on the numerology selected. In fact, the number of slots per subframe is defined as 2^{μ} and the slot duration as $1/2^{\mu}$ ms. Finally, each slot is composed of 14 OFDM symbols, with a symbol duration of $1/(14 \cdot 2^{\mu})$ ms.

An overview of the 5G radio frame in the time domain with different numerologies when using Release 15 is depicted in Figure 2.8.

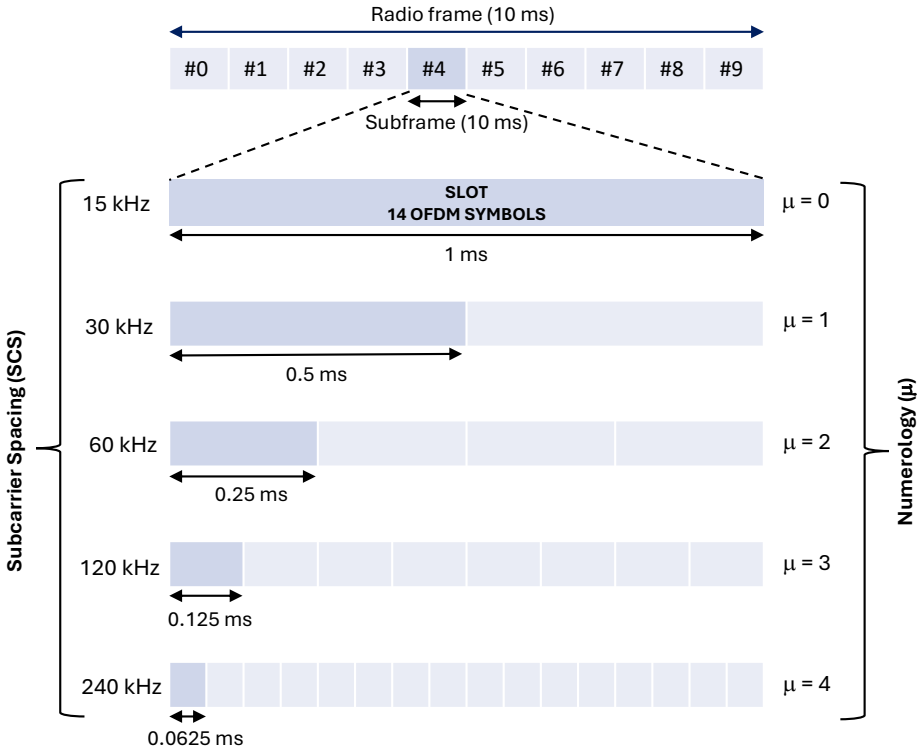


Figure 2.8: 5G numerologies scheme in the time domain (Release 15/16).

2.1.4 Cellular IoT

The 3GPP introduced in Release 13 two Machine Type Communications (MTC) solutions over cellular networks, commonly known as CIoT. In particular, these solutions were LTE for Machine Type Communications (LTE-M) and NarrowBand Internet of Things (NB-IoT), both based on LTE technology.

These technologies emerged to cover MTC use cases, characterized by low throughput requirements, support of massive connections, low power consumption, coverage

enhancements and low cost devices [83]. While LTE-M is intended for mid-range IoT applications with support of voice and video services, NB-IoT provides very deep coverage and support ultra-low-cost devices.

Although these technologies were introduced in Release 13, enhancements have been made in following releases to include new features. An important remark in Release 16 is that CIoT devices are allowed to connect to the 5GC by using a ng-eNB. This implies the support of 5G NAS messages and the 5G security framework, except data integrity protection.

The 3GPP has initiated Release 17 activities on NR Reduced Capability (RedCap) devices, also namely NR-Light [84]. The approach of NR RedCap devices is to address use cases (wearables, video surveillance, industrial IoT) in IoT with requirements that cannot be met using NB-IoT or LTE-M. This devices will offer lower cost, lower complexity, and longer battery life than NR eMBB and wider coverage than URLLC. In this thesis, the focus is set on LPWA technologies and therefore, a brief description of LTE-M and NB-IoT technologies is provided below.

LTE-M

As previously mentioned, LTE-M [85] was first introduced in Release 13 within a new UE category, namely Cat-M1. Cat-M1 enables a coverage enhancement of 15 dB with respect to LTE. Regarding its radio design, Cat-M1 operates like LTE but with a reduced radio frequency bandwidth of 1.08 MHz, which is equivalent to 6 PRBs. Cat-M1 was designed to operate with only one receive antenna, eight HARQ processes and a maximum Transport Block Size (TBS) of 1000 bits. Additional features from LTE that are supported in LTE-M are discontinuous reception, mobility, connection suspend/resume and data transmission via the CP.

LTE-M supports two Coverage Enhancement (CE) modes: CE mode A and CE mode B. Both CE modes enable coverage enhancement using repetition techniques for both data channels and control channels. CE mode A supports up to 32 repetitions, while CE Mode B supports up to 2048 repetitions. The default mode of operation for LTE-M is CE Mode A, which provides an efficient operation in coverage scenarios where moderate coverage enhancement is needed. On the other hand, CE Mode B is an optional extension which provides even further coverage enhancement at the expense of throughput and latency.

LTE-M category has evolved throughout following releases since its first definition.

In each release, enhancements in data rates, device capacities, and energy-efficient solutions have been integrated. In Release 14, support of high peak data rates, multicast transmission, voice enhancements and location services were introduced [86]. In addition, UE Cat-M2 was defined, which supports a radio frequency bandwidth of 5 MHz and higher data rates compared to Cat-M1. Moreover, the maximum TBS for uplink was increased to 2984 bits.

In Release 15, features such as increased spectral efficiency, sub-PRB resource allocation and transmission during the RA procedure were introduced to reduce the latency and power consumption [79]. In Release 16, 5G integration has been realized to share 5G capabilities, CE for non-bandwidth reduced low complexity, standalone development, and mobility improvements. Finally, Release 17 introduced a maximum downlink TBS of 1736 bits, 14 HARQ processes, and traffic management capabilities [87, 88].

NB-IoT

NB-IoT is a narrowband system which operates with a channel bandwidth of 180 kHz with support for multi-carrier operation [89]. NB-IoT is based on LTE technology, therefore, NB-IoT inherits part of its design such as channel codification, numerology, modulation scheme and higher protocols layers. Nevertheless, to reduce the complexity and the cost of these devices, some features are removed such as mobility in connected mode (handover).

NB-IoT supports three different operation modes:

- Stand-alone: using a dedicated carrier.
- In-band: using one PRB withing a normal LTE carrier.
- Guard-band: using unused resource blocks within an LTE carrier guard-band.

Regarding its radio design, in downlink, OFDMA is used with a SCS of 15 kHz over 12 subcarriers with 14 OFDM symbols. Same as LTE, the subframe duration is 1 ms. In the uplink, SC-FDMA is used, where two SCS are supported: 3.75 kHz and 15 kHz [90]. In addition, to reduce the PAPR in the uplink, the modulation is limited in the transmissions, where BPSK and QPSK schemes are adopted.

To operate with NB-IoT devices, only one antenna is necessary and one HARQ process is supported in Release 13 for uplink and downlink. Moreover, two classes of maximum output power are supported by NB-IoT devices: 20 and 23 dBm. In terms

of TBS, a maximum TBS size of 1000 bits for uplink and 680 bits for downlink is supported in Release 13.

NB-IoT supports a Maximum Coupling Loss (MCL) of 164 dB and uses the concept of repetitions and signal combining techniques to improve coverage extension [91]. To serve UEs with different coverage conditions, up to three CE configuration can be set in the network, and each UE will belong to a CE depending on its distance to the base station (see Figure 2.9). The CE is determined by the UE based on a Reference Signal Received Power (RSRP) threshold set by the network, where on each CE different transmission repetitions on physical channels, modulation and radio resources are used.

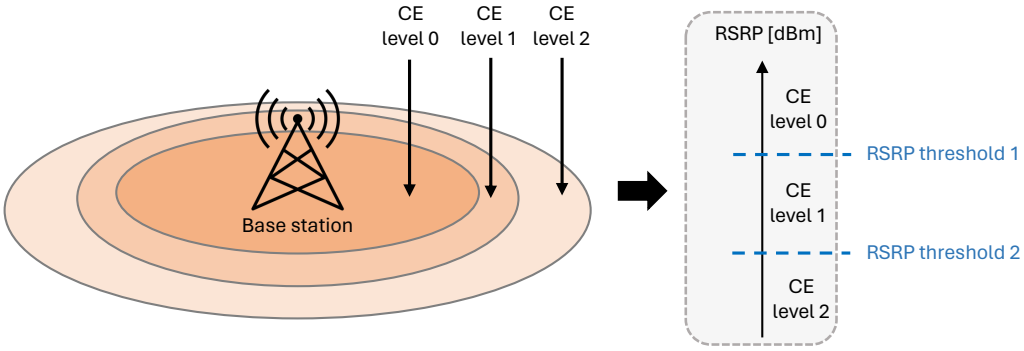


Figure 2.9: Relation between the CE levels and the RSRP thresholds.

Same as LTE-M, enhancements have been incorporated into NB-IoT which each subsequent release. The support of new bands, multicast transmission and positioning were introduced in Release 14. Furthermore, up to two HARQ processes are supported and the maximum TBS was enhanced to 2536 bits in the uplink and downlink [90]. A new NB-IoT category was also introduced, namely Cat NB2. Cat NB2 provides higher data rates and a new power class with a reduced output power of 14 dBm [86].

In Release 15, the focus was mainly set on enhancements on power consumption and latency reduction, with the introduction of data transmission during the RA procedure, Time Division Duplex (TDD) support and higher spectral efficiency [79]. Release 16 introduced coexistence with NR, improved energy and transmission efficiency, and scheduling enhancements. Finally, Release 17 enables data transmission in RRC Inactive state and introduced enhancements such as intra-UE multiplexing, positioning targeting factory automation, extended peak data rate, 16-QAM for uplink and downlink transmission, and time synchronization enhancements [87, 88]

Power saving techniques for CIoT

CIoT communications are usually characterized as sensors that transmit small data reporting the temperature, humidity, etc., with a low frequency. Due to the nature of these communications, it is important to ensure an efficient battery consumption, particularly while the devices are not transmitting any data, which is most of the time. This is quite important, since the International Telecommunication Union (ITU) and the 3GPP have defined a battery life requirement for CIoT devices in extreme coverage of beyond 10 years, with a desirable target of 15 years [92].

To address this, different power saving techniques have been introduced for CIoT devices (see Figure 2.10):

- Extended Discontinuous Reception (eDRX): defines a cycle where the UE monitors the Physical Downlink Control Channel (PDCCH) during a short period of time and sleeps the remaining time of the cycle. This mechanism is an extension of LTE Discontinuous Reception (DRX), where longer sleep periods are supported (DRX cycle is extended from 2.56 seconds to minutes or hours) [93]. The duration of the eDRX phase is defined by the active timer (T3324).
- Power Saving Mode (PSM): this feature was designed for CIoT devices to conserve more battery, where the UE enters in deep sleep mode. During the PSM, the device turns off its radio components completely, but maintains the registration in the network [94]. This means that there is no transmission or reception for any kind of channel or signal, and the UE is not reachable by the network. The advantage of this approach is that the UE can wake-up from PSM without reattaching the connection, thus, avoiding extra power consumption. The duration of the PSM is defined by the difference between the tracking area update timer (T3412) and the active timer (T3324) [95, 96].
- Release Assistance Indication (RAI): before the UE switches from RRC Connected state to RRC Idle state, it has to wait for receiving the RRC Release message from the network. If this message is not received, the UE has to wait until the expiry of an inactivity timer. To avoid this, the 3GPP introduced in Release 14 the RAI feature [97]. The RAI feature allows the UE to indicate to the network that it has no more uplink data or it does not expect to receive any data. This feature improves the battery by releasing the RRC connection without waiting for the inactivity timer expiration.

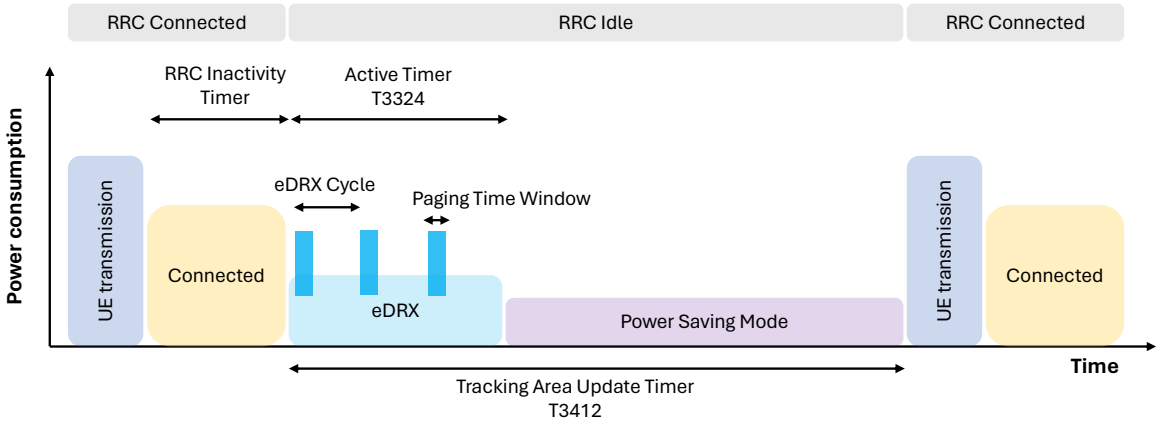


Figure 2.10: Overview of the RRC connection states and energy consumption with eDRX and PSM for a CIoT device.

CIoT signaling optimizations

In LTE and 5G, it is required to establish an RRC connection for the transmission of data from a UE to the network. This process is called Service Request (SR) and is shown in Figure 2.11. Since no RRC connection is active at the beginning, the first communication with the network is made using the RA procedure. The RA procedure consists of four steps: the preamble transmission (Msg1), the preamble response (Msg2), the connection establishment request (Msg3) and the connection establishment (Msg4). When receiving Msg4, the UE moves to RRC Connected state. After that, the AS security is configured. Once this process is finished, DRBs are created and the UE can transmit its data to the network. Finally, after an inactivity period, the UE receives a message from the network to release the connection and the UE returns to RRC Idle state. At the same time, DRBs and UE context are deleted in the CN. To further optimize this process for CIoT devices, two methods were introduced in Release 13 (see Figure 2.11): CP and UP CIoT optimizations.

The CP CIoT optimization consists in performing data transmission using the CP. The support of this mode is mandatory for NB-IoT devices and optional for LTE-M. In this case, data is encapsulated in NAS signaling messages that are sent to the CN. When using this procedure, the UE avoids the establishment of UP bearers and AS security each time it requires to send data.

The UP CIoT optimization is based on the concept of connection suspension and resume introduced in Release 13. This optimization requires a previous RRC connection establishment, where AS security and DRBs are created. Once this process is done, it is possible to suspend the connection and the UE moves to RRC Idle state. However,

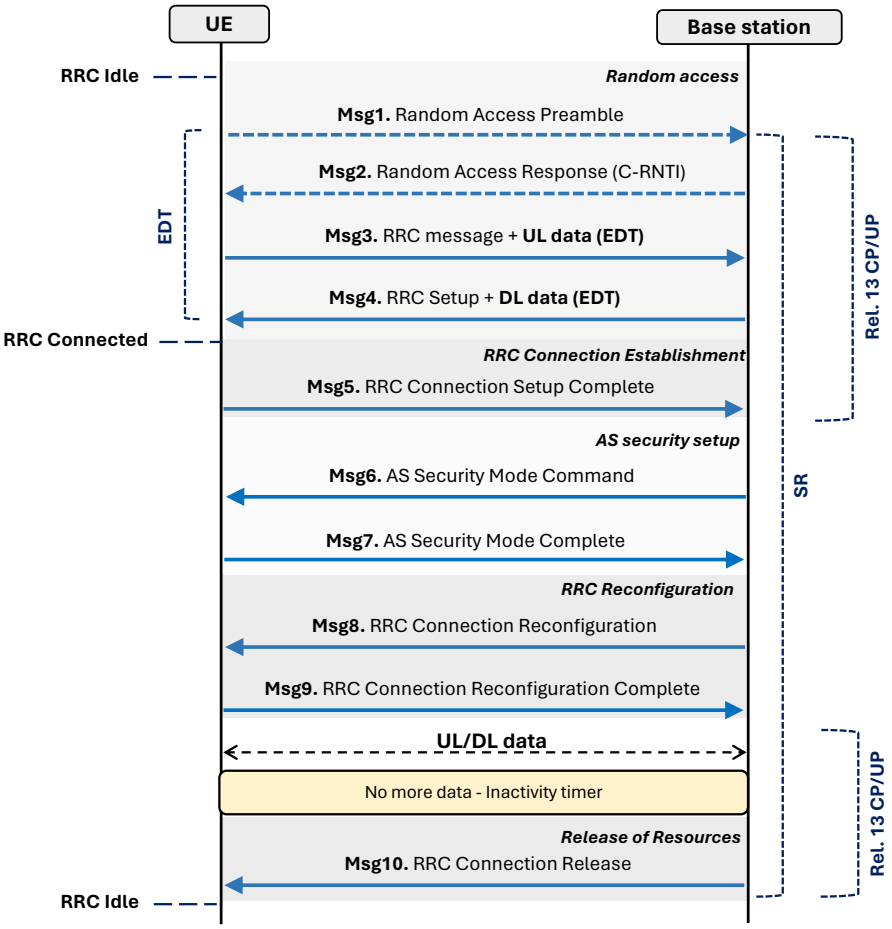


Figure 2.11: Signaling diagram of mobile originated data transport between the UE and the base station for SR, Release 13 CP/UP optimizations and EDT.

the suspension keeps the UE connection and security context in the entities involved (UE, base station and core). Therefore, when the UE needs to send data again, it can resume the previous context using for that a connection identifier provided in the suspension message. As the UE context is maintained in the network, it is not necessary to reconfigure the RRC connection with new DRBs and the AS security.

Although Release 13 CIoT optimizations reduce the signaling exchange between the UE and the network with respect to the SR procedure, a new mechanism was introduced in Release 15 to further reduce the latency and battery consumption of the UEs. This mechanism is known as EDT and is intended particularly for infrequent and small data transmissions. EDT allows the transmission of data during the RA procedure (see Figure 2.11) and is supported for the CP and UP. EDT was created to send uplink data in Msg3, without further need for the establishment of an RRC connection and a state change in the UE; significantly reducing both signaling and wake-up time in the UE. Moreover, the UE can also receive small data in Msg4 if

necessary.

To be able to use this optimization, a special preamble is used in Msg1, which lets the base station know that the UE has small data to transmit. Then, in Msg2, the base station returns a TBS, which indicates the maximum size of the Msg3 (RRC message and user data). For EDT, the maximum TBS allowed in Msg3 is 1000 bits, whereas the minimum is 328 bits [91]. On the other hand, the maximum TBS allowed in Msg4 is 680 bits. A more detailed description of EDT is provided in Chapter 6, along with a security analysis of this feature.

2.2 Multi-connectivity in 5G

Multi-connectivity consists in simultaneously establishing two or more links between the UE and the radio access nodes. In 5G, multi-connectivity inherits from two concepts introduced for LTE networks: Carrier Aggregation (CA) and Dual Connectivity (DC).

2.2.1 Carrier aggregation

CA was first introduced in Release 10 by the 3GPP for LTE networks [74]. In CA, two or more Component Carriers (CCs) are aggregated in order to support wider transmission bandwidths and thereby increase the bitrate. This aggregation is made from a single network node. Two types of CA are defined, which depend on the frequency of the aggregated CCs: (1) inter-band, and (2) intra-band with its two sub-modes, intra-band contiguous and intra-band non-contiguous. Inter-band means that the aggregated carriers are in different frequency bands while intra-band means that the aggregated CCs reside in the same frequency band. In the case of non-contiguous, however, the carriers are not co-located. The bandwidth of the aggregated CCs and the number of CCs used in downlink and uplink can be different, with a maximum of 16 CCs for 5G operation in both cases [80].

When CA is configured there are a number of serving cells, one for each CC. Although the different number of serving cells, the RRC connection is only handled by one cell, the Primary Cell (PCell), served by the Primary Component Carrier (PCC). The other CCs are all referred to as Secondary Component Carriers (SCCs), serving the Secondary Cells (SCells). The SCCs are added and removed as required, while the PCC is only changed in a handover procedure. A high-level diagram of CA with one UE that is served with two CCs in the same gNB is depicted in Figure 2.12.

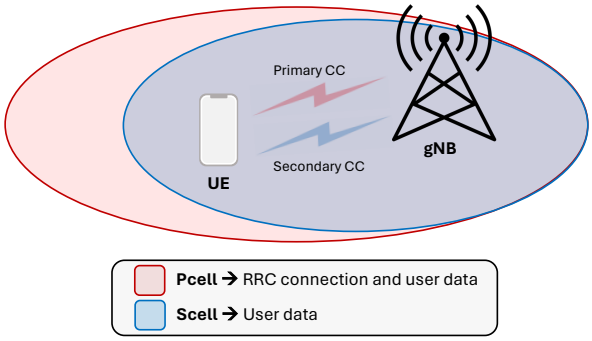


Figure 2.12: High-level CA diagram in 5G.

When using CA, the user traffic is split between the CCs in the MAC layer, and this layer must be able to handle scheduling on a number of CCs. For each serving cell, one HARQ entity is required. Also, one Transport Block (TB) is generated per TTI for each serving cell in the absence of spatial multiplexing [80]. Figure 2.13 shows the 5G layer-2 structure for downlink with CA configured and highlights the main changes.

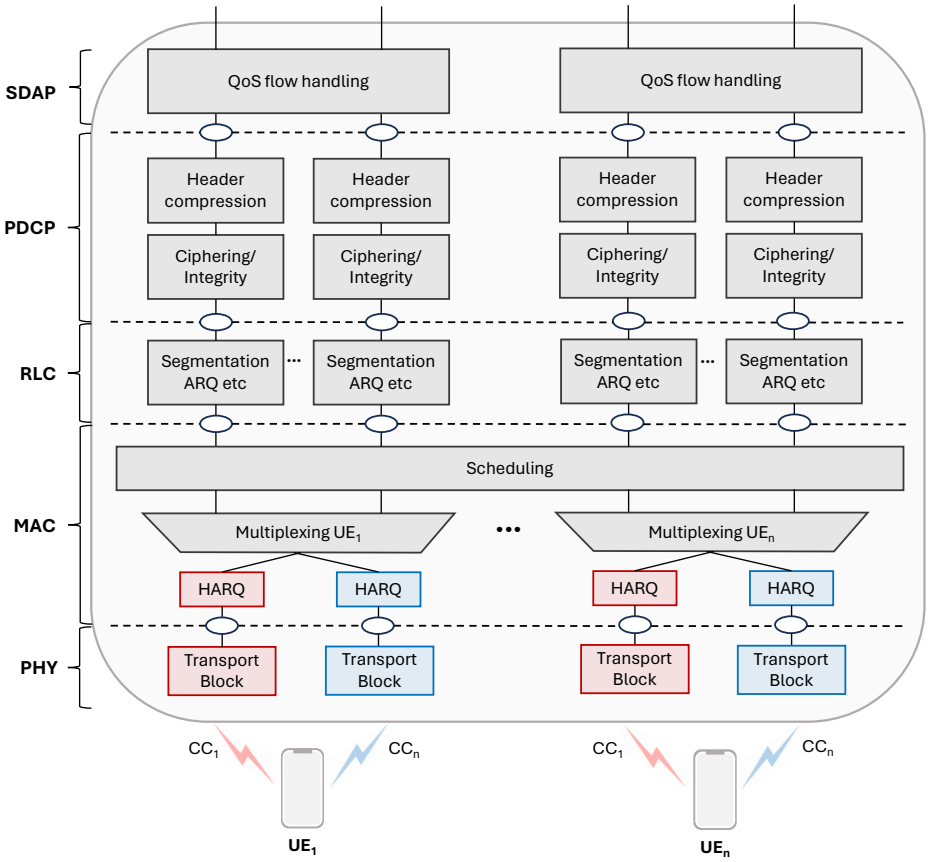


Figure 2.13: Layer-2 structure for downlink with CA configured in 5G.

2.2.2 Dual connectivity

DC feature was first introduced in Release 12 for LTE [74] and allowed UEs to simultaneously receive and send data from two eNBs that are connected via a non-ideal backhaul. In particular, DC allows to aggregate different CCs using two different network nodes, one acting as a Master eNB (MeNB) and the other one as a Secondary eNB (SeNB). The MeNB is in charge of the signaling between the E-UTRAN and the EPC, also managing the DC signaling. The DC signaling between the MeNB and the SeNB is performed via the X2 interface. This feature was first introduced to boost LTE throughput using different network nodes [98].

Release 15 introduced the support of DC with NR and LTE nodes as an extension of existing DC in LTE. In fact, this type of DC was specified for 5G NSA networks and it is known as Multi-Radio Dual Connectivity (MR-DC) [80]. In MR-DC, the UE is connected to one eNB that acts as a Master Node (MN), carrying the signaling between the UE and the EPC; and to a gNB that acts as a Secondary Node (SN). Moreover, in Release 16 DC was introduced for 5G SA deployments, namely as NR-NR DC. In NR-NR DC (see Figure 2.14), the UE is connected to two gNBs, one acting as a MN and another as a SN, and both connected via a non-ideal backhaul over the Xn interface.

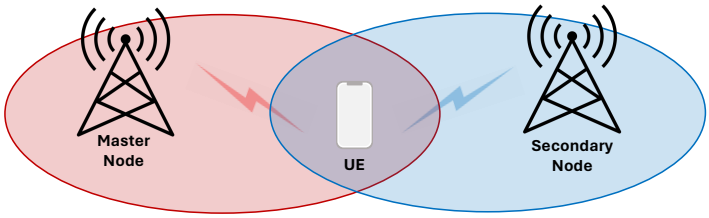


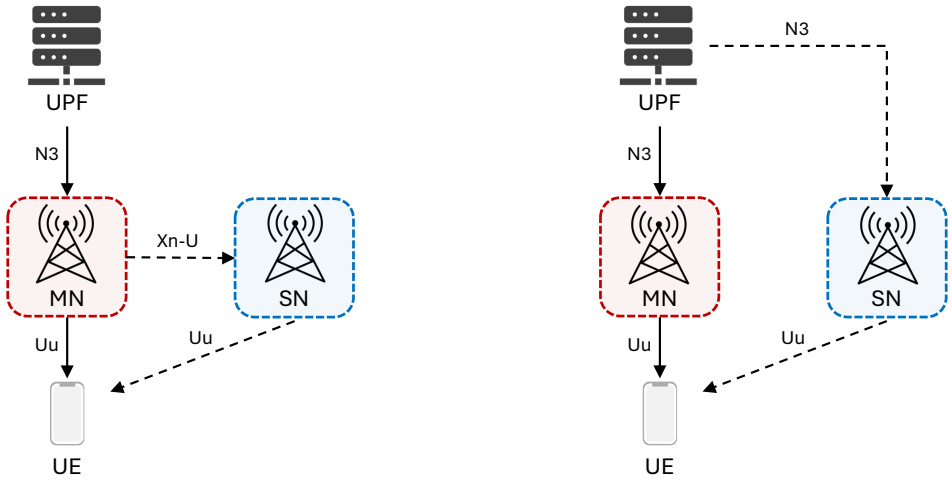
Figure 2.14: Dual Connectivity in 5G.

The 3GPP has defined two UP architectures for DC (see Figure 2.15). In the first architecture, the UP is split in the MN. When using this architecture, the UP data is transferred to the SN over the Xn-U interface. In the second architecture, both MN and SN have a UP connection to the UPF.

In DC, two different radio bearers exist:

1. Direct bearer: uses radio resources from one node. Direct bearers are divided into Master Cell Group (MCG) and Secondary Cell Group (SCG), depending on which node are located, MN or SN.

2. Split bearer: uses radio resources from both, MN and SN.



(a) UP split in the MN. (b) UP split in the CN.

Figure 2.15: 5G Dual Connectivity UP architectures.

Signaling Radio Bearer (SRB) are always configured as MCG, while DRBs can be configured as MCG, SCG or split bearers. Contrary to CA, in DC the user data is split at PDCP layer of the MN. Figure 2.16 shows the layer-2 structure for 5G DC in the downlink direction.

Packet duplication

One of the main changes introduced in the split in MR-DC with respect to LTE DC was the possibility of duplicating user data to further increase the reliability. This DC approach is known as Packet Duplication (PD). When PD is activated, the duplication is made in the PDCP entity of the MN. In the receiver, the PDCP entity is responsible for detecting and removing duplicated packets. Figure 2.17 depicts how the UP bearers can be split in a DC architecture, also including data duplication.

When using PD, the PDCP entity of the MN duplicates the PDU and adds the same sequence number in the PDCP header of both. This avoids performing twice functions such as ciphering, integrity and header compression. Then, the packet is sent from the MN to the SN via the Xn-U interface and the packet will undergo through independent RLC, MAC and Physical (PHY) layers as it can be seen in Figure 2.16.

Multiple copies of the packet are received on the receiver side, and the first successfully received packet is forwarded to the higher layers and the duplicated packets received later are discarded based on the PDCP sequence number.

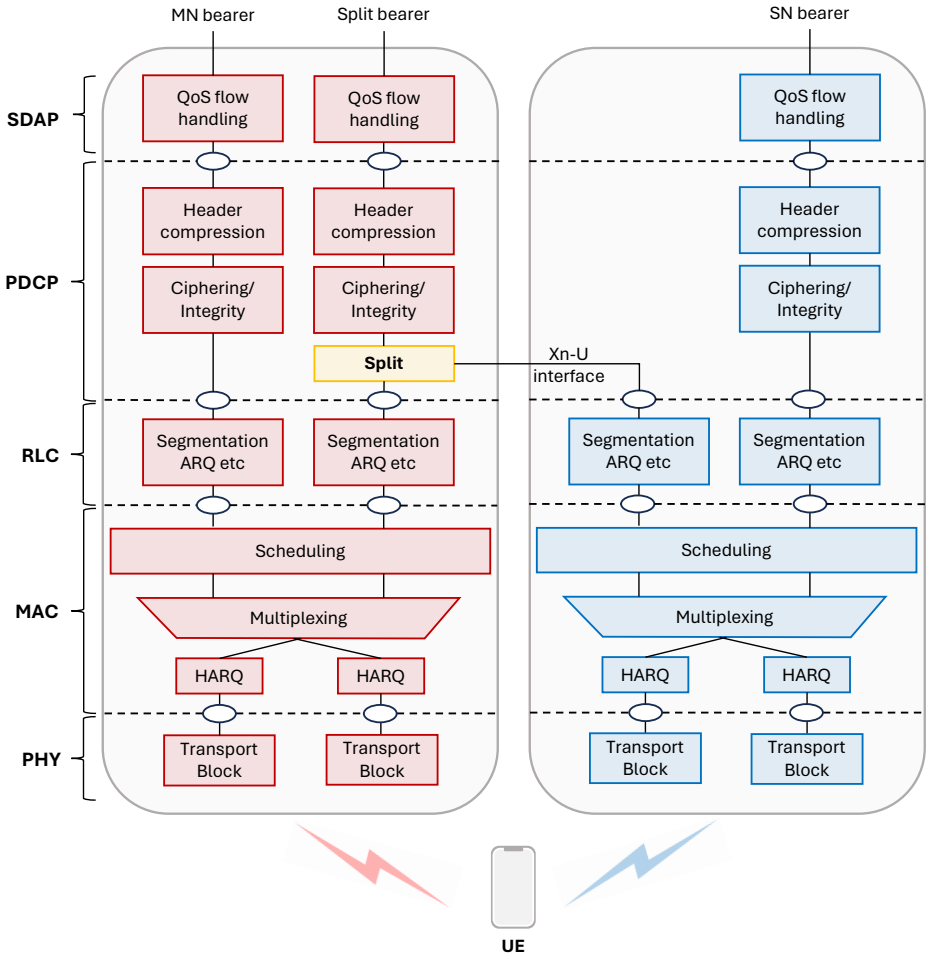


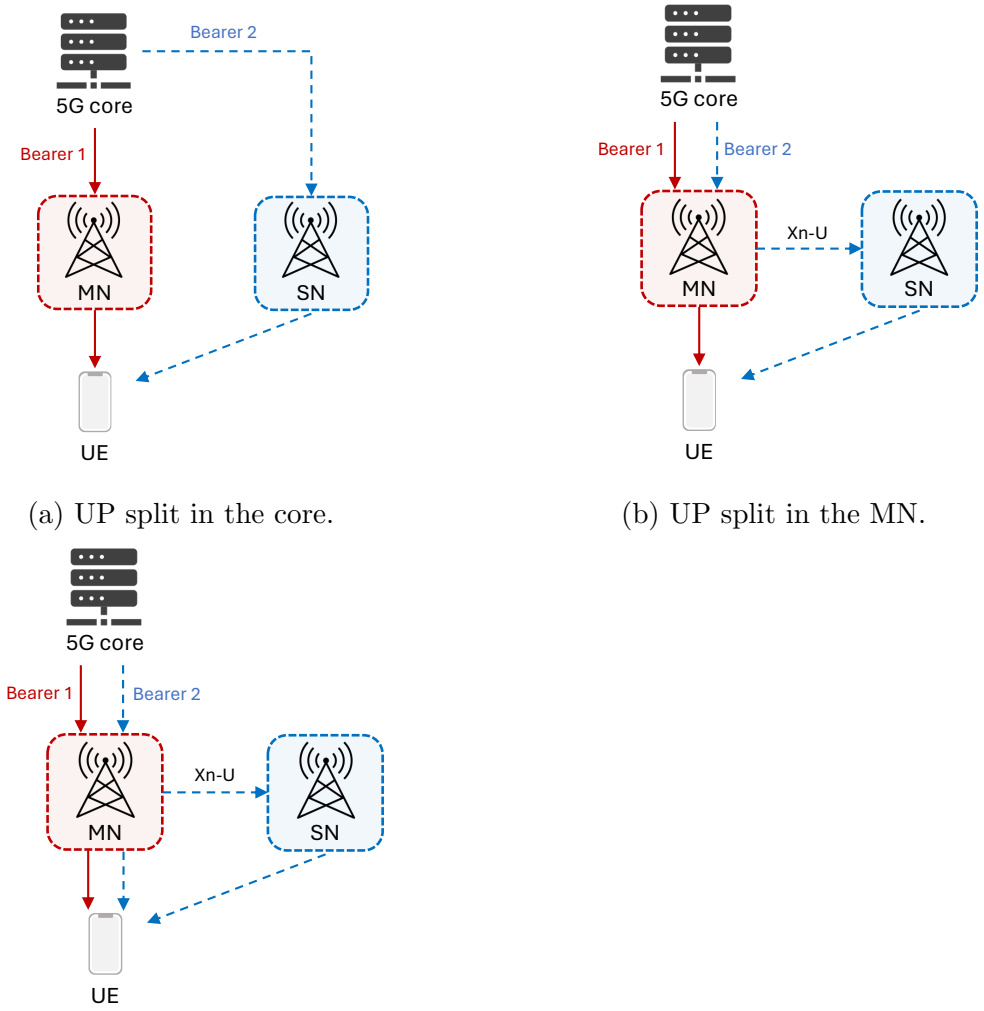
Figure 2.16: Layer-2 structure for 5G DC.

2.2.3 Multi-connectivity benefits and challenges

The use of multi-connectivity techniques offers different benefits, as demonstrated during the last few years by several research works, including [99]: (1) improved reliability, sending redundant data using different links, which also reduce the packet loss rate; (2) improved data rate, combining multiple data streams from different links into a single data stream; (3) service segregation, by segregating services with different requirements to different links; (4) mobility robustness, reducing the interruption time and the amount of signaling required.

Nevertheless, despite the benefits previously mentioned, some challenges are also present when using multi-connectivity [99]:

- Flow control: under- or over-utilized links might be created by an inadequate flow control logic, which results in a degradation of the service (i.e., out-of-order



(a) UP split in the core.

(b) UP split in the MN.

(c) Data duplication in the MN.

Figure 2.17: 5G DC UP bearers split architectures.

packet arrivals or poor overall system performance). The solution of this issue is to use a dynamic control of multi-connectivity that takes into account radio link conditions and radio resources instead of a static approach.

- Packet reordering: packets may arrive out of order due to different radio link conditions and communication path delays. To address this issue, the 3GPP has defined a reordering method for DC and MR-DC, which uses a static reordering timeout [100]. A special care should be taken in the decision of this timeout value, where aspects such as backhaul latency, radio link conditions, traffic type and QoS requirements should be considered.

- Multi-connectivity operation management: the decision of when to use multi-connectivity instead of single connectivity, the CCs involved or which base sta-

tions should be used is not trivial. Therefore, this decision is of a great importance since it has an impact on the overall system performance.

- Number of network nodes: only two network nodes are considered in the current standard. The use of a higher number of network nodes could provide more versatility for traffic aggregation if one of the nodes fails.

2.3 Security in 5G

2.3.1 Security architecture

The security architecture of 5G is divided into two domains [101]: the subscriber and the network domain. The subscriber domain is composed of the UE, while the network domain is composed of two elements, the home network and the serving network. Each of them contains different modules and subsystems, with the most important for the security aspects depicted in Figure 2.18.

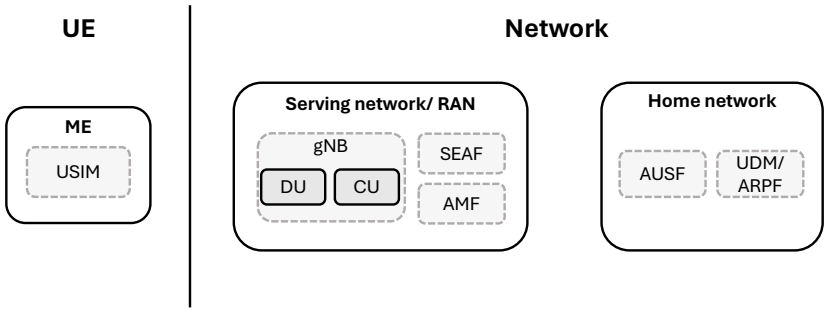


Figure 2.18: 5G security domains and submodules.

The UE contains the Mobile Equipment (ME) of the subscriber, and it is equipped with a Universal Subscriber Identity Module (USIM), which has cryptographic capabilities and stores the subscriber's credentials provided by the network operator.

The home network belongs to the subscribers' operator, manages subscriber information at the UDM and is in charge of verifying subscribers' authentication requests, using the Authentication credential Repository and Processing Function (ARPF) and the AUSF.

On the other hand, the serving network receives and stores the anchor key in the SEcurity Anchor Function (SEAF), and connects the UE with the home network, providing access to the UEs through the gNBs. It also manages the registration, mobility and reachability through the AMF. The gNB functionality is split into two

functional units: the Distributed Unit (DU), which contains the physical layer and the antenna; and the Central Unit (CU), which controls different DUs.

2.3.2 Security procedures between the UE and the 5G network

The security procedures between the UE and the 5G network are performed during the UE registration in the network, which allows the UE to transmit data if successfully registered. Before the UE being able to securely communicate, 5G requires an authentication process. This authentication is mandatory and is named primary authentication. The purpose of the primary authentication is to enable mutual authentication between the UE and the network and provide keying material that can be used between the UE and the serving network in subsequent security procedures [101]. For the primary authentication, the 3GPP proposes a novel Authentication and Key Agreement (AKA) protocol, namely 5G-AKA [102]. Alternatively, the previous EAP-AKA' from LTE can still be used [103]. While these two protocols share similarities, differences exist in the key derivation and the inclusion of new messages. The details of 5G-AKA protocol is described in detail in Chapter 6.

Once the primary authentication is done, the UE and the network share an anchor key called K_{SEAF}. From this anchor key, session keys for the communication between the subscriber and the home network are derived, as depicted in Figure 2.19. However, this authentication is implicit between the parties (UE, serving network and home network) according to the 3GPP. Therefore, upon successful completion of the primary authentication, a Security Mode Command (SMC) procedure is initiated by the AMF with the UE and at the end of this procedure, both are mutually authenticated.

Security Mode Command procedure

The SMC procedure is implemented for NAS and AS to establish the security of these domains, that is, the ciphering/integrity algorithms to be used and to derive the keys from K_{SEAF}. This procedure also checks the security capabilities of the UE to prevent bidding-down attacks [104].

The first one to execute after the primary authentication is the NAS SMC procedure (see Figure 2.20), in which the NAS security context is established between the UE and the AMF.

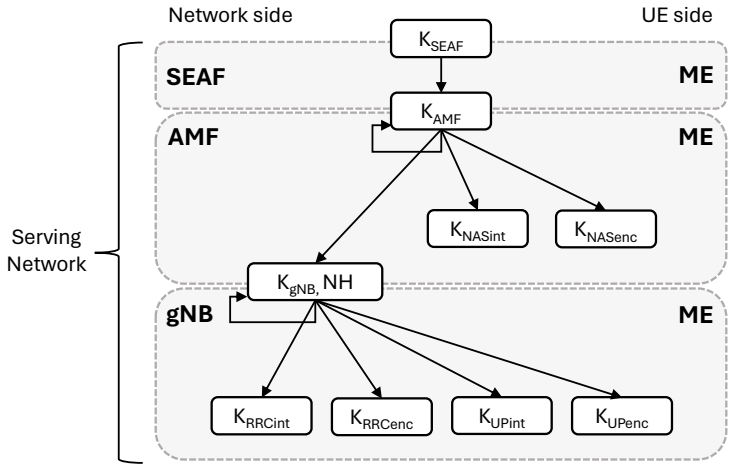


Figure 2.19: Key hierarchy generation.

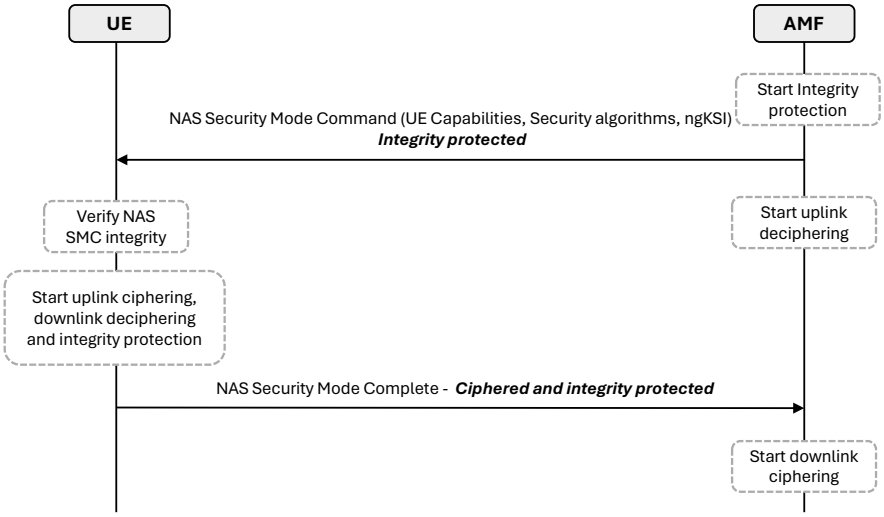


Figure 2.20: NAS Security Mode Command procedure.

The procedure consists of two messages and is initiated by the AMF. Before sending the first message, the AMF activates the NAS integrity protection. Then, the "NAS Security Mode Command" message is sent from the AMF to the UE. This message is integrity protected with K_{NASint} and contains the UE security capabilities (previously transmitted by the UE in the "NAS Registration Request" message), the selected NAS algorithms and the ngKSI for identifying the K_{AMF}. Upon reception of this message, the UE verifies its content. This includes checking that the UE security capabilities sent by the AMF match the ones stored in the UE to ensure that these were not modified by an attacker and verifying the integrity protection using the indicated NAS integrity algorithm and NAS integrity key based on the K_{AMF} indicated by the ngKSI. If the verification of the integrity of the message is successful, the UE starts NAS integrity protection and ciphering/deciphering with the security context indicated by the ngKSI

and sends the "NAS Security Mode Complete" message to the AMF ciphered and integrity protected with K_{NASenc} and K_{NASint}. Finally, the AMF deciphers and check the integrity protection of the "NAS Security Mode Complete" message using the key (K_{NASenc}, K_{NASint}) and algorithm indicated in the "NAS Security Mode Command" message and activates NAS downlink ciphering.

Once the NAS SMC procedure is successfully executed, the AS SMC procedure is triggered by the gNB (see Figure 2.21). Similar to the NAS SMC procedure, the AS SMC procedure aims to negotiate RRC and UP security algorithms and activate RRC security. First, the gNB sends the "AS Security Mode Command" message, which contains the selected RRC and UP ciphering and integrity algorithms and is integrity protected with K_{RRCint} (based on the current K_{gNB}). The UE then verifies the integrity of the message and if successful, starts RRC integrity protection and RRC downlink deciphering. Moreover, the UE sends the "AS Security Mode Complete" message to the gNB, which is integrity protected with the selected RRC algorithm indicated in previous message and the key K_{RRCint}. Finally, the gNB verifies the message sent by the UE and RRC uplink deciphering starts at the gNB.

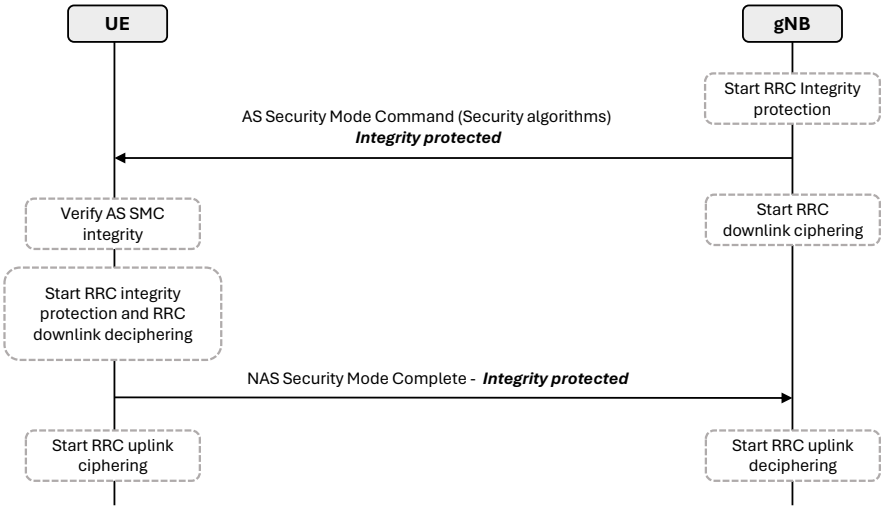
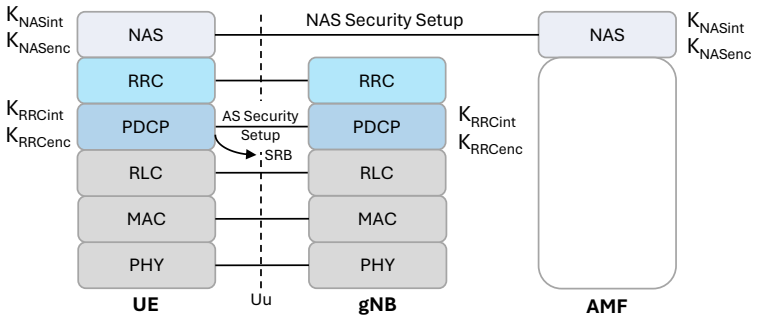


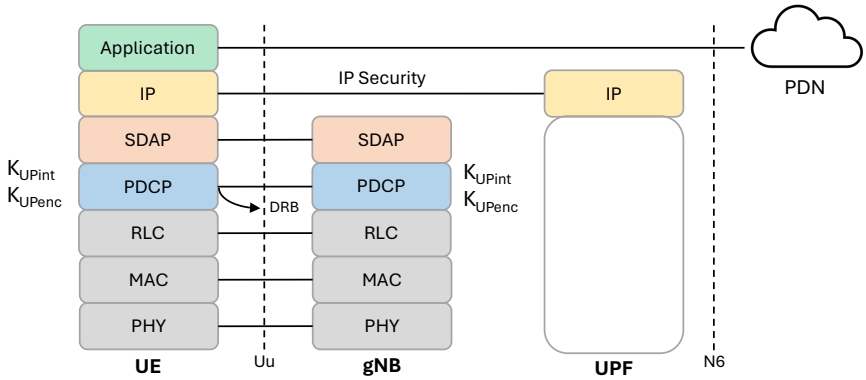
Figure 2.21: AS Security Mode Command procedure.

Ciphering and integrity protection of UP downlink and uplink, at the UE and the gNB, starts when configuring the DRBs, with the "RRC Connection Reconfiguration" and "RRC Connection Reconfiguration Complete" messages [101].

At the end of both procedures (NAS and AS SMC procedures), the UE and the network shares the CP and UP keys that are used to securely communicate the UE with the network and Figure 2.22 summarizes the layer where the key is used and the security contexts.



(a) Control Plane.



(b) User Plane.

Figure 2.22: Control and user plane keys and security contexts.

To summarize all the security procedures performed in the 5G registration, Figure 2.23 depicts the messages exchanged between the UE and the network for establishing a 5G communication. It comprises six phases:

1. The UE gets physical access to the gNB by using the RA procedure [105].
2. UE is authenticated with the network (primary authentication) using the 5G-AKA protocol and K_{SEAF} is derived.
3. NAS security context is created with NAS SMC procedure.
4. AS security context is created with AS SMC procedure.
5. RRC Connection Reconfiguration procedure is performed between the gNB and the UE to add DRBs and activate the UP security.
6. Finally, data exchanged in the network is ciphered and integrity protected using the keys derived in previous phases and the communication between the UE and network is secure.

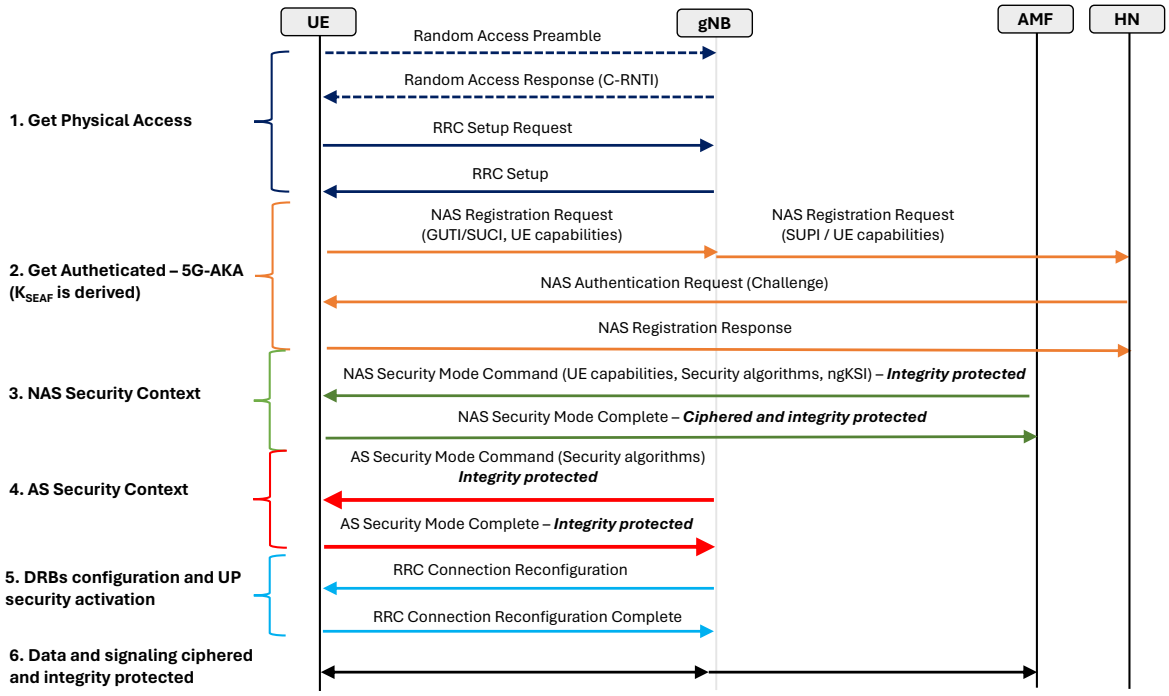


Figure 2.23: 5G registration and security initialization process.

2.3.3 Threat model and main attacks

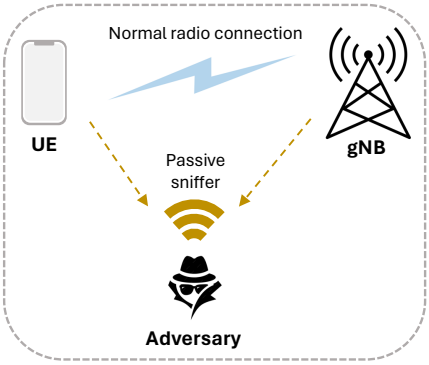
In general, threat models assume that the UE and the serving network are connected over an untrusted wireless channel, whereas serving network and home network communicate using a trusted channel [101]. Under this model, for the UE-Serving network channel, the ability of adversaries is usually modeled using the Dolev-Yao (DY) model [106]. In the DY model, the network is controlled by the adversary; where passive adversaries can eavesdrop on the communication and active adversaries can also intercept, inject, manipulate or drop messages. Thus, attacks on the radio interface can be classified into three different categories depending on the attacker capabilities [107] (see Figure 2.24):

- Passive attacker: plays an eavesdropper role that has the ability to receive, save and decode radio signals within a specific range and further extract concerned information without being noticed. Thus, a passive attacker can passively sniff the transmissions between the UE and the base station in the air interface.

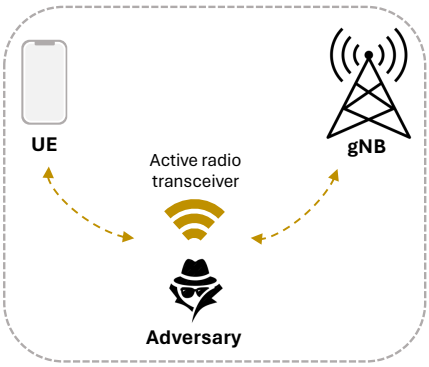
- Active attacker: in addition to the ability of a passive attacker, the active attacker can also send radio signals (e.g., spoofed signals or noise) into the open wireless channels. This type of adversary can launch radio jamming attacks, set up fake

base stations, or impersonate a UE towards the cellular network.

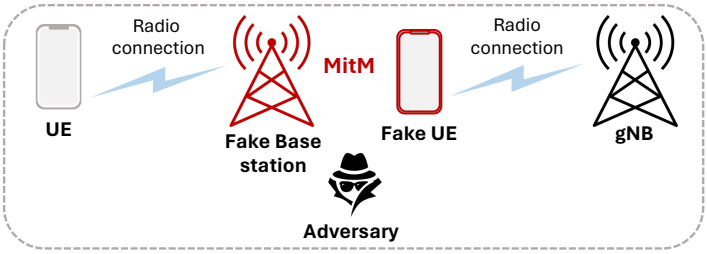
- Man-in-the-middle (MitM) attacker: it is considered as an online-version of the active attacker, where the attacker simultaneously impersonates a UE towards the network and a base station towards the UE. Therefore, the MitM attacker can establish and maintain an attacker-controlled relay transmission between the UE and the network.



(a) Passive attacker.



(b) Active attacker.



(c) MitM attacker.

Figure 2.24: Attacker models with different capabilities.

Depending on the aim of the attack, these can be classified into four main categories:

- Traceability: the adversary is able to determine the participation of a device in a specific communication and thus infer certain information about that device, i.e., location, type of exchanged information, communication frequency, etc.
- Impersonation: the adversary manages to impersonate one of the parties and communicate with the other on behalf of it.
- Denial/Degradation of Service (DoS): the adversary aims to compromise the availability of the system by interrupting temporarily or completely the service, or decreasing its performance.

- Bidding-down: the adversary tries to make UE and network entities believe that the other side does not support a security feature, even when both sides do support it. By indicating that a certain function, or a version of a function, is not supported, another function is used that may already have known vulnerabilities and exploits.

Although there are many attack variants depending on the attacker capability and aim of the attack [107], Table 2.2 summarizes the most important.

Attack	Description	Attacker	Type	Victim
Eavesdropping	An adversary could decode the essential UE information and network configuration details by sniffing the RAN	Passive	Traceability	UE/Network
RAN spoofing	An adversary is spoofing the RAN signals by transmitting a fake signal meant to pretend as an actual signal	Active	Impersonation	UE
Radio jamming	An adversary could disrupt the communication by deliberately jamming, blocking, or creating interference with the authorized wireless network	Active	DoS	UE/Network
Signaling storm	The adversary uses standard mechanism of the network CP to cause DoS, e.g., flooding the network with registration requests or the RA procedure	Active	DoS	Network
Replay attacks	The adversary first intercepts legitimate messages sent by one of the parties and later replays these messages with no or slight modifications to the other party	Active/ MitM	Impersonation/ DoS	UE/Network

Table 2.2: Summary of main existing threats and attacks against 5G network.

Part II

Publications

Chapter 3

Research outline

This chapter is structured in two sections. The first section describes the publications that support this thesis and associates them with the identified challenges and the thesis objectives. For each publication, their contributions to the state of the art are highlighted.

The second section presents the research methodology followed during the development of this thesis. This section also indicates the tools and equipment used in the research. For more details on the implementations made with these tools, refer to Appendix A.

3.1 Description of the publications

This section outlines the outcomes (research papers) arising from this thesis. These papers address the challenges identified and the objectives established in Section 1.2. Figure 3.1 illustrates the relationship between the challenges, the objectives and the corresponding outcomes. Each publication is represented as an individual block in the figure, indicating the chapter of this thesis in which it is included.

A brief summary of each of the papers that support this thesis is provided in the following subsections.

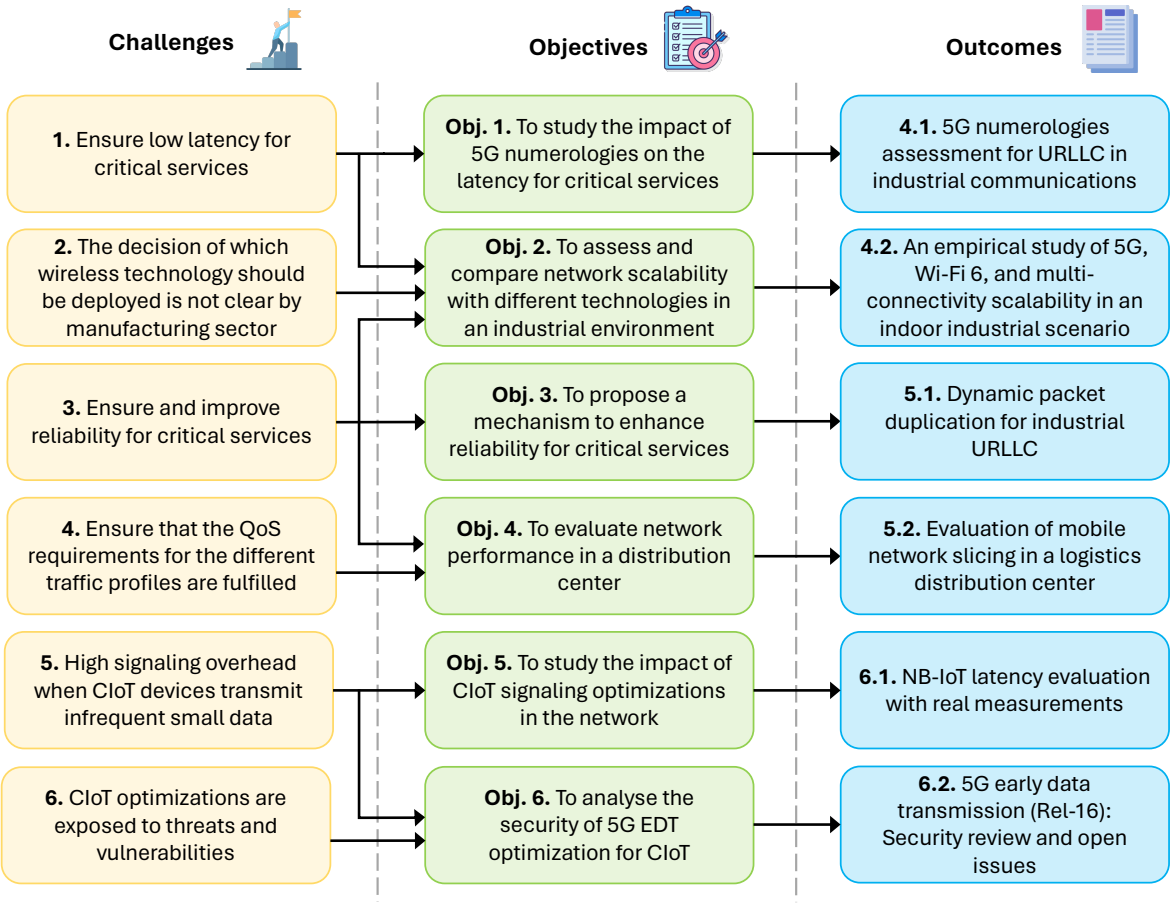


Figure 3.1: Challenges, objectives and outcomes.

3.1.1 5G numerologies assessment for URLLC in industrial communications

The advent of the 5G network has facilitated the introduction of novel features, enabling the development of new use cases and services. One of these features is the numerology, which allows a faster resource allocation process due to the use of shorter time slots. This feature is of particular importance for latency-constrained services such as those employed in the operation of AGVs, as it enables a reduction in the latency of their communications.

However, in industrial scenarios, the main challenge arises from the presence of concrete walls and large metallic machinery and structures, which can result in interference and multi-path propagation. Consequently, selecting an appropriate numerology is a challenging task, and it should be adapted to the radio conditions experienced.

Therefore, the first article presented in Chapter 4 is focused on assessing the impact of the numerology on the delay experienced at the radio link for a remote-control service

(AGVs communication), thus covering Obj. 1 of this thesis. More specifically, this study encompasses the assessment with varying packet sizes and channel conditions in a simulated factory environment, with a particular focus on identifying and analysing the outliers.

The results demonstrate that the assumption that a higher numerology leads to lower delay is not always true, particularly in NLOS conditions. In such cases, an intermediate numerology may be more suitable for this type of service.

3.1.2 An empirical study of 5G, Wi-Fi 6, and multi-connectivity scalability in an indoor industrial scenario

The manufacturing sector is adopting Industry 4.0 to enhance flexibility and reduce installation costs through the use of wireless connectivity. However, the question remains as to which wireless technology should be deployed in the factory to fulfil the requirements for next-generation applications such as Autonomous Mobile Robots (AMRs). While Wi-Fi technology is the most prevalent and easily deployed, the 5G network has been designed to support these industrial needs. It is therefore important to compare both technologies from a performance point of view, especially under different load conditions and with different number of devices. The use of multi-connectivity with different radio access technologies is also considered as a key enabler to fulfil the requirements of the most critical real-time applications.

Therefore, the second article presented in Chapter 4 is focused on the empirical assessment and comparison of the network scalability of 5G, Wi-Fi 6, and multi-connectivity in terms of latency and packet loss, thus covering Obj. 2 of this thesis. The work was carried out in the "5G Smart Production Lab" in Aalborg (Denmark), where different measurement campaigns were performed for different scenarios (static and mobility) and packet sizes.

The results obtained showed lower latencies with Wi-Fi in general, but large tails in the latency distribution, with a higher packet loss compared to 5G. On the other hand, 5G latency is very consistent with bounded tails, and low packet loss is obtained. With regard to scalability, 5G scales better than Wi-Fi, the latter being very affected by the number of devices transmitting data. Finally, the multi-connectivity solution showed an improved reliability and lower latencies in all evaluated cases.

3.1.3 Dynamic packet duplication for industrial URLLC

This work follows the line started with the first publication of Chapter 4. That is, when selecting an appropriate numerology to reduce the latency, the second step is to enhance the reliability for critical communications. One of the ways to improve the reliability of these communications is the use of multi-connectivity, particularly with the PD approach. Nevertheless, this solution comes at a cost in terms of redundancy, which can lead to an inappropriate use of network resources.

Therefore, to reduce the wastage of network resources, the first article of Chapter 5 proposes a dynamic PD algorithm based on ML, which determines whether PD is required at a specific data transmission to successfully send a critical message (Obj. 3). In particular, a latency estimator based on Random Forest (RF) was trained and evaluated, which decides when to duplicate a packet based on a latency threshold. The methodology presented was evaluated in a 5G simulator and the network performance was compared to different approaches: no duplication and a pure static PD.

The evaluation results demonstrated that the proposed dynamic PD algorithm reduced the number of duplicated packets sent by 81% while maintaining the same level of latency (i.e., the latency below the threshold) as a static PD technique. This reduction in the number of duplicated packets results in a more efficient usage of the network resources.

3.1.4 Evaluation of mobile network slicing in a logistics distribution center

The second article included in Chapter 5 addresses the problem of optimizing network resources for the different traffic profiles involved within a logistics distribution center scenario. In particular, these traffic profiles correspond to eMBB, URLLC, and mMTC, with distinct requirements in terms of latency, reliability, throughput, etc.

Specifically, this article first introduce a developed novel open-source simulator based on the ns-3 platform, with a realistic representation of a distribution center scenario, where different logistics activities are present. The communications of these activities have been modeled and used to estimate the performance of the different traffic profiles. As a result, the developed simulator serves as the foundation for evaluating the 5G network performance on smart logistics scenarios (Obj. 4).

Secondly, under the developed simulator, this work evaluates and compares the role of two 5G NS strategies in smart logistics: the use of a static slice with a balance division of network resources and the use of a dynamic slice that adapts the resources based on the traffic load, depending on the activity taken place. More specifically, this work evaluates these strategies in terms of QoS for the different traffic profiles, resulting in the following metrics: throughput for eMBB traffic, reliability for URLLC traffic, and the RA channel for mMTC traffic.

The results obtained show that a dynamic slice makes a more efficient usage of the network resources, improving the QoS for the different traffic profiles, even when there is a traffic peak on a specific profile. This improvement ranges from 6.48% to 95.65%, depending on the specific traffic profile and the evaluated metric.

3.1.5 NB-IoT latency evaluation with real measurements

Many optimizations have been proposed by the 3GPP for CIoT devices in order to improve the battery life and reduce the signaling exchange in the network. These optimizations started with the arrival of the Release 13, where the transmission via the CP was introduced. This optimization allowed to transmit data using the CP instead of the UP, thus avoiding the establishment of DRBs of the UP.

Moreover, with the arrival of Release 15, EDT optimization was introduced to support infrequent small data transmissions, supporting both the CP and UP transmission modes. The latter optimization allows the transmission of data during the RA procedure, with a significant reduction in the signaling exchange between the UE and the network, and without the need of an RRC state change (i.e., the UE transmits data in RRC Idle state).

Thus, the fist article of Chapter 6 is focused on the assessment and comparison of the aforementioned CIoT optimizations proposed by the 3GPP via the CP in terms of latency performance using the NB-IoT technology, covering Obj. 5 of this thesis. In particular, in this work a measurement campaign was performed with Amarisoft equipment (AMARI Crowdcell and AMARI UE Simbox) under different packet sizes and coverage levels.

The evaluation results showed lower latencies for EDT, particularly in the case of small packets, where a reduced TB is used, thereby being more efficient from a network perspective. Furthermore, it was demonstrated that EDT, in contrast to Release 13

optimization, fulfils the 3GPP latency requirement (10 seconds) for extreme coverage.

3.1.6 5G early data transmission (Rel-16): Security review and open issues

This section presents the second of the works carried out in relation to Chapter 6 of this thesis. In this case, this work extends the line started in the first publication of Chapter 6 by offering an in-depth description of the EDT optimization along with a security analysis of this mechanism. Thus, this work covers Obj. 6 of this thesis.

As mentioned above, EDT optimization was introduced in Release 15 to allow the transmission of data during the RA procedure. This optimization, intended particularly for infrequent and small data transmissions, aims to reduce the latency and the power consumption of the UEs. Nonetheless, despite the importance of this novelty and the general agreement about its effectiveness, there are few works in the literature that provide insight into its implementation and analyze the advantages and disadvantages of its two different implementation options (CP and UP).

Moreover, although security is recognized as a crucial aspect for the correct deployment of this technology, the literature lacks a review of the security issues and features of this mechanism. As a consequence of such a lack of works and the complexity of mobile network protocols, there is a divide between security experts and EDT researchers, that prevents the easy development of security schemes.

To overcome this important gap, this article offers a tutorial of EDT and its security, analyzing its main vulnerabilities and concluding with a set of recommendations for researchers and manufacturers. In particular, due to the simplifications in the protocols done by EDT, vulnerabilities such as packet injection, replay attacks and injection of fake values to disable EDT have been found.

3.2 Research methodology

The contributions reported in this thesis were conducted following a structured research methodology composed of different stages. Figure 3.2 depicts the different stages, which are described below.

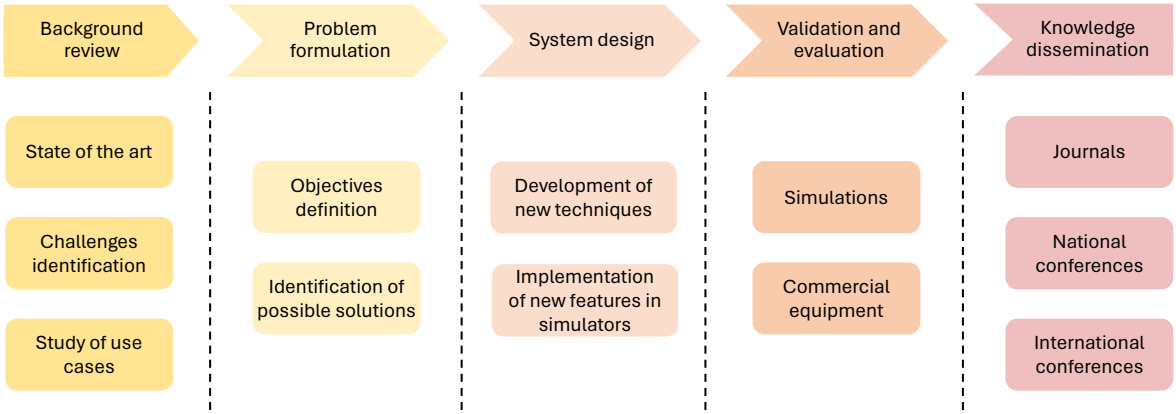


Figure 3.2: Research methodology.

1. **Background review**
   In the first stage of the research methodology, an exhaustive review of the background in the field of cellular network was performed, to clearly define which problems need to be solved. That is, the existing literature of the performance of cellular network focused on Industry 4.0 was reviewed. This resulted into the definition of the main challenges to be addressed and the study of different use cases in a factory.

2. **Problem formulation**
   In the second stage, the problem formulation is carried out for each challenge previously defined in the first stage. This stage comprises the definition of the objectives and the approaches to solve them.

3. **System design**
   The third stage of the methodology consisted in the development of the system and new techniques to overcome the challenges and objectives of the thesis, including the design and the implementation of new features in simulators.

4. **Validation and evaluation**
   Once the system has been designed, the proposed solutions and optimizations

were evaluated and validated either via simulations and with commercial equipment, in terms of network performance indicators:

- Simulations: For those works that require a controlled environment for the validation and evaluation of the features and methods designed, simulations were performed in the ns-3 simulator [108]. In particular, the 5G-LENA [109] module of the ns-3 simulator has been used in this thesis. 5G-LENA is an open-source module that provides a 5G NSA network and closely follows the 3GPP NR specifications, including features such as numerology support, frequency division multiplexing of numerology, beamforming, among others. Under the ns-3 framework and this module, many features were implemented focused on the particular environment of this thesis, which is the industrial scenario. Features such as the industrial channel and propagation loss in all its variants (3GPP 38.901) [110], the 5G DC feature with PD approach, slices with dedicated resources and assignation according to traffic requirements, a distribution center scenario with a realistic representation including its activities and applications, and the RRC Idle state, among others. A more in-depth detail of these contributions made to the simulator is provided in Appendix A. This resulted in a developed open-source simulator based on the ns-3 platform and the 5G-LENA module that can be found in [111].
- Commercial equipment: This thesis also evaluated the performance of the cellular network with different testbeds done with commercial equipment. In particular, Amarisoft equipment such as AMARI Callbox Classic [112] and AMARI UE Simbox [113] were used to evaluate the latency performance of CIoT signaling optimizations for NB-IoT under different radio conditions. On the other hand, measurement campaigns were performed in the "5G Smart Production Lab" [114] in Aalborg (Denmark), comparing the scalability performance of 5G, Wi-Fi 6, and multi-connectivity in terms of latency and packet loss in an indoor industrial scenario. These testbeds are described in detail in Appendix A.

During this phase it is necessary to analyze the results obtained from an statistical point of view. That is, to identify any unexpected effects that were not previously considered and, if necessary, make readjustment or reformulate the hypothesis. To this end, Python libraries and tools such as Scikit-learn [115–117], Pandas [118] or Numpy [119] have been used for data pre-processing.

5. **Knowledge dissemination**

   Finally, the most relevant results obtained during the thesis have been published in high impact journals and presented at national and international conferences.

Chapter 4

Performance evaluation

sensors

MDPI

Communication

5G Numerologies Assessment for URLLC in Industrial Communications

David Segura *, Emil J. Khatib, Jorge Munilla and Raquel Barco

Department of Communications Engineering, University of Malaga, 29071 Málaga, Spain; emil@uma.es (E.J.K.); munilla@ic.uma.es (J.M.); rbm@ic.uma.es (R.B.)

* Correspondence: dsr@ic.uma.es

Abstract: The fifth-generation (5G) network is presented as one of the main options for Industry 4.0 connectivity. Ultra-Reliable and Low Latency Communications (URLLC) is the 5G service category used by critical mechanisms, with a millisecond end-to-end delay and reduced probability of failure. 5G defines new numerologies, together with mini-slots for a faster scheduling. The main challenge of this is to select the appropriate numerology according to radio conditions. This fact is very important in industrial scenarios, where the fundamental problems are interference and multipath propagation, due to the presence of concrete walls and large metallic machinery and structures. Therefore, this paper is focused on analyzing the impact of the numerology selection on the delay experienced at radio link level for a remote-control service. The study, which has been carried out in a simulated cellular factory environment, has been performed for different packet sizes and channel conditions, focusing on outliers. Evaluation results show that not always a higher numerology, with a shorter slot duration, is appropriate for this type of service, particularly under Non-Line-of-Sight (NLOS) conditions.

Keywords: 5G; numerology; URLLC; Industry 4.0; Industrial IoT

Citation: Segura, D.; Khatib, E.J.; Munilla, J.; Barco, R. 5G Numerologies Assessment for URLLC in Industrial Communications. *Sensors* **2021**, *21*, 2489. https://doi.org/10.3390/s21072489

Academic Editors: Petros Bithas and Andrei Gurtov

Received: 11 February 2021
Accepted: 31 March 2021
Published: 3 April 2021

1. Introduction

Traditionally, wired connections have been used in industrial networks. These networks connect the programmable logic controllers (PLC), i.e., the computers that control the machines, with each other and with the manufacturing execution system (MES). The MES usually contains process monitoring software, as well as alarm monitoring, and constitutes the interface between the PLCs and the enterprise resource planning (ERP), which allows a global coordination at executive level. The irruption of wireless technologies in industry enables new applications, as well as lower installation costs. At this time, the second, third and fourth generation (2G, 3G, 4G) networks coexist in commercial deployments and can cover some of the industry needs in a basic way, although not at the scale required for the most advanced applications. The fifth-generation (5G) network is a wide area network (WAN) that supports all communication profiles that occur in industrial scenarios.

In the Industry 4.0 paradigm, agility is a key objective in the design of factories. Some of the main technologies that enable such agility in factories are the following: rearrangeable modules in production lines, automated guided vehicles, autonomous robots, connected worker solutions and even drones. All these applications are critical and have a common requirement: a low latency. There are other Industry 4.0 applications with different sets of requirements, such as a high bandwidth, low power consumption or very high reliability, but this paper focuses on the problem of latency.

In recent years, there has been a huge involvement of the 3rd Generation Partnership Project (3GPP) members to define the fifth-generation access technology of mobile networks, better known as 5G New Radio (NR) [1]. The main objective is to provide flexibility to be able to work with a wide variety of bands and different use cases. 5G defines three types of services according to their requirements:

- Enhanced Mobile BroadBand (eMBB): high speed connections (up to 20 Gbps) and high traffic density.
- Massive Machine-Type Communications (mMTC): presence of many machine-type devices that through sporadic connections, exchange short messages over the network. It focuses primarily on the Internet of Things (IoT).
- Ultra-Reliable and Low Latency Communications (URLLC): critical communications, short and with very restrictive needs in latency and reliability. In particular, in 5G it is expected to reach a maximum transmission time of 1 ms at the user plane with a packet loss rate of 10^{-5} for a packet size of 32 bytes [2].

There are several approaches followed to achieve such requirements for URLLC. One technique is the reduction of the time-slot duration by means of a higher numerology [3]. These new numerologies have been defined in 5G, which are determined by a SubCarrier Spacing (SCS) and a cyclic prefix. Another solution consists of eliminating steps in the connection protocols to reduce the access time, known as Grant-free transmission [4]. In Pedersen et al. [5], some changes in the radio resource scheduler are proposed to allow multiplexing of eMBB and URLLC services, allowing a latency below 1 millisecond in the case of URLLC. In Rao and Vrzic [6], the authors study packet duplication over independent radio links as a means of achieving high reliability and low latency. In order to provide these independent links, several carriers are used that may come from the same or different base stations, where the user will be connected simultaneously. This is the concept of multi-connectivity, resulting from extending 4G dual-connectivity functionality to more than two base stations. In Khatib et al. [7], the authors study multi-connectivity technique for URLLC and the cost in throughput for other services, as eMBB services. In Patriciello et al. [8], the impact of the end-to-end delay is studied based on the choice of a numerology under Line-of-Sight (LOS) conditions in an urban macro scenario. However, the impact of the channel condition on the delay is not evaluated.

The aim of this work is to analyze the impact of the 5G numerology selection on the delay experienced in the radio link, more specifically, at Packet Data Convergence Protocol (PDCP) layer for a remote-control service, which needs a low latency target. In contrast to the studies described above, which are centered only under LOS conditions, an evaluation of different numerologies under LOS and NLOS (Non-Line-of-Sight) conditions is included, the last case being very frequent in industrial scenarios, where the interference and multipath propagation increases. The hypothesis in this paper is that although in LOS conditions a higher numerology implies a lower delay, this may not be fulfilled under NLOS conditions given its lower robustness. This will be a very important consideration in the design of 5G-based communications systems for URLLC in industrial scenarios.

The remainder of this paper is organized as follows. A brief description of new numerologies is presented in Section 2. The simulator alongside the scenario and the metric to evaluate the numerologies is described in Section 3. Results are shown in Section 4. Finally, conclusions are drawn in Section 5.

2. 5G New Radio Access Technology

New Radio (NR) access technology is based on a flexible orthogonal frequency division multiplexing (OFDM) system, which allows operating in a wide range of bands, addressing different use cases and operating under multiple spectrum access [3]. Regarding the waveform, OFDM with cyclic prefix is used as the downlink waveform for NR. In contrast to Long-Term Evolution (LTE), OFDM can also be used in the NR uplink and Direct Fourier Transform spread OFDM (DFT-s-OFDM), the last one with the aim of minimizing the Peak-to-Average-Power-Ratio (PAPR) [9]. NR Release-15 allows frequencies up to 52.60 GHz, defining two frequency ranges (FR): FR1 (410 MHz–7.125 GHz) and FR2 (24.25 GHz–52.60 GHz). Higher frequencies are considered for Release 16, still undefined. The maximum available bandwidth per component carrier is limited to 400 MHz and the maximum number of component carriers is 16.

2.1. Numerology Concept and Frame Structure

The NR frame structure can adopt different numerologies. A numerology is defined by a SubCarrier Spacing (SCS) and a cyclic prefix (normal or extended) [9]. Numerology (μ) can take values from 0 to 4, defining the SCS as $15 \cdot 2^{\mu}$ kHz and the slot duration as $1/2^{\mu}$ ms, where high values of SCS are used at high frequencies. The maximum SCS value that can be reached is 240 kHz for $\mu = 4$. However, not all numerologies are suitable for a frequency range. In the case of synchronization (PSS, SSS, PBCH), $\mu = \{0, 1\}$ is used in FR1 and $\mu = \{3, 4\}$ in FR2. On the other hand, in the case of data channels (PDSCH, PUSCH, among others), only $\mu = \{0, 1, 2\}$ is supported in FR1 and $\mu = \{2, 3\}$ in FR2 [10]. The number of subcarriers in NR is 12 for all numerologies. In order to maintain compatibility with LTE, frame duration is fixed at 10 ms and subframe duration at 1 ms. The number of slots per subframe is defined as 2^{μ}, depending on the selected configuration. Therefore, as numerology increases, there are more slots available but with shorter duration. One slot is composed by 14 OFDM symbols, so the OFDM symbol duration is $1/(14 \cdot 2^{\mu})$ ms. Table 1 shows a summary of the characteristics for each numerology. An important remark is that $\mu = 0$ corresponds to the legacy LTE configuration.

Table 1. Numerologies defined in 5G.

μ	SCS (kHz)	Slots per Subframe	Slot Duration (ms)	Symbol Duration (µs)	Symbols per Slot
0	15	1	1	71.42	14
1	30	2	0.5	35.71	14
2	60	4	0.25	17.85	14
3	120	8	0.125	8.92	14
4	240	16	0.0625	4.46	14

2.2. Mini-Slots

The scheduler usually performs transmissions at slot level. NR Release-15 defines the possibility of transmitting only a portion of a slot, with minimum value of two and one OFDM symbol in downlink and uplink, respectively. This is known as a mini-slot. These very short transmissions are used in situations that require very low latency, such as URLLC services.

2.3. Bandwidth Part

Another feature included for NR is the Bandwidth Part (BWP) concept. BWP enable more flexibility in how resources are assigned in a given carrier. With BWP, a carrier can be subdivided and used for different purposes. Each BWP has its own parameters including bandwidth and numerology. Bandwidth is configured for each user equipment (UE) depending on its capabilities to support a maximum supported bandwidth and therefore, several UEs that have different capabilities can be served on a single broadband NR bearer. Moreover, multiple BWPs with different numerologies can be multiplexed within an NR bearer to support different types of services, as Figure 1 shows. Finally, an adaptation of the BWP based on changes between BWP with the same bandwidth and/or numerologies is also supported, with a single BWP being active at one time.

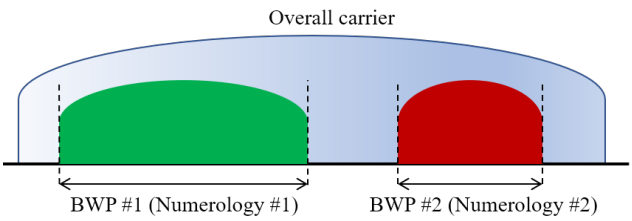


Figure 1. Frequency division multiplexing of numerologies.

3. Methodology

In this section, the simulator used to carry out the evaluation performance of the different 5G numerologies is presented. Furthermore, the scenario simulated is presented as well as the method and metric used to evaluate the delay.

3.1. Simulator

To simulate a 5G mobile network, ns-3 has been used, a free and open-source network software simulator, very popular in research [11]. In particular, to recreate 5G cellular communication, there are two extended modules, which are based on an evolution of the ns-3/LENA module for LTE networks [12]:

- Millimeter-wave (mmWave) module [13]: implements the full 3GPP protocol, where the physical and media access control (MAC) layers are own implementations to support a new mmWave-based channel along with beamforming techniques and antenna models. The MAC layer supports time division duplex (TDD), and the scheduler is based on time division multiple access (TDMA). The rest of the upper layers are based on the functionalities of the LTE module, but with extensions such as dual connectivity and low latency in the radio link control (RLC) layer. Finally, it should be noted that the frame structure is not-based at slot level.
- 5G-LENA module [14]: this module is based on the mmWave module, focusing on the new 3GPP NR specifications and includes numerology support, frequency division multiplexing of numerology and an OFDMA-based scheduler. Unlike the mmWave module, the frame structure in the time domain and the scheduler have slot granularity, adapted as indicated by the standard for each numerology.

For this study, the 5G-LENA module has been selected to carry out the simulations, due to the slot granularity in the frame structure, as explained above.

3.2. Simulation Scenario

The simulation scenario is shown in Figure 2. This scenario consists of a subsection of an indoor industry scenario. This subsection represents a stock storage area, where one automated guided vehicle (AGV) is moving to transport stock, as it is one of the main functions of AGVs. First, there is a remote host that is connected via a 100 Gb/s point-to-point connection to the Evolved Packet Core (EPC). This connection does not present propagation delay. Attending to the radio access network (RAN), a single 5G picocell with a height of 10 m is used, which will be shared by several users. There are six UEs connected to the picocell, where five of them generate background traffic to emulate a loaded cell environment. To do this, the remote host sends User Data Protocol (UDP) packet flows of 750 Mb/s for each one, with the aim of congesting the cell. These UEs have a fixed position and close to the next generation NodeB (gNB), therefore, they will have LOS conditions. On the other hand, the remaining UE has no fixed position, it moves and tries to emulate a remotely controlled AGV, which needs low latency. In this case, the remote host sends UDP packets with a periodicity of 100 ms and fixed size, as indicated in the simulation. Simulations have been made with packet sizes of 64 and 1000 bytes. These packet sizes represent two use cases when the remote host sends orders for controlling the AGV. The first one, with packet size of 64 bytes, corresponds to a packet that contains a unique order to the AGV. The second one, corresponds to a packet with several multiplexed orders. These packet sizes have been selected to evaluate the delay distribution obtained by each numerology.

To evaluate the NLOS condition, the AGV enters a room constructed of concrete with windows, with a height of 6 m. Inside, there are several concrete blocks with a height of 3 m that represent pallets and stock storage. These blocks are represented in Figure 2 as rectangles. In this figure, the movement of the AGV is also detailed, where t denotes the time in seconds during the simulation.

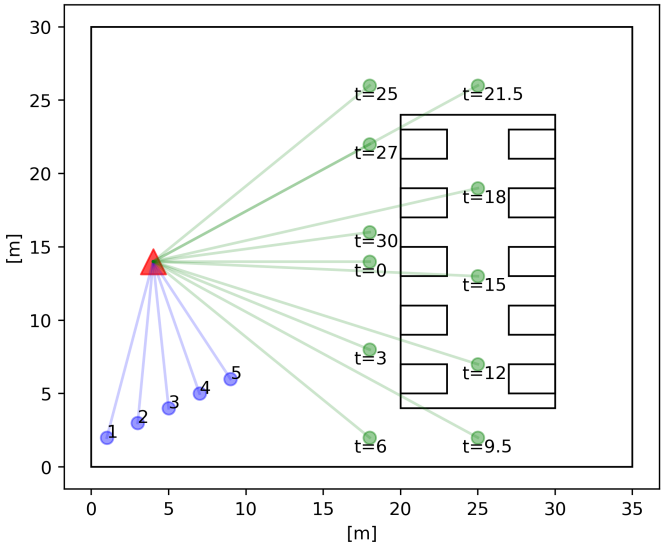


Figure 2. Simulation scenario. The red triangle represents the gNB position, the blue dots are the UEs that emulate the cell load, while the green dots represent the moving AGV. *t* denotes the time in seconds during the simulation, where the AGV is moving with a speed of 2 m/s.

3.3. Simulation Parameters

We compare NR numerologies, from 0 to 4, and analyze the UDP end-to-end delay at PDCP layer, for the AGV remote-control use case, under full load condition with different packet sizes and channel condition. This delay is measured from the instant the gNB sends the PDCP protocol data unit (PDU) to the RLC layer until this PDU is received in the UE at PDCP level. Once received, the delay is calculated using the timestamps attached in the packet header. The main configuration parameters of the simulations are shown in Table 2.

We repeat the same simulations using 40 different random seeds for each packet size and channel condition, in order to obtain statistically significant results. Then, we aggregate all the results with different seeds in a boxplot.

Table 2. Main configuration parameters.

Parameter	Value
Channel and propagation loss model	3GPP 38.900
Channel condition	LOS and NLOS
System Bandwidth	200 MHz
Center frequency	28 GHz
Scenario	Indoor
Transmission Direction	Downlink
Modulation	Adaptive
Scheduler	Round-Robin
UE height	1.5 m
gNB height	10 m

4. Evaluation Results

This section shows the results obtained for each numerology and packet size over several iterations. Although the standard defines the use of different numerologies for each FR, as mentioned in Section 2, this study will analyze all numerologies in FR2 and if its application within this range for URLLC services is feasible. The motivation for choosing FR2 for study is that millimeter-wave bands can potentially boost capacity, reduce latency and provide a higher bandwidth. The analysis will focus on outliers, to better understand the behavior of the tail in the distribution of latency. Since URLLC communications are critical, it is important to understand and be able to reduce these outliers.

4.1. Results with Packet Size of 64 Bytes

Figures 3 and 4 show the delay experienced by the packets when the AGV is under LOS and NLOS conditions, respectively. Under LOS conditions, it is observed that the higher μ is, the lower the delay. This was expected, since as μ increases, the slot duration is short, so the scheduling operation is faster. In this case, the median values for $\mu = \{0, 1, 2\}$ are 3.456, 1.778 and 0.956 ms, respectively. On the other hand, for $\mu = \{3, 4\}$ the median values are 0.536 and 0.336 ms, respectively. However, outliers exist for $\mu = 4$, reaching values above 10 ms.

Outliers are originated by two main factors. The first one is that as the AGV moves away from the gNB, the received signal-to-interference noise ratio (SINR) decreases, causing a more robust modulation selection. The selection of the modulation coding scheme (MCS) is done based on channel quality indicator (CQI) as shown in Figure 5. Upon a packet reception at the UE side, the UE measures the average SINR received for each packet chunk and then, based on this measure, the UE selects a CQI value, which is a scalar value from 0 to 15 that indicates how good or bad was the reception. Afterwards, the UE sends the CQI value to the gNB. When the gNB receives the CQI, it updates the MCS to be used in the next allocation for this UE according to CQI value. A robust MCS produces an increase in the delay, as OFDM symbols carry less bits, so it will be necessary more symbols to be allocated. The second one is related to the cell load level, where upon the arrival of a packet at MAC layer it may be the case of not having enough resources to allocate all the data in the current slot, having to wait for the next slot to allocate the rest of the data. The reason this occurs is because all traffic is treated fairly, i.e., there are no preferences for one traffic over another in the scheduler decision.

To check the effect of traffic background, we performed a simulation without that traffic that will help to understand why outliers occur with traffic background. Figures 6 and 7 show the delay distribution when there is no traffic background in LOS conditions for packet sizes of 64 and 1000 bytes. As it can be seen, a higher μ provides a lower latency in both cases. Also, outliers are clearly reduced, since they are originated only by the changes of the modulation, due to the interference and SINR decrease. On the one hand, with 64 bytes, a packet arrival at MAC layer will always have enough resources, since these are not shared with other users. Therefore, the impact of the scheduler in allocating the resources between the different traffics is higher than the fact of transmitting with a more robust modulation. On the other hand, with 1000 bytes, the modulation scheme selected will have a higher impact on the delay for $\mu = \{3, 4\}$, since the symbol duration is short and, if a robust modulation is selected, the OFDM symbol will carry less bits, so more symbols will be needed to allocate all the data. We do not repeat simulations without traffic background for the rest of the cases, since the trend is similar and this is an ideal case, cause in a real environment the network will not be empty.

Going back to Figure 3, $\mu = 3$ maintains the delay more stable, as 25% and 75% percentile are very close to the median. There is a remarkable asymmetry in the values for each numerology, i.e., the 25% percentile is very close to the median, contrary to the 75% percentile. As numerology decreases, this distance goes further. This indicates that the values above the median present higher variation and that the time waiting for the next slot allocation is higher as μ decreases, due to having a slot with a longer duration. Thus, the extra delay introduced by waiting for next slots allocation will affect more for lower numerologies.

In the case of NLOS conditions, in Figure 4, it is shown that the delay experienced in the numerologies suffers more alterations, increasing, due to propagation losses and signal reflections. This is reflected in the median values obtained for each numerology, higher than in LOS condition. Again, a higher μ implies a lower delay, although there are outliers for $\mu = 4$ between 10 and 15 ms. Moreover, it is observed that for $\mu = 3$ the values do not increase significantly, and they remain stable and low, as 25% and 75% percentile are close to the median. However, for the rest of numerologies, there are major changes in

the delay values. As numerology decreases, there is much more variation in the delay and, in contrast to LOS, the 25% and 75% percentile tends to be symmetric about the median.

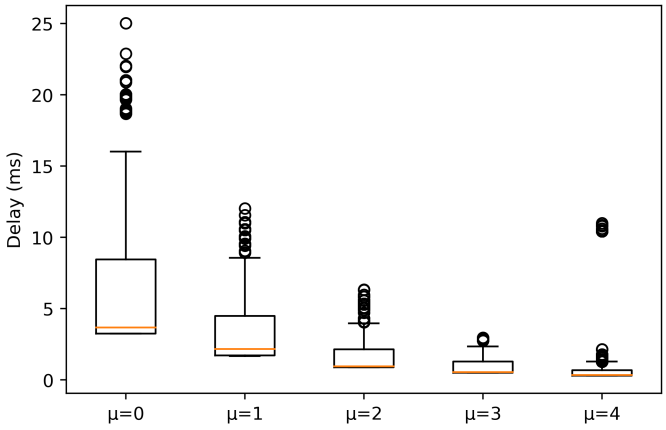


Figure 3. Experienced delay for packets with a size of 64 bytes in LOS conditions.

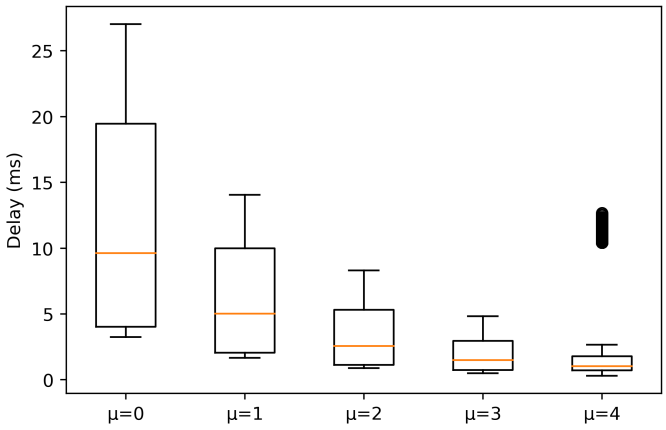


Figure 4. Experienced delay for packets with a size of 64 bytes in NLOS conditions.

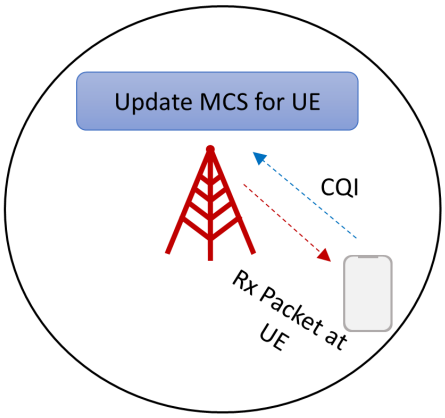


Figure 5. Modulation Coding Scheme (MCS) selection for downlink (DL) transmission.

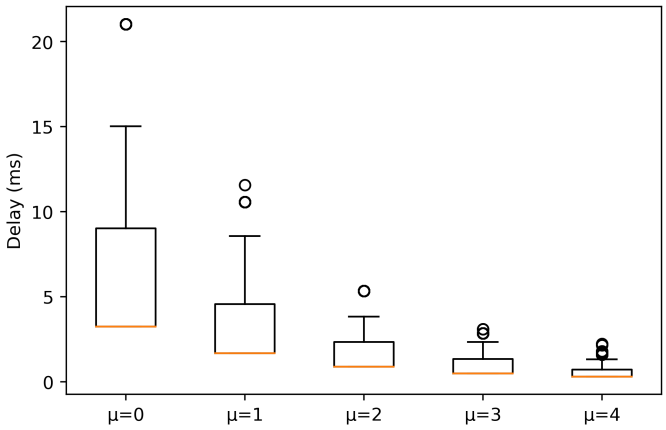


Figure 6. Experienced delay for packets with a size of 64 bytes in LOS conditions without background traffic.

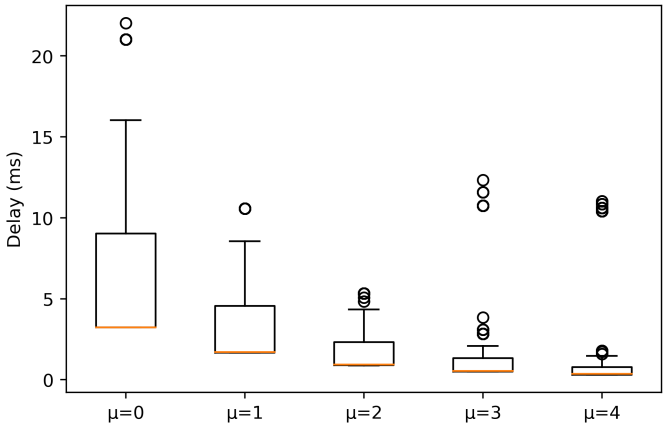


Figure 7. Experienced delay for packets with a size of 1000 bytes in LOS conditions without background traffic.

4.2. Results with Packet Size of 1000 Bytes

In the case of 1000 bytes packets size, the results obtained are shown in Figures 8 and 9. As it can be observed, in LOS conditions, the trend of the values is similar to the case of packets with a size of 64 bytes for $\mu = \{0, 1\}$, obtaining a median value of 3.669 and 2.169 ms, respectively. The main difference is that for $\mu = \{2, 3, 4\}$, as numerology increases, although the median value decreases, there are outliers that differ more from the 75% percentile. These outlier values have a higher impact on the delay as μ increases. With shorter slots, the packet information cannot be scheduled in a single slot, more slots are needed to allocate all the information. Also, the UEs with traffic flows of 750 Mb/s accentuate this delay, as they also need resources that cannot be allocated in one slot. Thus, for next slots allocation, this will occur again, increasing the system delay.

On the other hand, under NLOS conditions, significant differences are observed. A remarkable difference is a high increase in the median for all numerologies, being more accentuated for $\mu = \{3, 4\}$. Please note that for $\mu = 4$, the 25% and 75% percentile tends to be symmetric about the median. However, outliers exist, reaching values above 25 ms and below 5 ms. On the other hand, for $\mu = 3$, the 25% and 75% percentile tends to be asymmetric about the median. This clearly indicates that the data below the median present higher variation. Under this condition, $\mu = 2$ obtains a median value of 6.041 ms, although there are outliers between 25–30 ms. Also, $\mu = 1$ presents a similar behavior as $\mu = 2$, but with a higher median value (9.062 ms). Finally, for $\mu = 0$, it can be observed that the 25% and 75% percentile tends to be symmetric about the median and the median value is the highest. This indicates that this numerology is not suitable for AGV control, due to

the delay distribution, where the 25% percentile is around 10 ms, which is not desirable for URLLC.

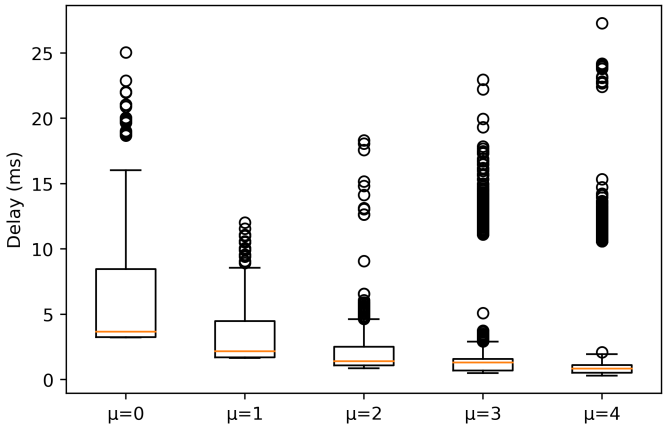


Figure 8. Experienced delay for packets with a size of 1000 bytes in LOS conditions.

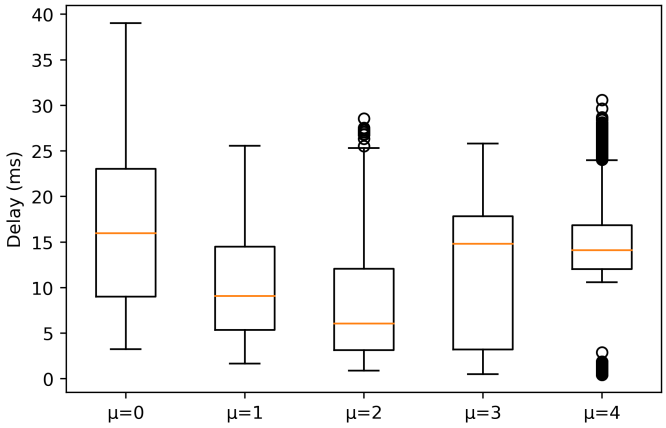


Figure 9. Experienced delay for packets with a size of 1000 bytes in NLOS conditions.

4.3. Result Discussion

On the one hand, it has been proven that with a higher μ in LOS conditions, the packet delay is lower, so shorter slots produce an apparent improvement, especially, with a small packet size. However, for $\mu = 4$ the delay seems to be more unstable than $\mu = 3$, due to a very short OFDM symbol duration. As adaptive modulation is used (the AGV distances itself from the gNB even though it has LOS), it may be the case that a more robust modulation is selected. The following occurs: if a more robust modulation is used, less information fits in a symbol and if that symbol has a very short duration in time (due to a very high numerology), more symbols are needed to be able to schedule all the packet information. Thus, the delay increases. The same happens when the packet size is increased for $\mu = \{2, 3, 4\}$, as more symbols are required to transmit the packet data.

On the other hand, under NLOS conditions, a higher μ is not always suitable. This conclusion should be taken into account when using it for URLLC. Detection of LOS/NLOS, could help to select the μ according to radio conditions. With a small packet size, an improvement in the delay at high μ values is achieved, with more stable values for $\mu = 3$ than $\mu = 4$. When the packet size increases, better results are obtained with $\mu = 2$ than $\mu = \{3, 4\}$. In this case it is observed that an intermediate value for μ is more efficient than a high value under a loaded cell condition. This reflects that there is a balance between throughput and delay purposes, cause the slot duration is reduced but not too much, so the delay will be reduced without affecting too much in terms of throughput, as one OFDM symbol is large enough to be able to transmit the data and the queue size will be reduced in the scheduler.

5. Conclusions

In this paper, a comparison of the different numerologies proposed in the 3GPP NR standard in an industrial scenario is presented. It has been proven that not always a higher numerology provides a lower delay; it will depend on packet size and channel conditions. When a low packet size is selected, the premise that with higher μ the delay is lower is fulfilled under LOS and NLOS conditions, except for $\mu = 4$ which presents outliers above 10 ms and a far distance from the median. This indicates that a very high slot time reduction cannot be suitable under high cell load conditions.

On the other hand, when the packet size increases, higher values of μ increase the delay. This is because when the slot is reduced, the information that can be scheduled in a single slot is also reduced, and the rest of the data must be allocated in other slots. This is important in industrial scenarios, where NLOS conditions are very common. Consequently, it will be necessary to complement the numerology selection with other mechanisms to service URLLC applications and reduce outliers, such as preemptive scheduling and resource reservation.

As a continuation of this work, multi-link connectivity will be investigated, which will provide higher reliability and, in certain cases, lower latency together with a good numerology selection.

Author Contributions: Conceptualization, D.S., E.J.K. and J.M.; methodology, D.S. and E.J.K.; software, D.S.; validation, D.S. and E.J.K.; formal analysis, D.S., E.J.K. and J.M.; investigation, D.S. and E.J.K.; writing—original draft preparation, D.S.; writing—review and editing, D.S., E.J.K. and R.B.; visualization, D.S.; supervision, R.B.; funding acquisition, R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially funded by Junta de Andalucía (projects EDEL4.0:UMA18-FEDERJA-172 and PENTA:PY18-4647).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

4G	Fourth generation
5G	Fifth generation
3GPP	Third Generation Partnership Project
AGV	Automated Guided Vehicle
BWP	Bandwidth Part
CQI	Channel Quality Indicator
DFT-s-OFDM	Direct Fourier Transform spread OFDM
eMBB	Enhanced Mobile BroadBand
EPC	Evolved Packet Core
ERP	Enterprise Resource Planning
FR	Frequency Range
gNB	Next generation NodeB
IoT	Internet of Things
LOS	Line-of-Sight
LTE	Long-Term Evolution

MAC	Media access control
MCS	Modulation Coding Scheme
MES	Manufacturing Execution System
mMTC	Massive Machine-Type Communications
mmWave	Millimeter-wave
NLOS	Non-Line-of-Sight
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
PAPR	Peak-to-Average-Power-Ratio
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PLC	Programmable Logic Controller
RAN	Radio Access Network
RLC	Radio Link Control
SCS	SubCarrier Spacing
SINR	Signal-to-interference noise ratio
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
UDP	User Datagram Protocol
UE	User Equipment
URLLC	Ultra-Reliable and Low Latency Communications
WAN	Wide Area Network

References

1. 3GPP. TR 38.912, Study on New Radio (NR) Access Technology. V15.0.0, Rel-15. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3059 (accessed on 18 January 2021).
2. 3GPP. TR 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies. V14.3.0, Rel-14. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2996 (accessed on 18 January 2021).
3. Zaidi, A.A.; Baldemair, R.; Tullberg, H.; Bjorkegren, H.; Sundstrom, L.; Medbo, J.; Kilinc, C.; Da Silva, I. Waveform and Numerology to Support 5G Services and Requirements. *IEEE Commun. Mag.* **2016**, *54*, 90–98. [CrossRef]
4. Jacobsen, T.; Abreu, R.; Berardinelli, G.; Pedersen, K.; Mogensen, P.; Kovacs, I.Z.; Madsen, T.K. System Level Analysis of Uplink Grant-Free Transmission for URLLC. In Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 4–8 December 2017; pp. 1–6.
5. Pedersen, K.; Pocovi, G.; Steiner, J.; Maeder, A. Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations. *IEEE Commun. Mag.* **2018**, *56*, 210–217. [CrossRef]
6. Rao, J.; Vrzic, S. Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance Analysis. *IEEE Netw.* **2018**, *32*, 32–40. [CrossRef]
7. Khatib, E.J.; Wassie, D.A.; Berardinelli, G.; Rodriguez, I.; Mogensen, P. Multi-Connectivity for Ultra-Reliable Communication in Industrial Scenarios. In Proceedings of the 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, 28 April–1 May 2019; pp. 1–6.
8. Patriciello, N.; Lagen, S.; Giupponi, L.; Bojovic, B. 5G New Radio Numerologies and their Impact on the End-To-End Latency. In Proceedings of the 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Barcelona, Spain, 17–19 September 2018; pp. 1–6.
9. 3GPP. TR 21.915, Release Description; Release 15. V15.0.0, Rel-15. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3389 (accessed on 18 January 2021).
10. 3GPP. TS 38.300, NR; NR and NG-RAN Overall Description. Stage-2. V15.11.0, Rel-15. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191 (accessed on 18 January 2021).
11. NS-3-A Discrete-Event Network Simulator for Internet Systems. Available online: https://www.nsnam.org/ (accessed on 18 January 2021).
12. LENA (LTE-EPC Network Simulator) Module for ns-3. Available online: https://www.nsnam.org/docs/models/html/lte-design.html (accessed on 18 January 2021).
13. Mezzavilla, M.; Zhang, M.; Polese, M.; Ford, R.; Dutta, S.; Rangan, S.; Zorzi, M. End-to-End Simulation of 5G mmWave Networks. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2237–2263. [CrossRef]
14. Patriciello, N.; Lagen, S.; Bojovic, B.; Giupponi, L. An E2E simulator for 5G NR networks. *Simul. Model. Pract. Theory* **2019**, *96*, 101933. [CrossRef]

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 May 2024, accepted 20 May 2024, date of publication 23 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3404870

RESEARCH ARTICLE

An Empirical Study of 5G, Wi-Fi 6, and Multi-Connectivity Scalability in an Indoor Industrial Scenario

DAVID SEGURA[1], SEBASTIAN BRO DAMSGAARD[2], AKIF KABACI[2], PREBEN MOGENSEN[2], EMIL J. KHATIB[1], (Member, IEEE), AND RAQUEL BARCO[1]

[1]Telecommunication Research Institute (TELMA), University of Málaga, 29010 Málaga, Spain
[2]Wireless Communication Networks Section, Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark

Corresponding author: David Segura (dsr@ic.uma.es)

This work was supported in part by the European Union-NextGenerationEU within the framework of the project Massive AI for the Open RadIo b5G/6G Network (MAORI), in part by the II Plan Propio de Investigación y Transferencia de la Universidad de Málaga, and in part by the Danish Innovation Fond as part of the 5G-ROBOT Grand Solution Project.

ABSTRACT Industry 4.0 is being adopted by the manufacturing sector to improve the flexibility and reduce installation costs by the use of wireless connectivity. There is an open question of which wireless technology deployment should be used in the factory to fulfil the requirements for next-generation applications such as autonomous mobile robots. Wi-Fi technology is the most extended and easy to deploy, while the fifth generation of mobile networks (5G) has been designed to support these industrial needs. Therefore, it is important to compare both technologies from a performance point of view, especially under different load conditions and number of devices. The use of multi-connectivity between 5G and Wi-Fi can also be an option to fulfil the requirements for the most critical real-time applications. In this paper, we empirically measure the scalability of 5G, Wi-Fi and multi-connectivity in the ''Aalborg University 5G Smart Production Lab'' and compare them in terms of latency and packet loss with different packet sizes. We found that in general Wi-Fi obtains lower latencies but large tails in the distribution, with a higher packet loss compared to 5G. On the other hand, 5G latency is very consistent with bounded tails, and low packet loss is obtained. In terms of scalability, 5G scales better than Wi-Fi, the latter being very affected by the number of devices transmitting data. Finally, multi-connectivity showed an improved reliability and lower latencies in all evaluated cases.

INDEX TERMS 5G, Wi-Fi, Industry 4.0, multi-connectivity, latency, scalability, packet loss.

I. INTRODUCTION

Currently, the industrial sector is facing its fourth revolution known as Industry 4.0 [1]. This new era aims to improve the efficiency and productivity of the factories with the use of novel technologies such as Artificial Intelligence (AI), Big Data, cyber-physical systems (CPS), and the Internet of Things (IoT). Industry 4.0 is characterized by the interconnection of numerous machines involved in manufacturing to collect data, control the production and manage the machinery. One important step of Industry 4.0 is to establish reliable and ubiquitous stationary and mobile networks for this type of communications, especially for the most critical applications involved in the factory.

Traditionally, wired connections have been used in industrial networks to connect different elements such as Programmable Logic Controllers (PLC), due to their reliability and determinism. However, wired communications are costly in terms of installation and maintenance and cannot cover new use cases, such as mobility in factories. Moreover, Industry 4.0 focuses on flexibility, re-configurable modules and the use of Autonomous Mobile Robots (AMRs) [2]. As a result, the industrial sector is starting to adopt wireless networks such as Wi-Fi and the fifth generation of mobile networks (5G) to achieve automation and flexibility on the factories [3]. In 2023, wireless deployments have experienced

The associate editor coordinating the review of this manuscript and approving it for publication was Adao Silva.

a growth of 22% [4] and accounts for 8% of new connected devices in industry. Although Wi-Fi is still the most extended wireless technology in factories, due to its simplicity and easy deployment, factories can also take advantage of cellular networks. With the arrival of 5G, one of the main focus on the design of this technology has been to support industrial communications requirements. This can be achieved thanks to the handover and Quality of Service (QoS) support, and the use of new 5G features such as numerology [5], network slicing [6] or Packet Duplication (PD) [7].

The use of multi-connectivity [8] between different technologies can also be an option to fulfil the requirements for the most critical real-time applications in the factories by improving the reliability and reducing the latency for the users.

Many of the applications present in Industry 4.0 have very strict requirements in terms of QoS. Therefore, there is a need for an assessment of the performance of the main network access technologies and commercial equipment present in such scenarios, with a special focus on applications such as AMRs and PLCs, where critical communications often take place. The aim of this paper is to compare the performance of 5G, Wi-Fi and multi-connectivity in an indoor industrial scenario. For this, we empirically measure the scalability performance of these technologies and multi-connectivity in the "Aalborg University 5G Smart Production Lab" [2] and compare them in terms of high percentile round-trip time (RTT) latency and packet loss. Moreover, two different scenarios have been considered, one with stationary terminals and another one with mobile terminals with different packet sizes.

We expect that the results of all of these measurements will provide a global vision of which technology suits better the manufacturing sector, depending on the type of applications and use cases involved in their factories.

The remainder of this paper is organized as follows. In Section II, an overview of the related works assessing the network in an industrial scenario is presented. Section III explains the methodology along with the scenario, setup and metrics to evaluate the performance of the network. Results are shown in Section IV, along with an overview of the system limitations. Finally, conclusions are drawn in Section V.

II. RELATED WORK

Evaluating the network performance with commercial equipment is very important since it provides a clear vision of the real performance obtained. Mostly, simulators are used to test the network performance under different conditions. However, the performance obtained via simulators sometimes is far from reality, as the wireless channel may not be accurate (e.g., with the standard) or some processes may be simplified. In this Section different works in the literature are analyzed where measurement campaigns have been performed in industrial scenarios with wireless technologies.

The latency performance has been one of the most addressed topics in the literature. In fact, since the adoption of wireless connectivity in the industrial sector and the emergence of new use cases with low latency requirements and high reliability, this topic has gained a high importance to determine which wireless technology is the most appropriate in industrial environments and if they can fulfil these requirements. In [9] and [10], 5G Non-Public Network (NPN) solutions are evaluated in terms of baseline Key Performance Indicators (KPIs). A comparison between 5G NPN and Public Network (PN) is performed in [11], where the network performance was evaluated in terms of latency, throughput and packet loss using one device. A framework for the integration of 5G in industry was proposed in [12], where the authors also evaluated the control-loop latency performance for the use case of controlling an AMR in the mobility case. These measurements were performed with Wi-Fi and 5G with only one device attached to the network. In [13], the latency performance of the 5G network was evaluated. Specifically, the uplink and downlink latency was measured with different packet sizes and inter-packet arrivals, with one user equipment. The authors of [14] evaluated the handover performance of Wi-Fi 6 in an indoor industrial environment; the 802.11r roaming functionality was evaluated for a mobility use case, using an AMR with some stationary background devices that transmitted traffic to load the network. The quality of experience (QoE) and throughput of the 5G network was evaluated in [15]. The results obtained by the authors indicate that the relationship between network performance and QoE in industrial settings is complex, due to a time-variant dependency. In [16] and [17], the authors compared the performance between Wi-Fi and Citizenship Broadband Radio System (CBRS) on the unlicensed spectrum of the USA using the Long-Term Evolution (LTE) radio network. In particular, they focused on the evaluation of different KPIs such as the average latency, the throughput and the packet loss under different loads.

Multi-connectivity consists in establishing two or more links between a user and two or more radio access nodes, which are typically uncorrelated links. For instance, the two links can use different channels, different networks or even different network access technologies, such as cellular and Wi-Fi. Multi-connectivity is often adopted for improving communication aspects such as latency, reliability and throughput. In the literature, different multi-connectivity schemes have been tested in industrial scenarios [8], [18], [19], [20]. In [18], the authors studied multi-connectivity for Ultra-Reliable and Low Latency Communications (URLLC) and the cost in throughput for other services such as Enhanced Mobile BroadBand (eMBB) services. A comparison between LTE and Wi-Fi technologies was done in [8], where different multi-connectivity schemes were evaluated (load balancing, PD and packet splitting). A multi-connectivity solution for Wi-Fi was evaluated in [19], where a device is composed of two Wi-Fi cards, each of them connected to different Access Points (APs) and coordinated by a smart Layer-4 scheduling mechanism. This work focused on the latency performance for the mobility case when using the PD

and best path switching solutions in an indoor factory. On the other hand, the authors of [20] presented a novel multi-connectivity solution that takes into account the QoS to dynamically select the links for PD. This scheme was evaluated with Wi-Fi 6 in terms of latency and throughput. Finally, the authors of [21] compares the performance of multi-Radio Access Technology (RAT) with Wi-Fi 6, LiFi and 5G. In particular, their multi-connectivity approach used was Multi-Path Transmission Control Protocol (MPTCP), which consists of splitting data flows into small flows and sending them over different interfaces to improve throughput. However, the scalability of the network was not considered (measurements were performed with one device) and the evaluated scenario was a museum.

Despite the different empirical measurements performed in the literature in industrial scenarios, we have not found any paper that takes into account the scalability of the network in terms of latency and packet loss. In fact, previous works usually take into account the performance of the network with only one device attached to the network. Also, multi-connectivity performance with a PD approach between 5G and Wi-Fi has not been addressed yet with a real implementation. Therefore, this paper tries to fill this gap by assessing the scalability of 5G, Wi-Fi and multi-connectivity between both technologies in an indoor industrial scenario in terms of latency and packet loss. For this, we used different packet sizes and use cases (stationary and mobility).

III. METHODOLOGY

A. SCENARIO AND NETWORK CONFIGURATION

The different measurements have been performed inside the "Aalborg University 5G Smart Production Lab" [2]. This lab consists of a small-scale industrial factory environment of approximately 1250 m^2 composed of two halls (see Figure 1) and a wide range of industrial manufacturing and production equipment, such as welding machines, robotics arms, production lines, etc. The dimensions of the halls are as follows: one measures 40 × 15 × 6 cubic meters, while the other measures 32 × 20 × 6 cubic meters. Approximately, 20% of the entire area is occupied by clutter, with a clutter height ranging from 1 to 3 meters. The lab is also equipped with different network technologies such as NPN 5G Stand-Alone (SA) and PN 5G Non-Stand-Alone (NSA), Wi-Fi 6, LTE and ultra-wide band (UWB). In this paper, the focus is set on 5G SA and Wi-Fi 6 technologies.

The 5G SA network is operated in collaboration with Telenor Denmark using Nokia equipment, more specifically, it is equipped with an in-house Nokia Mxie 5G SA core, a Nokia AirScale baseband unit and 3 Nokia AirScale indoor Radio (ASiR). The network operates in band N78 (3.7 GHz) with a bandwidth of 100 MHz and is configured as Time Division Duplex (TDD) with an UL/DL slot ratio of 3/7. In this deployment, all base stations (BS) transmit with a maximum power of 23 dBm and have the same configuration (i.e., emit the same cell), therefore, handovers will not occur during mobility.

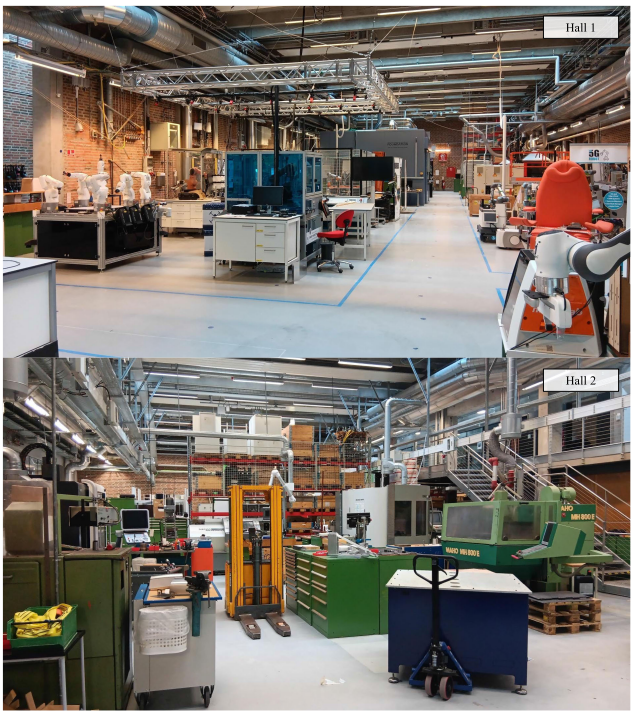


FIGURE 1. Overview of the Aalborg University 5G Smart Production Lab, including details on the two industrial halls.

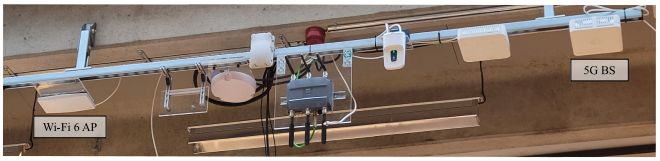


FIGURE 2. Overview of the different operational wireless network deployments.

The Wi-Fi 6 network is composed of three CISCO MR36 AP [22], distributed within the lab and operating in the 5 GHz band. The CISCO MR36 AP supports 2 × 2 Multi-User Multiple-Input Multiple-Output (MU-MIMO) and uplink/downlink Orthogonal Frequency Division Multiple Access (OFDMA) for more efficient transmission to multiple clients with up to 1024-Quadrature Amplitude Modulation (QAM) coding support. It also supports Basic Service Set (BSS) coloring which enables spatial reuse and reduces co-channel interference. Each AP transmits with a power of 20 dBm and is configured with a bandwidth of 20 MHz. To ensure that they do not interfere with each other, a dedicated channel is used on each AP (channels 132, 136 and 140), therefore, BSS coloring feature is not used. For roaming between APs when mobility, we enabled and used the IEEE 802.11r roaming functionality [23].

For both networks, the ASiRs/APs are mounted in the ceiling, approximately 6 meters above the ground and they are positioned to cover roughly 1/3 of the factory floor, as shown in Figure 2.

B. SETUP

The User Equipment (UE) used to perform the measurements is shown in Figure 3. It is composed of an Intel

FIGURE 3. Picture of the equipment used to evaluate the network performance.

NUC5i3MYHE [24], equipped with an Intel M2 Wi-Fi 6 AX200 card, running Arch Linux with kernel version 6.2.2. The Wi-Fi 6 adapter has been configured with the following features: uplink/downlink OFDMA, up to 1024 QAM coding and Target Wake Time (TWT) [25]. Regarding the 5G connection, a Simcom SIM8202G-M2 5G modem has been used [26] configured with 4 antennas, with MIMO 2 × 2. This modem is connected to the NUC through a M2 to USB3 adapter.

Two different scenarios have been considered: stationary and mobility. The stationary scenario represents the connectivity of a PLC in a production line. On the other hand, the mobility scenario represents a use case of an AMR that moves within the factory floor to transport goods/pallets.

To evaluate the scalability of the network, the number of devices is increased from 1 up to 10 in steps of 3 devices. One device was used to transmit data and measure the network performance, whereas the rest of the devices acted as background devices. All background devices were stationary and transmitted a constant bit rate.

Regarding data transmission, two different packet sizes have been considered in this study: 64 and 1250 bytes. The small packet size represents short control messages exchanged in the network, whereas the high packet size represents use cases such as video-operated remote control [27].

The measurement campaigns were performed for single connectivity with 5G SA and Wi-Fi 6, and multi-connectivity between both technologies. In this case, we used the PD [7] solution which consists on duplicating the data and sending it through each available link. This could improve the reliability and also reduce the latency when one of the links experiences poor channel conditions, and the data could be successfully transmitted through the other link. To test this feature in our setup, we used a multi-connectivity tunneling tool [28], developed at Aalborg University. This tool duplicates the packets at Layer 3 and sends it over Internet Protocol (IP) in Layer 4 (User Datagram Protocol, UDP) packets through 5G SA and Wi-Fi 6.

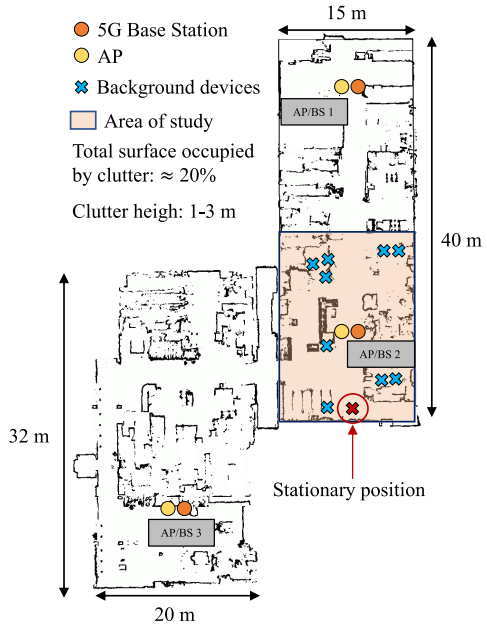


FIGURE 4. LiDAR floor plan of the lab and stationary setup. 5G BS and AP locations are marked with an orange and yellow circle. Background devices location are marked with a blue cross while the measuring device location is marked with a red cross.

The detailed information about the stationary and mobility setup is described below.

1) STATIONARY

Figure 4 shows the Light Detection and Ranging (LiDAR) floor plan of the lab and the stationary setup, where the location of the 5G BSs/APs are highlighted in circular markers. For the stationary case, the focus was set on the light orange area depicted in the figure, where all devices were placed and connected to AP/BS 2. Wi-Fi devices were forced to be connected to AP 2 by configuring the BSS Identifier (BSSID) in the connection profile.

The stationary position for the measuring device is marked with a red cross, while stationary background devices are marked with a blue cross. The mean distance from the devices to the AP/BS 2 is approximately 10 meters (in the range of 5 to 15 meters). Furthermore, regarding the density, up to ten devices were deployed in an area of 20 × 15 squared meters. At the beginning, only the measuring device was connected to the 5G SA and Wi-Fi 6 network. Then, background devices were added to the network to increase the number of devices for the different measurements.

2) MOBILITY

The mobility setup is depicted in Figure 5. Unlike the stationary setup, here background devices are placed throughout the factory floor. The maximum number of background devices per AP/BS was set to 3 to maintain consistency in the number of devices during the movement path. Similar to the stationary case, background devices were forced to be connected to the specific AP located where they were placed by configuring the BSSID in the connection profile. The

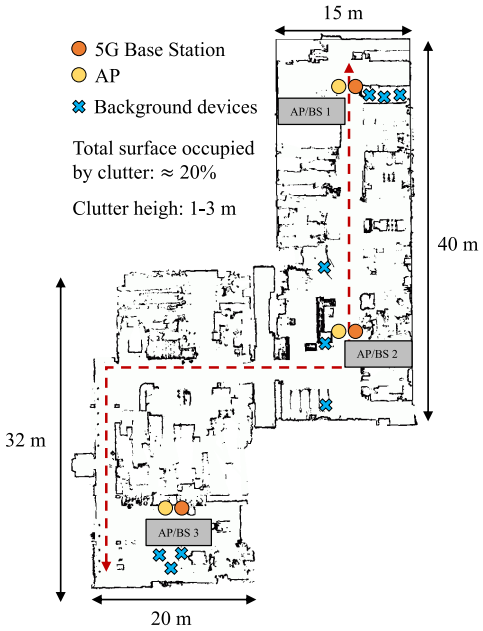


FIGURE 5. LiDAR floor plan of the lab and mobility setup. 5G BS and AP locations are marked with an orange and yellow circle. Background devices location are marked with a blue cross. AMR route is marked as a red dashed line.

FIGURE 6. MiR200 AMR within the 5G Smart Production Lab.

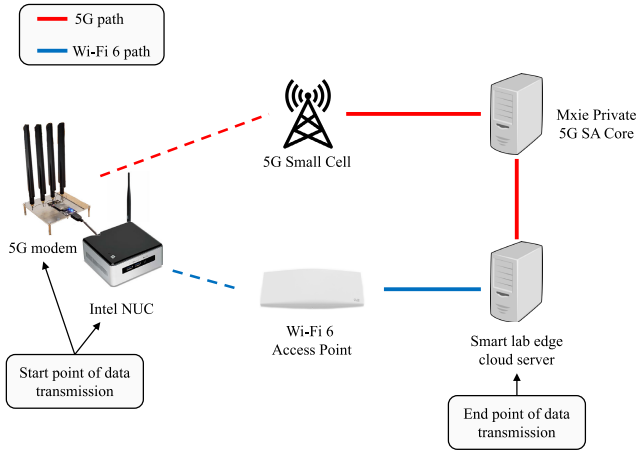


FIGURE 7. Diagram showing the data path between the measuring device and the Smart lab edge cloud server.

mean distance from the stationary background devices to the corresponding AP/BS is approximately 8 meters (in the range of 5 to 12 meters). Furthermore, regarding the density, up to six devices were stationary deployed in an area of 40 × 15 squared meters (hall 1) and up to three devices in an area of 32 × 20 squared meters (hall 2).

Mobility measurements were performed using a MiR200 AMR [29], with the 5G modem and Intel NUC placed on top, as shown in Figure 6. The MiR200 is designed for smaller transport tasks within the industry and logistics, such as transport of goods. The robot navigates using LiDAR, encoders and inertial measurement units with a payload of up to 200 kg. The use of this robot allowed to perform different reproducible mobility tests, which guarantees a consistency on the measurements. During the measurements, the AMR navigates within the 5G Smart Production Lab following the path marked with a red dashed line in Figure 5, with a speed of 1 m/s in a loop.

Similar to the stationary case, at the beginning only the measuring device was connected to the 5G SA and Wi-Fi 6 network. Then, background devices were added to the network to increase the number of devices for the different measurements. However, in this case, they were added proportionally to the APs/BSs, that is, the number of devices was increased by one on each AP/BS.

C. METRICS

In this study, the following metrics have been considered:

- **Latency**: The ping tool was used to measure the RTT of a packet sent from the UE to our edge-cloud server, and back. A diagram of the path of the packets is shown in Figure 7. This tool was configured to transmit Internet Control Message Protocol (ICMP) packets with a periodicity of 10 ms, with packet sizes of 64 and 1250 bytes, and a preload of 100 packets. The preload helps to maintain the transmission periodicity when long delays occur. The periodicity of 10 ms ensures that the modem does not enter power saving mode between requests, which could negatively impact the measurement campaigns. To obtain statistic results, we run ping until it transmits more than one million packets. For a real-time application in a factory scenario, the RTT latency should be less than 100 ms [27].
- **Packet loss**: Based on the packet statistics from the latency results, the number of lost packets is counted for each measurement campaign. This is done by reading the output of the ping tool after a completed measurement, which includes the number of ICMP packets transmitted (request) and received (replies).

Then, based on the difference, the packet loss is obtained.

IV. RESULTS AND DISCUSSION

In this section, the results obtained throughout the different measurement campaigns for the stationary and mobility cases are presented. Latency results are shown as a Complementary Cumulative Distribution Function (CCDF), whereas packet loss statistics are summarized in a table.

A. STATIONARY

1) 64 BYTES

Figure 8 shows CCDF plots of the latency measurements when using a packet size of 64 bytes and Table 1 summarizes the key values.

As it can be seen, the latency distribution with 5G SA is very stable, not exceeding 14.8 ms with 1 device. Obviously, when adding more devices to the network, the latency is increased as expected. In this case, the network needs to manage different data traffic and this is done in 5G by assigning different resources (time slots) to the users. This can be observed on the median values, which suffer an increase in the range of 0.3 ms to 0.8 ms. However, a similar trend is observed in the tails, obtaining a 99.99%-ile (10^{-4}) value of 12.6 ms, 17.5 ms, 19.8 ms and 23.8 ms with 1, 4, 7 and 10 devices, respectively.

When using Wi-Fi 6, it is observed that in general the latency is lower compared to 5G SA. This can be seen on the median values obtained, which ranges from 3.1 ms to 5.3 ms when increasing the number of devices. Another aspect observed is that higher latency tails are obtained compared to 5G SA, even with only one device connected. In terms of network scalability, contrary to 5G SA, the latency with Wi-Fi 6 is clearly affected when adding more devices, especially with 7 and 10 devices, obtaining latency values above 100 ms. The high values on the latency are expected when increasing the number of devices, since with Wi-Fi the devices compete for the channel to transmit data, using Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) at the Medium Access Control (MAC) layer. Therefore, whenever a device wants to transmit data, it first needs to listen and ensure that nobody is transmitting data. Otherwise, it will wait for a random period of time (backoff time) and check again if the channel is clear. As the number of devices increases, the probability of waiting to transmit data is also increased. In this case, a 99.99%-ile (10^{-4}) value of 21.1 ms, 24.1 ms, 88 ms and 127 ms is obtained with 1, 4, 7 and 10 devices, respectively.

When multi-connectivity is applied, it is observed that the tails are reduced and the trend is similar to 5G SA. This reduction is due to the fact that the packet is always sent duplicated via two radio links (5G SA and Wi-Fi 6) and the latency obtained will be the best of these two links. For the same reason, the CCDFs on the first part of the distribution are similar to Wi-Fi 6 with a slight offset. This offset is due to the multi-connectivity tool, that adds an extra overhead on the packets. In general, multi-connectivity takes advantage of both networks and reduces the latency values in all cases evaluated, as it can be seen in the maximum and 99.99%-ile (10^{-4}) values in Table 1.

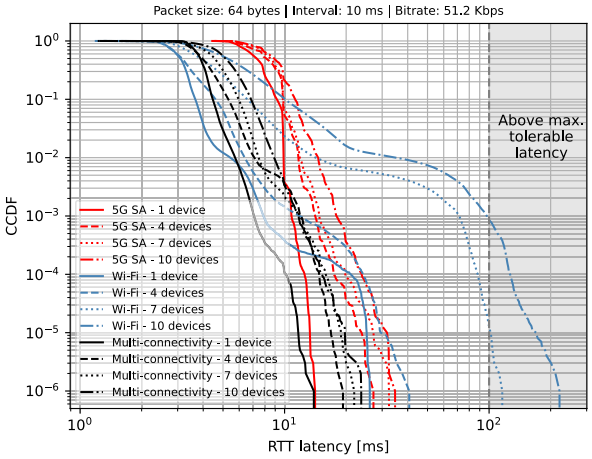


FIGURE 8. Latency CCDF obtained for the stationary case with a packet size of 64 bytes.

TABLE 1. Latency [ms] obtained for the stationary case with a packet size of 64 bytes.

Deployment	Median	Max.	99.9%-ile	99.99%-ile
5G SA - 1 device	7.3	14.8	10.9	12.6
5G SA - 4 devices	8.1	31.6	14.2	17.5
5G SA - 7 devices	8.6	34.2	14.8	19.8
5G SA - 10 devices	8.9	42.2	18	23.8
Wi-Fi 6 - 1 device	3.1	26.2	7.9	21.1
Wi-Fi 6 - 4 devices	3.2	51.8	11.9	24.1
Wi-Fi 6 - 7 devices	3.9	133	71.3	88
Wi-Fi 6 - 10 devices	5.3	231	97.6	127
Multi-connectivity - 1 device	3.9	14.6	7.1	10.2
Multi-connectivity - 4 devices	4	19.8	12	14.8
Multi-connectivity - 7 devices	4.5	22.7	11.9	15.9
Multi-connectivity - 10 devices	5.1	25.7	11.7	15.2

Table 2 summarizes the packet statistics, which include the number of packets sent, received and lost. In general, a low packet loss is obtained with both technologies in all cases. In 5G SA only 1 or 2 packets are lost, whereas Wi-Fi 6 suffers a slightly higher packet loss, reaching 5 in some cases. As expected, multi-connectivity reduces the packet loss to 0.

2) 1250 BYTES

Figure 9 shows CCDF plots of the latency measurements when using a packet size of 1250 bytes and Table 3 summarizes the key values.

In this case, a higher latency is noticeable when using 5G SA, with the CCDFs shifted to the right in comparison to the previous case with a smaller packet size. This clearly indicates that the packet size has a high influence on the latency values. Since the packet size is higher, more resources are necessary to transmit all data, that is, more slots need to be assigned to the users in the scheduler. Consequently, the

TABLE 2. Packet statistics for the stationary case with a packet size of 64 bytes.

Deployment	Sent	Received	Lost (%)
5G SA - 1 device	1001000	1000999	0.0001
5G SA - 4 devices	1001000	1000999	0.0001
5G SA - 7 devices	1001000	1000998	0.0002
5G SA - 10 devices	1001000	1000999	0.0001
Wi-Fi 6 - 1 device	1001000	1000995	0.0005
Wi-Fi 6 - 4 devices	1001000	1000995	0.0005
Wi-Fi 6 - 7 devices	1001000	1000997	0.0003
Wi-Fi 6 - 10 devices	1001000	1000995	0.0005
Multi-connectivity - 1 device	1001000	1001000	0
Multi-connectivity - 4 devices	1001000	1001000	0
Multi-connectivity - 7 devices	1001000	1001000	0
Multi-connectivity - 10 devices	1001000	1001000	0

TABLE 3. Latency [ms] obtained for the stationary case with a packet size of 1250 bytes.

Deployment	Median	Max.	99.9%-ile	99.99%-ile
5G SA - 1 device	12.9	44.7	25.7	27.6
5G SA - 4 devices	14.7	44.7	25.8	29.5
5G SA - 7 devices	15	41.6	27.2	31.3
5G SA - 10 devices	17.2	53.3	29.8	34.4
Wi-Fi 6 - 1 device	2.2	29.6	8.8	20.8
Wi-Fi 6 - 4 devices	2.4	61.1	16.5	30.1
Wi-Fi 6 - 7 devices	3	111	68.3	85.2
Wi-Fi 6 - 10 devices	5.8	231	99.7	128
Multi-connectivity - 1 device	2.4	26.8	6.7	14.1
Multi-connectivity - 4 devices	2.6	28.8	11.7	18.9
Multi-connectivity - 7 devices	3.2	32.7	14.9	22
Multi-connectivity - 10 devices	5.3	35.8	22.4	25.9

latency will increase. Taking a look on the median values, an increment of more than 6 ms is obtained. Despite that, a similar trend in the tails is observed when connecting more devices to the network, obtaining a 99.99%-ile (10^{-4}) value of 27.6 ms, 29.5 ms, 31.3 ms and 34.4 ms with 1, 4, 7 and 10 devices, respectively.

On the other hand, with Wi-Fi 6, a similar behaviour to case with a small packet size is observed. This occurs due to Wi-Fi trying to send all data from the buffer on each transmission opportunity. Therefore, the packet size is not clearly affected but the number of devices is. Moreover, a slight reduction in the median values is observed in all cases except with 10 devices and this is because of using a more efficient Modulation Coding Scheme (MCS) coding on the data transmission. In this case, similar tails are obtained with a 99.99%-ile (10^{-4}) value of 20.8 ms 30.1 ms 85.2 ms and 128 ms with 1, 4, 7 and 10 devices, respectively.

When using multi-connectivity, we observed a similar trend on the CCDFs. The main change is that large tails are obtained, but this is due to the fact that the latency with 5G SA is higher because of the packet size. Therefore, the potential gains of multi-connectivity in terms of the tails are reduced in this case. Moreover, it is observed that multi-connectivity obtains lower tails in all cases (even with 10 devices) than 5G SA with only one device.

Taking a look at the packet statistics in Table 4, a similar packet loss is obtained with the 5G SA network. On the other hand, with Wi-Fi 6, a slight increase on the packet loss is obtained and this can be related to the packet size, since it is

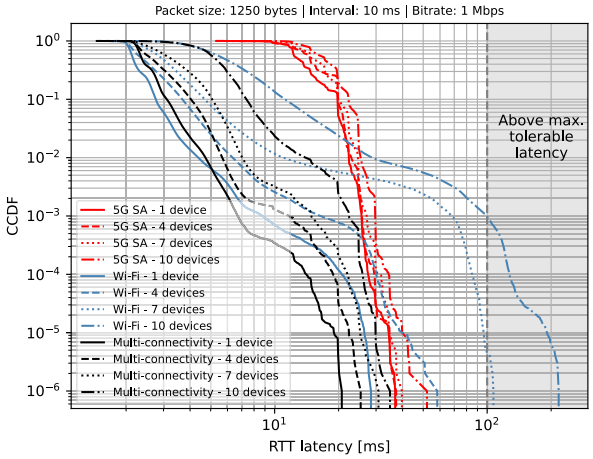


FIGURE 9. Latency CCDF obtained for the stationary case with a packet size of 1250 bytes.

TABLE 4. Packet statistics for the stationary case with a packet size of 1250 bytes.

Deployment	Sent	Received	Lost (%)
5G SA - 1 device	1001000	1000997	0.0003
5G SA - 4 devices	1001000	1001000	0
5G SA - 7 devices	1001000	1000996	0.0004
5G SA - 10 devices	1001000	1001000	0
Wi-Fi 6 - 1 device	1001000	1000994	0.0006
Wi-Fi 6 - 4 devices	1001000	1000992	0.0008
Wi-Fi 6 - 7 devices	1001000	1000984	0.0016
Wi-Fi 6 - 10 devices	1001000	1000981	0.0019
Multi-connectivity - 1 device	1001000	1001000	0
Multi-connectivity - 4 devices	1001000	1001000	0
Multi-connectivity - 7 devices	1001000	1001000	0
Multi-connectivity - 10 devices	1001000	1001000	0

higher and the time transmitting data over the channel is also higher, so the probability of failure (i.e., having bit errors) increases. Another aspect observed with Wi-Fi 6 is that as the number of devices increases, the packet loss is also increased. Finally, when applying multi-connectivity, no packet loss is obtained in any of the evaluated cases.

B. MOBILITY

1) 64 BYTES

Figure 10 shows CCDF plots of the latency measurements when using a packet size of 64 bytes and Table 5 summarizes the key values.

When mobility is introduced, a higher variation on the latency distribution is observed. This is expected, since the channel varies during the movement on the path, with changing reflections and propagation loss. This also causes the use of different MCS during data transmission, which may have an impact on the latency if a more robust MCS is selected.

In the case of 5G SA, a similar trend is observed on the CCDFs with respect to the stationary case when increasing the number of devices. In this particular case, a similar latency distribution is observed with 1 and 4 devices, whereas there is a gap in the latency tails with 7 and 10 devices. The median

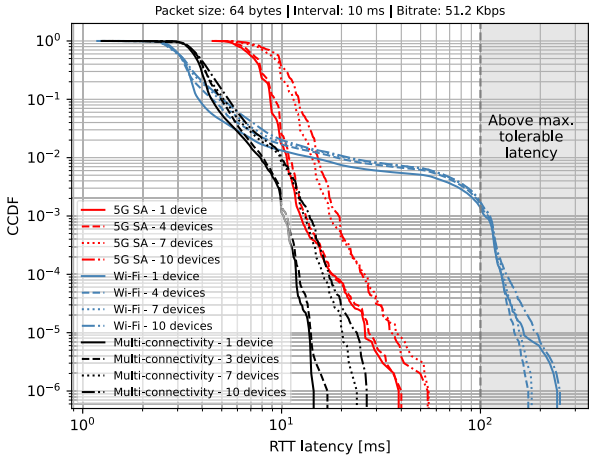


FIGURE 10. Latency CCDF obtained for the mobility case with a packet size of 64 bytes.

TABLE 5. Latency [ms] obtained for the mobility case with a packet size of 64 bytes.

Deployment	Median	Max.	99.9%-ile	99.99%-ile
5G SA - 1 device	7.3	43.2	12.5	17.9
5G SA - 4 devices	7.6	41.9	13.3	18.1
5G SA - 7 devices	9.2	63.7	17.4	26.2
5G SA - 10 devices	9.8	64.2	19.4	25.4
Wi-Fi 6 - 1 device	3.2	259	106	128
Wi-Fi 6 - 4 devices	3.1	182	110	128
Wi-Fi 6 - 7 devices	3.2	197	112	130
Wi-Fi 6 - 10 devices	3.2	266	112	135.2
Multi-connectivity - 1 device	3.7	14.6	10.3	11.9
Multi-connectivity - 4 devices	3.9	21.2	10.7	12.3
Multi-connectivity - 7 devices	3.8	31.8	13.1	16.2
Multi-connectivity - 10 devices	3.9	27.3	13.9	17.4

values obtained are slightly higher compared to the stationary case, which is expected due to the varying channel conditions during the path. Moreover, the tails on the distribution are also higher, obtaining a 99.99%-ile (10^{-4}) value of 17.9 ms, 18.1 ms, 26.2 ms and 25.4 ms with 1, 4, 7 and 10 devices, respectively.

On the other hand, with Wi-Fi 6 a clear difference in the tails of the latency distribution is observed. The high increase on the latency is caused due to Wi-Fi roaming between the APs along the movement of the AMR in the scenario. Consequently, latencies above 100 ms are obtained. In this case, although Wi-Fi 6 obtains a lower median value than 5G SA, the tails in the distribution are higher, with a 99.99%-ile (10^{-4}) value above 120 ms in all cases evaluated.

Finally, when using multi-connectivity, the tails are reduced and they converge to a similar trend respect to 5G SA, since when Wi-Fi 6 signal drops, it experiences higher latencies than 5G SA, especially in the roaming case between APs. A 99.99%-ile value of 11.9 ms, 12.3 ms, 16.2 ms and 17.4 ms is obtained with 1, 4, 7 and 10 devices.

In terms of packet loss statistics (see Table 6), the 5G SA network obtains a low packet loss, similar to the stationary case. On the other hand, Wi-Fi 6 obtains a high number

TABLE 6. Packet statistics for the mobility case with a packet size of 64 bytes.

Deployment	Sent	Received	Lost (%)
5G SA - 1 device	1000002	999999	0.0003
5G SA - 4 devices	1001000	1000999	0.0001
5G SA - 7 devices	1001000	1000998	0.0002
5G SA - 10 devices	1001000	1000998	0.0002
Wi-Fi 6 - 1 device	1001000	999136	0.18621
Wi-Fi 6 - 4 devices	1001000	999047	0.1951
Wi-Fi 6 - 7 devices	1001000	999049	0.19491
Wi-Fi 6 - 10 devices	1001000	999008	0.199
Multi-connectivity - 1 device	1001000	1001000	0
Multi-connectivity - 4 devices	1001000	1001000	0
Multi-connectivity - 7 devices	1001000	1001000	0
Multi-connectivity - 10 devices	1001000	1001000	0

of packet losses. Again, with multi-connectivity, the packet losses are reduced to 0, since the packet is sent duplicated over both interfaces and the reliability is increased (i.e., when Wi-Fi 6 AP roaming).

2) 1250 BYTES

Figure 11 shows CCDF plots of the latency measurements when using a packet size of 1250 bytes and Table 7 summarizes the key values.

When a high packet size is used in the mobility case, again, a higher variation on the latency values is obtained in all cases evaluated, which makes sense due to signal reflections and multi-path propagation.

A similar trend is observed in 5G-SA when increasing the number of devices, however, same as in the previous case with a small packet size, there is a higher step in the CCDF when increasing the number of devices from 4 to 7. In general, it is observed that 5G SA latency is very stable, particularly in the tails of the distribution. In this case, as expected, the 99.99%-ile (10^{-4}) value of the tails has increased, being 26.4 ms, 28.7 ms, 37.4 ms and 40.7 ms with 1, 4, 7 and 10 devices, respectively.

A similar latency distribution is observed with Wi-Fi 6 compared to when using a packet size of 64 bytes. As previously mentioned, this occurs due to Wi-Fi trying to send all available data in the buffer whenever the device has a transmission opportunity. Therefore, the packet size does not have a high impact on the latency if it does not exceed the Maximum Transmission Unit (MTU), configured as 1500 bytes, in which case packet fragmentation will occur. Again, the median values are lower when using a high packet size, since a more efficient MCS is used when transmitting data.

Finally, same as in the previous cases, the use of multi-connectivity reduces drastically the latency tails in the distribution, obtaining median values similar to Wi-Fi 6 and a 99.99%-ile values lower than 5G SA, as shown in Table 7.

Looking at Table 8, which contains the packet statistics, a slight packet loss is observed with 5G SA. In the case of Wi-Fi 6, higher packet loss were obtained, with values above 1800, due to roaming between APs and by the fact that the

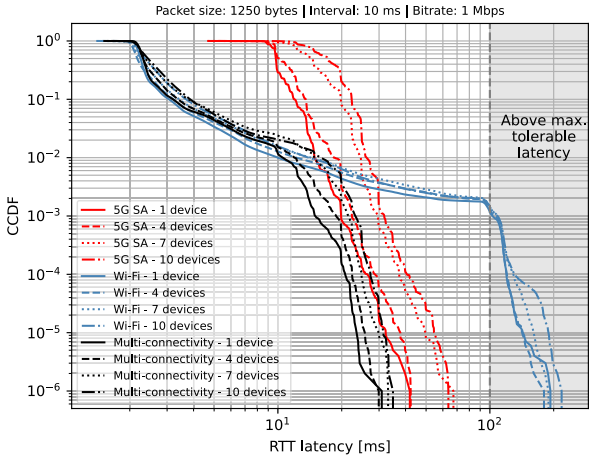


FIGURE 11. Latency CCDF obtained for the mobility case with a packet size of 1250 bytes.

TABLE 7. Latency [ms] obtained for the mobility case with a packet size of 1250 bytes.

Deployment	Median	Max.	99.9%-ile	99.99%-ile
5G SA - 1 device	9.8	42.3	19.9	26.4
5G SA - 4 devices	10.3	44.7	22.4	28.7
5G SA - 7 devices	14.6	75.7	29.7	37.4
5G SA - 10 devices	17.2	64.8	31.8	40.7
Wi-Fi 6 - 1 device	2.2	197	103	120
Wi-Fi 6 - 4 devices	2.2	193	106	121
Wi-Fi 6 - 7 devices	2.2	197	109	129.2
Wi-Fi 6 - 10 devices	2.4	226	109	130.2
Multi-connectivity - 1 device	2.3	33.6	15.1	20.8
Multi-connectivity - 4 devices	2.3	34.9	18.9	22.4
Multi-connectivity - 7 devices	2.4	33.3	21.6	24.8
Multi-connectivity - 10 devices	2.4	36.8	23.2	27.3

TABLE 8. Packet statistics for the mobility case with a packet size of 1250 bytes.

Deployment	Sent	Received	Lost (%)
5G SA - 1 device	1001000	1000991	0.0009
5G SA - 4 devices	1001000	1000992	0.0008
5G SA - 7 devices	1001000	1000988	0.0012
5G SA - 10 devices	1001000	1000997	0.0003
Wi-Fi 6 - 1 device	1001000	999166	0.18322
Wi-Fi 6 - 4 devices	1001000	999053	0.19451
Wi-Fi 6 - 7 devices	1001000	998753	0.22448
Wi-Fi 6 - 10 devices	1001000	998685	0.23127
Multi-connectivity - 1 device	1001000	1001000	0
Multi-connectivity - 4 devices	1001000	1001000	0
Multi-connectivity - 7 devices	1001000	1001000	0
Multi-connectivity - 10 devices	1001000	1001000	0

time transmitting data is longer, and therefore, the probability of having errors during the transmission is also increased.

C. SYSTEM LIMITATIONS

This study does not consider the impact of interfering devices on network performance. This will be addressed in subsequent steps, and the work carried out in this paper will serve as a baseline for comparison of network performance with and without interference.

For this reason, as the factory scenario is surrounded by multiple laboratories with different Wi-Fi networks deployed, a bandwidth of 20 MHz was used on each AP, since we have 60 MHz of spectrum dedicated for our APs. In particular, we have available channels 132, 136 and 140. Nevertheless, the results should not be significantly altered when using a higher bandwidth on each AP, as the maximum bitrate of each device is 1 Mbps and it is far from the measured capacity limit on each AP with this configuration (200 Mbps).

Conversely, in this study, up to 10 devices were employed due to the availability of commercial equipment for both technologies, rather than due to network/capacity limitations.

Finally, the multi-connectivity solution evaluated in this paper duplicates and transmits all packets over Wi-Fi 6 and 5G interfaces. As a future step, we will implement a dynamic duplication process that based on network metrics, will determine whether to duplicate or not the packet in order to improve network efficiency and reduce resource wastage.

V. CONCLUSION

In this paper, an empirical comparison of 5G SA, Wi-Fi 6 and multi-connectivity between both technologies have been performed in an indoor industrial scenario. Particularly, the focus of this paper has been to study the latency performance and packet loss with different packet sizes for stationary and mobility cases in terms of network scalability.

From the measurement campaign performed in this paper, the following conclusions can be derived:

- In general, Wi-Fi 6 produces lower latencies but large tails in the distribution, particularly in the mobility case due to APs roaming. On the other hand, with 5G-SA, the latency distribution is very stable with bounded tails for stationary and mobility cases.
- The packet size has an impact on the latency with 5G SA, obtaining higher latencies and tails when increasing the value. No impact on the latency is observed with Wi-Fi 6. Moreover, the packet size has a noticeable impact on packet losses with Wi-Fi 6, whereas with 5G SA the impact is negligible.
- In terms of network scalability, 5G SA performs better than Wi-Fi 6. An offset to higher values on the latency distribution is observed with 5G SA, whereas Wi-Fi 6 increments the tails as the number of devices increases.
- Multi-connectivity improves the latency distribution in all evaluated cases. This feature is specially useful in the mobility case, due to Wi-Fi 6 APs roaming. As the number of devices increases, multi-connectivity becomes necessary to reduce the latency tails.
- 5G SA is more reliable than Wi-Fi 6 in terms of packet losses, particularly in the mobility case. When increasing the number of devices, packet losses are also increased with Wi-Fi 6 while for 5G SA it does not seem to be affected. Multi-connectivity improves the reliability, with no packet losses obtained in any of the cases evaluated in this study.

The selected technology will vary by industrial sector and business use case. Companies should base their decision on a trade-off between the expected performance for their use cases and the economic cost. Some companies may opt for a low-cost Wi-Fi 6 technology, even if it comes at the expense of performance; or for a reliable technology such as 5G at the expense of a higher cost. However, for the most rigorous latency and reliability requirements, we recommend the multi-connectivity solution, as it can guarantee a low latency and high reliability, although this implies a higher cost.

REFERENCES

[1] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Bus. Inf. Syst. Eng.*, vol. 6, no. 4, pp. 239–242, Jun. 2014, doi: 10.1007/s12599-014-0334-4.

[2] I. Rodriguez, R. S. Mogensen, A. Schjørring, M. Razzaghpour, R. Maldonado, G. Berardinelli, R. Adeogun, P. H. Christensen, P. Mogensen, O. Madsen, C. Møller, G. Pocovi, T. Kolding, C. Rosa, B. Jørgensen, and S. Barbera, "5G swarm production: Advanced industrial manufacturing concepts enabled by wireless automation," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 48–54, Jan. 2021, doi: 10.1109/MCOM.001.2000560.

[3] E. J. Oughton, W. Lehr, K. Katsaros, I. Selinis, D. Bubley, and J. Kusuma, "Revisiting wireless internet connectivity: 5G vs Wi-Fi 6," *Telecommun. Policy*, vol. 45, no. 5, Jun. 2021, Art. no. 102127, doi: 10.1016/j.telpol.2021.102127.

[4] HMS. *Continued Growth for Industrial Ethernet and Wireless Networks*. Accessed: Sep. 14, 2023. [Online]. Available: https://www.hms-networks.com/news-and-insights/news-from-hms/2023/05/05/industrial-network-market-shares-2023

[5] D. Segura, E. J. Khatib, J. Munilla, and R. Barco, "5G numerologies assessment for URLLC in industrial communications," *Sensors*, vol. 21, no. 7, p. 2489, Apr. 2021, doi: 10.3390/s21072489.

[6] S. Zhang, "An overview of network slicing for 5G," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 111–117, Jun. 2019, doi: 10.1109/MWC.2019.1800234.

[7] D. Segura, E. J. Khatib, and R. Barco, "Dynamic packet duplication for industrial URLLC," *Sensors*, vol. 22, no. 2, p. 587, Jan. 2022, doi: 10.3390/s22020587.

[8] M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Evaluation of multi-connectivity schemes for URLLC traffic over WiFi and LTE," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–7, doi: 10.1109/WCNC45663.2020.9120829.

[9] S. Senk, S. A. W. Itting, J. Gabriel, C. Lehmann, T. Hoeschele, F. H. P. Fitzek, and M. Reisslein, "5G NSA and SA campus network testbeds for evaluating industrial automation," in *Proc. Eur. Wireless; 26th Eur. Wireless Conf.*, Nov. 2021, pp. 1–8.

[10] J. Rischke, P. Sossalla, S. Itting, F. H. P. Fitzek, and M. Reisslein, "5G campus networks: A first measurement study," *IEEE Access*, vol. 9, pp. 121786–121803, 2021, doi: 10.1109/ACCESS.2021.3108423.

[11] S. B. Damsgaard, D. Segura, M. F. Andersen, S. A. Markussen, S. Barbera, I. Rodríguez, and P. Mogensen, "Commercial 5G NPN and PN deployment options for industrial manufacturing: An empirical study of performance and complexity tradeoffs," in *Proc. IEEE 34th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2023, pp. 1–7, doi: 10.1109/pimrc56721.2023.10293869.

[12] I. Rodriguez, R. S. Mogensen, A. Fink, T. Raunholt, S. Markussen, P. H. Christensen, G. Berardinelli, P. Mogensen, C. Schou, and O. Madsen, "An experimental framework for 5G wireless system integration into industry 4.0 applications," *Energies*, vol. 14, no. 15, p. 4444, Jul. 2021, doi: 10.3390/en14154444.

[13] J. Ansari, C. Andersson, P. de Bruin, J. Farkas, L. Grosjean, J. Sachs, J. Torsner, B. Varga, D. Harutyunyan, N. König, and R. H. Schmitt, "Performance of 5G trials for industrial automation," *Electronics*, vol. 11, no. 3, p. 412, Jan. 2022, doi: 10.3390/electronics11030412.

[14] A. Fink, R. S. Mogensen, I. Rodriguez, T. Kolding, A. Karstensena, and G. Pocovi, "Empirical performance evaluation of EnterpriseWi-fi for IIoT applications requiring mobility," in *Proc. Eur. Wireless ; 26th Eur. Wireless Conf.*, Nov. 2021, pp. 1–8.

[15] W. Tärneberg, O. Hamsis, J. Hedlund, K. Brunnström, E. Fitzgerald, A. Johnsson, V. Berggren, M. Kihl, A. Rao, R. Steinert, and C. Kilinc, "Towards intelligent industry 4.0 5G networks: A first throughput and QoE measurement campaign," in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2020, pp. 1–6, doi: 10.23919/SoftCOM50211.2020.9238299.

[16] V. Sathya, L. Zhang, and M. Yavuz, "A comparative measurement study of commercial WLAN and 5G LAN systems," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2022, pp. 1–7, doi: 10.1109/VTC2022-Fall57202.2022.10013019.

[17] V. Sathya, L. Zhang, M. Goyal, and M. Yavuz, "Warehouse deployment: A comparative measurement study of commercial Wi-Fi and CBRS systems," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2023, pp. 242–248, doi: 10.1109/ICNC57223.2023.10074584.

[18] E. J. Khatib, D. A. Wassie, G. Berardinelli, I. Rodriguez, and P. Mogensen, "Multi-connectivity for ultra-reliable communication in industrial scenarios," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, Apr. 2019, pp. 1–6, doi: 10.1109/VTCSPRING.2019.8746357.

[19] A. Fink, R. S. Mogensen, I. Rodriguez, T. Kolding, A. Karstensen, and G. Pocovi, "Radio-aware multi-connectivity solutions based on layer-4 scheduling for Wi-Fi in IIoT scenarios," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 1821–1826, doi: 10.1109/WCNC51071.2022.9771995.

[20] R. S. Mogensen, S. B. Damsgaard, I. Rodriguez, G. Berardinelli, A. E. Fink, T. E. Kolding, and G. Pocovi, "A novel QoS-aware multi-connectivity scheme for wireless IIoT," *IEEE Access*, vol. 10, pp. 104123–104134, 2022, doi: 10.1109/ACCESS.2022.3210340.

[21] A. Emami, H. Frank, W. He, A. Bravalheri, A.-C. Nicolaescu, H. Li, H. Falaki, S. Yan, R. Nejabati, and D. Simeonidou, "Multi–RAT enhanced private wireless networks with intent-based network management automation," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2023, pp. 1789–1794, doi: 10.1109/gcwkshps58843.2023.10464742.

[22] *Cloud Managed WiFi 6 Indoor Access Point | 802.11ax | MR36 | Cisco Meraki*. Accessed: Sep. 14, 2023. [Online]. Available: https://meraki.cisco.com/product/wi-fi/indoor-access-points/mr36/

[23] CISCO. *What is 802.11r? Why is This Important?*. Accessed: Sep. 14, 2023. [Online]. Available: https://blogs.cisco.com/networking/what-is-802-11r-why-is-this-important

[24] *Intel NUC Kit NUC5i3MYHE*. Accessed: Sep. 14, 2023. [Online]. Available: https://www.intel.co.uk/content/www/uk/en/products/sku/84860/intel-nuc-kit-nuc5i3myhe/specifications.html

[25] Intel. *Intel Wi-Fi 6 AX200 (Gig+) Module*. Accessed: Apr. 8, 2024. [Online]. Available: https://cdrdv2.intel.com/v1/dl/getcontent/607009

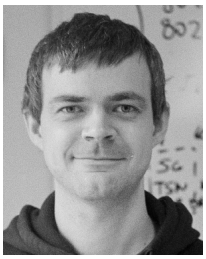
[26] *Simcom SIM8202G-M2*. Accessed: Sep. 14, 2023. [Online]. Available: https://www.simcom.com/product/SIM8202X_M2.html

[27] *Service Requirements for Cyber-Physical Control Applications in Vertical Domains*, document TS 22.104, 3GPP, Sep. 2021.

[28] *MPCONN—The Open Source Multi-Path Connectivity Tool*. Accessed: Sep. 14, 2023. [Online]. Available: https://github.com/drblah/mpconn

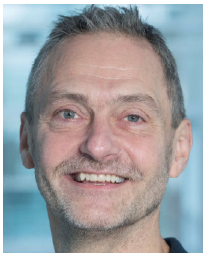
[29] *MiR200 Data Sheet*. Accessed: Sep. 14, 2023. [Online]. Available: https://htlelec.com/pdf/mir200.pdf

DAVID SEGURA received the B.Sc. degree in telematics engineering and the M.Sc. degree in telematics and telecommunication networks from the University of Málaga, Spain, in 2019 and 2020, respectively, where he is pursuing the Ph.D. degree in cellular communications. In 2019, he started to work as a Researcher with the Communication Engineering Department, University of Málaga. His research interests include wireless communication for Industry 4.0 and security.



SEBASTIAN BRO DAMSGAARD received the B.Sc. degree in internet technologies and computer systems and the M.Sc. degree in networks and distributed systems from Aalborg University, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree, focusing on 5G enabled autonomous mobile robotic systems. Previously, he worked in support and deployment of IT systems for manufacturing with Grundfos. His research interests include the application of wireless communication for Industry 4.0, (edge) cloud computing for industrial applications, and optimizing communication for use in mobile manufacturing equipment.

AKIF KABACI received the B.Sc. degree in electronics and communication engineering, the B.Sc. degree in control and automation engineering, and the M.Sc. degree in telecommunication engineering from Istanbul Technical University, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree with Aalborg University, focusing on unlicensed band telecommunication systems integration to smart factories. Previously, he was a Research Assistant and a Research and Development Engineer with defence industry, mainly focusing on physical layer design. His research interests mainly include wireless communication for Industry 4.0, optimization problems in communications, and physical layer design.



PREBEN MOGENSEN received the M.Sc. and Ph.D. degrees from Aalborg University, in 1988 and 1996, respectively. Since 1995, he has been a part-time associated with Nokia in various research positions and has made contributions from 2G to 5G cellular technologies. He has been with Aalborg University, since graduation in 1988. In 2000, he became a Full Professor with Aalborg University, where he is currently leading the Wireless Communication Networks Section, Department of Electronic Systems. He is also a Principal Scientist with Nokia Standardization Research. He has coauthored over 400 papers in various domains of wireless communication and his Google Scholar H-index is 71. His current research interests include industrial and critical use cases for 5G, 5G evolution, and 6G. He is a Bell Labs Fellow.

EMIL J. KHATIB (Member, IEEE) received the Ph.D. degree in machine learning, big data analytics, and knowledge acquisition applied to the troubleshooting in cellular networks, in 2017. He is a Postdoctoral Juan de la Cierva Fellow with the University of Málaga. He has participated in several national and international projects related to Industry 4.0 projects. Currently, he is working on the topic of security and localization in industrial scenarios.

RAQUEL BARCO is a Full Professor of telecommunication engineering with the University of Málaga. Before joining the university, she was with Telefonica, Madrid, Spain, and the European Space Agency (ESA), Darmstadt, Germany. As a researcher, she is specialized in mobile communication networks and smart-cities. She has led projects funded by several million euros. She has published more than 100 papers in high impact journals and conferences, authored five patents, and received several research awards.

Chapter 5

Optimization

sensors

MDPI

Article

Dynamic Packet Duplication for Industrial URLLC

David Segura *, Emil J. Khatib and Raquel Barco

Instituto Universitario de Investigación en Telecomunicación (TELMA), Universidad de Málaga, CEI Andalucía TECH E.T.S.I. Telecomunicación, Bulevar Louis Pasteur 35, 29010 Malaga, Spain; emil@uma.es (E.J.K.); rbm@ic.uma.es (R.B.)

* Correspondence: dsr@ic.uma.es

Abstract: The fifth-generation (5G) network is presented as one of the main options for Industry 4.0 connectivity. To comply with critical messages, 5G offers the Ultra-Reliable and Low latency Communications (URLLC) service category with a millisecond end-to-end delay and reduced probability of failure. There are several approaches to achieve these requirements; however, these come at a cost in terms of redundancy, particularly the solutions based on multi-connectivity, such as Packet Duplication (PD). Specifically, this paper proposes a Machine Learning (ML) method to predict whether PD is required at a specific data transmission to successfully send a URLLC message. This paper is focused on reducing the resource usage with respect to pure static PD. The concept was evaluated on a 5G simulator, comparing between single connection, static PD and PD with the proposed prediction model. The evaluation results show that the prediction model reduced the number of packets sent with PD by 81% while maintaining the same level of latency as a static PD technique, which derives from a more efficient usage of the network resources.

Keywords: 5G; machine learning; prediction; Industry 4.0; URLLC; multi-connectivity

Citation: Segura, D.; Khatib, E.J.; Barco, R. Dynamic Packet Duplication for Industrial URLLC. *Sensors* **2022**, *22*, 587. https://doi.org/10.3390/s22020587

Academic Editors: Paolo Bellavista and Jose F. Monserrat

Received: 24 November 2021
Accepted: 11 January 2022
Published: 13 January 2022

1. Introduction

Wired communications have been widely used in industrial scenarios as new applications of automation and Artificial Intelligence (AI) with high mobility are rolled out. However, wired communications are costly in terms of installation and maintenance and cannot cover new use cases, such as mobility in factories. As a result, the whole industry is shifting towards more flexible and adaptable scenarios, resulting in the Industry 4.0 paradigm. Industry 4.0 is the fourth industrial revolution to improve the flexibility of production and distribution processes where wireless networks have an important part.

The advances in the fields of robotics, AI and Machine Learning (ML) converge in Industry 4.0 to adapt production to new customer demands, such as an increased customization, reduced costs and lower environmental impact [1]. Wireless networks are a major enabler of flexibility in Industry 4.0, allowing easy reconfiguration of production lines, mobile appliances, such as robots both within and beyond the bounds of factories.

The fifth-generation (5G) radio technology, which has been standardized by the 3rd Generation Partnership Project (3GPP) members, aims to provide more flexibility to support new services and applications. Unlike previous technologies, that were focused on traditional mobile broadband, in 5G new services categories have been defined according to their requirements:

- Enhanced Mobile BroadBand (eMBB): this service category is an evolution of traditional mobile broadband, with higher data rates (up to 20 Gbps) and bandwidth. It is similar to the traditional use of networks by users, such as web browsing or streaming multimedia content.
- Massive Machine-Type Communications (mMTC): this service category covers massive connection of devices, with a sporadic and lower volume of data exchange over the network. It is mainly focused on the Internet of Things (IoT).

- Ultra-Reliable and Low Latency Communications (URLLC): this service category aims to cover critical communications, where short messages are exchanged with requirements of lower latency and higher reliability. The latency requirement varies from 1 to 15 ms, depending on the application itself. However, in 5G, it is expected to reach a maximum latency of 1 ms with a reliability target of $1 - 10^{-5}$ for a packet size of 32 bytes at the user plane [2].

URLLC can support use cases, such as closed loop control, hazard sensors, robot automatization, augmented/virtual reality, drone communications and mobile eHealth. The latency requirement varies from 1 to 15 ms, while the reliability targets can range between $1 - 10^{-5}$ and $1 - 10^{-9}$ [3]. These use cases need advanced radio features and resource cost techniques to fulfill highly demanding latency and reliability targets. There are several approaches to achieve such requirements. One technique is the reduction of the time-slot duration by means of a higher numerology [4,5] and by changing the radio resource scheduler [6].

Another solution consists of eliminating steps in the connection protocols to reduce the access time—known as Grant-free transmission [7]. Multi-connectivity [8,9] has been proposed for the sake of achieving high reliability and low latency. In particular, for the sake of achieving high reliability, many solutions have been proposed. The studies in [9,10] are focused on a static threshold (i.e., Reference Signal Received Power (RSRP) or Channel Quality Indicator (CQI)) that determines the dual connectivity range or Packet Duplication (PD) activation.

In [11], eMBB/URLLC multiplexing through preemptive URLLC puncturing was studied, where a Deep Learning Link Adaptation was proposed for eMBB users to maximize the throughput, while ensuring reliability for eMBB users due to corruption of the packets derived from URLLC puncturing. In [12], the PD architecture for carrier aggregation and the dual connectivity approach is presented. The study focuses on how many links are necessary to fulfill reliability and latency requirements for URLLC. A conservative link adaptation algorithm for URLLC is proposed in [13], where the base station keeps statistics of received CQI reports and calculates the maximum channel quality degradation over a time window.

Moreover, in [14] increasing the number of CQI and Modulation Coding Scheme (MCS) values is proposed to reduce the gap between the Signal to Interference plus Noise Ratio (SINR) threshold values and, hence, the radio resource wastage. In [15], two methods for enhancing the efficiency of data duplication over dual connectivity are proposed. In the first method, the duplicates of packets already successfully delivered to the User Equipment (UE) are promptly dropped from the transmission queues of the involved base stations using an uplink indication provided by the UE.

In the second method, the duplication is only performed when the primary base station receives a negative acknowledgement (NACK) associated with the transmission. A Deep Reinforcement Learning method was proposed in [16] to decide which secondary legs to use to duplicate and transmit the packet for the UE within the dual connectivity range. The studies above specifically cover Urban Macro (UMa) and Urban Micro (UMi) scenarios, but not the industrial scenario.

Although these techniques have shown their effectiveness in reducing latency and packet loss, they come at an additional cost, which derives from a less efficient usage of the network resources. The usage of these techniques does not add any benefit if the baseline network can guarantee the requirements of the services/applications at a given moment. Prediction may allow the base station to decide if a redundant technique is required.

This paper proposes a method to predict the End-to-End (E2E) latency as a Service Key Performance Indicator (S-KPI) [17] by observing the network conditions right before a critical transmission in the downlink and, based on that prediction, activate the PD technique (dynamic algorithm) [12,18] in industrial scenarios. The approach for the prediction model followed in this paper is based on the solution shown in [19], where Radio Access

Network (RAN) Key Performance Indicators (KPIs), such as RSRP and Reference Signal Received Quality (RSRQ), are used to predict the S-KPIs using an estimator based on ML.

In particular, the authors in [19] applied the estimator for video Key Quality Indicators (KQIs), such as the average buffer size or number of stalls. Nevertheless, the methodology is independent to the final application and ML algorithm used. In this paper, the same concept is adopted for the prediction of the E2E latency S-KPI, assuming that the main contribution of the latency is at the RAN level. As in 5G, it is expected that Mobile Edge Computing (MEC) [20] will play a central role for URLLC, this assumption is justifiable, since the path of the packets will be limited to the RAN.

The advantage of this approach is that RAN KPIs can be easily measured by the network terminals and infrastructure elements. In this paper, the precision of the estimator is evaluated. The dynamic PD approach using the predictor was evaluated, comparing the E2E latency performance with single connection and a static PD (that is, always duplicating the packet regardless of network conditions) techniques and the resource consumption when using a static vs. a dynamic PD approach (based on the proposed latency predictor).

The remainder of this paper is organized as follows. In Section 2, the materials are described: first, a brief description of cellular networks in Industry 4.0 and critical applications is given in Section 2.1; and in Section 2.2, multi-connectivity in 5G is presented. Then, in Section 3, the proposed system is described. In Section 4, the simulated implementation details are described. The results are shown in Section 5. Finally, our conclusions are drawn in Section 6.

2. Background

2.1. Industrial Networks

2.1.1. Wireless Connectivity in Industry

Wired connections have been widely used in industrial networks, such as ProfiNET, EtherCAT and the set of Time Sensitive Networks (TSN) protocols. Nevertheless, wired infrastructures are costly in terms of installation and maintenance. These are also not suitable for novel Industry 4.0 use cases, such as mobile robots.

In wireless technologies, there is a division between two types of networks that enables different applications. The division is regarding Local Area Networks (LAN) and Wide Area Networks (WAN). LANs have a coverage range of up to 100 m; covering areas, such as rooms or even full factories. The main wireless LAN technologies are based on the IEEE 802.11 family—commonly named WiFi. Moreover, there exist customized solutions for factories, based on IEEE 802.15.1 and 802.15.4, such as Wireless Interface to Sensors and Actuators (WISA) and WirelessHART. These technologies operate in an unlicensed spectrum and suffer from poor scalability [21].

On the other hand, WANs provide a higher coverage range, from distances of a few kilometers up to whole countries. The most extended and known wireless WANs are the 3GPP-based technologies (GSM, GPRS, EDGE, UMTS, LTE and 5G). As cellular networks operate at a licensed spectrum, a better performance can be obtained. Cellular networks provide ubiquitous connectivity beyond the limits of the factory.

This means that production that is geographically distributed can be monitored and managed using a single network. As WANs may be provided as a service by cellular network operators, they do not imply the acquisition, installation and maintenance of infrastructure, allowing multi-tenancy and, therefore, resulting in lower costs [22] and improved connectivity.

Wireless networks generally suffer from a higher latency and packet loss probability than wired networks. For this reason, the recent generations of cellular networks have introduced different optimizations for reducing the latency and also improving reliability.

For 5G, Industry 4.0 is one of the main development verticals, where its applications have been explored for a special network design. One of the main novelties in 5G is the introduction of MEC [20]. MEC consists of moving the application servers to the network

edge, thus, reducing the path that a packet must travel. This approach reduces the latency for the devices and the network load beyond the RAN.

2.1.2. Critical Applications in Industry 4.0

In the Industry 4.0 paradigm, agility is a key objective in the design of factories. Agility means the flexibility of the system to changing requirements by replacing or improving separated modules. Some of the main technologies that allow such agility in factories are the following:

- Rearrangeable modules in production lines [23]: traditionally, production lines have been made up of static modules that perform specific operations. These modules, each controlled by a Programmable Logic Controller (PLC), are interconnected via wired to the Manufacturing Execution System (MES). By enabling the mobility of these modules, new combinations of elements into new types of production lines are possible.
- Automated Guided Vehicle (AGV) [24]: it is common that vehicles driven by workers perform tasks, such as moving stocks and supplies in factories. In the Industry 4.0 paradigm, due to the customization of production, these kinds of movements increase exponentially. It is harder to provide supplies in batches; therefore, smaller vehicles are required with an increase in the number and variety of trips. To achieve this without increasing the workload, AGVs do this without the need for human drivers.
- Drones [25]: drones are a new category of vehicle that enables novel possibilities in factories. Applications, such as emergency assistance, surveillance or rapid point-to-point delivery can be highly optimized with these vehicles.
- Autonomous robots [26]: robots have been extensively adopted in industry since commercial variants have been available. Nevertheless, early iterations of robotics technologies were limited in the number of tasks that they could perform and depended strongly on operators programming them correctly. Currently, AI and ML, along with Simultaneous Location and Mapping (SLAM) and navigation technologies, are enabling novel functionalities on robots that are much more autonomous and perform tasks that were previously reserved for workers.
- Connected workers solutions [27]: the development of consumer electronics in the last years has had a higher impact in the professional area. Gadgets, such as Augmented Reality (AR) glasses, tablets, haptic interfaces and sensors have shown a productivity boost in factories.

All of these technologies rely on wireless connectivity. Some of them, such as production lines, are already mature technologies where a mobility component is added in the Industry 4.0 paradigm. In that cases, wired connections need to be changed to wireless [28].

2.2. 5G Multi-Connectivity Overview

Multi-connectivity in 5G New Radio (NR) inherits from the Long Term Evolution (LTE) Dual Connectivity (DC) concept. LTE DC was first specified in Release 12 [29] and allows UE to simultaneously send/receive data from different evolved NodeBs (eNBs). The data split is performed at the Packet Data Convergence Protocol (PDCP) layer of the transmitting eNB. At the receiving side, the information is decoded from lower layers, and it is combined at the PDCP layer on the receiver. This process allows boosting the throughput [30].

In Release 15, multi-RAT DC was specified for DC operation with NR and LTE nodes [31]. Not only data split but also PD at the PDCP layer is introduced. PD allows the same packet to be transmitted by different nodes, thus, improving the reliability. The nodes are commonly known as the Master Node (MN) and Secondary Node (SN) and are interconnected via a Xn interface. MN is in charge of activating/deactivating PD via Radio Resource Control (RRC) signaling [12]. If the different links are spatially uncorrelated, transmitting the duplicate packet can compensate poor channel conditions. This is very important in industrial scenarios, where the fundamental problems are interference and

multipath propagation, due to the presence of concrete walls and large metallic machinery and structures.

Moreover, NR-NR DC for standalone deployments was standardized in Release 16 [31], in which a UE is connected to one gNB that acts as a MN and another gNB that acts as a SN. In this paper, NR-NR DC with the PD approach is assumed for the downlink direction.

Packet Duplication for URLLC

PD is a multi-connectivity solution that improves reliability by increasing redundancy of the transmission. When PD is activated, the PDCP entity in the MN is responsible for PD, whereas the PDCP entity in the receiver is responsible for detecting and removing duplicated packets. The PDCP entity duplicates the packet data unit (PDU) to avoid twice performing functions, such as ciphering, header compression, integrity protection etc. This PDCP PDU has the same sequence number in both.

Then, the packet is forwarded by MN to the SN via Xn-U interface for transmission to the UE. The packet will undergo through independent Radio Link Control (RLC), Medium Access Control (MAC) and physical layer processing at each gNB. This implies that the packet can be transmitted at different time intervals and over different frequency resources and that physical transmission aspects, such as beamforming, MCS, ACK/NACK signaling and the Hybrid Automatic Repeat Request (HARQ) mechanism, are independent.

On the receiver side, multiple copies of the packet are received, and the UE will forward the first successfully received packet to the higher layers and remove duplicated packets received later, based on the PDCP sequence number. Figure 1 shows a 5G NR-NR PD scheme for downlink transmission.

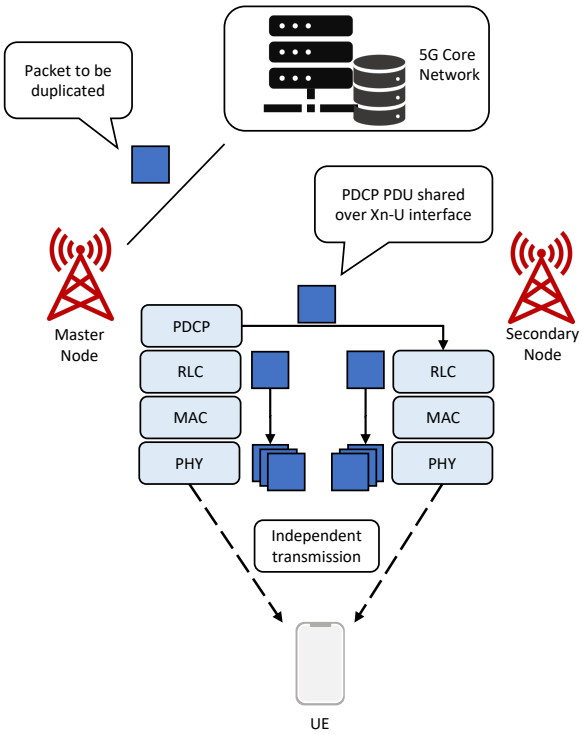


Figure 1. A downlink packet duplication scheme in a NR-NR DC scenario.

As PD improves reliability, it becomes a suitable mechanism for URLLC, which demands a higher reliability and low latency. Not only to improve reliability but also to reduce latency, as independent transmissions are performed by two different gNBs, as mentioned before. In this case, latency reduction will be dependent on the best link. Reliability and latency for URLLC are dependent on each other, that is, high reliability is

consequential only if packets are received within the latency constraint. As a consequence, PD may be used to improve both the reliability and latency.

Nevertheless, PD operates at the cost of wasting resources on SN. Thus, it is not efficient to always send packets duplicated, as there could be situations where the conditions are fulfilled by the primary node. For that reason, it is important to provide a dynamic mechanism for MN that controls the activation/deactivation of PD for URLLC devices, in order to reduce the resource costs at SN, which may affect other UEs attached on SN.

In this paper, we propose the use of a ML predictor, which, based on channel conditions, predicts the E2E latency for URLLC devices in order to activate/deactivate duplication for a packet transmission.

3. Proposal

This section describes the proposed solution in order to assess reliability for URLLC communications by using a dynamic packet duplication algorithm based on ML to reduce the resource consumption.

3.1. System Description

The objective of the system described in this paper is to provide a prediction of the E2E latency to assess the need for using PD solution. The proposed system provides an estimator for downlink transmissions that is trained offline in a server running in the network edge to reduce the computing workload and memory requirements at the master node.

The end devices may be installed in production line elements, AGVs, drones etc. These terminals may need to receive messages with a certain guarantee of latency and reliability. The reliability and latency can be improved by using multi-connectivity, that is, duplicating the packet via two paths, with the cost of dramatically increasing the amount of radio resources spent per transmission.

An alternative way, as presented in this paper, is to first predict whether the network will be able to provide that guarantee without using PD. The MN, right before transmitting, can estimate if a regular transmission, that is, without extra resources, will be sufficient or if PD technique needs to be established. When URLLC devices are not involved in any processor-intensive task, they will measure the latency and report the measurements along with the KPIs to the system where a ML method will update the estimation model.

Figure 2 shows the overall architecture of the system. There are three main domains: the device, which contains the sample collection modules; the server, which runs the ML algorithm and stores a dataset with solved trained cases; and the MN, which contains the estimator. The device needs to receive URLLC messages, and therefore KPIs and S-KPI are forwarded to the server, which runs the ML algorithm and is located at the network edge with computational resources and access to a non-restrictive power source.

The server will collect the data gathered by the devices and generate the ML parameters (prediction model) for the estimator. Finally, the MN uses the estimator to predict the E2E latency of URLLC devices. Based on the prediction, MN will activate/deactivate PD for the current packet transmission.

The functions of the different modules of the system are explained below:

- KPI monitor: collects the KPIs from the radio interface at regular intervals.
- S-KPI monitor: reads the information from the URLLC device and measures the latency.
- Training data collector: joins the data generated by the KPI monitor and the S-KPI monitor. The data joined is used as input to train the ML model.
- Estimator: performs the task of estimating the S-KPI. The primary inputs are the current KPIs (such as SINR, MCS, HARQ feedback etc.) as measured by the MN. The output is the estimation of the E2E latency. This module also has a secondary input that consist in the estimation model extracted from the ML.

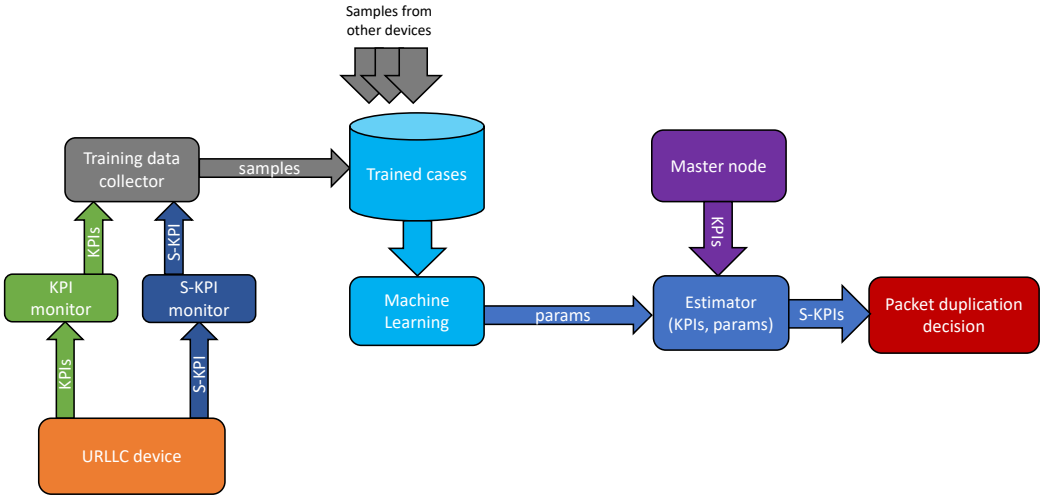


Figure 2. Block diagram of the system.

3.2. KPI to S-KPI Mapping

In cellular networks, performance monitoring (PM) indicators are collected from different points of the network: gNBs, radio interfaces, core network etc. UE traces can also be collected, representing the radio PM indicators collected by the terminals. The PMs are sent to a centralized location where they are analyzed.

Monitoring the network provides many alternatives for detection of problems, analysis of the performance, among others. Nevertheless, these KPIs contains information from lower layers of the network, but not on the performance at application layer in the UEs. Since the development of 5G is centered on the E2E Quality of Service (QoS), this approach has gained importance in recent years. S-KPIs [17] measure these magnitudes. S-KPIs are specific to the final application; that is, for a video transmission, relevant S-KPIs include the average buffer size or the number of stalling, while, for delay-sensitive applications, the main S-KPI is the E2E latency.

The main inconvenience of S-KPI is that they are difficult to measure. They require special procedures at the application layer of the device. The E2E latency can only be measured a posteriori, thus, becoming useless to decide whether a special transmission, such as PD can be performed at a given time.

To address this, a KPI to S-KPI estimator can be used [19]. This technique allows an estimation of the S-KPI based on the available KPIs, which are easy to obtain. Since the KPIs that can be obtained are only a subset of the variables that influence the value of the S-KPI, the estimator will always have a margin for error, and the formula for the mapping will not be trivial. In [19], the approach for creating such estimator is with ML techniques.

In this paper, the hypothesis is that, through this mapping, the E2E latency can be obtained a priori by monitoring the following KPIs:

- Signal to Interference plus Noise Ratio (SINR): includes all the usable signals in the computation. It is used by some vendors to better determine the CQI to adapt the modulation.
- Modulation index: indicates the modulation index used from the table of the MCS when performing a packet transmission. A higher index selects a more efficient modulation, with a higher spectral efficiency and code rate. Otherwise, a lower modulation index selects a more robust modulation, with a lower spectral efficiency and code rate.
- Reception Success: indicates if a packet has been decoded successfully at the receiver or not. This is used to determine if a packet has suffered a HARQ retransmission, since the NACK message is indicated by the receiver to the base station.

The hypothesis assumes that the RAN network is the main contributor to the variable part of the latency. This is due to the jitter being higher at Radio Access Technologies

(RATs) [32] as opposed to the wired networks present in the trunk [33]. When using MEC, the processing is done with the gNB or after a negligible network path, and thus the trunk network component is canceled completely. Latency does not depend directly on the measured KPIs but instead on factors, such as the network load. Nevertheless, in this paper, it is assumed that these KPIs represent the overall status of the RAN connection.

Since the relations between the measured KPIs and the factors that determine the latency are complex, ML is commonly used to find and exploit such relations. In this paper, ML generates the prediction model that encompasses the complex relations among the cited KPIs and the latency.

3.3. Random Forests

In this paper, random forests [34] were selected for the implementation of the ML part of the system. The choice of this method is based on the computation simplicity once the model has been trained, since fast prediction for URLLC services must be performed.

Random forests are an ensemble method commonly used to resolve several types of ML learning problems, such as classification and regression. A random forest consists of a set of decision trees. Each decision tree takes the input to the forest and returns an estimated value. The structure of the tree is created in the training process. On each tree, a decision is performed by comparing the input with a threshold. Based on the output of the comparison, a new comparison is performed with a different input and threshold, which determines the prediction of the tree.

To improve the accuracy, each decision tree output is aggregated. For regression, the aggregation method is typically the average of the output of all the trees. A summarized scheme of the random forests prediction is shown in Figure 3.

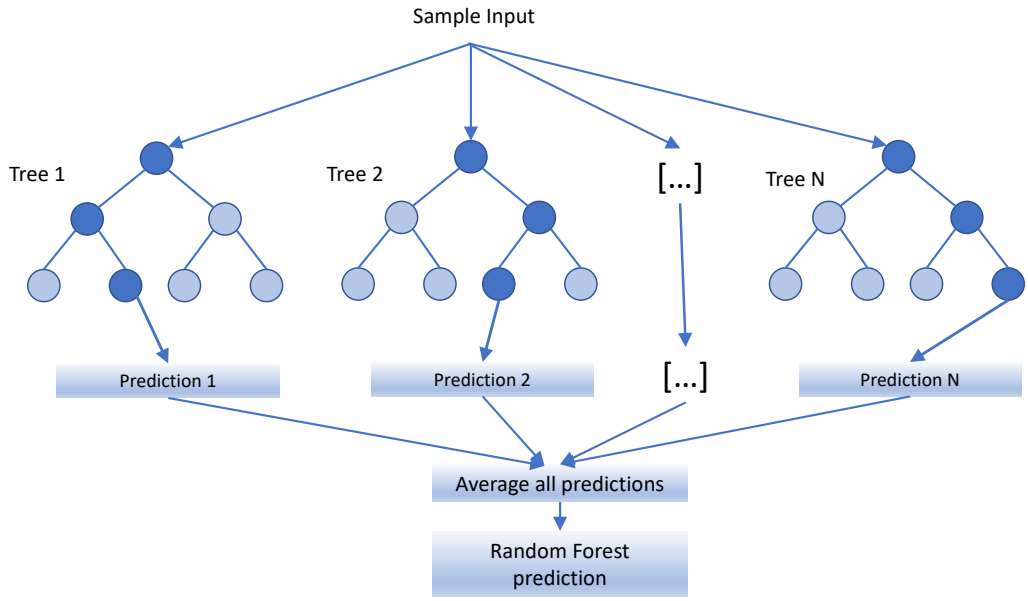


Figure 3. Random forest prediction scheme.

Moreover, several parameters must be adjusted before starting the execution, related to the performance of both the decision trees and the complete set that builds the forest. Some of the most important configuration parameters are the following:

- Number of decision trees: this establishes the number of trees that constitutes the forest. This must be chosen in relation to the input dataset to avoid overhead.
- Bootstrap: this parameter decides how each tree is built independently. If it is not activated, the complete dataset is used for each tree. Otherwise, the initial dataset is divided into subsets of dataset for each tree.

- Division criterion: this defines the quality of a split according to the condition set in the node. The most used criteria for regression are the squared error and absolute error.
- Maximum leaves per tree: this sets the maximum depth of the tree in the forest.
- Maximum samples to split: this determines the maximum number of samples to consider in order to choose the condition that determines the split.
- Minimum samples to split: this determines the minimum samples needed to consider a new split.

A training method is applied over a set of labeled samples, (vectors with inputs to the system and the expected output) to obtain the trees. The training process [34] consists of randomly selecting a subset of features and a subset of training samples for each tree and then training it using the CART [35] algorithm without pruning. The random selection of samples contributes in avoiding overfitting. Once the model has been trained, an evaluation phase is performed, where only the values of the different inputs are considered, and the algorithm provides the output. Then, the output of the model and the original are compared in order to estimate the accuracy of the model.

In the ML scheme proposed in this paper, the labeled samples are collected in the UEs and transmitted to a server located at the network edge, where different datasets are collected. The collection of different datasets helps the KPI to S-KPI mapping by capturing the effects of contextual variables. The ML process is executed in a server, and the estimator is used by the MN to predict the S-KPI. In particular, the inputs of the algorithm are the SINR, the modulation index and the reception success, as previously indicated in Section 3.2. The output will be the E2E latency experienced by the URLLC devices.

3.4. Implementation Considerations

While the measurements in this paper were performed on simulations, it is important to discuss the issues that arise when this implementation is ported to the real world, in a real MN and stock UEs. Specifically, in this subsection, the aspects on data collection requirements, the need for retraining and the hardware and software requirements both on the MN and the UE are discussed.

Regarding the data, the ML algorithm requires data on KPIs and S-KPIs. KPIs are easily collected by the Network Monitoring System (NMS), which is part of the maintenance systems of the MN. The S-KPIs are harder to obtain, since they require the collection of data in the UEs. Therefore, to obtain real values, a measurement campaign must be done, including UEs with the appropriate software for capturing and transmitting the S-KPIs.

Another important aspect is the need for retraining the model to adapt it to changes in the environment. The validity of the model depends on variations in the behavior of the features that are used in it, which determine the relation between the independent and dependent variables. Second-order effects, such as contextual factors that are not included in the model (e.g., geometry of the buildings, weather conditions, power supply variations etc.). These effects are out of the scope of the tests in this paper; however, on real tests, the required frequency of retraining would be an important parameter to study.

To implement the proposed functionality in reality, some hardware and software requirements are posed over both the MN elements and the UEs. Specifically, in the MN, the software for data collection (KPIs and S-KPIs), as well as the ML algorithm and the estimator (described in Figure 2) must be implemented in some element of the network. An important aspect to take into account is the latency of the decision. There are two possible implementations:

- Implement the data collection and ML stages in the network core. The main advantage is the availability of large datasets that add diversity to the final model. Another advantage of this method is that cloud computing resources can be used better. This is even more important when looking ahead to future 6G networks, where network elements in the core network for ML and AI are envisioned.
- Implement everything in the network edge. In this case, to gain diversity, a Federated Learning (FL) mechanism can be used to share model parameters between different

agents. FL is the collaborative learning, which trains and updates the model through the joint effort of multiple servers that are deployed in a decentralized manner within the network.

In both of these cases, the estimator will run in the network edge to minimize latency in the decision, the KPI collection will be done in the network edge due to the nature of the task, and the S-KPI collection will be done in the terminals. All these software elements must be supported by the appropriate hardware equipment. While ML is hardware intensive, the estimator has a low demand in resources.

Regarding the UE, the hardware and software requirements are quite low. Specifically, the devices must implement a functionality for capturing S-KPI information and sending it to the MN (to wherever the data collection function is implemented; either in the edge or the core). While the requirement is quite low in terms of hardware and software, it has some complications in the form of privacy and confidentiality issues. Therefore, it is expected that the use of UEs with the appropriate software is limited to a reduced set of devices acting over a predefined period of time.

Ownership of the network will also play a central role for this aspect; if the user (in the case of industry, the owner of the factory) is also the owner of the network equipment where the data is collected (edge or core), then it is likely that the rate of penetration of the S-KPI probes is higher in the installed base of UEs and more stable in time. Therefore, this would also be an advantage of the system architecture where the data collection and machine learning is done in the edge, and the model parameters are then shared with FL.

4. Tests

The proposed scheme was tested in a simulated 5G network, using ns-3, which is a free and open-source network simulator that is very popular in research [36]. In particular, the 5G-LENA module [37] was selected to conduct the simulations. This module focuses on the new 3GPP NR specifications and includes numerology support, frequency division multiplexing of numerology and an OFDMA-based scheduler. It also includes beamforming and HARQ feedback implementation.

In this section, the simulation scenario along with the KPI recollection phase to train the model are described.

4.1. Simulation Scenario

The scenario consists in an indoor factory, with an area of 4800 m^2 and a height of 10 meters. The scenario is based on the one proposed in 3GPP 38.901 (Table 7.8-7) [38], which was used to calibrate the indoor factory scenario defined in Release 16. In particular, the Indoor Factory with Dense clutter and High base station (InF-DH) [38] scenario was selected with a clutter height of 6 m and clutter density of 80%.

Attending to the RAN part, there are two picocells with a height of 8 m, which are interconnected via Xn interface, with a base station distance of 50 m. Both gNBs operate with a frequency of 3.7 GHz and a bandwidth of 20 MHz. One transmission/reception omnidirectional antenna was used in both, picocells and UEs, with 23 dBm as downlink transmission power. Figure 4 shows the distribution of the scenario simulated.

The slot length configured is set to 0.25 ms, which corresponds to numerology 2, as defined in the standard [39]; whereas the number of HARQ retransmissions attempts was set to a maximum of 1, due to URLLC latency constraints. Moreover, the link adaptation used at gNB for MCS selection is an error model-based, where the MCS is selected to meet a target transport Block Error Rate (BLER). The MCS table used is Table 1 (up to 64-QAM) from 3GPP 38.214 [40]. The delay for the scheduling procedure was set as following:

- The packet processing from MAC to PHY layer is fixed at two slots. This is a delay between the control/data acquisition from the RLC layer by the MAC layer and the moment at which the data is available to go over the air.

- The transport block decode latency is set to 100 microseconds at UE and gNB. It is a delay between the data acquisition from the air by the PHY layer and the moment at which the data block is available to process at the MAC layer.
- The processing delay needed to decode Downlink Control Information (DCI) and decode downlink data is set to 0 slots.
- The processing delay needed from the end of downlink data reception to the earliest possible start of the corresponding ACK/NACK transmission is set to 1 slot.

On each base station, there are 15 UEs attached, whose positions are fixed and randomly selected at the beginning of the simulation. For each UE, a Constant Bit Rate (CBR) flow of 1 Mbps is sent by the gNB in downlink direction.

On the other hand, there is a URLLC device that is connected to both gNBs simultaneously, one acting as a MN and the other one as a SN. The initial position of the URLLC device is random over the scenario, using a uniform distribution. Then, when the simulation begins, the URLLC device selects a random direction (where all angles are equiprobable, since a uniform distribution is used) over 360 degrees and maintains that direction during 10 s with a speed of 2 m/s.

When the timer expires, a new direction is selected. The URLLC device represents an AGV that is remotely controlled by the network. To do this, the network sends short commands, in this case, UDP packets with a periodicity of 10 ms and size of 64 bytes.

In this paper, a MEC [20] is considered to allocate the URLLC service at the network edge. Thus, the main contribution of the latency comes from the RAN. Table 1 shows the main configuration parameters of the simulations.

Table 1. The main configuration parameters.

Parameter	Value
Channel and propagation loss model	3GPP 38.901
System bandwidth	20 MHz
Center frequency	3.7 GHz
Numerology	2
Scenario	InF-DH
Transmission direction	Downlink
Modulation	Adaptive
Scheduler	Round-Robin
UE height	1.5 m
gNB height	8 m
Transmission power	23 dBm
Xn interface delay	100 µs
MAC to PHY delay	2 slots
Transport block decode latency	100 µs
HARQ feedback delay	1 slot
HARQ retranmission attempts	1
Packet size	64 bytes
Packet interval	10 ms

4.2. KPIs Recollection

The first part of the test consists in obtaining a dataset from URLLC devices—that is, a collection of KPIs in order to train the ML part. To do this, a simulation was performed, where an URLLC device moves from the entire scenario and recollects different KPIs. This movement along with the scenario setup is shown in Figure 4.

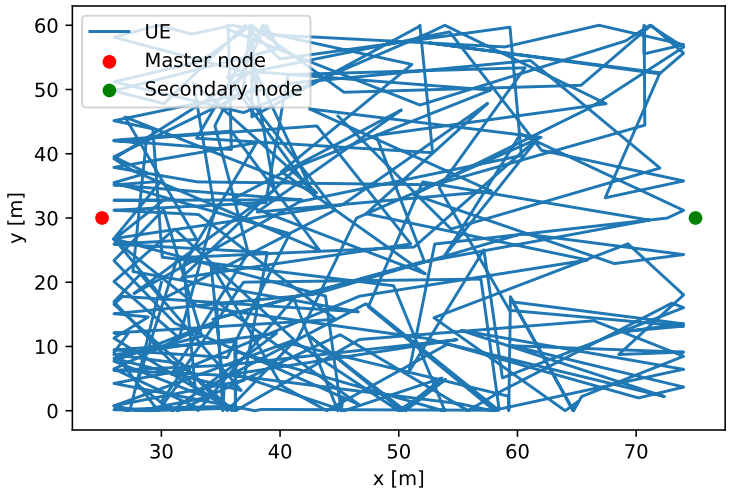


Figure 4. UE movement over the entire scenario.

Upon a packet reception at UE, it knows the SINR received if the packet was not decoded successfully and the modulation index used in the transmission. These KPIs along with the latency measured are the inputs to train the ML model. Based on these samples, it is possible to train the model when there are sufficient samples.

Once the ML model was trained, the latency predictor is used by the MN. That is, based on the actual conditions, such as SINR, modulation index and HARQ feedback of the previous packet, the estimator predicts the E2E latency. Since URLLC transmission intervals are very short, we assume that the channel conditions are the same between the current packet and the previous (since $\Delta T > T_{coh}$, $T_{coh} \approx \frac{\lambda/2}{v} \approx 20$ ms is the coherence time during which channel conditions are stable, where λ is the wavelength and v is the speed), which is why HARQ feedback from the previous packet was selected as an input.

Based on the output of the estimator, the MN decides whether to duplicate or not the actual packet taking into account a latency threshold. In this case, the latency threshold was set at 2 ms.

5. Results and Discussion

This section shows the results obtained by the test explained in Section 4. The first part of the test consists in obtaining a collection of KPIs in order to train the ML part. This is explained in Section 4.2. Once the ML model has been trained, the second part of the test consists in evaluating the accuracy of the predictor, that is, the accuracy of predicting the latency by MN.

Moreover, new simulations upon the scenario presented in the previous section have been done, comparing the latency performance and resources consumption between not using PD (single connection), always duplicating the packet and the dynamic packet duplication based on the prediction model (which has been trained before, as explained in Section 4.2), where the duplication is performed when the latency predicted by the model is higher than a threshold. Upon this test, latency threshold of 2 ms was chosen.

5.1. Prediction Results

This subsection shows the results obtained by the predictor when using the test samples, to measure the accuracy of the predictor. In the test, there are 59,940 samples—packets transmitted by the MN to the URLLC device. Figure 5 shows the distribution of the samples. In general, these distributions show that, as expected, with worse radio conditions, a higher latency is experienced. This is because, with high attenuation, retransmissions at lower layers increase.

The latency with SINR at intermediate values seems to be below or above 2 ms. The increase on the latency is due to HARQ retransmissions that occur. These retransmissions occur when there is a suddenly drop of the SINR, commonly in industrial scenarios, and the modulation index used is still higher (that drop of the SINR was not expected).

This effect is clearly visible in the latency of the modulation index, where robust modulation decrease the probability of HARQ retransmission. Nevertheless, the higher the modulation index is, the higher the probability of retransmission is, as SINR drops occurs. Finally, the latency when the packet decodification was not successful at the PHY layer at UE (that is, a retransmission needs to be performed) is higher, whereas when the reception was successful when the latency was lower as expected.

Table 2 shows the results of the estimator in the validation part. In particular, the false positive rate, the false negative rate and the success rate are given for the predictor for latency. As mentioned above, the latency predictor estimates whether the latency will be higher than 2 ms.

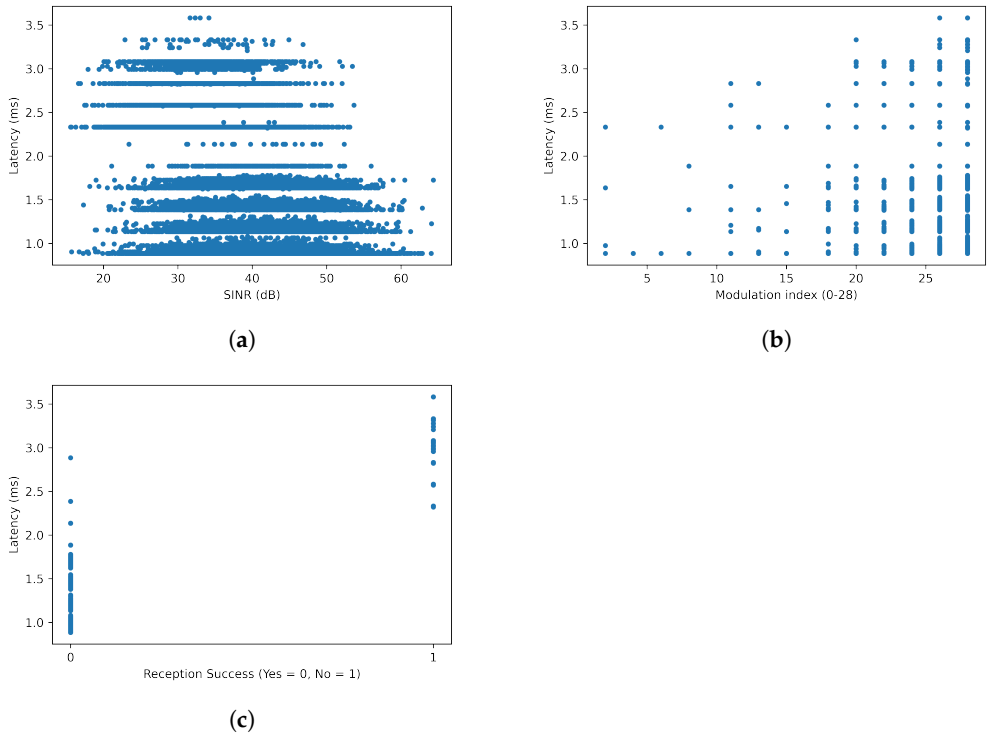


Figure 5. Latency samples. (**a**) SINR, (**b**) Modulation index and (**c**) Reception Success.

The false positive rate indicates the proportion of times where the actual latency was lower than the threshold but the estimator predicted it would be higher. In this case, the false positive rate is 0.0041%. The false negative rate indicates the proportion of times where the actual latency was higher than the threshold and the estimator failed at predicting this. Ideally, this value should be close to zero. In the results, the latency estimator has a 0.0615% of error rate. The success rate indicates the percent of times where the value of latency was correctly predicted to be either below or above the threshold. In the results, the latency estimator has a 99.9849% of success rate.

Table 2. Prediction results.

S-KPI	False Positive Rate	False Negative Rate	Success Rate
Latency	0.0041%	0.0615%	99.9849%

5.2. Packet Duplication Results

This subsection presents the results obtained when the ML algorithm was trained and the PD approach is used. In particular, a latency comparison between not using PD (single connection), always duplicating the packet (always PD) and the proposed dynamic PD algorithm is performed. Furthermore, for the PD techniques, a comparison of the resource consumption and latency obtained is performed.

Figure 6 shows the empirical cumulative distribution function (ECDF) of the latency obtained when PD is not used (single connection), when always duplicating the packet and when activating PD based on the latency prediction using Random Forest, which is the system proposed in this paper. First of all, as it can be seen, the random forest and always PD distributions are similar, and they converge at 1.75 ms. The probability of receiving lower values of latency (below 1 ms) is 70%, 62% and 55% for always PD, random forest and single connection, respectively. There is a remarkable difference on the probability of receiving a latency below 2 ms between both PD methods and by transmitting via single connection. In this case, there is a gap of 20% on the probability.

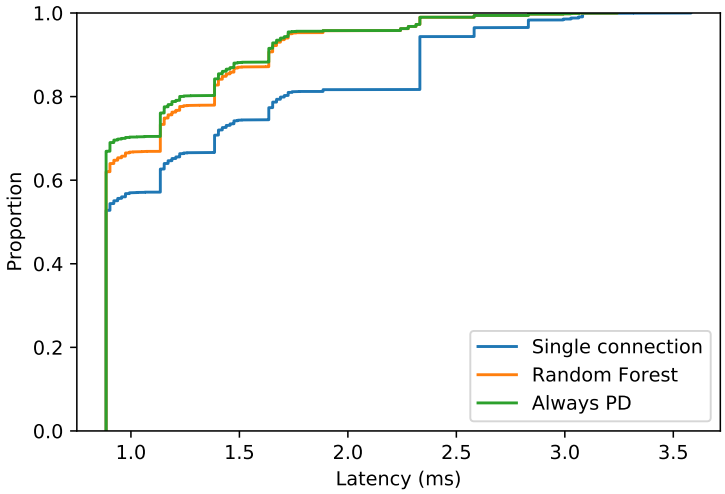


Figure 6. ECDF of the latency received.

The latency below the threshold rate when using PD and not using PD is shown in Table 3, where PD techniques demonstrate that improve the reliability by duplicating the packet incoming for the UE.

Table 3. Latency below the threshold rate for the different techniques.

Technique	Latency below Threshold Rate
Single connection	81.6549%
Always PD	95.7891%
PD via Random Forest	95.7541%

The latency gain distribution when PD is activated by the prediction model is shown in Figure 7. The distribution shows that, when the estimator predicts that latency will be higher than the threshold, in general, there is a latency gain compared to sending over a single link. The latency gain helps to improve the reliability of URLLC communications. The vertical dashed line represents the 75% percentile. In this case, the probability of reducing the latency more than 1 ms is 75%. Otherwise, the probability of not reducing the latency so far, that is, between 0.1 and 0.2 ms, is below 20%.

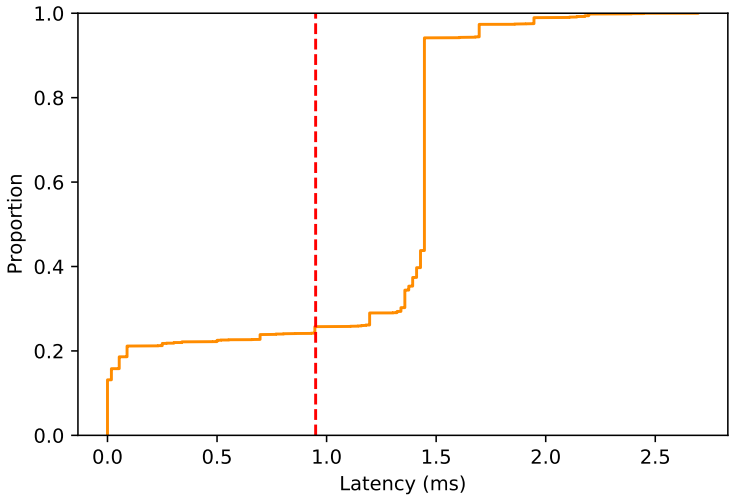


Figure 7. ECDF of the latency gain when the predictor activates PD.

Table 4 shows the results obtained when using a static PD technique, that is, always duplicating the packet transmitted and the results obtained when using PD only when the latency prediction is above the threshold. In particular, the number of packets that have been sent duplicated, the latency below the threshold rate, the average (packet) latency reduction rate and the PD reduction rate.

Taking into account the number of packets duplicated, it is remarkable that, when using random forest, the number of packets duplicated is lower, specifically, the PD reduction rate is 81.0211%. Both techniques present a similar latency below the threshold rate, with 95.7891% for always PD and 95.7541% when using random forest. That clearly indicates that a dynamic PD is more suitable, guaranteeing the same level of low latency but without wasting extra resources.

The average (packet) latency reduction rate measures the proportion of times where latency was reduced when duplicating the packet. In the results, the average (packet) latency reduction rate is better when duplicating via random forest than always duplicating, obtaining a latency reduction rate of 86.5506% for the system proposed and 25.0917% for a static duplication. The lower rate when always duplicating is due to sending the packet via two path when conditions are favorable at MN—that is, the latency constraint is fulfilled by MN link.

Table 4. Comparison results between static and dynamic PD.

PD Technique	Number of Packets Duplicated	Latency below Threshold Rate	Average (Packet) Latency Reduction Rate	PD Reduction
Always PD	59,940	95.7891%	25.0917%	Not applicable
PD via Random Forest	11,376	95.7541%	86.5506%	81.0211%

6. Conclusions

Many solutions have been proposed for URLLC connectivity in order to achieve high reliability and low latency; however, these typically come at a cost in terms of resource usage. This paper proposes a scheme to predict the E2E latency in order to determine if the PD technique is required for downlink transmissions (dynamic PD) and, thus, reduce the resource wasting.

The proposed algorithm uses a ML approach, where an estimator is running in the MN. The algorithm was implemented and tested in a 5G simulator, showing high accuracy for determining whether or not to activate PD for a packet transmission.

Moreover, the PD technique was evaluated using the predictor (once the model is already trained). The results obtained show that the proposed dynamic PD based on the

E2E latency prediction is more efficient than always duplicating, obtaining a high PD reduction rate (81%) and maintaining the same level of latency below the threshold.

In addition, when activating PD based on the predictor, the latency reduction is higher than when using always duplicating technique, that is, there are so many unnecessary packets duplicated, where the latency constraint can be fulfilled by a single connection. A comparison between PD techniques and a single connection was performed, where PD techniques demonstrate that help to achieve URLLC latency constraint.

As a future line of work, some of the implementation aspects will be studied. A major milestone toward the real implementation of the system will be the study of FL for the creation and enrichment of a model. Another important aspect will be the study of the required model update frequency, which may also be reduced due to FL.

Author Contributions: Conceptualization, D.S. and E.J.K.; methodology, D.S. and E.J.K.; software, D.S.; validation, D.S. and E.J.K.; formal analysis, D.S. and E.J.K.; investigation, D.S. and E.J.K.; writing—original draft preparation, D.S. and E.J.K.; writing—review and editing, D.S., E.J.K. and R.B.; visualization, D.S.; supervision, R.B.; project administration, R.B.; funding acquisition, R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially funded by Junta de Andalucía (Consejería de Transformación Económica, Industria, Conocimiento y Universidades, Proyecto de Excelencia PENTA, P18-FR-4647 and EDEL4.0, UMA18-FEDERJA-172) and ERDF.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lasi, H.; Fettke, P.; Kemper, H.G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [CrossRef]
2. 3GPP. TR 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies; V14.3.0, Rel-14. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2996 (accessed on 28 September 2021).
3. 5G Americas. New Services & Applications with 5G Ultra-Reliable Low Latency Communications; 5G Americas White Paper, November 2018. Available online: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_URLLLC_White_Paper_Final__updateJW.pdf (accessed on 21 December 2021).
4. Zaidi, A.A.; Baldemair, R.; Tullberg, H.; Bjorkegren, H.; Sundstrom, L.; Medbo, J.; Kilinc, C.; Da Silva, I. Waveform and Numerology to Support 5G Services and Requirements. *IEEE Commun. Mag.* **2016**, *54*, 90–98. [CrossRef]
5. Segura, D.; Khatib, E.J.; Munilla, J.; Barco, R. 5G Numerologies Assessment for URLLC in Industrial Communications. *Sensors* **2021**, *21*, 2489. [CrossRef] [PubMed]
6. Pedersen, K.; Pocovi, G.; Steiner, J.; Maeder, A. Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations. *IEEE Commun. Mag.* **2018**, *56*, 210–217. [CrossRef]
7. Jacobsen, T.; Abreu, R.; Berardinelli, G.; Pedersen, K.; Mogensen, P.; Kovacs, I.Z.; Madsen, T.K. System Level Analysis of Uplink Grant-Free Transmission for URLLC. In Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 4–8 December 2017; pp. 1–6.
8. Khatib, E.J.; Wassie, D.A.; Berardinelli, G.; Rodriguez, I.; Mogensen, P. Multi-Connectivity for Ultra-Reliable Communication in Industrial Scenarios. In Proceedings of the 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, 28 April–1 May 2019; pp. 1–6.
9. Mahmood, N.H.; Lopez, M.; Laselva, D.; Pedersen, K.; Berardinelli, G. Reliability Oriented Dual Connectivity for URLLC services in 5G New Radio. In Proceedings of the 2018 15th International Symposium on Wireless Communication Systems (ISWCS), Lisbon, Portugal, 28–31 August 2018; pp. 1–6.
10. Rayavarapu, S.M.; Amuru, S.D.; Kiran, K. Dynamic Control of Packet Duplication in 5G-NR Dual Connectivity Architecture. In Proceedings of the 2020 International Conference on COMmunication Systems NETworkS (COMSNETS), Bengaluru, India, 7–11 January 2020; pp. 539–542.
11. Huang, Y.; Hou, Y.T.; Lou, W. A Deep-Learning-based Link Adaptation Design for eMBB/URLLC Multiplexing in 5G NR. In Proceedings of the 2021 IEEE Conference on Computer Communications (INFOCOM), Vancouver, BC, Canada, 10–13 May 2021; pp. 1–10.

12. Rao, J.; Vrzic, S. Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance Analysis. *IEEE Netw.* **2018**, *32*, 32–40. [CrossRef]
13. Belogaev, A.; Khorov, E.; Krasilov, A.; Shmelkin, D.; Tang, S. Conservative Link Adaptation for Ultra Reliable Low Latency Communications. In Proceedings of the 2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Sochi, Russia, 3–6 August 2019; pp. 1–5.
14. Khan, J.; Jacob, L. Link Adaptation for Multi-connectivity Enabled 5G URLLC: Challenges and Solutions. In Proceedings of the 2021 International Conference on COMmunication Systems and NETworkS (COMSNETS), Bangalore, India, 5–9 February 2021; pp. 148–152.
15. Centenaro, M.; Laselva, D.; Steiner, J.; Pedersen, K.; Mogensen, P. Resource-Efficient Dual Connectivity for Ultra-Reliable Low-Latency Communication. In Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 25–28 May 2020; pp. 1–5.
16. Zhao, Q.; Paris, S.; Veijalainen, T.; Ali, S. Hierarchical Multi-Objective Deep Reinforcement Learning for Packet Duplication in Multi-Connectivity for URLLC. In Proceedings of the 2021 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit), Porto, Portugal, 8–11 June 2021; pp. 142–147.
17. Lorca, J.; Solana, B.; Barco, R.; Herrera-Garcia, A.; Palacios, D.; Fortes, S.; Demestichas, P.; Kosmatos, E.; Georgakopoulos, A.; Stavroulaki, V.; et al. Deliverable D2.1: Scenarios, KPIs, Use Cases and Baseline System Evaluation. Technical Report, E2E-Aware Optimizations and Advancements for Network Edge of 5G New Radio (ONE5G). 2017. Available online: https://one5g.eu/wp-content/uploads/2017/12/ONE5G_D2.1_finalversion.pdf (accessed on 28 September 2021).
18. Aijaz, A. Packet Duplication in Dual Connectivity Enabled 5G Wireless Networks: Overview and Challenges. *IEEE Commun. Stand. Mag.* **2019**, *3*, 20–28. [CrossRef]
19. Herrera-Garcia, A.; Fortes, S.; Baena, E.; Mendoza, J.; Baena, C.; Barco, R. Modeling of Key Quality Indicators for End-to-End Network Management: Preparing for 5G. *IEEE Veh. Technol. Mag.* **2019**, *14*, 76–84. [CrossRef]
20. Hu, Y.C.; Patel, M.; Sabella, D.; Sprecher, N.; Young, V. Mobile Edge Computing—A key technology towards 5G. *ETSI White Pap.* **2015**, *11*, 1–16.
21. Hasan, S.; Ben-David, Y.; Bittman, M.; Raghavan, B. The Challenges of Scaling WISPs. In Proceedings of the 2015 Annual Symposium on Computing for Development (DEV'15). Association for Computing Machinery, New York, NY, USA, 1–2 December 2015; pp. 3–11.
22. Rostami, A. Private 5G Networks for Vertical Industries: Deployment and Operation Models. In Proceedings of the 2019 IEEE 2nd 5G World Forum (5GWF), Dresden, Germany, 30 September–2 October 2019; pp. 433–439.
23. Weyer, S.; Schmitt, M.; Ohmer, M.; Gorecky, D. Towards Industry 4.0—Standardization as the crucial challenge for highly modular, multi-vendor production systems. *IFAC-PapersOnLine* **2015**, *48*, 579–584. [CrossRef]
24. Mehami, J.; Nawi, M.; Zhong, R.Y. Smart automated guided vehicles for manufacturing in the context of Industry 4.0. *Procedia Manuf.* **2018**, *26*, 1077–1086. [CrossRef]
25. Fernández-Caramés, T.M.; Blanco-Novoa, O.; Froiz-Míguez, I.; Fraga-Lamas, P. Towards an Autonomous Industry 4.0 Warehouse: A UAV and Blockchain-Based System for Inventory and Traceability Applications in Big Data-Driven Supply Chain Management. *Sensors* **2019**, *19*, 2394. [CrossRef] [PubMed]
26. Gonzalez, A.G.; Alves, M.V.; Viana, G.S.; Carvalho, L.K.; Basilio, J.C. Supervisory control-based navigation architecture: A new framework for autonomous robots in industry 4.0 environments. *IEEE Trans. Ind. Inform.* **2017**, *14*, 1732–1743. [CrossRef]
27. Paelke, V. Augmented reality in the smart factory: Supporting workers in an industry 4.0. environment. In Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA), Barcelona, Spain, 16–19 September 2014; pp. 1–4.
28. Mogensen, R.S.; Rodriguez, I.; Berardinelli, G.; Fink, A.; Marcker, R.; Markussen, S.; Raunholt, T.; Kolding, T.; Pocovi, G.; Barbera, S. Implementation and Trial Evaluation of a Wireless Manufacturing Execution System for Industry 4.0. In Proceedings of the IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 22–25 September 2019; pp. 1–7.
29. 3GPP. TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2; V12.10.0, Rel-12. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2430 (accessed on 17 September 2021).
30. Rosa, C.; Pedersen, K.; Wang, H.; Michaelsen, P.-H.; Barbera, S.; Malkamäki, E.; Henttonen, T.; Sébire, B. Dual connectivity for LTE small cell evolution: Functionality and performance aspects. *IEEE Commun. Mag.* **2016**, *54*, 137–143. [CrossRef]
31. 3GPP. TS 37.340, Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-Connectivity; Stage 2; V16.0.0, Rel-16. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3198 (accessed on 17 September 2021).
32. Kassim, M.; Rahman, R.A.; Aziz, M.A.A.; Idris, A.; Yusof, M.I. Performance analysis of VoIP over 3G and 4G LTE network. In Proceedings of the 2017 International Conference on Electrical, Electronics and System Engineering (ICEESE), Kanazawa, Japan, 9–10 November 2017; pp. 37–41.
33. Alderisi, G.; Iannizzotto, G.; Bello, L.L. Towards IEEE 802.1 Ethernet AVB for Advanced Driver Assistance Systems: A preliminary assessment. In Proceedings of the 2012 IEEE 17th International Conference on Emerging Technologies Factory Automation (ETFA 2012), Krakow, Poland, 17–21 September 2012; pp. 1–4.
34. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
35. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

36. NS-3-A Discrete-Event Network Simulator for Internet Systems. Available online: https://www.nsnam.org/ (accessed on 20 September 2021).
37. Patriciello, N.; Lagen, S.; Bojovic, B.; Giupponi, L. An E2E simulator for 5G NR networks. *Simul. Model. Pract. Theory* **2019**, *96*, 101933. [CrossRef]
38. 3GPP. TR 38.901, Study on Channel Model for Frequencies from 0.5 to 100 GHz; V16.1.0, Rel-16. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3173 (accessed on 20 September 2021).
39. 3GPP. TR 21.915, Release Description; Release 15; V15.0.0, Rel-15. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3389 (accessed on 22 December 2021).
40. 3GPP. TS 38.214, NR; Physical Layer Procedures for Data, Release 16; V16.7.0, Rel-16. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216 (accessed on 22 December 2021).

Evaluation of Mobile Network Slicing in a Logistics Distribution Center

David Segura, Emil J. Khatib, *Member, IEEE,* Raquel Barco

***Abstract*—Logistics is a key economic sector where any optimization that reduces costs or improves service has a great impact on society at large. Network Slicing (NS) is a technique that allows the creation of different independent networks with different dedicated resources on a shared physical infrastructure. This is particularly useful in scenarios where different applications with different requirements coexist. In this paper, a novel open-source simulator based on NS-3 has been developed with a realistic representation of a distribution center scenario, including the logistics activities that take place there. Under this developed simulator, the role of two 5G NS strategies in Smart Logistics is studied: the use of a static slice with a balance division of network resources and the use of a dynamic slice. These strategies have been evaluated in terms of Quality of Service (QoS) for different traffic profiles via simulations. Results show that a dynamic slice makes a more efficient usage of the network resources, improving the QoS for the different traffic profiles, even when there is a traffic peak. This improvement ranges from 6.48% to 95.65%, depending on the specific traffic profile and the evaluated metric.**

***Index Terms*—5G, Industry 4.0, Logistics, Mobile Networks, Network Optimization, Network Slicing, simulator**

I. Introduction

In the last years, the emergence of networks and mobile communications has led to the development of new solutions that have revolutionized logistics. Wireless networks are one of the key enablers of Smart Logistics [1], which allow the supply of products in small or individual batches with Just-In-Time delivery, reverse logistics, continuous feedback to clients, etc. Distribution centers [2] are a key element in the Smart Logistics supply chain, replacing traditional warehouses with lean nodes that act more as post offices where products spend few hours before being shipped in the next transport mean.

To support the Smart Logistics supply chain, there are numerous Industry 4.0 applications. For instance, Automated Guided Vehicles (AGVs [3]) are used within distribution centers to move packages between different points; and Smart Tags [4] are used to track parcels at all times within a distribution center or even during transportation between different centers. The different applications running in Smart Logistics will have different requirements depending on the criticality of the messages, their size and the number of devices transmitting a message simultaneously. In the fifth-generation (5G) of cellular networks, three main traffic profiles are defined [5]:

- Enhanced Mobile Broadband (eMBB): messages that require a high bandwidth. Typically associated with multimedia applications, such as Augmented and Virtual Reality (AR/VR [6]).
- Ultra Reliable Low Latency Communications (URLLC): messages that require a very high reliability and very low latency. Normally, mission critical messages belong to this category, such as AGV navigation systems.
- Massive Machine Type Communications (mMTC): short, infrequent messages with low requirements for reliability and latency, but with a massive density of devices and a need for low power consumption. Applications such as Smart Tags belong to this category.

These traffic profiles are all present in Smart Logistics [7]. To allow such profiles with conflicting requirements to coexist on a single wireless network, 5G introduces Network Slicing (NS [5]). With NS, the resources (physical machines, software, radio spectrum, etc.) are divided dynamically into independent sets with optimized configurations. A slice for each class of service can then be defined such that its requirements are met without negatively affecting other service classes.

Ranging from high-level studies of wireless applications in Smart Logistics [1], [8], to specific use-cases [9]–[11], the topic of the application of 5G technologies on logistics is gaining an increasing interest from the research community. Several studies [7], [12]–[16] are centered in optimizing 5G networks specifically for Smart Logistics, analyzing the specific particularities of the applications and the different environments where the processes take place.

The use of NS [17] has been widely agreed as a promising technique to accommodate diverse services that occur in industrial scenarios, such as distribution centers. In [18], the architecture and requirements of NS for smart factory is presented, along with the different challenges and implementation aspects. In [19], the authors perform an evaluation of the bandwidth utilization and the number of connected users, using different NS strategies. A proactive NS for Smart Logistics is proposed in [7] to adapt the radio resources into the different slices. This method is based on a Big Data prediction module to predict the traffic within a base station and then divide the resources proportionally among the expected traffic profiles.

This paper provides a study of network optimization in logistics scenarios through simulations. To the best of the authors knowledge, there are no practical studies covering the optimization of 5G technology in logistics scenarios. The use of 5G technology in logistics has been previously proposed at a high level, but without providing network performance results under this particular scenario for the different traffic profiles

The authors are with Telecommunication Research Institute (TELMA), Universidad de Málaga, E.T.S. Ingeniería de Telecomunicación, Bulevar Louis Pasteur 35, 29010, Málaga (Spain) (e-mail: dsr@ic.uma.es, emil@uma.es, rbm@ic.uma.es). (Corresponding author: Emil J. Khatib.)

(eMBB, URLLC and mMTC). Therefore, to cover this need, the key contributions of this paper are the following:

1) **Development of a novel open-source simulator.** A novel open-source simulator based on NS-3[1] has been developed with a realistic representation of a distribution center scenario, where several different logistics activities are done. The communications of these activities have been modeled and used to estimate the performance of the different traffic profiles. This simulator serves as the foundation for studying the impact of different NS strategies on Smart Logistics. Moreover, a Python-based open-source simulator[2] has been developed to evaluate the performance of the random-access procedure.
2) **Comparison of two 5G NS strategies.** This study focuses on comparing the effectiveness of two 5G NS strategies in the context of Smart Logistics: a static slice with balanced resource allocation and a dynamic slice.
3) **Performance evaluation via simulations.** Through simulations in the developed simulator, the paper evaluates the Quality of Service (QoS) provided by these NS strategies across various traffic profiles. In particular, the focus on this paper has been set on the following metrics: throughput for eMBB traffic, reliability for URLLC traffic; and the random-access channel for mMTC traffic.

The remainder of this paper is organized as follows. In Section II a description of Smart Logistics along with its main scenario and applications is given. In Section III a brief description of wireless connectivity in Industry 4.0 is given. In Section IV, the different 5G technologies that allow optimizing the network are explained. In Section V, the floorplan of the simulated distribution center is described. The simulator and the enhancements made to it along with network parameters are described in Section VI. The results and discussion are shown in Section VII, along with the limitations and assumptions made in this study; and finally, the conclusions and future work are summarized in Section VIII.

II. Background

Smart Logistics [1] emerges from the adoption of lean principles and the application of advanced Information Technology (IT) systems, with the objective of responding to new demands from the public. Compared to traditional logistics, Smart Logistics optimizes the processes for small product batches, where economies of scale cannot be applied easily; reduces the delivery times and allows reverse logistics. These principles apply to the full logistics chain, from the manufacturer to the consumer. To achieve this agility, the whole structure of distribution centers, transport vehicles and control and monitoring systems is changed with respect to traditional logistics. The need to support very small batches in an agile manner requires that in Smart Logistics, each parcel, delivery vehicle, storage facility resource, etc., is tracked in real time with a much higher precision, to allow for better planning. Such a sophisticated monitoring and control system can only be achieved with wireless connectivity, especially with cellular networks such as 5G, that allow connectivity within the distribution center and along the whole logistics chain.

The main scenario in Smart Logistics are the distribution centers [2], which are large buildings with two main differentiated areas: the receiving and shipping docks (which may be the same or independent), and the storage area. The receiving and shipping docks are areas adjacent to doors adapted for loading and unloading trucks, where the equipment and personnel work to unload incoming and load outgoing goods. In Fig. 1 this area can be seen in the right side, with the door for loading/unloading trucks in the background.

Advanced systems such as palletizing machines, AGVs [3] and AR assistance for workers [6] will operate in these areas. The storage area (left side in Fig. 1) consists mainly of racks where products are stocked for short periods of time between arriving and being redirected to the next means of transport. In this area, products will be localized with systems such as Smart Tags [4] and stored/retrieved with AGVs.

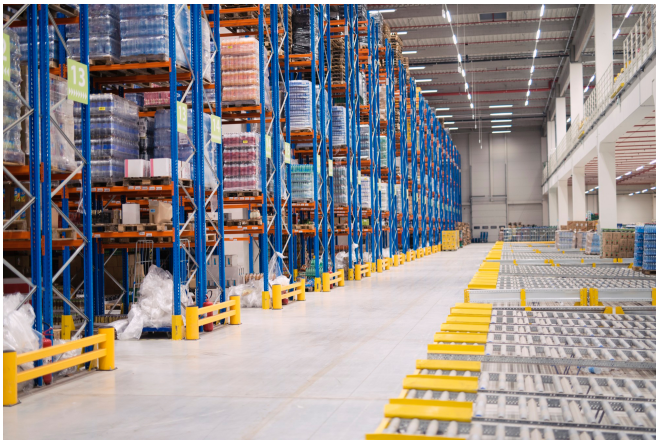


Fig. 1. Example of a distribution center.

The different applications that will run within a distribution center will have different requirements. For instance, AGV communications can be considered an URLLC profile, since large delays may cause malfunctions such as collisions between different vehicles or even with human workers. For this reason, reliability must also be very high and can be improved with multi-connectivity and packet duplication approach [20]. AR/VR [21] will require the uplink transmission of live video feeds and downlink transmission of complex 3D objects, so they will need a very high bandwidth (up to 50 Mbps) and ideally an end-to-end latency (including processing of 3D objects) below 20 ms to avoid dizziness [6]. This matches the eMBB traffic profile. Smart Tags, on the other hand, have the main priority of conserving energy and having a very high coverage to allow connectivity within very cluttered environments (such as racks or pallets of parcels), with low bandwidth, latency and reliability requirements.

In a distribution center, there will be some particularities that differentiate the wireless network from other scenarios. First, the combination of services depends greatly on the activity taking place at a certain time in the center. For

[1]https://github.com/dsr96/5g-simulator
[2]https://github.com/dsr96/ra-simulator

instance, when unloading an incoming container, the receiving dock will be populated by workers (possibly wearing and making active use of AR/VR glasses) and AGVs with URLLC control messages assisting them. After that period, products will be moved to specific places in the storage, again with a combination of workers and AVGs. A similar activity will be present when loading containers. Between loading and unloading (or in areas that are far from the docks in large distribution centers), only Smart Tags will be active sending periodic readings, with the occasional worker or AGV passing by. Smart Tags will be active at all times, creating a massive background traffic with low priority, but with tight energy and coverage requirements. Second, the high clutter caused by racks, machines and parcels will make a distribution center a very harsh environment for propagation, similar to the conditions found in a factory [22].

To provide connectivity for these applications, several wireless solutions are available, as it will be explained in the next section. Nevertheless, only 5G supports all traffic profiles that occur in a distribution center, using techniques such as NS.

III. Wireless connectivity in Industry 4.0

Traditionally, wired connections have been used in industrial networks to connect different elements such as Programmable Logic Controllers (PLC) [23], that is, the computers that control the machines, with each other or with the Manufacturing Execution System (MES) [24]. Process monitoring as well as alarm monitoring are usually contained in the MES, constituting an interface between the PLCs and the Enterprise Resource Planning (ERP) [25], which allows a global coordination at executive level.

With the arrival of the Industry 4.0 paradigm [26], it is intended to obtain more flexibility and the support of new use-cases. To provide this, wireless connectivity has gain an increasing usage in factories, which can also reduce the installation costs.

Wireless technologies are divided into two types of networks that enable different applications: Local Area Networks (LAN) and Wide Area Networks (WAN). LANs have a coverage range of up to 100 meters and can cover an area of a distribution center or many rooms. On the other hand, WANs provide a higher coverage range, from distances of a few kilometers up to whole countries.

Regarding LANs, the IEEE 802.11 family (WiFi) [27], [28] is the most common technology used in some industrial deployments, due to its low cost and wide availability of components. Also, customized solutions for factories and based on IEEE 802.15.4 have been used, such as WirelessHART [29], WIA-PA [30], ZigBee [31], ISA100.11a [32] and IETF 6LoWPAN [33]. On the other hand, regarding WANs, technologies such as SigFox [34] and LoRaWAN [35] have been used in manufacturing plants for power limited devices.

Moreover, cellular networks (GSM/GPRS, Long Term Evolution, LTE) have also been applied for specific applications, such as sensors [36] and robotics [37]. These cellular networks were not design for Industrial Internet of Things (IIoT). Therefore, technologies such as NB-IoT, Cat-M1 and EC-GSM [38] were designed, focusing on energy saving and high coverage range with a low data rate.

The main problem of the non-cellular technologies mentioned above is that they are insufficient for logistics needs, since they are either not wide area networks, which limits the interoperability, or they do not offer the necessary data transmission capacity. Likewise, the actual technologies based on cellular networks such as NB-IoT, Cat-M1 and EC-GSM only fulfills the needs of coverage. Moreover, although LTE was designed as a broadband access technology, it cannot fulfill the requirements of extreme industrial applications, such as AR/VR. Apart from that, the latency and high reliability requirements for the new use-cases in factories (i.e., AGVs navigation or closed loop control) cannot be achieved with LTE. In other words, none of these technologies is capable of providing a service that covers the three main traffic profiles that occur in a distribution center (eMBB, URLLC and mMTC). For further information about the specific requirements for logistics applications, we refer the reader to [7].

Only 5G supports all the traffic profiles, using techniques such as NS [7]. Since 5G is a WAN, it support communications over the whole logistics chain, not limited to the distribution center. With Non-Public Networks [39], the logistics operators can have custom connectivity with a private network supported over a public operator hardware, with the corresponding reduction in ownership and expertise costs.

IV. 5G technologies

A. Numerologies in 5G

5G New Radio (NR) introduced in Release 15 a new frame structure to provide flexibility and the adoption of new use-cases such as critical communications. The frame structure in 5G NR can adopt different numerologies. A numerology (μ) is defined by a cyclic prefix (normal or extended) and a SubCarrier Spacing (SCS).

Four new numerologies have been defined in 5G NR, which ranges from 0 to 4, where $\mu = 0$ corresponds to LTE configuration. A numerology defines the SCS as $15 \cdot 2^{\mu}$ kHz and the slot duration as $1/2^{\mu}$ ms. As numerology increases, shorter slots are used, but they are wider in frequency, so a higher SCS is necessary. The standard states that not all numerologies are suitable for a determined frequency range (FR) and its use is divided into synchronization and data channels [40].

For synchronization channels, $\mu = \{0, 1\}$ is used in FR1 (sub-6 GHz bands) and $\mu = \{3, 4\}$ in FR2 (millimeter wave bands). On the other hand, for data channels, $\mu = \{0, 1, 2\}$ is supported in FR1 and $\mu = \{2, 3\}$ in FR2. The number of subcarriers in 5G NR is 12 for all numerologies. Same as LTE, in 5G NR the frame duration is fixed at 10 ms and subframe duration at 1 ms. Depending on the selected configuration, the number of slots per subframe is defined as 2^{μ}. Finally, one slot is composed by 14 Orthogonal Frequency Division Multiplexing (OFDM) symbols, where the symbol duration is defined as $1/(14 \cdot 2^{\mu})$ ms. Fig. 2 shows a summary of the characteristics for each numerology in the time domain.

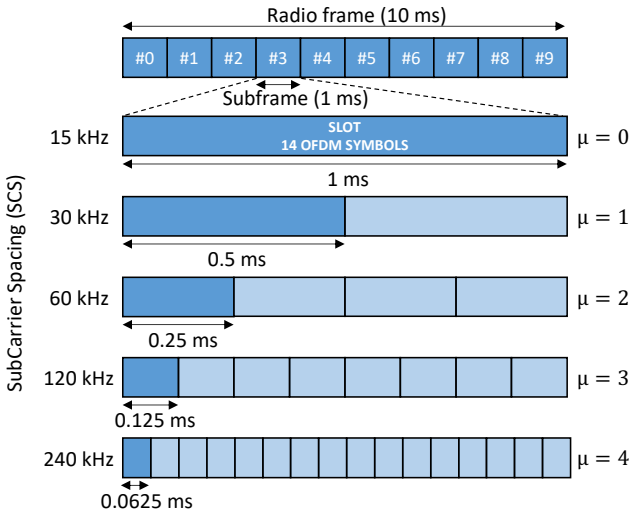


Fig. 2. 5G numerologies scheme in the time domain.

The numerology is one of the main accepted solutions to reduce the latency, since the slots are shorter as μ increases. The scheduler normally works at slot level, so a decrease in the slot duration makes the resource allocation faster. However, this comes at the cost of network efficiency, since lower numerologies are better for capacity (eMBB traffic), that is, throughput. Also, bandwidth and packet size are two factors that affect when selecting the numerology. If the numerology selected is very high and the packet size is larger, this can lead to an unexpected increase in latency [22], especially in poor radio conditions.

B. Network slicing

The different types of traffic profile (URLLC, eMBB and mMTC) have different requirements that can be achieved by optimizing network parameters; but these optimizations may cause conflicts. This is the case of numerology, as explained in Section IV-A. To solve this problem, NS [5], [7] has been proposed to assign one slice per type of traffic; each slice being optimized independently without affecting others.

Network slicing enables the creation of multiple independent virtual end-to-end networks on a shared physical infrastructure. Each network slice can have different network resources allocated to it, such as portions of spectrum supported by a network access point [41]–[43]. Accordingly, network slices that use different network resources can be effectively isolated from one another. This means that issues with one network slice are unlikely to impact another network slice.

Different network slices may be associated with different use cases, services, or applications. One possible approach for NS in the Radio Access Network (RAN) is the division of resources in time and frequency, as described in [5] and depicted in Fig. 3. Under this approach, resources are assigned to each type of service attending to their needs. For instance, the resources assigned to the URLLC service will use a high numerology (i.e., shorter transmission time interval) to reduce the latency, with the possibility of using redundant channels in time and frequency, while mMTC will use narrowband channels with low numerology to better adapt to poor radio conditions. The eMBB slice will use wideband channels and opportunistically reuse frequencies that are not used by URLLC.

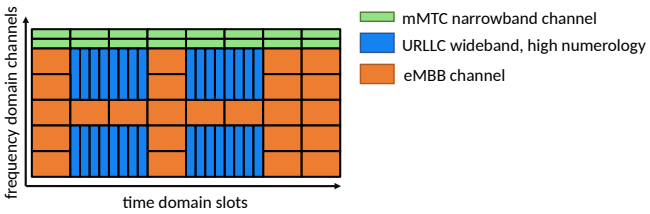


Fig. 3. Example of NS division for different service types.

While NS can be used in this manner to provide optimal channels to all services, resources may be wasted if they are not assigned proportionally. In these cases, NS allows to dynamically reshape the resource assignation. This can even be done proactively, using external data sources to predict when a specific profile will need more resources. In distribution centers [7], data sources such as the truck schedule (combined with traffic information), the registry of online transactions or the current inventory can be used to estimate the composition of the traffic and therefore assign resources to the different profiles.

C. Random access procedure

The random access (RA) procedure is used for the User Equipments (UEs) to start communicating with the base station in cellular communications. It was first introduced in Release 8 and updated for 5G NR in Release 15. To handle this communication, two different RA procedures are defined in the standard [44]:

- Contention-based: where the UEs selects a randomly preamble from a pool of preambles to request a network access. This procedure is susceptible to collisions, therefore, it is used for delay-tolerant access (e.g., for the initial Radio Resource Control, RRC, connection establishment).
- Contention-free: where the base station allocates dedicated resources (i.e., dedicated preamble) to the UEs, avoiding a preamble conflict. The preamble is allocated via RRC signaling or physical layer, and this procedure is used for delay-constrained access that requires a high probability of success (e.g., starting communication with the target base station in handover).

In the contention-based RA procedure, there are four messages involved between the UE and the next generation NodeB (gNB), as depicted in Fig. 4:

- Msg1: the UE sends through the Physical Random Access Channel (PRACH) a preamble randomly selected among a list broadcast periodically by the gNB in the System Information Block (SIB) [45], [46]. There are 64 preambles available for the RA, but not all of them are available for contention-based (some of them are reserved

for contention-free access). A collision will occur in case that multiple UEs transmit the same preamble in the same RA slot. Once the UE sends the preamble, it waits a time window to receive a response from the gNB (Msg2). The duration of this window is broadcast by the gNB and has a maximum value of 10 ms [46]. In case that the timer expires, the UE performs a new access attempt if the number of attempts is less than *preambleTransMax*, which defines the maximum allowed value [46].

- Msg2: once the Msg1 is received at the gNB, it replies with a Random Access Response (RAR) message over the Physical Downlink Shared Channel (PDSCH) with a Random Access Radio Network Temporary Identifier (RA-RNTI) and a temporary Cell RNTI (C-RNTI), providing an uplink resource grant and time alignment to be used to transmit the next message (Msg3) [45]. The RA-RNTI identifies the preamble sent in Msg1, so the UE that transmitted that preamble is informed that it has been heard; and the C-RNTI is used by the UE to identify itself in the next steps. If the same preamble was selected by two or more UEs, the collides UEs will wait a random backoff time (according to the Backoff Indicator parameter, BI, attached to the RAR) before retrying a new access attempt (Msg1).
- Scheduled Transmission (Msg3): the UE starts sending it request over the Physical Uplink Shared Channel (PUSCH) along with the temporary C-RNTI [44], using the grant received on Msg2. The signaling message and the information associated will vary depending on the particular request: initial RRC connection setup, reestablishment of the RRC connection, etc.
- Contention Resolution (Msg4): upon reception of the connection request, the gNB replies with a contention resolution message. If a UE does not receive this message, it declares a contention resolution failure and the UE will perform a new access attempt, as previously explained in Msg1. If the counter reaches the maximum value (*preambleTransMax*), the UE will indicate a random access failure to the upper layers.

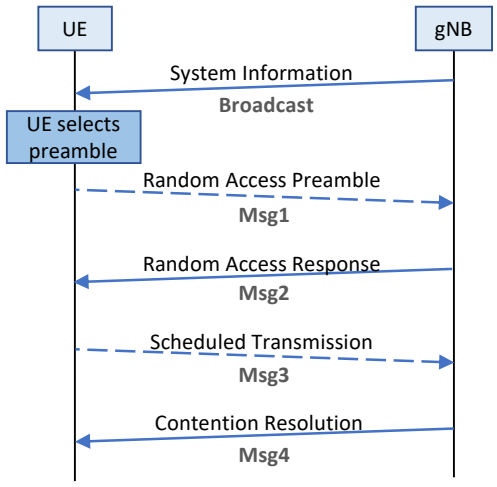


Fig. 4. Contention-based RA procedure.

This procedure uses the RACH, which is formed by a periodic sequence of allocated resources in the time-frequency domain, namely RA slots. These slots are reserved in the uplink channel for the transmission of the different access requests. The periodicity of the RA slots is broadcast by the gNB in the SIB, in particular, it is defined by the PRACH Configuration Index parameter [46]. This periodicity ranges from 1 RA slot every 2 frames to a maximum of 1 RA slot per subframe [47]. Fig. 5 shows an example of different PRACH Configuration Index values, assuming $\mu = 0$ (same as LTE).

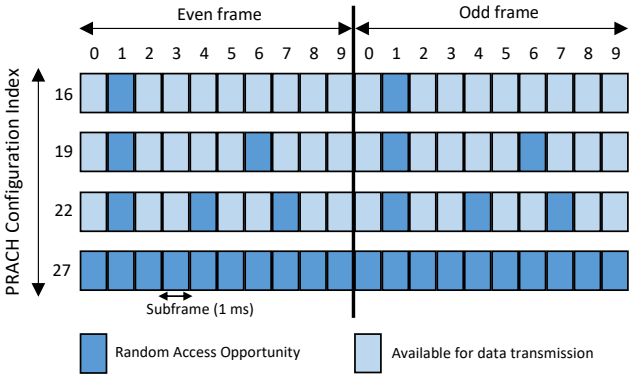


Fig. 5. Example of different PRACH Configuration Index with $\mu = 0$.

Since RA slots use uplink resources, it is important to maintain a trade-off between the amount of resources dedicated for the RA and the amount of resources available for uplink data transmissions.

V. Scenario

In this Section, the simulation setting is described, including the floorplan of the scenario, the activity taking place over this floorplan and the Industry 4.0 applications being used in these activities.

Floorplan

In this paper, a realistic scenario of a small distribution center is reproduced, in order to simulate the behavior of a 5G network supporting different Industry 4.0 applications. The floorplan of the setting is shown in Fig. 6. The setting is that of a small distribution center with a capacity for a single truck, a shared receiving and shipping dock and a small storage area. The storage area has 7 racks of 10 meters of length, 2 meters of depth and 6 meters of height, separated 2 meters from each other. The distribution center also counts with a single robotic palletizing machine, several AGVs, and is serviced by workers with AR/VR glasses. The floorplan in Fig. 6 also includes a truck trailer that is serviced during the simulation, measuring 12 meters in length, 2.5 meters in width and 3 meters of height (approximately the standard shipping container sizes). The distribution center is also served by one gNB (marked as the magenta triangle).

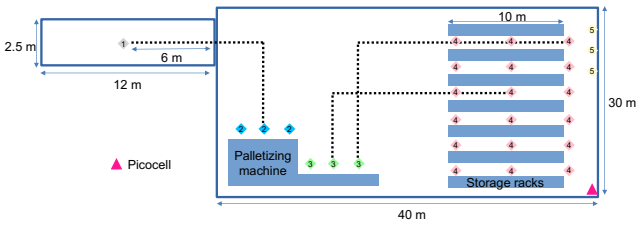


Fig. 6. Floorplan of the simulated scenario.

Activity

The developed simulator can also represent the activity taking place in a distribution center. Fig. 6 shows some numbered marks, each representing one type of location where specific activities take place:

1) Truck trailer loading/unloading: AGVs move pallets to/from this point.
2) Palletizing machine (pallet side): pallets with parcels are either received from or sent to the truck (type 1 point).
3) Palletizing machine (parcel side): parcels are handled to be palletized and shipped or are processed after being depalletized to be stored in the storage racks.
4) Storage racks: points where parcels are placed on or retrieved from the storage racks.
5) Worker access: points where human workers access the installation.

The spatial distribution of the aforementioned areas impacts network demand differently. For instance, loading and unloading areas (left side in Fig. 6) increase network demand only when receiving an incoming truck, as pallets with parcels are moved from the truck to the palletizing machine area and vice versa. On the other hand, when parcels are stored or retrieved, the main activity takes place in the storage area (right side in Fig. 6). In this area, AGVs are used to transport parcels, Smart Tags are active and the storage area is also populated by different workers, thereby increasing network demand.

The activity of the different applications is organized around these points. Specifically, four activities are emulated:

- Loading/unloading the container: AGV moves between point 1 and a randomly selected point 2. Approximately 12 trips per hour take place with 3 AGVs. When this activity is performed, URLLC traffic increases, as AGVs start sending control messages to a fleet manager on a server.
- Storing/retrieving parcels: each trip consists of an AGV moving between a randomly selected point 3 and another randomly selected point 4. There are 10 AGVs doing this and each one does 1 round trip per minute. This activity is similar to the previous one, but the network impact is higher, as more AGVs are involved on this task.
- Human worker activity: workers occasionally retrieve or place parcels on the storage areas, so they perform round trips between a randomly selected point 5 and point 4. The round trip always starts at a point 5. There are 5 workers and each one performs 12 round trips per hour. This activity creates a high impact on eMBB traffic profile, as workers' AR/VR glasses download 3D objects, resulting in a high bitrate.
- Smart Tag activity: parcels that are stored in the racks send updates of their location with mMTC. There are 1000 parcels in total. Smart Tags increase mMTC traffic profile, producing a high overhead on the network (specially on the RACH), as they simultaneously transmit short periodic updates.

In this paper, three different time intervals are considered as represented in Fig. 7, where in each interval there is one traffic profile that has a high load. More specify, the time intervals correspond to different situations that are performed in the distribution center:

- Time interval 1: corresponds to the activity of loading/unloading the container with 3 AGVs, human worker activity with medium traffic load and a massive Smart Tag activity (1000 simultaneous arrivals).
- Time interval 2: corresponds to the activity of storing/retrieving parcels with 10 AGVs, human worker activity with medium traffic load and medium Smart Tag activity (500 simultaneous arrivals).
- Time interval 3: workers with high traffic load, 3 AGVs storing/retrieving parcels with low traffic load and low Smart Tag activity (250 simultaneous arrivals).

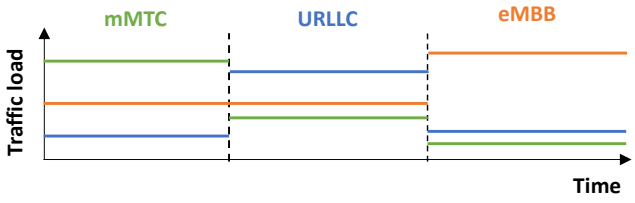


Fig. 7. Simulated traffic load over time.

Applications

The applications that run in the simulated distribution center are the following:

- AGV messaging: downlink URLLC messages which are generated at a rate of 10 messages per second per AGV, with 64 bytes of data. These messages are critical and will require a low latency and high reliability.
- AR/VR headsets of workers: each worker has a full-buffer downlink video stream of 8 Mbps for a high traffic activity and 4 Mbps for a medium traffic activity, with 1000 byte packets.
- Smart Tags: parcels that are processed in the distribution center will be equipped with Smart Tags that regularly send their information to a remote server. These uplink messages occur once per hour, with 64 bytes of data.

VI. Simulator

To evaluate the performance of the different traffic profiles and their effects when using a static and dynamic NS, two developed open-source simulators have been used. The first one is used to study the mobile network communications

within a distribution center, while the second one is used to evaluate the random access procedure for mMTC devices. Throughout this section, both simulators and their parameters will be described in more details.

A. 5G simulator

To study the impact of different slices compared to a static slice, a simulator based on the NS-3 platform has been used, using the 5G-LENA module [48]. This module provides a 5G Non-Standalone network and focuses on the new 3GPP NR specifications, which includes features such as numerology support, frequency division multiplexing of numerology, beamforming, among others.

Simulator enhancements

To the aim of evaluating network performance in a distribution center, different features have been included in a developed open-source simulator [49] based on the NS-3 framework and the 5G-LENA module, and they are summarized in Fig. 8. A description of these enhancements is provided below:

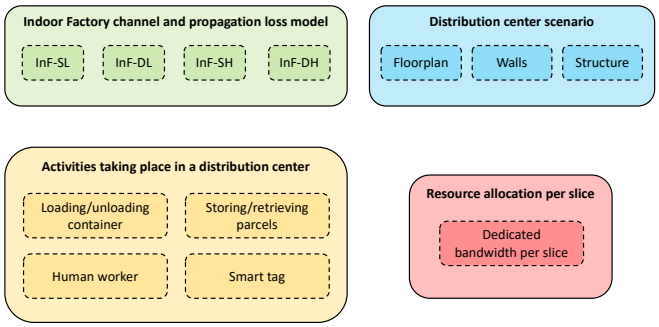


Fig. 8. Enhancements to NS-3 framework and 5G-LENA module.

- **Indoor Factory channel and propagation loss model**. Currently, the channel and propagation loss models provided by the 5G-LENA module are the Urban Macro (UMa), Urban Micro (UMi) street canyon, Rural Macro (RMa) and Indoor-Office, as specified in 3GPP 38.901 standard. However, it does not include the industrial scenario. Therefore, we included the 3GPP Indoor Factory (InF) scenario in all its variants (InF-SL, InF-DL, InF-SH, InF-DH), as specified in the standard [50]. It also includes the outdoor-to-indoor (O2I) penetration loss.
- **Distribution center scenario**. A distribution center scenario has been developed, as specified in Section V. This scenario has been created by using the NS-3 buildings module and introducing walls and structures with metal and concrete materials. The penetration loss in the signal due to walls and structure was also included, following the 3GPP standard.
- **Activities in a distribution center**. The different activities that take place in a distribution center, described in Section V, have been developed. The NS-3 mobility module has been used to provide movement around the floorplan for the different devices.
- **Resource allocation per slice.** By default, the resources dedicated for the different slices are equally divided into the total bandwidth. We added the possibility of select how many resources are dedicated for each slice, and each traffic profile is assigned to a slice according to their requirements.

Network configuration

The gNB operates with a frequency of 3.7 GHz and a total bandwidth of 20 MHz. One transmission/reception omnidirectional antenna is used in both, gNB and UEs, with 15 dBm as downlink transmission power. Regarding numerology, $\mu = 0$ has been configured for eMBB and mMTC, whereas $\mu = 2$ has been used for URLLC, which is the highest μ supported for data channels in FR1, as previously explained in Section IV-A.

Attending to the slices, two different configurations have been used:

- Static NS: a static slice (baseline) for each type of service with a balance division of network resources.
- Dynamic NS: network resources will be appropriately distributed between the different slices according to the traffic load, depending on the activity taking place. This slice will try to maximize the performance of the traffic profile with a peak in each interval and minimize the degradation for the other traffic profiles.

The total bandwidth is divided into three slices, one for each traffic profile. For a static NS, the bandwidth division is configured as 55% for eMBB, 30% for URLLC and 15% for mMTC. On the other hand, when using a dynamic NS, the bandwidth is divided dynamically depending on the time interval:

- Time interval 1: 40% for eMBB, 30% for URLLC and 30% for mMTC.
- Time interval 2: 30% for eMBB, 60% for URLLC and 10% for mMTC.
- Time interval 3: 70% for eMBB, 20% for URLLC and 10% for mMTC.

The network parameters alongside traffic parameters on each time interval are summarized in Table I.

B. PRACH simulator for mMTC

One of the main drawbacks of the 5G-LENA module [48] in NS-3 is that all devices are connected at the beginning of the simulation to the base station in an ideal way. That is, no real RA procedure is performed, the signaling is ideal and it does not consume any radio resources. Furthermore, it does not allow to configure different PRACH Configuration Index values (by default preambles can be sent on any system frame number and subframe number). This is an important aspect that must be taken into account, since the main problem of mMTC comes from the saturation of the RACH as the number of simultaneous devices increases, which results on higher collisions when transmitting the preambles and may cause a device to block if it reaches the maximum allowed RA preamble attempts.

TABLE I
SUMMARY OF SIMULATION PARAMETERS.

Network parameters			
Parameter	Value		
Channel and propagation loss model	3GPP 38.901, InF-DH [50]		
Total bandwidth	20 MHz		
Frequency	3.7 GHz		
Numerology (μ)	0 for eMBB and mMTC; 2 for URLLC		
Transmission direction	UL for mMTC; DL for eMBB and URLLC		
Modulation	Adaptive		
Scheduler	Round-Robin		
gNB height	10 m		
gNB transmission power	15 dBm		
UE power control	3GPP 38.213 [51]		
MAC to PHY delay	2 slots		
Transport block decode latency	100 μs		
HARQ feedback delay	1 slot		
Traffic parameters			
Traffic profile	Parameter	Interval	Value
eMBB	Message size	All	1000 bytes
	Stream rate	1, 2	4 Mbps
		3	8 Mbps
	Number of devices	All	5
URLLC	Message size	All	64 bytes
	Message rate		10 Hz
	Number of devices	1, 3	3
		2	10
mMTC	Message size	All	64 bytes
	Message rate		1 per hour
	Number of devices		1000
	Simultaneous arrivals	1	1000
		2	500
		3	250

To overcome this drawback, a new open-source simulator for the RACH has been developed [52] to evaluate the performance of mMTC traffic profile, following the 3GPP standard [45]–[47] behavior. This simulator has been implemented on Python and enables to configure different 3GPP parameters for the RACH. In particular, the parameters that can be modified are the following:

- PRACH Configuration Index: defines the periodicity of the RA slots. The periodicity ranges between a maximum of one RA slot per subframe to a minimum of one RA slot every two frames.
- Number of available preambles: corresponds to the number of preambles reserved for the contention-based procedure.
- *preambleTransMax*: maximum number of preamble attempts for a device before declaring RA failure.
- RAR Window Size: time window to monitor RA response.
- Backoff Indicator: random backoff that is used by the UEs to wait a time when a preamble collision occurs before retrying a new access attempt. This backoff is intended to disperse the access attempts and thus, reduce the probability of preamble collision.

Table II summarizes the different RACH parameters used in this paper with a static and dynamic slice to evaluate the performance of mMTC. Note that for each time interval (previously defined in Section VI-A), the static slice maintains the same parameters, whereas the dynamic slice changes the resources dedicated for the RA procedure (PRACH Configuration Index parameter).

TABLE II
RACH PARAMETERS FOR MMTC.

Parameter	Interval	Static NS	Dynamic NS
PRACH Configuration Index [a]	1	19	22
	2		22
	3		16
Number of available preambles [b]	All	60	
preambleTransMax [c]	All	10	
RAR Window Size [c]	All	5 ms	
Backoff Indicator [b]	All	20 ms	

[a] Refer to 3GPP TS 38.211 [47] for all possible values.
[b] Refer to 3GPP TS 38.321 [45] for all possible values.
[c] Refer to 3GPP TS 38.331 [46] for all possible values.

C. Metrics

In this paper, the following metrics have been considered for eMBB and URLLC:

- Throughput (eMBB): average throughput measure as the quantity of bytes received b_r divided into the simulation time t_s for N devices, as denoted in the following equation:

$$Throughput = \frac{1}{N}\sum_{i=0}^{N}\frac{b_r}{t_s} \quad (1)$$

  Note that t_s for the throughput calculation is the time elapsed between the first packet transmitted in the transmitter and the latest packet received in the receiver. The ideal situation is when the throughput is equal to the application sent rate, which means that the maximum QoS is achieved.
- Reliability (URLLC): average reliability, defined as the number of packets received p_r whose latency l is below a threshold th divided into the total packets transmitted p_t for N devices and it is calculated as follows:

$$Reliability = \frac{1}{N}\sum_{i=0}^{N}\frac{p_r}{p_t} \quad \forall\, l < th \quad (2)$$

  In this particular case, the latency threshold th has been set to 5 ms and it is measured at Packet Data Convergence Protocol (PDCP) layer in the downlink side. Given that the end-to-end latency requirement for remote-control of Mobile Robots should be below 10 ms, the threshold has been set to half of this value. Moreover, the reliability requirement for Mobile Robots must be above $1 - 10^{-3}$ (99.9%) [53].

On the other hand, three different metrics have been considered for the evaluation of mMTC with the developed RACH simulator:

- Blocking probability: probability that a device reaches the maximum number of transmission attempts (*preambleTransMax*) and is unable to complete an access process.
- Average number of preamble retransmissions: measure the average number of preamble retransmissions required to have a success access.
- Access delay: time elapsed between the transmission of the first preamble and the reception of the Random Access Response (Msg2) by the mMTC device. Only for devices that do not reach the maximum number of transmission attempts.

VII. Results and Discussion

This section presents the results obtained for the evaluation of each traffic profile and time interval within a distribution center environment, as defined in Section V and VI.

To obtain statistic results, 25 iterations with different seeds have been simulated in the NS-3 simulator, with a duration of 600 seconds in each interval. On the other hand, for the mMTC RACH evaluation, 1000 iterations with different seeds have been performed to obtain the different metrics.

A. URLLC

Fig. 9 shows the reliability obtained by the AGVs in each time interval and also the combined reliability during all intervals.

Looking at the first time interval, where URLLC traffic is low, it can be noted that the reliability obtained is very similar in both cases (static and dynamic slice). More specify, for a static slice a reliability of $5 \cdot 10^{-4}$ is obtained; whereas for a dynamic slice the reliability is $4.8 \cdot 10^{-4}$, with an improvement of 4.37%. The similarity is due to the use of the same percentage of slice in both cases.

In the time interval 2, where URLLC traffic is very high, it is observed a clearly reduction in the reliability with a static slice compared to a dynamic slice. The static slice obtains a reliability of 10^{-2} whereas the dynamic slice obtains an improvement of 98.78% on the reliability, with a value of $1.3 \cdot 10^{-4}$. Since a high increment on the traffic of the AGVs is done under this time interval due to the activity taking place, the static slice with a fixed size of 30% of the total bandwidth (6 MHz) does not provide enough resources for this critical service. Consequently, the latency values are higher (i.e., the probability of receiving packets above the threshold is increased) and therefore, the reliability decreases. On the other hand, since the dynamic slice has incremented its size to 60% of the total bandwidth (12 MHz), the reliability requirement is guaranteed with a good level (near 10^{-4}), although the number of AGVs has increased from 3 to 10.

During the time interval 3, similar to the time interval 1, the intensity of the traffic of the AGVs is low. Under this time interval, the static slice obtains a better reliability ($1.9 \cdot 10^{-4}$), since the slice size is higher (30% vs 20%). On the other hand, the dynamic slice obtains a degradation of 30.61% on the reliability, with a value of $6.2 \cdot 10^{-4}$. The fact that the size of the dynamic slice is lower than the static is due to adjust the slice size to maximize the peak traffic profile (eMBB), which will be discussed later. Note that despite using the same amount of resources for the static slice in time intervals 1 and 3, the reliability is better under time interval 3. This is mainly due to the activity taking place and the distance with the gNB (see Fig. 6). In the time interval 1, the AGVs move between points 1 and 2 (load and unload the container); whereas in time interval 3, the AGVs move between points 3 and 4 (storing/retrieving parcels), which are closer to the base station. Consequently, the propagation losses are higher for those AGVs in the time interval 1.

Taking the values of all time intervals combined, it is clearly noticeable that the dynamic slice performs better (95.65% improvement) and obtains a reliability of $2.9 \cdot 10^{-4}$ compared to $6.6 \cdot 10^{-3}$, which is obtained with the static slice.

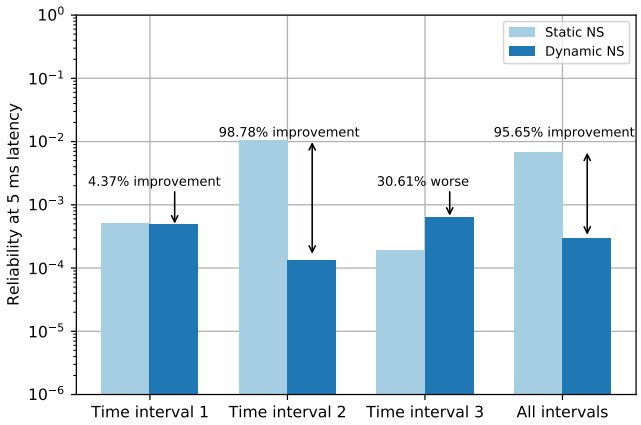


Fig. 9. URLLC reliability at 5 ms latency with a static and dynamic slice for each time interval.

B. eMBB

The average throughput obtained by the workers in each time interval is depicted in Fig. 10, where the horizontal lines represent the application sent rate in each time interval.

In the time interval 1, it can be seen that the throughput obtained for both slices (static and dynamic) is equal to the application sent rate, obtaining a value of 4 Mbps. Although the dynamic slice has a reduced bandwidth (40% of total bandwidth) than the static slice (55% of total bandwidth), the same throughput is received. This clearly denotes that there is a resource wastage with the static slice; whereas the slice reduction maintains the throughput and this reduction can be used to increase the mMTC slice size, which is the peak traffic in this interval.

Looking at the second time interval, again, the received throughput is equal to the application sent rate for a static slice. However, since the dynamic slice size has been reduced to 35% of total bandwidth, a throughput of 3.75 Mbps is obtained. Upon this case, the throughput degradation is minimum (6.18%) and the resources reduced for eMBB are used to increase the URLLC slice size, which improves the URLLC reliability, as previously seen.

During the time interval 3, the throughput obtained is 6.45 Mbps for a static slice and 7.64 Mbps for a dynamic slice.

Under this interval, both slices suffer a throughput degradation, not achieving the application sent rate (8 Mbps). Due to the eMBB traffic peak, more resources are needed, so a static slice cannot fulfill this traffic demand, since its bandwidth is fixed. That is the reason why the throughput is dropped more. On the other hand, although the dynamic slice bandwidth has been increased to 70% of total bandwidth (14 MHz), the throughput received is slightly reduced but it is very close to the application sent rate, obtaining a throughput improvement of 18.55% with respect to a static slice.

Taking the values of all time intervals combined, the dynamic slice improves the throughput by 6.48%, despite the slightly reduction in the time interval 2. In particular, the throughput value obtained is 4.81 Mbps and 5.13 Mbps for a static and dynamic slice, respectively. Note that the throughput of the dynamic slice is very close to the average sent rate (5.33 Mbps).

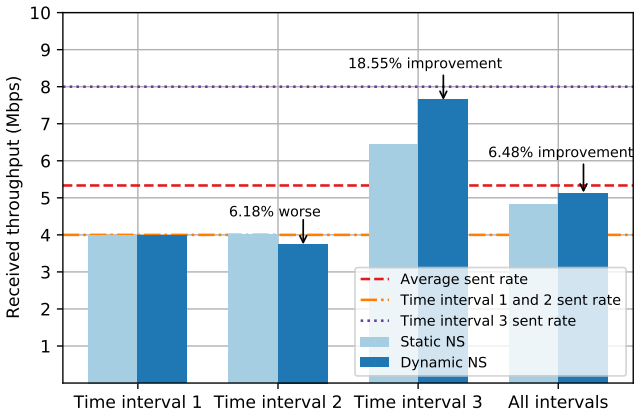


Fig. 10. Throughput obtained by the workers with a static and dynamic slice for each time interval.

C. mMTC

The mMTC evaluation has been performed with the developed RACH simulator, as explained in Section VI-B. Fig. 11 shows the blocking probability obtained by the Smart Tags during the different time intervals.

During the time interval 1, since there are 1000 simultaneous arrivals, the collisions are very high. A blocking probability of 0.91 is obtained with a static slice, whereas the blocking probability for a dynamic slice is 0.62. Since during this interval the Smart Tags activity is very high, more resources for the RA are dedicated with the dynamic slice (3 RA slots per frame) compared to the static slice (2 RA slots per frame), thus reducing the blocking probability by 31.29%.

In the time interval 2, since the traffic intensity has decreased (500 simultaneous arrivals), there is no blocking probability with the dynamic slice (it maintains the same RA slots per frame as previous interval), whereas the static slice obtains a blocking probability of 0.08.

The same trend is observed in the time interval 3, where the traffic intensity is low (250 simultaneous arrivals). In this case, there is no blocking probability for the static slice, whereas the dynamic slice (with only one RA slot per frame) obtains a blocking probability of 0.006, which is insignificant. Although the number of RA slots has been decreased for the dynamic slice, the blocking probability is insignificant, but there are more resources available for uplink data transmissions. Therefore, the dynamic slice is more efficient during this interval.

Taking a look in the combination of all intervals, it is clearly visible that the dynamic slice performs better with a reduced blocking probability of 36.22%, obtaining a probability value of 0.21 compared to 0.33 which is obtained with a static slice.

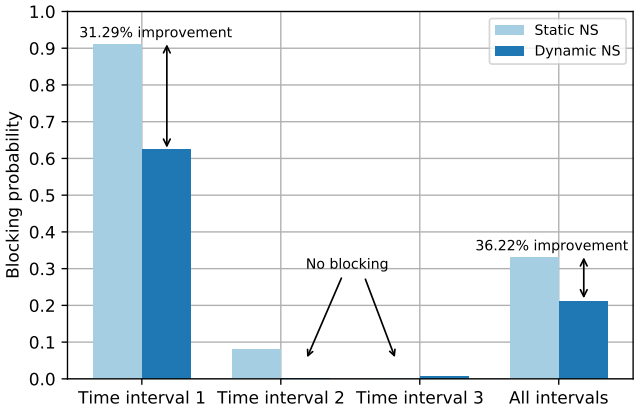


Fig. 11. mMTC blocking probability with a static and dynamic slice for each time interval.

The average number of preamble retransmissions needed to have a successful network access is shown in Fig. 12. As it can be seen, in general, the value is decreased as the time interval is increased, since the simultaneous arrivals are reduced.

In the time interval 1, the average number of preamble retransmissions is close to the maximum allowed value (*preambleTransMax*), which is 10. A value of 8.34 and 6.97 is obtained for a static and dynamic slice, respectively. In this case, the dynamic slice obtains an improvement of 16.4%.

On the other hand, in the time interval 2, these values are decreased, since there are less simultaneous arrivals and also the blocking probability is very low. The static slice obtains a value of 5.43, whereas the dynamic slice obtains a value of 3.21, which results in a reduction of 40.82%.

Finally, during the time interval 3, unlike the previous time intervals, it is observed that the dynamic slice obtains a higher value than the static slice, with an increased value of 51.68%. This is mainly due to the use of less RA slots, which results in more accumulated request on each RA opportunity, and consequently, more collisions occur. Upon this case, the static slice obtains a value of 2.21, whereas the dynamic slice obtains a value of 4.27.

With the combination of all intervals, although in the interval 3 the static slice performs better, it is compensated on time intervals 1 and 2, so the dynamic slice obtains an improvement of 9.51% in combination, with a value of 4.82. On the other hand, the static slice obtains a value of 5.32, which is closer to the value obtained with the dynamic slice.

Finally, the average access delay is shown in Fig. 12, where the whiskers represent the upper and lower deviation. Note that

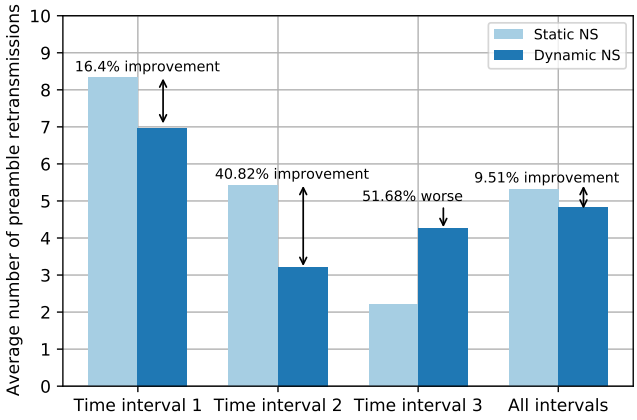


Fig. 12. mMTC average number of preamble retransmissions with a static and dynamic slice for each time interval.

the access delay is only measured for the devices that are not blocked.

In the time interval 1, the static slice obtains an average access delay of 160.92 ms with a lower deviation of 9.07 ms and upper deviation of 11.09 ms. On the other hand, the dynamic slice obtains an average access delay of 124.25 ms with a lower and upper deviation of 6.35 ms and 7.33 ms, respectively. Since more collisions occur with the static slice, achieving a high blocking probability (see Fig. 11), this increments more the delay of the devices that have a successful access. In this case, the dynamic slice reduces the average access delay by 22.78%.

During the time interval 2, the static slice obtains an average access delay of 100.84 ms with a lower deviation of 9.42 ms and upper deviation of 10.6 ms. On the other hand, the dynamic slice obtains an average access delay of 65.56 ms with a lower and upper deviation of 5.47 ms and 5.07 ms, respectively. Although the blocking probability is very similar with both slices during this time interval (see Fig. 11), the dynamic slice obtains an average access delay reduction of 34.99%. This reduction is mainly due to the use of more RA slots per frame, which increment the RA opportunities and the devices can achieve a faster network access.

Taking a look in the time interval 3, it can be seen a drastic increment in the delay with a dynamic slice (23.59%), whereas the static slice reduces the delay. As previously commented, the delay is highly dependent on the periodicity of the RA slots (PRACH Configuration Index value). The higher the periodicity is, the lower the delay is. That is the main reason why the static slice (with 2 RA slots per frame) performs better than the dynamic slice (with 1 RA slot per frame). Upon this interval, the static slice obtains an average access delay of 53.58 ms, with a lower and upper deviation of 5.02 ms and 5.42 ms, respectively. On the other hand, the dynamic slice obtains an average access delay of 94.51 ms, with a lower deviation of 12.21 ms and upper deviation of 13.33 ms.

With the combination of all intervals, it is noted that the access delay is reduced by 9.83% when using a dynamic slice. Despite the higher delay obtained during the time interval 3, it is compensated with a lower delay when the number of simultaneous arrivals is higher (time intervals 1 and 2).

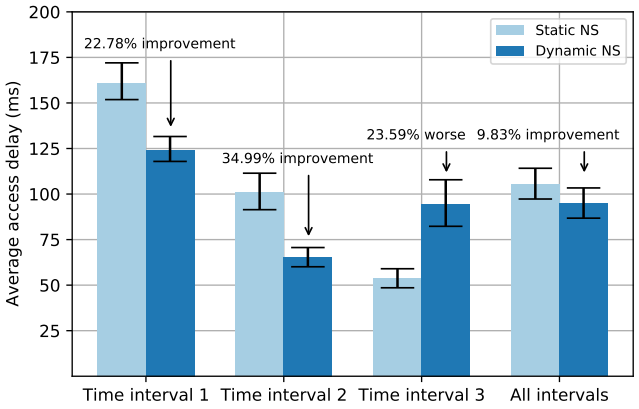


Fig. 13. mMTC average access delay with a static and dynamic slice for each time interval.

Discussion

On the one hand, it has been proven that a dynamic slice improves the overall QoS for the different traffic profiles compared to a static slice. Since the static slice has a fixed size, it guarantees the QoS only in certain intervals, but not in all. When there is a traffic peak, the static slice does not fulfill the requirements of that traffic, that is, there are not sufficient resources for this eventually traffic intensity. As a consequence, the QoS decreases and this is especially important for critical services (e.g., AGVs navigation), since a lower QoS could cause malfunction or accidents within the distribution center.

On the other hand, a dynamic slice adapts better to the changes on the traffic intensity, maximizing the QoS of the traffic profile with a peak intensity. When this occur, a slightly degradation of the service is achieved in the other slices, but in combination of all time intervals, the QoS is improved, since this degradation is negligible compared to the gain obtained when the traffic intensity of this slice has a peak.

Limitations and assumptions of this study

One of the current limitations of this study is that it is not possible to modify the resources assigned to the slices in real-time when simulating, due to simulator constraints. More specifically, this limitation arises from the channel and propagation loss model, which does not allow run-time modifications of frequency- and time-related physical parameters, such as system bandwidth, central carrier frequency or symbol length. Consequently, for the assessment of network performance in this paper, separate simulations were performed for the different time intervals. This is an area that will be developed in the future to allow proactive algorithms that run in real-time and decide the resource allocation of the slices based on actual network conditions.

Conversely, for the evaluation of the RACH, it is assumed that all devices transmit a preamble simultaneously at the

beginning, and that devices with successful preamble transmission do not send additional preambles subsequently.

VIII. Conclusions

The objective of this paper has been to design a novel open-source simulator based on the NS-3 platform and the 5G-LENA module that provides a realistic representation of a distribution center scenario, along with the activities that take place with the aim of assessing network performance.

Under this developed simulator, we evaluated two network slicing strategies using the 5G network in a logistics distribution center: the use of a static slice with a balanced division of network resources and the use of a slice that dynamically adjust its size depending on the traffic activity taken place.

The results show that a dynamic slice in fact results in an improved QoS, especially, under high traffic load. This improvement ranges from 6.48% to 95.65%, depending on the specific traffic profile and the evaluated metric. A static slice performs well when the traffic load is low, since there are sufficient resources to cover the traffic requirements. However, when there is a traffic peak, the static slice does not have enough resources, thus, a reduced QoS is obtained.

While this study offers a clear view of the potential gains that can be obtained with a dynamic slice in a logistics distribution center, there are some features that remain unexplored yet, such as a scalability analysis and a study of linking the wireless performance with the production performance. This will be the subject of future work in upcoming experiments with the developed simulator.

Acknowledgments

This work has been funded by Ministerio de Asuntos Económicos y Transformación Digital y la Unión Europea - NextGenerationEU within the framework "Recuperación, Transformación y Resiliencia y el Mecanismo de Recuperación y Resiliencia" under project MAORI; and by Junta de Andalucía (Consejería de Transformación Económica, Industria, Conocimiento y Universidades, Proyecto de Excelencia) under project PENTA. This work is also partially supported by the "II Plan propio de Investigación y Transferencia de la Universidad de Málaga".

References

[1] Y. Ding, M. Jin, S. Li, and D. Feng, "Smart logistics based on the internet of things technology: An overview," *Int. J. Logist. Res. Appl.*, vol. 24, no. 4, pp. 323–345, Apr. 2021.

[2] A. Alicke, J. Rachor, and A. Seyfert, "Supply chain 4.0–the next-generation digital supply chain, McKinsey & Company," *Supply Chain Management*, Oct. 2016.

[3] L. Kong, M. K. Khan, F. Wu, G. Chen, and P. Zeng, "Millimeter-wave wireless communications for IoT-cloud supported autonomous vehicles: Overview, design, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 62–68, Jan. 2017.

[4] N. Gligoric, S. Krco, L. Hakola, K. Vehmas, S. De, K. Moessner, K. Jansson, I. Polenz, and R. Van Kranenburg, "SmartTags: IoT product passport for circular economy based on printed sensors and unique item-level identifiers," *Sensors*, vol. 19, no. 3, p. 586, Jan. 2019.

[5] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sep. 2018.

[6] A. Baratè, G. Haus, L. A. Ludovico, E. Pagani, and N. Scarabottolo, "5G technology for augmented and virtual reality in education," in *Proc. of the Int. Conf. on Educ. and New Develop.*, Jun. 2019, pp. 512–516.

[7] E. J. Khatib and R. Barco, "Optimization of 5G networks for smart logistics," *Energies*, vol. 14, no. 6, p. 1758, Mar. 2021.

[8] Y. Song, F. R. Yu, L. Zhou, X. Yang, and Z. He, "Applications of the internet of things (IoT) in smart logistics: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4250–4274, Mar. 2020.

[9] Z. Yang, R. Wang, D. Wu, H. Wang, H. Song, and X. Ma, "Local trajectory privacy protection in 5G enabled industrial intelligent logistics," *IEEE Trans. Ind. Informat.*, vol. 18, no. 4, pp. 2868–2876, Apr. 2022.

[10] G. Li, "Development of cold chain logistics transportation system based on 5G network and internet of things system," *Microprocess. Microsyst.*, vol. 80, p. 103565, Feb. 2021.

[11] J. M. Marquez-Barja, S. Hadiwardoyo, B. Lannoo, W. Vandenberghe, E. Kenis, L. Deckers, M. C. Campodonico, K. dos Santos, R. Kusumakar, M. Klepper, and J. Vandenbossche, "Enhanced teleoperated transport and logistics: A 5G cross-border use case," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC) & 6G Summit*, Jun. 2021, pp. 229–234.

[12] J. Zhan, S. Dong, and W. Hu, "IoE-supported smart logistics network communication with optimization and security," *Sustain. Energy Technol. Assess.*, vol. 52, p. 102052, Aug. 2022.

[13] S. Iranmanesh, F. S. Abkenar, R. Raad, and A. Jamalipour, "Improving throughput of 5G cellular networks via 3D placement optimization of logistics drones," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1448–1460, Feb. 2021.

[14] M. Savic, M. Lukic, D. Danilovic, Z. Bodroski, D. Bajović, I. Mezei, D. Vukobratovic, S. Skrbic, and D. Jakovetić, "Deep learning anomaly detection for cellular IoT with applications in smart logistics," *IEEE Access*, vol. 9, pp. 59 406–59 419, 2021.

[15] J. Cheng, Y. Yang, X. Zou, and Y. Zuo, "5G in manufacturing: a literature review and future research," *The International Journal of Advanced Manufacturing Technology*, pp. 1–23, 2022.

[16] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, "URLLC and eMBB in 5G Industrial IoT: A survey," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1134–1163, 2022.

[17] S. Zhang, "An overview of network slicing for 5G," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.

[18] Y. Wu, H.-N. Dai, H. Wang, Z. Xiong, and S. Guo, "A survey of intelligent network slicing management for industrial IoT: Integrated approaches for smart transportation, smart energy, and smart factory," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1175–1211, 2022.

[19] T. Umagiliya, S. Wijethilaka, C. De Alwis, P. Porambage, and M. Liyanage, "Network slicing strategies for smart industry applications," in *2021 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2021, pp. 30–35.

[20] D. Segura, E. J. Khatib, and R. Barco, "Dynamic packet duplication for industrial URLLC," *Sensors*, vol. 22, no. 2, p. 587, 2022.

[21] F. Loch, F. Quint, and I. Brishtel, "Comparing video and augmented reality assistance in manual assembly," in *Proc. 12th Int. Conf. Intell. Environ (IE)*, Sep. 2016, pp. 147–150.

[22] D. Segura, E. J. Khatib, J. Munilla, and R. Barco, "5G numerologies assessment for URLLC in industrial communications," *Sensors*, vol. 21, no. 7, p. 2489, Apr. 2021.

[23] E. R. Alphonsus and M. O. Abdullah, "A review on the applications of programmable logic controllers (PLCs)," *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 1185–1205, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032116000551

[24] S. Mantravadi and C. Møller, "An overview of next-generation Manufacturing Execution Systems: How important is MES for Industry 4.0?" *Procedia Manufacturing*, vol. 30, pp. 588–595, 2019, digital Manufacturing Transforming Industry Towards Sustainable Growth. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2351978919301155

[25] V. Majstorovic, S. Stojadinovic, B. Lalic, and U. Marjanovic, "ERP in Industry 4.0 context," in *Advances in Production Management Systems. The Path to Digital Transformation and Innovation of Production Management Systems*. Springer International Publishing, 2020, pp. 287–294.

[26] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Business & information systems engineering*, vol. 6, no. 4, pp. 239–242, 2014.

[27] T. Adame, A. Bel, B. Bellalta, J. Barcelo, and M. Oliver, "IEEE 802.11AH: the WiFi approach for M2M communications," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 144–152, 2014.

[28] G. Cena, L. Seno, A. Valenzano, and C. Zunino, "On the performance of IEEE 802.11e wireless infrastructures for soft-real-time industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 6, no. 3, pp. 425–437, 2010.
[29] J. Song, S. Han, A. Mok, D. Chen, M. Lucas, M. Nixon, and W. Pratt, "WirelessHART: Applying wireless technology in real-time industrial process control," in *2008 IEEE Real-Time and Embedded Technology and Applications Symposium*, 2008, pp. 377–386.
[30] W. Liang, X. Zhang, Y. Xiao, F. Wang, P. Zeng, and H. Yu, "Survey and experiments of WIA-PA specification of industrial wireless network," *Wireless Communications and Mobile Computing*, vol. 11, no. 8, pp. 1197–1212, 2011.
[31] ZigBee Alliance, "Zigbee specification (document 053474r06, version 1)," 2012. [Online]. Available: https://zigbeealliance.org/wp-content/uploads/2019/11/docs-05-3474-21-0csg-zigbee-specification.pdf
[32] ISA, "100.11 a-2009: Wireless systems for industrial automation: Process control and related applications, ansi/isa," 2009. [Online]. Available: https://www.isa.org/products/ansi-isa-100-11a-2011-wireless-systems-for-industr
[33] G. Montenegro, J. Hui, D. Culler, and N. Kushalnagar, "Transmission of IPv6 Packets over IEEE 802.15.4 Networks," RFC 4944, Sep. 2007. [Online]. Available: https://www.rfc-editor.org/info/rfc4944
[34] SigFox, "SigFox System Description," LPWAN@IETF97, Nov 2016. [Online]. Available: https://datatracker.ietf.org/meeting/97/materials/slides-97-lpwan-25-sigfox-system-description-00.pdf
[35] LoRa Alliance, "LoRaWAN 1.1 Specification," 2017. [Online]. Available: https://lora-alliance.org/resource_hub/lorawan-specification-v1-1/
[36] A. R. Al-Ali, I. Zualkernan, and F. Aloul, "A mobile GPRS-sensors array for air pollution monitoring," *IEEE Sensors Journal*, vol. 10, no. 10, pp. 1666–1671, 2010.
[37] Y. Ma, C. H. Liu, M. Alhussein, Y. Zhang, and M. Chen, "Lte-based humanoid robotics system," *Microprocessors and Microsystems*, vol. 39, no. 8, pp. 1279–1284, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141933115001179
[38] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.
[39] J. Prados-Garzon, P. Ameigeiras, J. Ordonez-Lucena, P. Muñoz, O. Adamuz-Hinojosa, and D. Camps-Mur, "5G Non-Public Networks: Standardization, architectures and challenges," *IEEE Access*, vol. 9, pp. 153 893–153 908, 2021.
[40] *Release 15 Description; Summary of Rel-15 Work Items*, document TR 21.915, v15.0.0, 3GPP, Oct. 2019.
[41] *System architecture for the 5G System (5GS)*, document TS 23.501, V17.11.0, 3GPP, Jan. 2024.
[42] *Procedures for the 5G System (5GS)*, document TS 23.502, V17.11.0, 3GPP, Jan. 2024.
[43] *Policy and charging control framework for the 5G System (5GS); Stage 2*, document TS 23.503, V17.10.0, 3GPP, Sep. 2023.
[44] *NR; NR and NG-RAN Overall description; Stage-2*, document TS 38.300, V16.10.0, 3GPP, Oct. 2022.
[45] *NR; Medium Access Control (MAC) protocol specification*, document TS 38.321, V16.10.0, 3GPP, Oct. 2022.
[46] *NR; Radio Resource Control (RRC); Protocol specification*, document TS 38.331, V16.10.0, 3GPP, Oct. 2022.
[47] *NR; Physical channels and modulation*, document TS 38.211, V16.10.0, 3GPP, Jul. 2022.
[48] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simul. Model. Pract. Theory*, vol. 96, Nov. 2019, Art. no. 101933.
[49] 5G-simulator: Extended 5G-simulator based on NS-3 and 5G-LENA. [Online]. Available: https://github.com/dsr96/5g-simulator.Accessed:30/04/2024.
[50] *Study on Channel Model for Frequencies from 0.5 to 100 GHz*, document TR 38.901, V16.1.0, 3GPP, Nov. 2020.
[51] *NR; Physical layer procedures for control*, document TS 38.213, V16.0.0, 3GPP, Dec. 2019.
[52] ra-simulator - A random-access channel simulator for cellular networks. [Online]. Available: https://github.com/dsr96/ra-simulator.Accessed:30/04/2024.
[53] *Service requirements for cyber-physical control applications in vertical domains*, document TS 22.104, V17.7.0, 3GPP, Sep. 2021.

David Segura received his B.Sc. degree in Telematics Engineering in 2019 and his M.Sc. degree in Telematics and Telecommunication Networks in 2020 from the University of Malaga, Spain. In 2019, he started to work as a research with the Communication Engineering Department, University of Malaga, where he is pursuing a Ph.D. degree in the field of cellular communications. His research interests include wireless communication for Industry 4.0 and security.

Emil J. Khatib (Member, IEEE) is a postdoctoral Juan de la Cierva fellow in the University of Málaga. He got a Ph.D in 2017 on the topic of Machine Learning, Big Data analytics and Knowledge Acquisition applied to the troubleshooting in cellular networks. He has participated in several national and international projects related to Industry 4.0 projects. Currently he is working on the topic of security and localization in industrial scenarios.

Raquel Barco is Full Professor in Telecommunication Engineering at the University of Malaga. Before joining the university, she worked at Telefonica (Madrid, Spain) and at the European Space Agency (ESA) (Darmstadt, Germany). As researcher she is specialized in mobile communication networks and smart-cities, having led projects funded by several million euros, published more than 100 papers in high impact journals and conferences, authored 5 patents and received several research awards.

Chapter 6

Cellular IoT evaluation and security analysis

NB-IoT latency evaluation with real measurements

David Segura
Telecommunication Research Institute (TELMA)
University of Malaga
Malaga, Spain
dsr@ic.uma.es

Emil J. Khatib
Telecommunication Research Institute (TELMA)
University of Malaga
Malaga, Spain
emil@uma.es

Jorge Munilla
Department of Communication Engineering
University of Malaga
Malaga, Spain
munilla@ic.uma.es

Raquel Barco
Telecommunication Research Institute (TELMA)
University of Malaga
Malaga, Spain
rbm@ic.uma.es

2022 IEEE Workshop on Complexity in Engineering (COMPENG) | 978-1-7281-7124-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/COMPENG50184.2022.9905471

***Abstract*—In the 3GPP LTE Release 13, NB-IoT was standardized to provide wide-area connectivity for IoT. To optimize network signaling and power consumption, control plane (CP) optimization was introduced. In Release 15, to support infrequent small data transmissions, Early Data Transmission (EDT) was also included, in which the data are sent during the random access procedure. Thus, this paper analyses the latency performance of the different NB-IoT optimizations for the CP. The study, carried out in a real device, has been performed for different packet sizes and coverage levels. Evaluation results show lower latencies for EDT, particularly with small packets, where a reduced transport block is used, being more efficient from a network point of view. Additionally, we verify that EDT, unlike Release 13 optimization, fulfills 3GPP latency requirement for extreme coverage.**

***Index Terms*—5G, EDT, IoT, NB-IoT, latency, optimization**

I. Introduction

Massive machine type communications (mMTC) are characterized by the presence of a large number of devices that, through sporadic connections, exchange short messages over the mobile network. This type of service is commonly known as Internet of Things (IoT). This service category aims to achieve a higher battery life with a target of 10 years and a maximum latency of 10 seconds in extreme coverage level [1]. To cover these challenges in Celullar IoT (CIoT), the 3rd Generation Partnership Project (3GPP) introduced in Release 13 the Narrow-Band IoT (NB-IoT) technology [2], which is based on Long Term Evolution (LTE) architecture. NB-IoT was developed to provide better coverage, support of massive connections with low power consumption and low throughput. The main problem from the point of view of the network is that a very large number of devices with short messages have a large overhead in terms of establishing and closing connections. Therefore, many optimizations have been developed in order to reduce the signaling exchange between the user equipment (UE) and the base station, and also to increase the battery life.

In Release 13, the concept of suspension and reactivation of the connection for the user plane (UP) was introduced, where the Access Stratum (AS) UE context is stored. Based on this context, the UE can resume data and signaling radio bearers previously established, including the derivation of the security keys, reducing the signaling. Release 13 also introduced the possibility of transmitting small data through the control plane (CP) within Non Access Stratum (NAS) messages, commonly known as Data-over NAS (DoNAS). The CP solution allows sending data without the establishment of the data bearers and AS security.

Moreover, Release 15 introduced the possibility of transmitting user data during the random access (RA) procedure, known as Early Data Transmission (EDT). This optimization is focused on infrequent and small data transmissions, where the latency and battery consumption are mainly due to the connection establishment procedure with the network.

Some studies have already evaluated the latency performance of these optimizations. In [3], [4], EDT performance is evaluated for the UP and CP, assuming that the user data of the different traffic profiles evaluated can fit in the message 3 (Msg3), using an ideal Transport Block Size (TBS). On the other hand, in [5], EDT is evaluated for the UP, assuming a certain size for the Msg3 overhead. Finally, a mathematical model for estimating the latency of EDT is presented in [6], which takes into account the TBS and the modulation scheme used. However, none of these studies provide real measurements, but use analytical frameworks or simulators and they also assume different ideal situations, which do not correspond with the actual implementation of the 3GPP standard. In this paper, an evaluation and comparison of the latency of NB-IoT optimizations based on the CP is carried out for different packet sizes and coverage levels, using a real device, whose implementation is based on the most recent 3GPP specification (Release 16).

This work has been partially funded by the Ministerio de Asuntos Económicos y Transformación Digital y la Unión Europea - NextGenerationEU, en el marco del Plan de Recuperación, Transformación y Resiliencia y el Mecanismo de Recuperación y Resiliencia under project MAORI and, by the Junta de Andalucía and European Union (ERDF) under grant UMA-18-FEDERJA-172.

The remainder of this paper is organized as follows. A brief description of NB-IoT technology is presented in Section II. The different CP optimizations for the user data transmission are described in Section III. The methodology followed is described in Section IV. The results obtained are shown in Section V. Finally, conclusions are drawn in Section VI.

II. NB-IoT

NB-IoT technology was introduced as part of Release 13 to cover mMTC use cases [2]. These use cases are characterized by massive connection support, low complexity to provide low cost devices, low consumption and coverage enhancements [7]. NB-IoT is a narrowband system which operates with a channel bandwidth of 180 kHz and provides a maximum coupling loss (MCL) of 164 dB. Coverage enhancement (CE) is mainly achieved by using transmission repetitions on physical channels [8]. NB-IoT supports three different operation modes: in-band (using one physical resource block within a normal LTE carrier), guard-band (using unused resource blocks within a LTE carrier guard-band) and standalone (using a dedicated carrier).

NB-IoT is based on LTE technology, therefore, NB-IoT inherits part of its design such as channel codification, numerology, modulation scheme and higher protocols layers. Nevertheless, to reduce the complexity and the cost of these devices, some features are removed such as mobility in connected mode (handover). In the downlink, Orthogonal Frequency-Division Multiple Access (OFDMA) is used with a subcarrier spacing (SCS) of 15 kHz over 12 subcarriers with 14 OFDM symbols, where the subframe duration is 1 ms, same as LTE. On the other hand, in the uplink (UL), Single Carrier Frequency Division Multiple Access (SC-FDMA) is used, where the SCS can be 3.75 kHz or 15 kHz [8].

In addition, to reduce the peak-to-average power ratio (PAPR) in the UL, the modulation is limited in the transmissions, where $\pi/2$-binary phase-shift keying (BPSK) and $\pi/4$-quadrature (QPSK) schemes are adopted. To operate with NB-IoT devices, only one antenna is necessary and one Hybrid Automatic Repeat Request (HARQ) process is supported in Release 13 for UL and DL, whereas Release 14 supports up to two HARQ processes. Also, two classes of maximum output power are supported by NB-IoT devices, 20 dBm and 23 dBm.

In terms of TBS, Release 13 supports a maximum TBS size of 1000 bits for UL and 680 bits for DL, whereas in Release 14 the maximum TBS supported is increased to 2536 bits for both [8].

III. Data transmission optimization in CIoT

This section describes the different optimizations for the UL data transmission in CIoT, based on the CP, introduced by the 3GPP to improve the communications [9]. In this case, the UE is in Radio Resource Control (RRC) Idle state and, when it has data to send, the UE starts the connection establishment with the network to transmit the data.

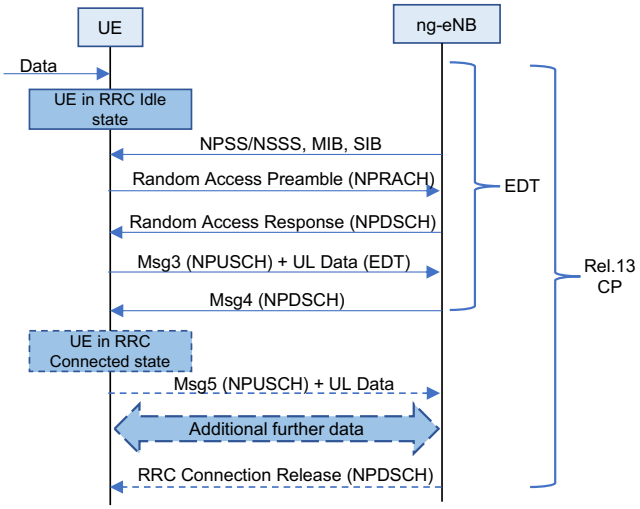


Fig. 1. Signaling messages exchange in the network with the different optimizations.

A. Control Plane CIoT Optimization

The CP optimization was standardized in Release 13 and consists of transmitting small data through the CP (instead of the UP), as shown in Fig. 1. The support of this feature is mandatory for NB-IoT devices. With this optimization, the user data are sent within NAS signaling messages to the core network in Msg5. When using this connection mode, the UE avoids the establishment of the UP bearers and the AS security each time it has data to transmit. In this case, the UE moves to RRC Connected state, so it can send more data if necessary. After a period of inactivity set by the network, the UE returns to RRC Idle state when receiving the RRC Connection Release message.

B. Control Plane CIoT Early Data Transmission

Release 15 introduced the possibility of transmitting data during the RA procedure, known as EDT. This optimization is focused on devices that transmits small and infrequent data, with the aim of improving the latency and battery consumption. Although this optimization has been proposed for the UP and the CP, this paper addresses the CP solution.

The RA procedure consists of four steps: the preamble transmission (1), the preamble response (2), the connection establishment request (3) and the connection establishment (4), as shown in Fig. 1. EDT was created to send UL data in Msg3 (within NAS messages), without further need for the establishment of an RRC connection and a state change in the UE; significantly reducing both signaling and wake-up time in the UE.

To be able to use this optimization, a special preamble is used in Msg1, which lets the base station know that the UE has small data to transmit. Then, in Msg2, the base station returns a TBS, which indicates the maximum size of the Msg3 (RRC message + user data). For EDT, the maximum TBS allowed is 1000 bits, whereas the minimum is 328 bits [10]. Finally,

TABLE I
NB-IoT CELL CONFIGURATION PARAMETERS

Parameter	**Value**			
Cell band	Band 7			
Cell bandwidth	180 kHz for both UL and DL			
Operation mode	Standalone			
Coverage levels		CE0	CE1	CE2
Coupling loss		144 dB	154 dB	164 dB
Repetitions	NPRACH	2	8	32
	NPDCCH	2	8	64
	NPUSCH	2	8	64
	NPDSCH	2	8	64
	Msg3 NPUSCH	2	8	128
Bandwidth for UL Tx		15 kHz	15 kHz	15 kHz
Number of subcarriers for NPUSCH		12	12	12
Number of subcarriers for Msg3		1	1	1
I_{MCS}	NPUSCH	4	4	4
	NPDSCH	5	5	5
EDT I_{MCS}	Msg3	7	5	3
NPRACH periodicity (ms)	Normal	80	160	320
	EDT	40	160	640

in Release 16, the support of EDT was also included in the fifth-generation (5G) core, using a next-generation eNodeB (ng-eNB) from the LTE Radio Access Network (RAN) [11].

IV. METHODOLOGY

This section describes the methodology used to evaluate the different data transmission optimizations for NB-IoT devices.

To evaluate these optimizations, AMARI Callbox Classic and AMARI UE Simbox solutions from Amarisoft [12] have been used. Both devices have a completely software-based network implementation, where different network elements are deployed in a virtualized way (base station and core in the Callbox; and UE terminal in the UE Simbox). This virtualization allows configuring different networks using the same physical device and thus, adding more flexibility. Regarding the Callbox, it allows the implementation of many LTE/New Radio (NR) network elements by different scripts, such as Mobility Management Entity/Access and Mobility Management Function, as well as a large number of protocols and interfaces of an LTE/NR network and thus creating a virtual core. Similarly, the software allows to create instances of eNodeB/ng-eNB/gNB, through which it is allowed to manage the software-defined radio card of the device. All of this is implemented on a PC running on top of the Linux operating system. Same as the Callbox, the UE Simbox allows a software implementation of a virtualized UE, where the different network elements of the UE are implemented, along with its protocol layers. In this case, the UE Simbox allows LTE, NB-IoT/LTE-M and NR devices. The entire implementation of both devices is based on the 3GPP standard with support up to Release 16.

In this study, the Callbox has been configured with a 5G core and one NB-IoT cell using a ng-eNB, since Release 16 allows LTE radio access IoT devices to connect to a 5G core via a ng-eNB. This implies the support of 5G NAS message transport and security framework, except for data integrity protection [9]. On the other hand, the UE Simbox has been configured with a category 2 NB-IoT device, with support of control plane optimizations of Release 13 and Release 16 (EDT). Also, one transmission antenna has been used in the base station and the UE, with two HARQ processes. Table I summarizes the different NB-IoT cell configuration parameters used for each CE: normal (CE0), robust (CE1) and extreme (CE2). The coupling loss is defined as the attenuation in the signal strength between the transmit antenna port and the receive antenna port. Also, the CE is determined by the UE based on a Reference Signal Received Power (RSRP) threshold set by the network, where on each CE different transmission repetitions on physical channels are used. These CE have been configured adjusting the antenna gain on the base station and UE.

This study is centered on infrequent UL data transmissions via the CP, carrying out a comparison of the latency performance between Release 13 and Release 16, using for that a real device based on the 3GPP standard, as explained above. For that, the UE sends an UL ping packet with a size of 12 and 60 bytes, where EDT has been configured with a maximum TBS for the Msg3 of 125 bytes and 73 bytes. The IP header (20 bytes) and ICMP header (8 bytes) are added to the size of these packets. Thus, at the application layer (including the overhead), the packets have a size of 40 and 88 bytes, respectively. To obtain statistical results, 50 UL data transmissions have been performed for each CE, connection mode, packet size and TBS for EDT.

Finally, the latency is defined as the time from when the UE has data to send at application layer from RRC Idle state and RA procedure is triggered until when the UL data are delivered at the base station. For infrequent transmissions, the 3GPP has defined a maximum latency of 10 seconds for a size of 105 bytes at the physical layer with 164 dB of MCL [1].

V. Results

The results obtained with the different optimizations and coverage levels are presented below. Fig. 2 shows the latency distribution obtained for the UL data transmission from the NB-IoT device, based on the connection mode and packet size at CE0. For EDT transmissions, the TBS indicates the maximum data size in bytes to transmit in Msg3.

When Release 13 optimization is used, it is observed that the packet size has a high influence on the latency, producing a higher fluctuation on the values, where the user data are sent along with the Msg5 and it does not have an imposed limitation of the TBS, since the UE is in RRC Connected state and it can continue sending more data if necessary. In this case, the 90th percentile value obtained with packet sizes of 40 and 88 bytes are 435.04 ms and 487.03 ms, respectively. On the contrary, with EDT optimization, the latency distribution for both packet sizes are very similar with a TBS of 125 bytes, obtaining a 90th percentile value of 263.05 ms and 238.89 ms with packet sizes of 40 and 88 bytes, respectively. This is because, when using EDT, the TBS indicated by the network must be sent completely, so if the user information is less than the TBS, padding is added to the Msg3 until complete it. This, logically, reduces the efficiency when the data size is small compared to the TBS. As a result, the use of a small TBS improves the efficiency of EDT. This can be observed when a TBS of 73 bytes is used with a packet size of 40 bytes, where the latency is slightly reduced, due to the fact of reducing the TBS, being more efficient and obtaining a 90th percentile value of 224.67 ms. The outliers existing in the latency distribution are originated by retransmissions in the RA procedure, prior to sending the data in Msg3.

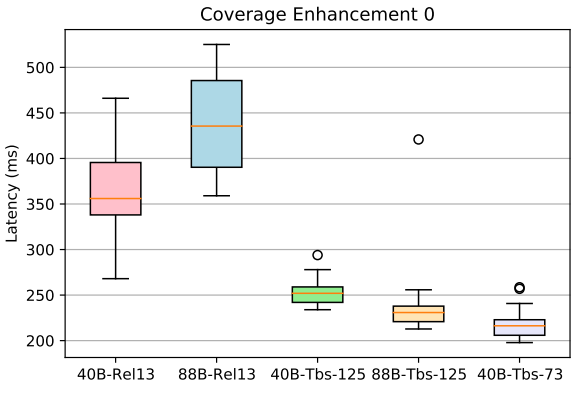


Fig. 2. Latency obtained with each connection mode at CE0.

Fig. 3 shows the latency distribution obtained at CE1 for each packet size and connection mode. In this case, unlike CE0, it is observed that the latency values have suffered a considerable increase. This is mainly due to the increase on the physical channel repetitions. Upon this case, the Release 13 optimization has suffered a more pronounced increase in the latency values, obtaining a 90th percentile value of 1146.65 ms and 1313.68 ms for packet sizes of 40 and 88 bytes, respectively. On the other hand, same as in CE0, the latency for EDT when the TBS value is fixed at 125 bytes is similar, with a 90th percentile value of 689.97 ms and 670.64 ms for 40 and 88 bytes, respectively. With a small TBS (73 bytes), the latency is reduced, obtaining a 90th percentile value of 615.17 ms. In both cases, the improvement of using a smaller TBS (73 bytes) is approximately 10% (the advantage is maintained with coverage changes). The fact that in Release 13 the latency increase is higher than in EDT is mainly because of transmitting in Msg5 instead of Msg3 and with a higher number of repetitions. Therefore, as the number of repetitions increases, the time transmitting the data also increases, which results in more latency and a higher impact on the battery of the UE. On the contrary, reducing the number of messages to transmit the user data, as EDT does, increases the efficiency and reduces the latency. Additionally, it also improves the battery life, since the time transmitting/receiving data decreases.

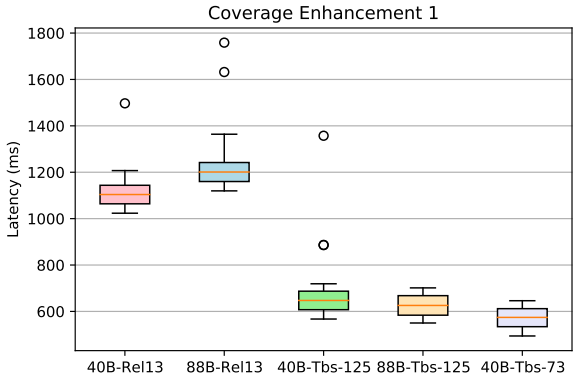


Fig. 3. Latency obtained with each connection mode at CE1.

Finally, the latency distribution at CE2 for each connection mode and packet size is represented in Fig. 4. Under these coverage conditions, it is observed that the latency obtained with Release 13 optimization is around 9 seconds for 40 bytes and exceeds 10 seconds for 88 bytes. For EDT, same as at CE0 and CE1, the latency obtained with a TBS of 125 bytes is similar for both packet sizes, obtaining a 90th percentile value of 5905.45 ms and 6063.97 ms for packet sizes of 40 and 88 bytes. However, unlike the previous cases, with extreme coverage, the latency reduction is increased when using a small TBS, obtaining a reduction in the 90th percentile of more than 25%, with a value of 4744.66 ms. Thus, the 3GPP latency requirement of 10 seconds is not fulfilled with Release 13 optimization, whereas it is with EDT. Finally, Table II collects the latency results obtained for the 90th percentile in all the cases evaluated.

TABLE II
90TH PERCENTILE LATENCY VALUES OBTAINED WITH EACH CONNECTION MODE

UL data at application layer (bytes)	UL data at MAC layer with EDT (bytes)	144 dB MCL			155 dB MCL			164 dB MCL		
		Rel. 13	Rel. 16 EDT		Rel. 13	Rel. 16 EDT		Rel. 13	Rel. 16 EDT	
			TBS (bytes)			TBS (bytes)			TBS (bytes)	
			125	73		125	73		125	73
		Latency (ms)								
40	70	435.04	263.05	224.67	1146.65	689.97	615.17	9128.67	5905.45	4744.66
88	118	487.03	238.89	✗	1313.68	670.64	✗	10422.29	6063.97	✗

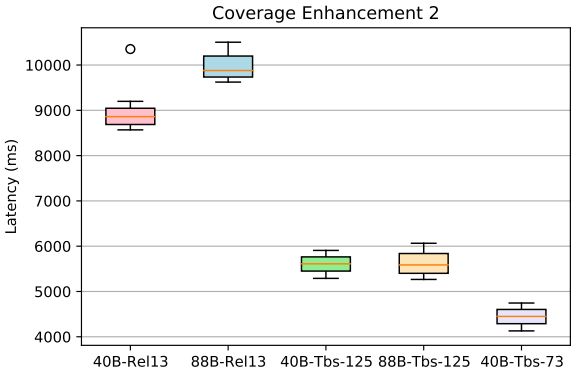


Fig. 4. Latency obtained with each connection mode at CE2.

VI. CONCLUSIONS

In this paper, the different connection modes via the CP for NB-IoT devices in terms of latency for infrequent transmissions have been evaluated, using real measurements. It has been proven that EDT reduces significantly the latency, especially at CE2 and when using a small packet size, where a small TBS can be used, being the transmission more efficient. In contrast, Release 13 CP optimization presents higher latency values, since the data transmission occurs later, not achieving the 3GPP latency requirement at CE2.

Therefore, it is concluded that the usage of EDT optimization is necessary to fulfill the 3GPP latency requirement for infrequent small data transmissions, in particular, with extreme coverage level (CE2). Network operators will need to find a balance in the selection of the TBS for EDT to: (1) guarantee the usage of EDT for the majority of devices; and (2) perform an efficient data transmission, avoiding the resource wastage.

REFERENCES

[1] *Study on scenarios and requirements for next generation access technologies*, document TR 38.913, Rel. 16, v16.0.0, 3GPP, Jul. 2020.

[2] *Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description*, document TS 36.201, Rel. 13, v13.2.0, 3GPP, Jun. 2016.

[3] *Evaluation for early data transmissions*, TSG-RAN WG2 #100, document R2-1713058, 3GPP, Nov. 2017.

[4] A. Hoglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.

[5] O. Liberg, J. Bergman, A. Höglund, T. Khan, G. A. Medina-Acosta, H. Rydén, A. Ratilainen, D. Sandberg, Y. Sui, T. Tirronen, and Y. P. E. Wang, "Narrowband internet of things 5G performance," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, doi: 10.1109/VTC-Fall.2019.8891588.

[6] F. J. Dian and R. Vahidnia, "A simplistic view on latency of random access in cellular Internet of Things," in *Proc. IEEE 11th Annu. Inf. Technol. Electron. Mob. Commun. Conf. (IEMCON)*, Nov. 2020, pp. 0391–0395.

[7] *Service requirements for Machine-Type Communications (MTC); Stage 1*, document TS 22.368, Rel. 13, v13.2.0, 3GPP, Dec. 2016.

[8] M. Kanj, V. Savaux, and M. Le Guen, "A tutorial on NB-IoT physical layer design," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2408–2446, 4th Quart., 2020.

[9] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*, document TS 36.300, Rel. 16, v16.5.0, 3GPP, Mar. 2021.

[10] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*, document TS 36.213, Rel. 16, v16.6.0, 3GPP, Jun. 2021.

[11] D. Segura, J. Munilla, E. J. Khatib, and R. Barco, "5G early data transmission (Rel-16): Security review and open issues," *IEEE Access*, vol. 10, pp. 93 289–93 308, Sep. 2022.

[12] "Amarisoft," https://www.amarisoft.com/, (accessed March 2022).

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2022, accepted 28 August 2022, date of publication 1 September 2022, date of current version 12 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3203722

TUTORIAL

5G Early Data Transmission (Rel-16): Security Review and Open Issues

DAVID SEGURA[1,2], JORGE MUNILLA[1], EMIL J. KHATIB[1,2], (Member, IEEE), AND RAQUEL BARCO[1,2]

[1]Department of Communications Engineering, University of Malaga, 29010 Málaga, Spain
[2]E.T.S.I. de Telecomunicación, Telecommunication Research Institute (TELMA), University of Malaga, 29010 Málaga, Spain

Corresponding author: Emil J. Khatib (emil@uma.es)

This work was supported in part by the Junta de Andalucía and European Union [European Regional Development Fund (ERDF)] under Grant UMA18-FEDERJA-172, and in part by the European Union-NextGenerationEU within the Framework of the Project "Massive AI for the Open RadIo b5G/6G Network (MAORI)."

ABSTRACT The fifth-generation technology is called to support the next generation of wireless services and realize the "Internet of Everything" through Machine-Type Communications and Cellular Internet of Things optimizations. As part of these optimizations, Release 15 introduced a new mechanism, known as Early Data Transmission (EDT), that allows the transmission of data during the Random Access procedure. This feature, intended particularly for infrequent and small data transmissions, aims to reduce the latency and the power consumption of user equipments. Nonetheless, despite the importance of this novelty and the general agreement about its effectiveness, there are few papers in the literature that provide insight into its implementation and analyze the advantages and disadvantages of its two different implementation options (Control and User Plane). Moreover, although security is recognized as a crucial aspect for the correct deployment of this technology, we have not found any paper that reviews the security issues and features of this mechanism. As a consequence of such a lack of papers and the complexity of mobile network protocols, there is a divide between security experts and EDT researchers, that prevents the easy development of security schemes. To overcome this important gap, this paper offers a tutorial of EDT and its security, analyzing its main vulnerabilities and concluding with a set of recommendations for researchers and manufacturers. In particular, due to the simplifications in the protocols done by EDT, vulnerabilities such as packet injection, replay attacks and injection of fake values to disable EDT have been found.

INDEX TERMS 5G mobile communication, Cellular Internet of Things (CIoT), early data transmission (EDT), massive machine-type communication (mMTC), security.

I. INTRODUCTION

The fifth-generation (5G) of cellular networks was designed by the 3rd Generation Partnership Project (3GPP [1]) having Internet of Things (IoT) as one of its main use-cases. While prior generations of 3GPP networks supported mainly end user communications, in 5G new requirements for Machine-Type Communications (MTC) were included in the design. As a result, three different kind of traffic profiles [2] are supported by the 5G network:

- Enhanced Mobile Broadband (eMBB): for applications with a very high bandwidth requirement (up to 100 Gbps). End user communications fall into this group, but also MTC applications that rely, for instance, on video feeds or high resolution images.
- Massive MTC (mMTC): for very high density of devices (10^6 devices/km^2), each with low requirements in bandwidth, latency or reliability. Typically, sensor networks fall into this category.
- Ultra-Reliable Low-Latency Communications (URLLC): for mission critical services, that require very low latency (down to 1 ms) and high reliability (above 99.999%). This traffic profile corresponds, for instance,

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

to safety systems or control of potentially hazardous machines.

In this paper, the focus is set on mMTC. While these kinds of communications are not usually critical (i.e., they pose no risk of personal damage), the malfunction of systems that rely on them (monitoring systems with many sensors, tracking of items, etc.) may cause great economic loss.

Securing mobile technologies is a major challenge. Privacy issues detected in earlier mobile network generations have increased the public distrust in this technology and security has been revealed to be a crucial issue that may derail or, at least, delay large-scale adoption. The 3GPP consortium, which designed the third-generation (3G) and fourth-generation (4G) standards and is now involved in the development of 5G, has already defined a security architecture for 5G systems [3]. While convenient, IoT devices can become a serious issue if not secured properly. Their wireless nature makes it more difficult to detect illicit activity, such as eavesdropping or spoofing messages. This puts users at risk of confidentiality breaches or sabotage, with a high potential for economic loss. Therefore, to reduce reluctance of stakeholders to using 5G solutions, one of the main points to address is to ensure security.

The main problem from the point of view of the network is that a very large number of devices with short messages have a large overhead in terms of establishing and closing connections. While this issue does not affect in terms of bandwidth requirements, it does put a pressure on the networks' access mechanisms. In fact, this overhead will cause congestion in traditional cellular networks [4]. To solve this, one obvious solution is to reduce the overhead of connections. Early Data Transmission (EDT [5]) was proposed in 3GPP Release 15, reducing significantly the amount of signaling needed to set up a connection. Specifically, EDT allows to transmit data during the Random Access (RA) procedure which is normally used to initiate the connection setup. Furthermore, EDT has served as a baseline for improving the RA procedure for URLLC, with the two-step RA procedure mechanism [6]. EDT has been described in [7], where the concept is explained in detail and a performance analysis is done, showing the gains in terms of latency and power saving. Also, the studies in [8], [9], [10], and [11] have already evaluated the performance of EDT, showing the gains when using this optimization. In [8] the battery life and latency of EDT are evaluated under different coverage levels. In [9] a latency evaluation of EDT is performed, whereas in [10] a mathematical model for estimating the latency of EDT is presented. Moreover, in [11] the authors present a battery life model that considers the cell load, where EDT is evaluated using this model. Although there are few studies covering the performance aspects of EDT, no study performs a global vision of this feature in terms of security and its associated vulnerabilities.

There remain several open questions about the usage of EDT, and this paper has the main objective of highlighting them and providing an answer based on the thorough study conducted by the authors. Firstly, there is a lack of research/technical papers that provide a global vision of this feature and, in particular, the security of EDT is a topic treated very tangentially, where no study performs an exhaustive analysis of EDT in terms of security and its vulnerabilities. While the documents of 3GPP standards on EDT [12], [13], [14], [15], [16] do provide a detailed technical description, they are often hard to read due to the excessive reliance on prior mechanisms, and thus, they fail at conveying a clear vision of the technology. This is a major roadblock in the research in security for EDT, since it poses an entry barrier for experts in security, who are normally not specialized in cellular technology and lack the required knowledge to comfortably understand and see the vulnerabilities of the standards. Additionally, there is no clear guideline either on when or why to use the two possible EDT modes: Control Plane (CP) and User Plane (UP), which will be later defined in more detail.

Therefore, this paper targets two different audiences: *i*) security experts who wish to know more about security of EDT and 5G in general; and *ii*) 5G experts needing a deeper knowledge about 5G security, EDT and its possible security shortcomings. Thus, it first does a thorough description of the security methods implemented in 5G that affect EDT. Second, an overview of existing Cellular IoT (CIoT) technologies is provided, and how these technologies serve as a base for EDT. Then, some security threats and open questions about implementation and security issues are discussed; and finally, a comparative between the two operation modes of EDT is drawn. To round up the main takeouts of the paper, a summary of recommendations derived from the analytical work of EDT security is done.

The rest of the paper is organized as follows. In Section II the security procedures in 5G are reviewed. In Section III the different existing 5G CIoT technologies are described. In Section IV, EDT is described in detail and in Section V the open security aspects of EDT are reviewed. In Section VI the UP/CP selection problem is described, and potential insights from the security standpoint are provided. Finally the conclusions are summarized in Section VII, along with recommendations for policy makers and vendors.

The main notations and abbreviations used in this paper are listed in Tables 1 and 2, respectively.

II. SECURITY OVERVIEW OF 5G

A. SECURITY ARCHITECTURE

The security architecture of 5G systems is described in the technical specification 3GPP TS 33.501 [3]. This architecture is divided into two domains: the Subscriber and the Network domain. In the Subscriber domain we find the User Equipment (UE), whereas in the Network domain there are two elements: the Home Network (HN) and the Serving Network (SN). Each of them contains different modules or

TABLE 1. List of main notations.

Notation	Description
K	Long-term secret key, stored in HN and UE
pk_{HN}	Public key of HN, stored in HN and UE
sk_{HN}	Secret key of the public key cryptosystem, stored in HN
K_{SEAF}	Anchor key, used to derive integrity and encryption keys
SQ_{UE}, SQ_{HN}	Sequence numbers, stored in UE and HN, respectively
SUPI	Subscription Permanent Identifier
SUCI	Subscription Concealed Identifier
GUTI	Global Unique Temporary Identifier of the UE
SN_{name}	Serving Network identifier
$\text{enc}_{\text{pk}_{\text{HN}}}(\cdot)$	Public key encryption with pk_{HN}
KDF	Key Derivation Function
f1-f5	One-way key functions
f1*, f5*	One-way key functions (used for re-sync)
SHA256	SHA hash function with output of 256 bits
ne	Random bitstring for public key cryptosystem
R	Random number
MAC	Message authentication code
AUTN	Authentication value
XRES, XRES*	Actual and computed Challenge response
HXRES, HXRES*	Hashed values of XRES and XRES*

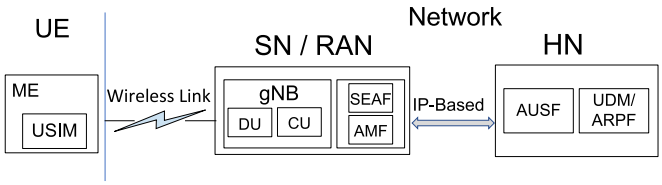


FIGURE 1. Subscriber (UE) and Network domains and submodules.

subsystems, and Fig. 1 sketches the most important for the security aspects.

The UE contains the Mobile Equipment (ME) of the subscriber, typically a smartphone or an IoT device, equipped with a Universal Subscriber Identity Module (USIM), which has cryptographic capabilities and stores the subscriber's credentials provided by the network operator. For IoT devices, traditional plastic (physical) cards are substituted for (virtual) embedded SIMS (eSIM), where credentials are remotely provisioned, to remove the cost of SIM card hardware and installation.

The HN belongs to the subscribers' operator, manages subscriber information at the Unified Data Management (UDM) and is in charge of verifying subscribers' authentication requests, using the Authentication Credential Repository and Processing Function (ARPF) and Authentication Server Function (AUSF).

Finally, the SN receives from HN and stores in SEcurity Anchor Function (SEAF) the anchor key, and connects the UE with the HN, providing wireless access to the UE through its base stations, called Next Generation NodeB (gNB). It manages registration, reachability and mobility, implemented in the Access and Mobility Management Function (AMF)

TABLE 2. List of main abbreviations (in order of appearance in the text).

Abbreviation	Description
5G	The fifth-generation
3GPP	3rd Generation Partnership Project
IoT	Internet of Things
eMBB	Enhanced Mobile Broadband
mMTC	Massive Machine-Type Communications
URLLC	Ultra-Reliable and Low-Latency Communications
EDT	Early Data Transmission
RA	Random Access
CP	Control Plane
UP	User Plane
CIoT	Cellular IoT
UE	User Equipment
HN	Home Network
SN	Serving Network
SEAF	Security Anchor Function
gNB	Next Generation NodeB
AMF	Access and Mobility Management Function
AS	Access Stratum
NAS	Non-Access Stratum
RRC	Radio Resource Control
PDCP	Packet Data Convergence Protocol
RLC	Radio Link Control
AKA	Authentication and Key Agreement
SRB	Signaling Radio Bearer
NCC	Next Hop Chaining Counter
RNTI	Radio Network Temporary Identifier
LTE	Long Term Evolution
LTE-M	LTE for Machine-Type Communication
NB-IoT	NarrowBand IoT
eDRX	Extended Discontinuous Reception
RAI	Release Assistance Indication
ng-eNB	Next Generation Evolved NodeB
SIB	System Information Block
RAR	Random Access Response
TBS	Transport Block Size
PDU	Protocol Data Unit
CCCH	Common Control Channel
HARQ	Hybrid Automatic Repeat Request
DRB	Data Radio Bearer
DTCH	Dedicated Traffic Channel
MIB	Master Information Block
SMF	Session Management Function

(similar to the Mobility Management Entity, MME, in 4G). gNB functionality is split into two functional units: a distributed unit (DU), which contains the antenna and the physical layer; and the centralized unit (CU), which controls one or several DU. SN and HN may belong to the same or different (like in roaming) operators.

Thus, for general threat models it is assumed that UE and SN are connected over an untrusted wireless channel, whereas SN and HN communicate using a trusted channel [3] (Clause 5.9.3). For the UE-SN channel, the ability (control) of adversaries is usually modeled using the Dolev–Yao (DY) model [17]. In this model, the network is controlled by the adversary; passive adversaries can eavesdrop on the communication, whereas active adversaries can also intercept, inject, manipulate or drop messages. Under this model, there are so many attack variants that their taxonomy is challenging and too vast to be described here [18], but we summarize the most important for the understanding of this paper. Firstly, replay attacks are one of the simplest attacks carried out by

an active adversary, and they consist of two phases. In a first phase, the adversary eavesdrops or intercepts legitimate messages sent by one of the parties, and later, the adversary replays these messages with no or slight modifications to the other party. Man-in-the-middle-attacks (MitM) can be considered an on-line version of replay attacks where the messages are intercepted and modified while the communication between the parties is taking place. Regarding the aim of the attack, we can highlight four: traceability, impersonation, Denial/Degradation of Service (DoS) attacks and bidding-down attacks. The first three as they are the most representative attacks against the three vertexes of information security: confidentiality (privacy), integrity and availability, respectively. And bidding-down attacks because it becomes relevant for systems where devices with different security capabilities interact, being the protection against these attacks explicitly pointed out as a security requirement by the standard (Clause 5.1.1 [3]). In bidding-down attacks the adversary tries to make UE and network entities believe that the other side does not support a security feature, even when both sides do support it. The goal is that the parties use weaker cryptographic systems. In traceability attacks, which are a particular case of privacy attacks, the adversary is able to determine the participation of a device in a specific communication and thus infer certain information about that device: location, communication frequency, type of exchanged information, etc. In impersonation attacks (aka spoofing), the adversary manages to impersonate one of the parties and communicate with the other on behalf of it. Finally, denial and degradation of service attacks aims to compromise the availability of the system by interrupting completely or temporarily the service or decreasing its performance.

B. STRATUMS, COMMUNICATION PLANES AND PROTOCOL LAYERS

Two stratums are distinguished in 5G: Access Stratum (AS) and Non-Access Stratum (NAS). AS refers to the communication between UE and gNB, whereas NAS refers to the communication between UE and HN. The limit between both stratums is located in AMF.

Similarly to Internet Protocols (IP) with control and data planes, communication in the 5G network is partitioned into CP and UP, where the CP is used to transmit signaling data and the UP the user data. The 5G protocol stack, as shown in Fig. 2, changes for the two planes: in the UP, the application layer is served by transport protocols such as Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) which run over an IP layer, whereas in the CP, the NAS and the Radio Resource Control (RRC) layers generate the signaling messages that are exchanged between the UE and the HN (NAS), and the UE and SN (AS), respectively. Next, for both planes, the information is processed by the Packet Data Convergence Protocol (PDCP) layer, which applies ciphering and integrity protection to the CP in the AS (not in the CP of the NAS) and to the user data (UP); the Radio Link Control (RLC) layer, which checks the correct delivery of the messages; and the Medium Access Control (MAC) layer, before being passed to the physical layer for transmission.

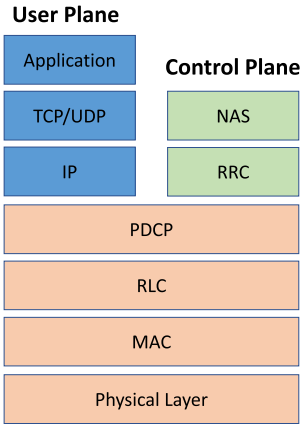


FIGURE 2. Air protocol stack.

C. 5G-AKA PROTOCOL

Prior to an IoT or any other device being able to securely communicate, 5G requires an authentication process. This first authentication is named primary authentication and is compulsory for all devices regardless of the service or network access they request (agnostic access network). The purpose of this primary authentication is to enable mutual authentication between the UE and the network, and provide keying material that can be used between the UE and the SN. A secondary authentication, intended for services provided over 5G and related to security domains outside of the serving network (e.g., access to corporate data), is also possible in 5G but it is optional. For the primary authentication, 3GPP proposes a new Authentication and Key Agreement (AKA) protocol, named 5G-AKA, although it is still possible to use the previous EAP-AKA'. They are quite similar but for the inclusion of some messages and slight changes in the key derivation. For more details, we refer the reader to [19] and Clause 6.1.3.1 of [3].

The 5G-AKA protocol employs the exclusive-or operation "$\oplus$" for one-time pad encryption, a public key encryption function $\text{enc}_{\text{pk}_{\text{HN}}}(\cdot)$ with public key pk_{HN} and secret key stored in HN, a key derivation function KDF based on the hash function SHA256 [20], and seven one-way keyed authentication and key generation functions: f1, f2, f3, f4, f5, and f1*, f5*. The standard does not specify an implementation for these keyed functions but only security requirements. In particular, they should be cryptographically secure and mutually independent. That is, without knowing the key, their outputs should be practically *indistinguishable* from independent random functions, and it should be infeasible to determine any part of the key, or the operator variant configuration field, by manipulating their inputs and examining their straightforward or combined outputs. 3GPP partners have developed an example of authentication and key generation functions, called the MILENAGE algorithms [21],

which can be used for those who do not want to develop their own functions. Regarding KDF, the final output, i.e., the derived key, is equal to the KDF computed on a string S, whose construction concatenating the input parameters and their lengths is described in [22], using a key K as follows: derived key = HMAC-SHA256(K, S) as defined in [20] and [23].

For executing the protocol, the USIM (located within the UE) stores privately the following elements:

- Subscription Permanent Identifier (SUPI), which uniquely identifies the UE.
- A long-term secret key K, which is different for each subscriber.
- The public key pk_{HN} of the HN.
- A sequence number SQ_{UE}, which is incremented with each execution of the protocol to prevent replay attacks.

HN stores the SUPI too and also the secret key of the public key cryptosystem sk_{HN} and its own sequence number SQ_{HN}. Randomized public key encryption (with ne a random bitstring) is used to conceal the subscriber identity: SUCI = $\text{enc}_{\text{pk}_{\text{HN}}}$(SUPI, ne), corresponding, in practice, to the "scheme-output" field within SUCI packets, which also includes information needed for routing and setting the protection scheme. Thus, the SUPI is never sent in the clear to prevent the most famous attack available on previous versions of cellular technologies: "International Mobile Subscriber Identity (IMSI) catcher" attacks [24], [25], which compromised the privacy of subscribers having access to their identities during transmissions. Additionally, a Global Unique Temporary Identifier (GUTI) can also be used to identify the UE [26]. This is a temporary identity assigned and sent to UE by HN (AMF) after a successful authentication (activation of NAS security) upon receiving Registration Request messages of type "initial registration", "mobility registration update" or "periodic registration update"; or Service Request message sent by the UE in response to a Paging message, although it is left to implementation to reassign GUTI more frequently (Clause 6.12.3 of [3]). GUTI identifies the AMF that allocated the (GUTI) and the UE within this AMF, and replaces the encrypted SUPI, thus avoiding a public key encryption and a random number generation.

Fig. 3 describes the flows of the 5G-AKA protocol [27], where || stands for concatenation. The protocol consists of three phases:

- The first phase is the initiation of the authentication procedure and the selection of the authentication method. The SEAF of SN may initiate an authentication with the UE during any procedure establishing a signaling connection with the UE. If UE does not have a valid GUTI, then it computes and sends the SUCI to SN, who relays it to HN. Otherwise, UE sends the GUTI to SN. If the SN (AMF) is able to obtain the corresponding SUCI, by looking it up in its database, then SN forwards it to HN. If not, it requests ("Identifier Request Message") the SUCI to UE and relays it to HN. SN includes its identifier (SN_{name}) in the message to HN. The HN (AUSF) then checks whether the SN is authorized and if so, obtains the SUCI: directly if it was sent or, otherwise, deconcealing it from the SUCI. Based on SUCI, the HN (UDM/ARPF) shall choose the authentication method.

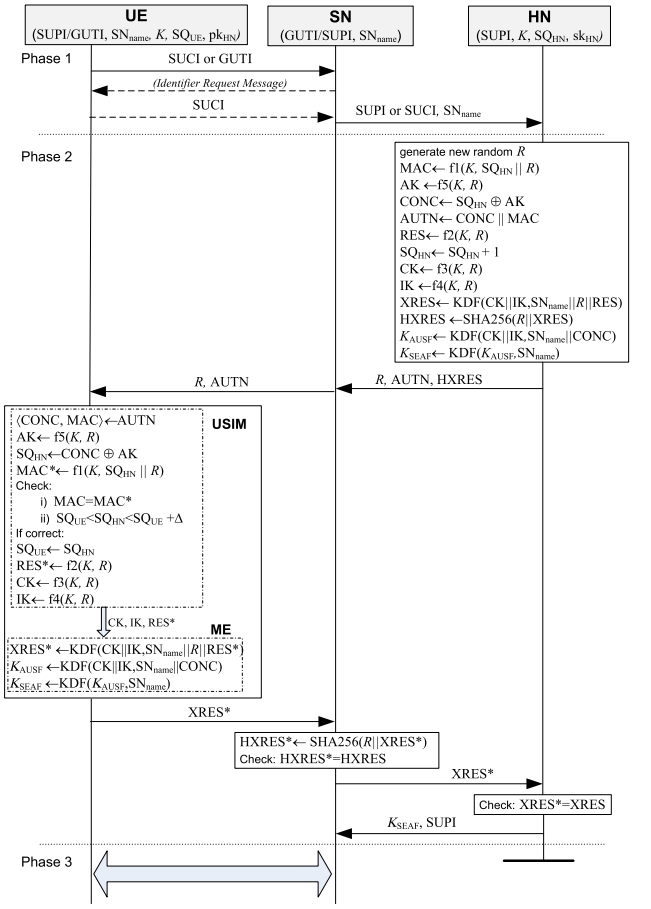


FIGURE 3. 5G-authentication and key agreement (AKA).

- The second phase is the actual authentication procedure, which is based on a conventional challenge–response mechanism. Thus, upon receiving a request for authentication, HN generates a random number R, which is used as a challenge; and an authentication value AUTN, which results of the concatenation of a message authentication code MAC, computed using f1 (MAC = f1(K, $\text{SQ}_{\text{HN}}||R$)), with another value CONC, computed using f5 (CONC = $\text{SQ}_{\text{HN}} \oplus$ f5(K, R)). CONC masks the sequence number SQ_{HN}, which is required to compute the MAC. It additionally computes the response to this challenge, XRES, and sends its hashed value, HXRES = SHA256(R||XRES), to SN, so that this can detect incorrect responses, even when it is not able to compute the correct ones. The computation of this response involves the long-term key K, the challenge R and the identifier of SN, to guarantee that both parties are communicating with the same SN, along with intermediate values RES, CK and IK computed as RES = f2(K, R), CK = f3(K, R) and IK = f4(K, R), respectively; XRES =

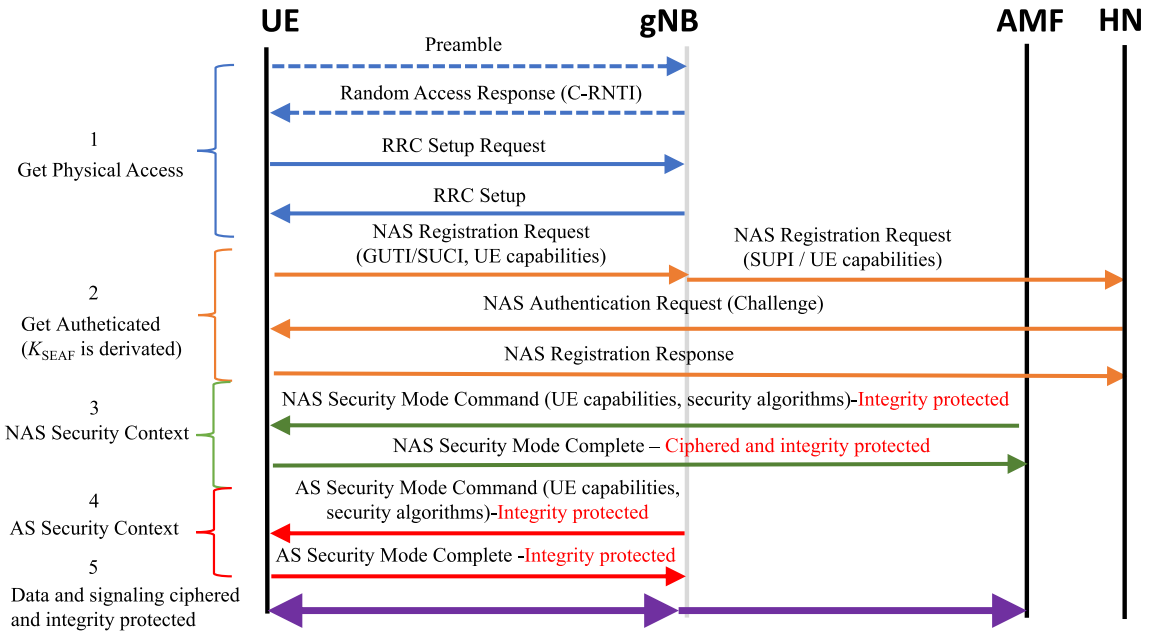


FIGURE 4. 5G-security initialization process.

KDF(CK||IK, SN_{name}||R||RES). UE receives the challenge and the authentication value: R, AUTN. Then, USIM retrieves SQ_{HN}, computes the MAC and makes a twofold check: (*i*) that the MAC is correct and therefore the authenticity of the message; and (*ii*) that SQ_{HN} has not been replayed ($SQ_{HN} > SQ_{UE}$) and that it is within a certain range ($SQ_{HN} < SQ_{UE} + \Delta$), to prevent de-synchronization attacks by forcing the counter to wrap around. If (*i*) fails, then a "MAC failure message" is sent. If (*i*) is correct but (*ii*) fails, then a "Synchronization failure message" is replied along with a re-sync token to inform to HN of the value SQ_{UE} in a hidden way; using f1* for authentication and f5* to mask and protect it from eavesdroppers. Otherwise, if (*i*) and (*ii*) are correct, USIM computes RES* (using f2), CK (using f3) and IK (using f4), and forwards it to ME, which computes and sends XRES*, and derives the anchor key K_{SEAF}. SN verifies that the hashed value of this response is correct and if so, forwards it to HN. Finally, HN verifies the response and if correct, sends K_{SEAF} to SN.

- Final phase. A successful 5G-AKA ends up with the derivation of the anchor key K_{SEAF} by both HN and UE, from which session keys for the communication between the subscriber and the HN are derived (see Section II-E). The authentication is implicit, since the standard does not specify any additional key confirmation query for K_{SEAF}, relying this on the correct execution, using the derived keys of the security procedure explained in the next section.

D. SECURITY MODE COMMAND PROCEDURE

As explained in the previous section, once 5G-AKA is successfully executed, UE and HN share the key K_{SEAF}. At this point, however, the parties (UE, SN and HN) are not mutually authenticated yet. This authentication is, according to the 3GPP standard, "implicit", provided through the successful use of keys derived from K_{SEAF} in subsequent procedures [3]. In particular, the implicit assurance provided to the UE that it is connected to a serving network that is authorized by HN is given by the use of the parameter SN_{name} in the derivation of K_{SEAF}; and UE and HN are mutually authenticated by executing the "Security Mode Command Procedure", which also checks the security capabilities of UE to prevent bidding-down attacks.

The "Security Mode Command Procedure" is implemented with NAS and AS security commands and using keys derived from K_{SEAF}. Fig. 4 sketches the security process for establishing a 5G communication. It comprises five phases:

1) The UE gets physical access to the gNB [12] (Clause 10.1.5.1).
2) The UE is authenticated using the AKA protocol described above and K_{SEAF} is computed. K_{SEAF} is used then to derive the K_{AMF} key. This key is used to derive the keys for integrity and encryption in the NAS and the RRC (AS) layer, and the UP.
3) NAS Security Mode Command procedure, where AMF sends the *Security Mode Command* message unciphered, but protects the integrity of the message with the 5G-NAS integrity key K_{NASint} and the selected integrity algorithm (indicated by the key set identifier, *ngKSI*, field within the message). Upon reception of this message, the UE shall check whether the integrity check of the message and the received "Replayed UE security capabilities" are correct. If so, the message is accepted and a NAS *Security Mode Complete* message is sent back to AMF integrity protected and ciphered using the selected ciphering algorithm with the key

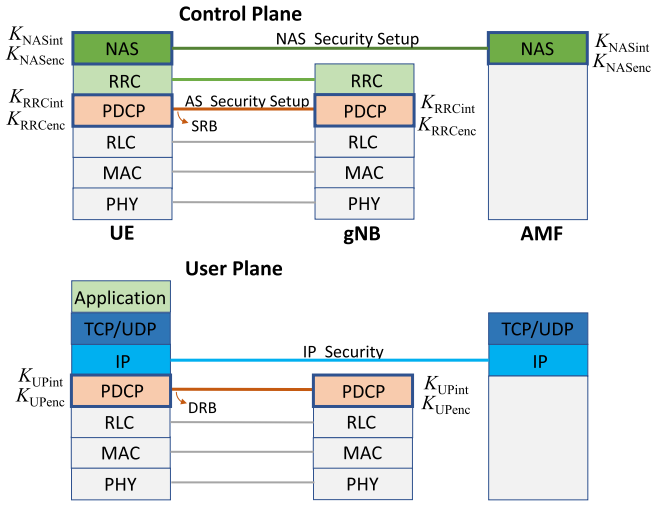


FIGURE 5. Control and user plane keys and security contexts.

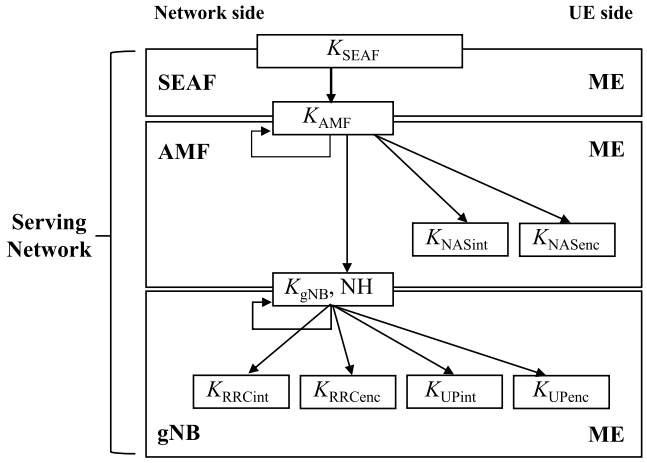


FIGURE 6. Key hierarchy generation (based on [3]).

K_{NASenc}, indicating that the NAS security context has been established. The security context comprises the keys and the algorithms to use for secure communication along with counters that prevent replay attacks.

4) The fourth step is similar to the previous one and aims to set the AS security context. UE and gNB derive the encryption and the ciphering keys for the RRC layer, K_{RRCint} and K_{RRCenc}, and the UP, K_{UPint} and K_{UPenc}. The network sends the *Security Mode Command* message applying integrity protection and using the only Signaling Radio Bearer (SRB) established, SRB1. The UE checks that the message is correct: integrity and capabilities, and if so, considers AS security to be activated, configures lower layers to apply SRB integrity and ciphering and submits the *Security Mode Complete* message.
5) Now, security contexts are established and data and signaling information are ciphered and integrity protected.

Fig. 5 summarizes the final situation with the keys used for the different planes and layers. NAS layer performs the ciphering and integrity protection of NAS signaling information, whereas PDCP performs ciphering and integrity protection of RRC signaling and user IP packets.

E. KEY DERIVATION AND UPDATES

Fig. 6 shows the key hierarchy generation from K_{SEAF} upon completion of the authentication process. Fig. 7 outlines the horizontal and vertical procedures for key derivation during context modifications and handover. Whenever an initial AS security needs to be established between UE and gNB, AMF and the UE shall derive, using the KDF (see Section II-C), a K_{gNB} and a Next Hop parameter (NH). A NH Chaining Counter (NCC) is associated with each K_{gNB} and NH. In the vertical procedure, NH is updated (NCC is incremented), whereas in the horizontal update, the new K_{gNB} is generated from its previous value, the current NH and the "PCI, ARFCN-DL" values, which correspond to the identifier of the Physical Cell and the code used to identify the absolute frequency of Synchronization Signal Block (SSB) of that cell.

F. SECURITY GOALS AND DESCRIBED ISSUES

The design of the 5G-security architecture identifies the following privacy and integrity security goals [3]:

SG1. Mutual authentication between UE and SN, and UE and HN. The authentication provided by 5G-AKA is implicit because the parties do not confirm the shared key [28]. This is then confirmed with the Security Mode Command procedure.
SG2. SN is authorized by HN. This is achieved by including the identity of SN as a parameter of the protocol.
SG3. Confidentiality of the keys (K_{SEAF}) even if the attacker learns session keys established in other sessions (previous or subsequent).
SG4. UE identities shall remain anonymous in the presence of a passive attacker in order to guarantee privacy. The identity of the UE is never sent in clear; SUPI and GUTI are used instead. GUTI is sent to a UE only after a successful activation of NAS security.
SG5. Unlinkability (user location confidentiality and user untraceability) against passive adversaries. An attacker cannot deduce the presence of a subscriber in a certain area or whether services are being delivered to a user by eavesdropping on the radio access link. Temporary identifiers 5G-Temporary Mobile Subscriber Identity (5G-TMSI) and Radio Network Temporary Identifier (RNTI) are used. 5G-S-TMSI is a shortened version of GUTI (generated simultaneously by AMF), reducing from 80 to 48 bits (AMF Set ID (10 bits) + AMF Pointer (6 bits) + 5G-TMSI (32 bits)) [26], and identifies the UE within the AMF. I-RNTI is used to identify a UE (and its context) in the cell within RRC signaling messages. Note that there are many other types of RNTI identifiers, distinguished by their prefix [29]: system information

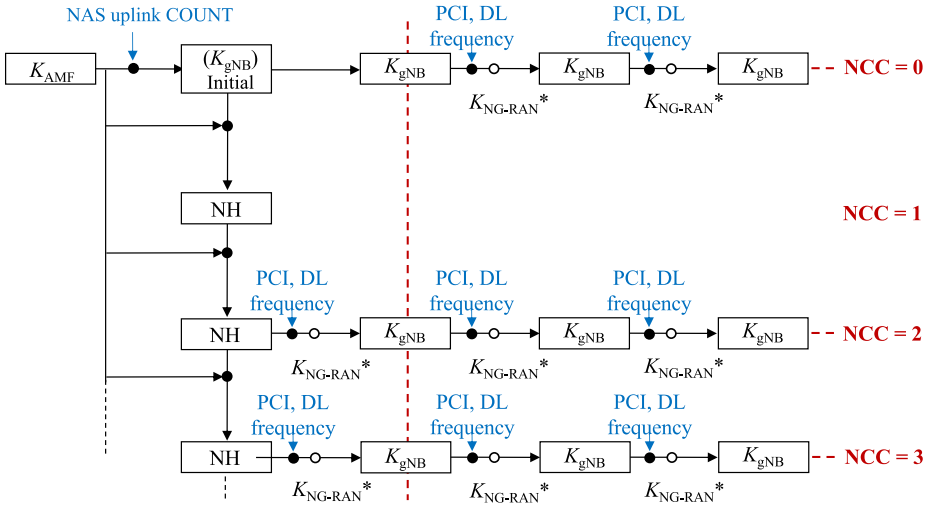


FIGURE 7. Horizontal/vertical key derivation (based on [3]).

(SI-RNTI), paging (P-RNTI), random access (RA-RNTI), etc.

Several privacy and integrity issues of the 5G-AKA protocol have been described in the literature [28], [30], highlighting the Linkability of AKA Failure Messages (LFM) attacks [31], [32], [33] as the most serious threat against privacy. LFM attacks exploit the fact that in 5G-AKA protocol, in the event of a failed authentication challenge, the reason for the failure is exposed (see Section II-C). Thus, in this attack, the adversary, after eavesdropping on a session of a target UE, acts as a fake base station (active attack) and replays the second message (authentication challenge: *R*, AUTN) to a UE: if the response of UE is *Sync_Failure* then the target is the same user, whereas if the response is *MAC_Failure*, then the target is some other user. This simple attack compromises subscription location, allowing, as an extension, user-traceability. This, nevertheless, does not contradict the fifth security goal described above, since it specifies that this protection is only required against passive adversaries.

III. CELLULAR IoT TECHNOLOGIES

In Release 13, 3GPP specified two different CIoT solutions to operate in licensed spectrum based on Long Term Evolution (LTE), LTE for MTC (LTE-M) and NarrowBand IoT (NB-IoT). The objective was to cover mMTC use cases. These use cases are characterized by having low requirements such as low complexity to provide low cost devices, support of massive connections, low power consumption and coverage enhancements [34]. LTE-M is intended for mid-range IoT applications and can support voice and video services, whereas NB-IoT can provide very deep coverage and support ultra-low-cost devices.

Both, LTE-M and NB-IoT, meet 5G mMTC requirements [9] which have been defined by the International Telecommunication Union (ITU) and 3GPP [35]. At the present time, operators are migrating from LTE to New Radio (NR) for mobile-broadband services, but they may need to provide service to deployed legacy mMTC devices. In these cases, it is important to enable efficient spectrum coexistence between 5G NR and LTE-based MTC. For that reason, an important aspect to highlight as part of Release 16 is the work in the coexistence of LTE-M and NB-IoT with 5G NR [36], [37]. Moreover, in Release 16, 3GPP introduced the support of CIoT in 5G system for NB-IoT and LTE-M with 5G system support, where the following features were included in the 5G core [38]: support for infrequent small data transmission (data over NAS), frequent small data communication (UP optimization), support of EDT and power saving functions.

Also, 3GPP has initiated Release 17 activities on reduced-capabilities NR devices, called NR-Light [39], to cover IoT use cases with higher requirements that cannot be provided by NB-IoT and LTE-M, such as higher reliability, lower latency and higher data rate. However, NR-Light does not overlap LTE-based CIoT, therefore, NB-IoT and LTE-M will continue to be developed in future releases as part of 5G. This paper is focused on low-power wide-area (LPWA) technologies in 5G, such as NB-IoT and LTE-M, since EDT feature has been defined for these CIoT technologies.

A. LTE-M

LTE-M was introduced as part of Release 13 within a new UE category (Cat-M1), that enables a coverage enhancement of 15 dB with respect to LTE.

Compared to LTE, Cat-M1 operates with a radio frequency bandwidth of 1.08 MHz, which is equivalent to 6 physical resource blocks (PRB), one receive antenna and a maximum transport block of 1000 bits [40]. Moreover, additional features that are supported are extended discontinuous reception (eDRX), mobility, RRC connection suspend/resume procedure and also data transmission via CP.

Coverage enhancement (CE) is achieved mainly through transmission repetitions. Two CE modes were introduced, mode A (up to 32 repetitions for data channels, and which support is mandatory) and mode B (up to 2048 repetitions for data channels) [40]. CE mode A provides small to medium coverage enhancement, whereas CE mode B provides deep coverage.

In Release 14, support of high peak data rates, multicast transmission, voice enhancements and location services were introduced [41]. Moreover, UE Cat-M2 was defined, which supports a radio frequency bandwidth of 5 MHz and higher data rates compared to Cat-M1.

In Release 15, latency and power consumption reduction techniques, such as increased spectral efficiency, sub-PRB resource allocation and early data transmission during the RA procedure, were introduced [42].

B. NB-IoT

NB-IoT technology was also introduced in Release 13, focusing on ultra-low-cost devices, with lower capacities than LTE-M and better coverage range [43]. Moreover, features supported in LTE-M, such as CP and UP CIoT optimizations are also supported by NB-IoT.

NB-IoT is a narrowband system which operates with a channel bandwidth of 180 kHz and also supports a maximum coupling loss (MCL) of 164 dB [44]. NB-IoT supports three different operation modes: in-band (using one PRB within a normal LTE carrier), guard-band (using unused resource blocks within a LTE carrier guard-band), and standalone (using a dedicated carrier).

NB-IoT, which is based on LTE, inherits part of its design, such as the channel coding and modulation scheme, numerology and higher layer protocols. However, in order to reduce the device cost and complexity, some functionalities are removed, such as mobility in RRC_CONNECTED state (i.e., handovers).

NB-IoT has been evolving and including new features to support additional uses. In Release 14, additional features were added, such as support for new bands, multicast transmission and positioning. Also, Cat NB2 was defined, which provides higher data rates and a new power class with a reduced output power of 14 dBm [41]. In Release 15, enhancements on power consumption and latency reduction, early data transmission during the RA procedure, Time Division Duplex (TDD) support and higher spectral efficiency, were introduced [42].

C. POWER SAVING MECHANISMS

Different power saving mechanisms have been introduced for CIoT devices to improve the battery consumption:

- Extended discontinuous reception (eDRX): it is a mechanism where the UE is allowed to stop monitoring the radio channel (e.g., physical downlink control channel, PDCCH) for a certain period of time, entering in a low power consumption mode or sleep mode [45]. This mechanism is an extension of LTE DRX mechanism, where longer sleep periods are supported (DRX cycle is extended from 2.56 seconds to minutes or hours [45], [46]). During the eDRX cycle, the UE is not reachable by the network until the cycle finishes and the UE monitors the PDCCH.
- Power saving mode (PSM): this feature was designed for CIoT devices to conserve more battery, where the UE enters in deep sleep mode. During the PSM, the device turns off its radio components completely, but maintains the registration in the network [45], [46]. This means that there is no transmission or reception for any kind of channel or signal, and the UE is not reachable by the network. The advantage of this approach is that the UE can wake-up from PSM without reattaching the connection, thus, avoiding extra power consumption. The duration of the PSM mode is defined by two timers (negotiated with the network): the activity timer (T3324) and the tracking area update timer (T3412) [40], [47].
- Release Assistance Indication (RAI): before the UE switches from RRC_CONNECTED state to RRC_IDLE state, it has to wait for receiving the RRC release message from the network. If this message is not received, the UE has to wait until the expiry of an inactivity timer. To avoid this, the 3GPP introduced in Release 14 the RAI feature [48]. The RAI feature allows the UE to indicate to the network that it has no more uplink data or it does not expect to receive any data. This feature improves the battery by releasing the RRC connection without waiting for the inactivity timer expiration.

Although the use of these mechanisms can improve the battery life, the RA procedure is triggered for each data transmission when the UE is in RRC_IDLE state, therefore, it is important to also improve the RA procedure for data transmission, as EDT does.

IV. EARLY DATA TRANSMISSION

Release 15 introduced the possibility of transmitting data during the RA procedure (that is, the first phase shown in Fig. 4). This feature, known as EDT, was introduced to reduce the UE power consumption, particularly for infrequent and small data transmissions.

EDT is defined for both, CP and UP CIoT optimizations [12]. CP-optimization uses the CP to transport user data by encapsulating them in NAS packets. UP-optimization is based on UP transport of user data. Fig. 8 shows the IP packets path over the network when using the CP or UP optimization. The type of optimization is negotiated with the AMF [49]. UE indicates (in the Preferred Network Behavior Procedure) the Network Behavior that supports and what it would prefer to use. AMF indicates (in the Supported Network Behavior) per Tracking Area Identity (TAI) list, if any of the optimizations, none or both are supported. However, for NB-IoT UEs that only support CP-optimization, the AMF

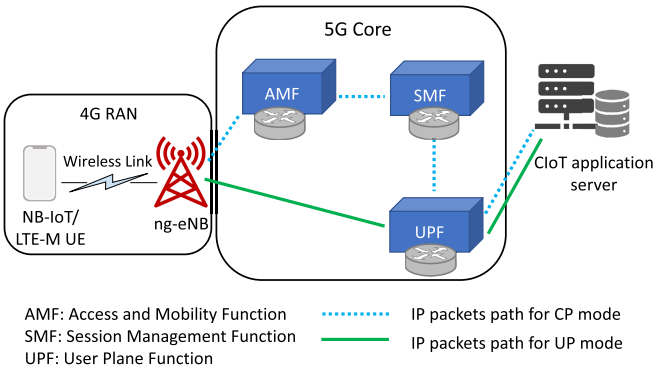


FIGURE 8. CP and UP IP packets path over the network.

shall include support for CP-optimization in the Registration Accept message [50].

In Release 16, the support of EDT was added to the 5G core network (5GC) when connected through a Next Generation Evolved NodeB (ng-eNB) using the 4G Radio Access Network (RAN). This implies the support of 5G NAS message transport and security framework, except data integrity protection. For more details, we refer the reader to Clause 24 of [12]. So far, no security enhancements have been done for EDT in Release 17 and 18. Throughout this paper, we will focus on Release 16 EDT.

A. LEGACY RA PROCEDURE

The RA procedure was first introduced in Release 8 and updated for 5G-NR in Release 15, and it is used for UEs to start communicating with the base station, as depicted in the first phase of Fig. 4. There are two types of RA procedure: non-contention-based (when the UE is in RRC_CONNECTED state, such as when starting communication with the target ng-eNB in a handover), and contention-based (when the UE is in RRC_IDLE state) [12]. In the contention-based RA procedure (which is the mode used in EDT), there are four messages interchanged between the UE and ng-eNB (see Fig. 9):

- Msg1: transmitted from the UE to the ng-eNB through the Physical Random Access Channel (PRACH). This message consists of a preamble randomly selected among a list previously transmitted by the ng-eNB in the System Information Block (SIB) [14], [15]. In case that multiple UEs transmitted the same preamble on the same RA slot, a collision will occur. On the receiver side, the ng-eNB may be able to detect the collision based on the different time of arrival, in which case it will ignore both UEs and the RA process is aborted. If the ng-eNB does not detect the collision (in case that both UEs are at a similar distance), the process will continue and will be resolved in subsequent steps.
- Msg2: upon reception of the Msg1, the ng-eNB answers with a Random Access Response (RAR) message in the Physical Downlink Shared Channel (PDSCH) with a RA-RNTI and a temporary *C-RNTI* [14], [15]. The RA-RNTI identifies the preamble sent in Msg1, so the UE that transmitted that preamble is informed that it has been heared, and the temporary *C-RNTI* is used by the UE to identify itself in the next steps. In case that an undetected collision occurs, the ng-eNB will send the RAR, which, on the receiver side will be delivered to all UEs that transmitted the same preamble on the same RA slot.
- Msg3: the UE starts its request, by sending an *RRC-ConnectionRequest* on the Physical Uplink Shared Channel (PUSCH), along with the temporary *C-RNTI* [12]. At this point, if an undetected collision occurred in Msg1, the different conflicting UEs will cause the ng-eNB to be unable to decode the message and therefore to detect the collision. Upon this case, each UE will retransmit Msg3 (no acknowledgment will be transmitted by the ng-eNB) for the maximum number of retransmissions allowed before declaring access failure and scheduling a new access attempt (i.e., a new preamble transmission, starting the process over again) [15].
- Msg4: the ng-eNB will send an *RRCConnectionSetup* message over the PDSCH indicating that the connection was completed and sending configuration parameters to the UE [12].

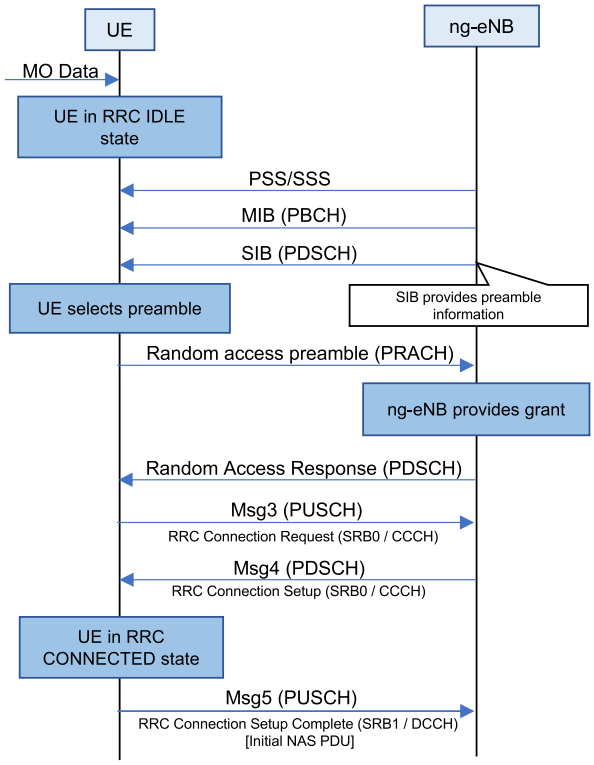


FIGURE 9. Legacy contention-based RA procedure.

Once the RA procedure is completed, one additional message is sent in the uplink (UL) over the PUSCH (Msg5), the *RRCConnectionSetupComplete*, containing the initial NAS message [13].

B. EDT BASIC OPERATION

When the UE is in RRC_IDLE state and needs to transmit data of a size equal or lower than the maximum Transport Block Size (TBS), it may use EDT. It will do so mainly to save on energy by reducing the number of interchanged messages (see Fig. 4). To transmit data over EDT, the following steps will be taken [12]:

1) The UE reads SIB-2 that contains all the information related to EDT, such as maximum TBS supported and a set of PRACH preambles reserved solely for EDT. In particular, the UE will read the RRC *edt-parameters* structure (*edt-TBS* field), which is enclosed in RACH configuration parameters [13].
2) The UE selects randomly an EDT PRACH preamble and transmits the selected preamble to the ng-eNB [15], announcing that the UE has small data to be transmitted using EDT.
3) The ng-eNB receives an EDT preamble and responds with a RAR (Msg2) that includes an UL grant (with the corresponding TBS) to be used for Msg3.
4) The UE transmits the user data and other necessary additional information, such as RRC message, in Msg3. The RRC message is different for the CP (see Section IV-C) and UP (see Section IV-D).
5) The ng-eNB receives Msg3 and decides whether to keep the UE in RRC_IDLE or move it to RRC_CONNECTED, depending on if there are more data available for transmission.
6) The ng-eNB sends Msg4 to the UE with user data if there are data available. The specific RRC message is indicated in Section IV-C for the CP and in Section IV-D for the UP.
7) The UE receives Msg4. If Msg4 indicates that the UE can be kept in RRC_IDLE, the procedure is completed. Otherwise, the UE moves to RRC_CONNECTED and sends an RRC message to the ng-eNB, indicating that the state transition has been completed.

EDT was created to send data in Msg3 and Msg4 steps, without further need for the establishment of an RRC connection and a state change in the UE [12]; significantly reducing both signaling and wake-up time in the UE. The most clear-cut example of a use case is metering, where devices have small amounts of data to be sent over long periods of time. For these devices it is very beneficial to just send small data packets in Msg3 and then return to a deep sleep mode, instead of going to connected mode before sending the data.

The maximum allowable *edt-TBS* is 1000 bits and minimum is 328. However, it is possible to use a value less than *edt-TBS* if *edt-SmallTBS-Enabled* field in RRC *edt-parameters* structure is set to true. The full details of the *edt-TBS* configuration field can be found in 3GPP TS 36.213, Table 16.3.3-2 [14]. Although it was discussed during its definition, segmentation was not introduced in EDT to keep the mechanism as simple as possible. For messages that are longer than the maximum TBS, the UE falls back to using the legacy mechanism (i.e., completing the connection and then sending the data) instead of sending data in Msg3 [12]. Similarly, if the upper layer downlink (DL) Protocol Data Unit (PDU) does not fit into the TBS used for Msg4 then the ng-eNB will trigger fall-back to (legacy) RRC connection establishment/resumption and will force the UE to enter RRC_CONNECTED state.

C. MOBILE ORIGINATED EDT FOR CP

Mobile Originated (MO)-EDT for CP aims to send the user data in NAS messages within RRC signaling (Msg3 and Msg4), for UL and DL respectively, so user data can be delivered before Msg5 and UE will remain in RRC_IDLE state.

For this purpose, two new RRC messages are introduced, as specified in [12]. UL user data are transmitted in a NAS message within *RRCEarlyDataRequest* message (Msg3), whereas for DL, user data are transmitted in a NAS message within *RRCEarlyDataComplete* message (Msg4).

Msg4 signals to the UE whether a state change is required, that is, if DL data are available and do not fit in Msg4, the ng-eNB sends a non-EDT message, so the UE moves to RRC_CONNECTED state and the connection establishment procedure continues with *RRCConnectionSetup* message. Upon this case, no early DL data transmission is performed; i.e., as segmentation is not implemented, no data are included in Msg4 and the data are transmitted using a legacy connection. Otherwise, if *RRCEarlyDataComplete* is received, the UE remains in RRC_IDLE state. The signaling messages exchanged between UE and ng-eNB are represented in Fig. 10, where dashed lines represent the messages that are sent only in the case that UE changes to RRC_CONNECTED state.

Both Msg3 and Msg4 are transmitted via SRB0 using the Common Control Channel (CCCH), so there is no RLC feedback mechanism associated, as RLC Transparent Mode (TM) is used [13]. In addition, there is no Hybrid Automatic Repeat Request (HARQ) process associated when using the CCCH. For this reason, in EDT the UE and ng-eNB do not receive feedback on whether the data have been successfully delivered. This can be problematic in Msg4, which determines the UE RRC state. If neither *RRCEarlyDataComplete* nor *RRCConnectionSetup* is received in response to Msg3, the UE considers that the UL data transmission was not successful. It is up to the UE implementation how to handle this situation [13].

Regarding mobility, since the UE is in RRC_IDLE state and it does not have an AS security context stored, a change of serving ng-eNB does not affect the MO-EDT procedure for CP. No AS security context retrieval procedure is performed. The new serving ng-eNB will broadcast the EDT related information in SIB-2, such as reserved PRACH preambles and TBS size, among others; and the UE will follow the same steps as when it was in the serving area of the origin ng-eNB.

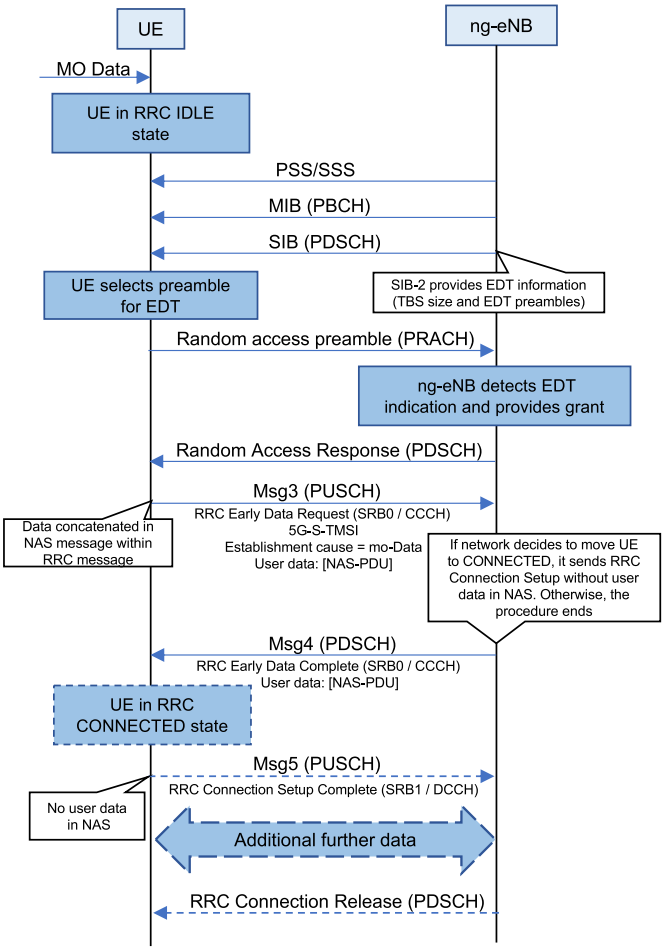


FIGURE 10. UE and ng-eNB signaling messages exchange when using MO-EDT procedure for CP.

D. MOBILE ORIGINATED EDT FOR UP

Similar to the MO-EDT procedure for CP, the MO-EDT for UP aims to transmit user data in Msg3 and Msg4 via UP, for UL and DL respectively, and reduce signaling exchange in the network.

For EDT, UP reuses the suspend/resume procedure introduced in Release 13, where the UE AS context is stored by the UE. Based on this context, the UE can resume data and signaling radio bearers previously established [12]. It also allows the UE to derive the new AS security keys based on the AS context. Therefore, the data are securely sent via UP. The only requirement is that the UE must have established an RRC connection previously (moving from RRC_IDLE to RRC_CONNECTED state), where AS context is created and later stored by the UE in RRC_IDLE state when receiving *RRCConnectionRelease* message with suspension indication. With this suspension, the UE stores the I-RNTI, the NCC and the *drb-continueROHC* (this field indicates whether to continue or reset the header compression protocol context for each Data Radio Bearer (DRB) configured with the header compression protocol) which are provided by the network in the RRC message [13].

When using EDT with UP, the UE must resume AS context before transmitting Msg3. This means that AS security functions and data bearers are active, so the UE can transmit UL data on Dedicated Traffic Channel (DTCH) multiplexed with an *RRCConnectionResumeRequest* message at MAC layer. RRC Msg3 is transmitted on CCCH using SRB0. As DRBs and AS security are resumed, DL data can also be forwarded to the UE on DTCH multiplexed with RRC Msg4 at MAC layer, which means that both are transmitted at the same time (signaling and data) [12]. In this case, unlike MO-EDT for CP, the procedure ends with the reception of the HARQ feedback (ARQ) acknowledging the successful DL transmission, due to DL data being sent on DTCH and RLC Acknowledge Mode (AM) being used.

If the *RRCConnectionRelease* message is received in response to Msg3, the UE remains in RRC_IDLE state and the procedure ends. The message includes suspension indication, the I-RNTI, the NCC and *drb-ContinueROHC* which are stored by the UE and used in the next transmission [13]. Otherwise, the UE is indicated to continue the resume procedure in the same way as Release 13 UP optimization, moving to RRC_CONNECTED state. The signaling messages exchanged between the UE and the ng-eNB are represented in a diagram in Fig. 11, where the dashed lines correspond to signaling messages that are only sent in case the UE is forced to move to RRC_CONNECTED state.

Regarding to mobility, when using MO-EDT for UP, an RRC connection can also be resumed in a new ng-eNB, different from the one where the connection was suspended, the old ng-eNB. The procedure is transparent for the UE, since the same Msg3 (*RRCConnectionResumeRequest*) is sent to the new ng-eNB. However, the new ng-eNB must retrieve the UE context from the old ng-eNB. Inter ng-eNB connection resumption is handled using context fetching, where the UE context is retrieved by the new ng-eNB over the Xn interface [12], [16]. The new ng-eNB locates the old ng-eNB using the I-RNTI, which is extracted from Msg3. When the new ng-eNB retrieves the UE context, the context is removed in the old ng-eNB. Finally, in Msg4, transmitted from the new ng-eNB to the UE in response to Msg3, new I-RNTI, NCC and *drb-continueROHC* are provided.

E. MOBILE TERMINATED EDT

Mobile Terminated (MT)-EDT is intended for a single DL data transmission during the RA procedure. MT-EDT is originated by the core network, when DL data for a UE arrives. The procedure is similar for both CP and UP.

The main difference with respect to MO-EDT is that there is a paging procedure before the MO-EDT signaling messages exchange between UE and ng-eNB. If the data can fit in a single DL transmission according to the UE category (the paging indication from the core network to the base station contains DL data size info), the paging message sent by the base station to the UE contains a MT-EDT indication. The UE will then initiate a MO-EDT procedure triggered by the MT-EDT indication. However, when using MT-EDT,

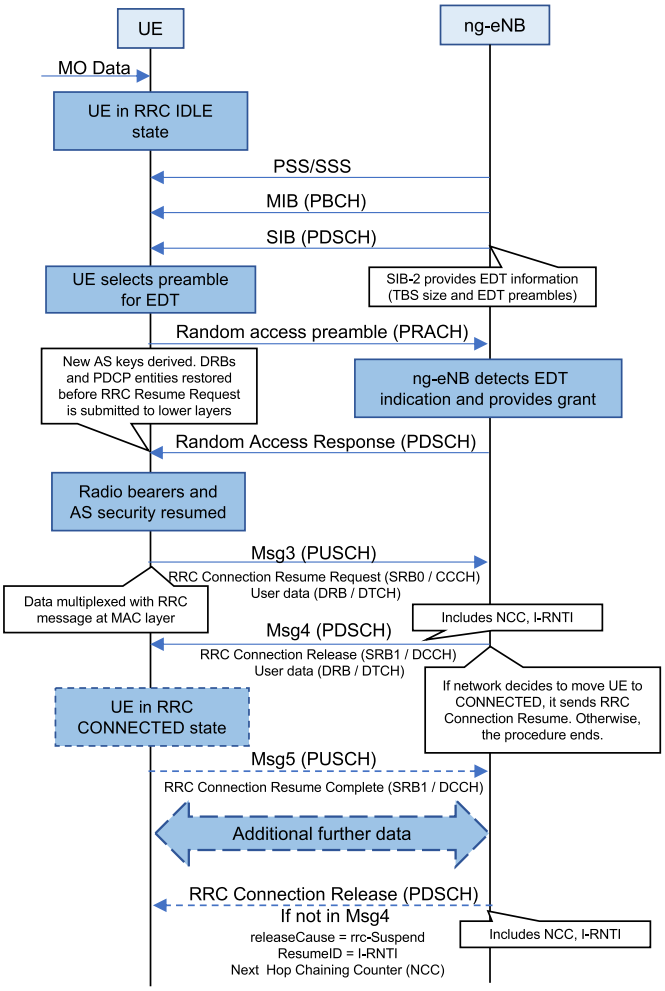


FIGURE 11. UE and ng-eNB signaling messages exchange when using MO-EDT procedure for UP.

there are some changes in the MO-EDT procedure for both, CP and UP [12]:

- When using MT-EDT for CP, no user data are sent in Msg3 and the DL data (that is, the data which fits in one single DL transmission) may optionally be included in Msg4 in case the network decides to move the UE from RRC_IDLE to RRC_CONNECTED state.
- On the other hand, when using MT-EDT for UP, the UE selects a RA preamble outside of the EDT PRACH pool. Also, the UE does not send any data in Msg3.

Up to this date, MT-EDT is only supported by 4G core. This feature has been discussed to be included also in 5GC [51], but there has not been a final agreement (see Section V-D for further discussion).

V. SECURITY DETAILS OF EDT

After describing the general EDT procedure in previous sections, a question that arises is whether the UE should remain in connected mode until receiving the response from the remote end (e.g., a server, or a peer in a P2P network). For some applications, a significant delay may occur between the request from the UE and the reply from the remote end. In this case, if the UE remains connected, it will not be able to return to an energy-saving mode (i.e., deep sleep), resulting in a higher battery consumption. This can become a serious threat for both EDT optimizations (CP and UP), since adversaries could make UEs stay in this state for a long time to drain the batteries. To prevent this, the UE may indicate AS/NAS RAI in Msg3 (see Section III-C). For EDT, RAI indicates to the network, in the DDX field for CP optimization and in the MAC subheaders for the UP optimization, whether the UE expects to receive a DL transmission in Msg4 upon transmitting Msg3 or if more data are going to be sent by the UE. In particular, the UE may indicate the following: "no further uplink and downlink data transmission" or "only a single downlink data transmission subsequent to the uplink transmission" [16]. Thus, the ng-eNB/AMF always sends a reply immediately, keeping the UE in RRC_IDLE state (in case that no DL transmission is expected or a single data transmission fitting in Msg4 is expected) or moving it to RRC_CONNECTED state (in case that DL data are available for the UE or the RAI indicates that the UE is expecting data). Table 3 summarizes the behavior of ng-eNB after receiving Msg3 depending on the RAI [16].

TABLE 3. RAI-dependent behavior of ng-eNB after receiving Msg3.

DL Data available	DL Data expected	Action
Yes	Yes	Depending of the size of DL data: Send Msg4 with DL data or *RRCConnectionResume*/*RRCConnectionSetup*
Yes	No	Send *RRCConnectionResume* for UP or *RRCConnectionSetup* for CP
No	Yes	Send *RRCConnectionResume* for UP or *RRCConnectionSetup* for CP
No	No	Send Msg4 with no DL data

Next sections compares the security mechanisms implemented in the CP and UP optimization modes and review their security issues when compared with the legacy mode. However, before particularizing for CP and UP modes, which use different security mechanisms, implemented in the NAS and AS stratum, respectively, next subsection describes a security issue shared by both modes, which is not present in the legacy mode.

A. INJECTION OF MASTER/SYSTEM INFORMATION MESSAGES

Master Information Block (MIB) and SIB messages do not have any integrity/authenticity protection, so potentially an adversary could inject fake values [52]. In particular, an adversary controlling a fake ng-eNB could broadcast fake TBSs that effectively disables the EDT procedure. If a fake TBS that is larger than the real one is broadcasted, then, UE could try to transmit UL data longer than the network can accept. It is true, nevertheless, that MIB/SIB are broadcast very frequently by the network so that these attacks would require the permanent presence of the fake station, which limit its effectiveness in practice.

B. CONTROL PLANE OPTIMIZATION

The 48-bit length 5G-S-TMSI is used in the RRC re-establishment procedure as the UE identifier (UE-ID) so that the ng-eNB can identify the user and restore de NAS security context. AS security context is not re-established (SRB0 does not have PDCP support and thus no AS security protection). The UE and the AMF perform integrity protection and ciphering for the user data by using NAS PDU integrity protection and ciphering. Fig. 12 sketches the fields of Msg3, indicating the application of integrity and ciphering with K_{NASint} and K_{NASenc}, respectively. The RAI parameter, described previously, is in the DDX field.

Next, we analyze security aspects that affect each of the three vertexes of the Confidentiality/Integrity/Availability (CIA) triad.

1) PROTECTION AGAINST TRACEABILITY ATTACKS (CONFIDENTIALITY)

In CP optimization, NAS security is used and user data are piggybacked on NAS messages. NAS-PDUs are protected with NAS security functions, where the data part is ciphered and the whole NAS message is integrity protected. However, due to not using AS security, the information related to RRC layer in Msg3 is not protected (see Fig. 12). In particular, the 5G-S-TMSI is sent in clear, which jeopardizes the privacy of UE. An adversary can access confidential information related to the position of UE (location privacy) and the frequency of its transmissions (information privacy). This breaks one of the security goals of 5G (SG5 in Section II-F).

5G-S-TMSI is a shortened form of the 5G-GUTI to enable more efficient radio signaling procedures (see Section II-F). As explained in Section II, the standard states that a new 5G-GUTI must be provided upon receiving the following messages from UE:

- Registration Request message of type ''initial registration'' or ''mobility registration update''.
- Registration Request message of type ''periodic registration update''.
- Service Request message (sent by the UE in response to a Paging message).

These are the minimums set by the standard; network equipment manufacturers are free to reassign 5G-GUTI more frequently. While this value is not updated, the UE can be traced, so it would seem trivial that a reallocation of the 5G-GUTI after each EDT transmission would solve the traceability problem. Nevertheless, it is clear that this solution is against the main objective of EDT (energy saving).

2) PROTECTION AGAINST REPLAY ATTACKS (INTEGRITY)

5G uses NAS counters for DL and UL messages to prevent replay attacks. The ''Sequence Number'' field in Msg3 (see Fig. 12) is used to construct these counters (8 least significant bits), along with a NAS overflow counter (16 most significant bits). Thus, the same NAS message, with the same counter will not be accepted twice by any of the parties. This, nevertheless, cannot prevent that a message that was sent by a legitimate UE and did not reach the ng-eNB is accepted some time after the actual transmission took place. Thus, an attacker can intercept a message from a UE and replay it later, before the UE communicates again with the network. The network verifies that the message is correct (integrity check is correctly generated) and accepts it. This has two consequences regarding security. First, the network is accepting now a value that was valid some time ago, but it may not be the case now (alarm, temperature, etc.). And second, the possible response of the application, transmitted by the network in Msg4, will be lost, preventing a potentially critical action to take place. This is due to the fact that CP optimization does not implement any procedure for the acknowledgment of Msg4 (cf. Section IV-C). The introduction of a fifth message for confirmation was discussed but eventually not included in the standard.

As the UE is aware that its message (Msg3) did not reach its destination, a possible solution would be that the UE increases the transmission frequency until the message reaches the network successfully. This retransmission frequency is up to the UE implementation policy and must be chosen carefully. An increase of this frequency due to the loss of Msg3, either because the UE is out of the range of the ng-eNB or it is being intercepted by an adversary, could lead to a faster drain of its battery (denial/degradation of service attack).

Thus, the solution for these replay attacks should be to include a fresh value which acted as a timestamp and linked the generated message with the current session. This, however, is quite challenging because NAS communication (along with NAS security) is re-established in Msg3 and, therefore, before this point, the network cannot send a fresh value for the NAS to prevent the replay attack.

3) PROTECTION AGAINST PACKET INJECTIONS (AVAILABILITY)

As already explained, AS security is not implemented in CP optimization and therefore, no AS security check is possible before passing the packets to the NAS layer. Messages can thus be injected and pass through the different layers until the sequence number and the NAS-MAC are checked. Note here that integrity protection must be checked for different (the next ones) NAS-overflow counters to prevent de-synchronization problems, so that the transmission of multiple forged messages would increase the computation load at the network side, leading to possible degradation of service attacks.

C. USER PLANE OPTIMIZATION

The *RRCConnectionResumeRequest* (see Fig. 13) sent by the UE to the ng-eNB includes its I-RNTI (used to identify the UE), the resume cause, and an authentication token MAC-I. Neither integrity protection nor ciphering applies for SRB0 [13], and the short resume MAC-I is calculated using the integrity key from the previous connection. A stored

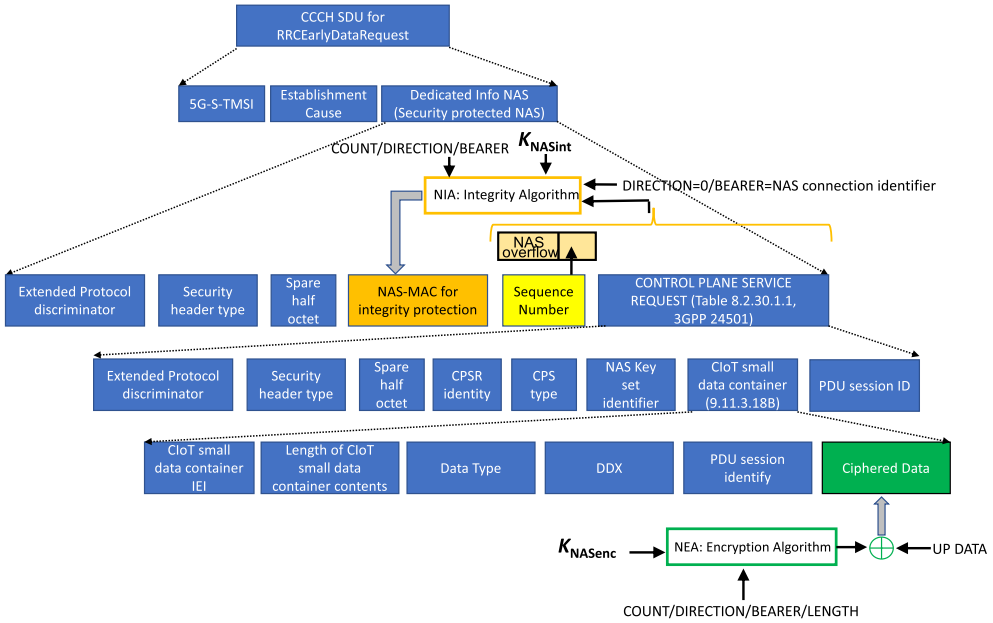


FIGURE 12. Message 3 of CP optimization.

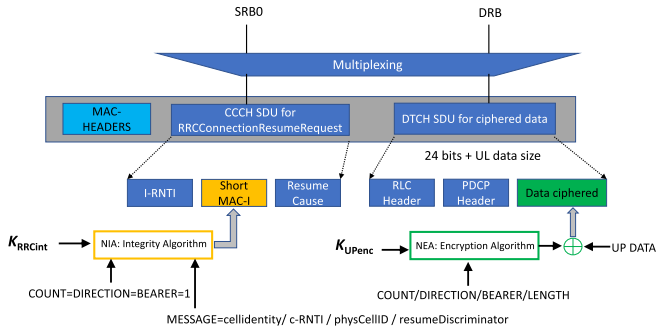


FIGURE 13. Message 3 of UP optimization.

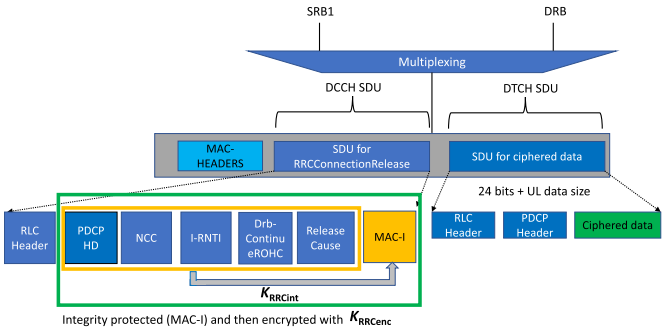


FIGURE 14. Message 4 of UP optimization.

NCC value, included in the previous Msg4 when the network moved UE to RRC_IDLE, is used by the UE, along with the K_{AMF} key to which the current k_{gNB} is associated, for vertical key derivation of an updated K_{gNB}. The UE then uses it to compute updated values for the K_{RRCint}, K_{RRCenc} and K_{UPenc}. The latter is used for ciphering user data and there is no integrity key since integrity protection is not supported, as previously explained in Section IV.

Msg4 is integrity protected and ciphered with the derived (updated) keys. In the case that UE is sent to fall back to the RRC_CONNECTED mode, the *RRCConnectionResume* message is integrity protected and ciphered with the derived keys (this message is not ciphered in the legacy mode because AS security is not activated but in EDT it is) and the UE ignores the NCC included in this message. That is, new keys are not derived.

Possible de-synchronization problems caused by the use of the previous or updated keys are prevented by forcing the UE to delete newly derived AS keys when ng-eNB rejects the connection (sending *RRCConnectionReject*) after receiving Msg3 from UE [3]. This could happen, for example, if the connection is resumed with a new ng-eNB that was not able to fetch the security context from the original ng-eNB. This, nevertheless, could not be enough to avoid certain synchronization problems, which will be discussed later.

In order to compare both modes, we analyze how UP optimization addresses the issues described for CP.

1) PROTECTION AGAINST TRACEABILITY ATTACKS (CONFIDENTIALITY)

A new identifier I-RNTI is ciphered (SRB1) and included in each Msg4 (see Fig. 14), so that UP optimization provides session unlinkability [53]. The UE can only be traced if communications are not completed and while these are not completed. Thus, the UP mode does not provide untraceability against active adversaries, who could interrupt the communication and traces the tag until UE manages to connect again with the network using the caught I-RNTI, but does against passive adversaries, as demanded by the standard.

2) PROTECTION AGAINST REPLAY ATTACKS (INTEGRITY)

Replay attacks are prevented in UP optimization by the updating of the keys. Nevertheless, this does not prevent the replay attack described for CP optimization where messages are

intercepted and replayed later before the UE communicates with the network. Thus, UP optimization is also subject to this kind of attacks but their effects are more limited. UP optimization implements, in the RLC layer, an acknowledgment mechanism for the Msg4; even when there is no Msg5. The procedure ends with the reception of the HARQ feedback (ARQ) acknowledging the successful DL transmission. Therefore, replay attacks manage to inject messages in the network which are delayed in time but, in contrast with the CP case, response messages (network reaction) are not considered as delivered by the network.

3) PROTECTION AGAINST PACKET INJECTIONS (AVAILABILITY)

ng-eNB uses short (16-bit) resume MAC-I (whose computation is discussed below, in Section V-C5) to check the integrity of the RRC message, but it does not check neither the integrity of the UL data nor that the message is not being replayed (as discussed above). As a consequence, an altered message, using one previously intercepted or in transit (MitM), which keeps MAC-I and modifies UL data, will be accepted by the ng-eNB. Such fields are the result of the encryption using K_{UPenc} and therefore an adversary, without knowing this key, cannot control the result of the decrypted message in the network side. However, as integrity is not applied, this message can only be rejected by the upper layers using semantic reasoning. That is, that the decrypted message results in an unacceptable value. Logically, if on the other hand, the decrypted message results, by chance, in an acceptable message to the upper layer, it will be accepted as valid. Alternatively to the interception of Msg3, an adversary could send his own forged Msg3 with random values of MAC-I. Short MAC-I is only 16 bits long, which makes the probability of randomly guessing the value not negligible.

The consequences of these injected messages span from simple degradation of the service, caused because messages are accepted, decrypted and processed through different layers of the protocol until they are detected in the application layer, to serious/catastrophic affectation of the service (denial of service) if the messages are eventually accepted as genuine. This latter case can only be prevented by implementing any kind of integrity check in the application layer.

4) RESISTANCE AGAINST DE-SYNCHRONIZATION ATTACKS (AVAILABILITY)

As explained above, synchronization mismatch between the updated keys in UE and ng-eNB are prevented by making the UE delete newly derived AS keys if it receives an *RRCConnectionReject* (ng-eNB rejects the connection) in response to the *RRCConnectionResumeRequest* message. The problem arises if any of the messages, Msg3 or Msg4, is intercepted and does not reach its destination. The UE should only confirm the updated values if Msg4 is received, and consequently, ng-eNB should only confirm the updated values if the ACK (RLC mechanism) is received after sending Msg4. Otherwise, the parties could get de-synchronized. However, this cannot prevent a de-synchronization attack where an adversary prevents the ACK to reach the ng-eNB. In this case, the UE will update the keys while the ng-eNB will not.

As a consequence of this mismatch between the keys stored in the UE and the ng-eNB, the next MAC-I will not be accepted and although the standard does not specify what happens in this case, it can be assumed that an *RRCConnectionSetup* message would be sent by the network in response to the *RRCConnectionResumeRequest*. When the UE receives this message, it discards the stored UE AS context, NCC and *resumeIdentity*. That is, the entire AS communication must be re-established, causing a degradation of the service.

5) MAC-I CRYPTOGRAPHIC DISCUSSION

To finish with the analysis of the security of the UP optimization, this section discusses the computation of MAC-I, which has some characteristics which may become a threat from a cryptographic point of view.

As described in Fig. 13, *cellidentity*, *C-RNTI*, *physCelID* and *resumeDiscriminator* are inputs for the MAC-I computation [13]. *physCellId* is the physical cell identity of the primary cell the UE was connected prior to suspension; *C-RNTI* is the identifier that the UE had in the primary cell the UE was connected prior to suspension and *resumeDiscriminator* is set to 1. *COUNT*, *BEARER* and *DIRECTION* inputs are set to binary ones, the key is the current (previous to updating) K_{RRCint} and the integrity algorithm is previously configured during the AS security procedure.

Taking a closer look at the inputs, it can be noted that if the UE remains connected to the same physical cell, all the parameters are constant. In effect, all the parameters are either related to the cell that UE is connected to or constant, but for *C-RNTI*. A temporary *C-RNTI* is assigned to UE by the network during the RA procedure, but according to [12] this temporary *C-RNTI* is promoted to *C-RNTI* for a UE which detects RA success and does not already have a *C-RNTI*; it is dropped by others. Thus, a UE which detects RA success and already has a *C-RNTI*, resumes using its *C-RNTI*. Thus, *C-RNTI* does not change either. Consequently, the only parameter that changes is the key, which is updated in each interaction.

As a result, a passive adversary that eavesdrops on the messages exchanged by a specific device, could store a set of encrypted outputs of identical and known inputs, computed using different integrity keys, resulting in a kind of ''known-plaintext'' attack [54].

D. MT DATA

As explained previously, EDT for mobile terminated data has been already discussed for inclusion in 5GC [51], but there has not been an agreement yet. The main reason for this is that it would require to reallocate a new 5G-GUTI during each MT-EDT procedure (for paging), which, as explained in Section V-B1, would compromise the main objective of EDT of simplifying the communication to save battery. In this

TABLE 4. Summary of the differences between UP and CP implementation.

	CP	UP
Support	Mandatory for NB-IoT	Optional
Implementation considerations	NAS level congestion control in AMF/SMF	No changes necessary in AMF/SMF
	DL data buffering in AMF	Uses the traditional way for data transmission
	Processing of NAS data PDU	
	Addition of tunneling protocol in AMF	

case, not only would not updating these identifiers affect the traceability (see Section V-B1) but also may open the possibility of degradation of service attack where adversaries aims to drain the batteries of the UEs by forging paging procedures.

VI. CP/UP SELECTION

Previous sections describe with detail both operation modes: UP and CP. The standards do not specify which one to use at a given moment, and leave it, as explained in Section IV, solely to the implemented capabilities and preferences of the devices and the network: UE and AMF. Firstly, in this respect, it should be noted that CP optimization may be more straightforward to implement than UP at the UE side, as this optimization is more extended. CP is mandatory for NB-IoT devices, both UE and AMF, whereas UP support is optional (in LTE-M, support of the control and user plane CIoT 5GS optimization are both optional at the UE and at the AMF). On the other hand, at the AMF side, the opposite is true; UP optimization is simpler to implement, since CP optimization requires changes in the AMF so that these are capable of managing user data in the control plane. That is, a NAS level congestion control applied in AMF and Session Management Function (SMF) [50] is required, which derives in an extra computation load in both; i.e., DL data buffering, processing of NAS data PDU and addition of a tunneling protocol in AMF [55]. Table 4 summarizes the CP and UP selection considerations. Apart from these implementation aspects, we next compare both modes pointing out the main differences between them and reviewing their advantages and disadvantages from efficiency and security point of views.

Table 5 summarizes the main differences between CP and UP optimizations. For the UP mode, AS security context is required and the data are sent multiplexed with RRC messages at MAC layer, whereas for the CP mode, the data are sent piggybacking a NAS message. Thus, UP optimization requires that UE has established an RRC connection previously, whereas for CP it is not necessary. AS context for UP must be fetched from the old ng-eNB for mobility, whereas for CP, this is not required, because the new EDT parameters

TABLE 5. Summary of the differences between UP and CP EDT.

	CP	UP
Layer for data transmission	NAS	AS
AS context required	No	Yes
Previous RRC connection establishment required	No	Yes
Context Fetching for mobility	No	Yes
Msg4 ACK	No	Yes
Data sent in Msg4 when move to CONNECTED	No	Yes

are provided from the new ng-eNB in SIB-2. Regarding Msg4, when using the UP solution, since user data are sent on DTCH, the transmitted DL data are acknowledged. In contrast to UP, the CP solution sends the user data on CCCH, which does not have an HARQ process associated and neither a RLC feedback mechanism (RLC TM is used). Finally, in case that Msg4 indicates that the UE has to move from RRC_IDLE to RRC_CONNECTED state, if UP solution is used, it is possible to also send DL data in Msg4. On the contrary, when using the CP solution, DL data cannot be sent in Msg4.

There are few performance evaluations of EDT which measure the network efficiency of UP/CP EDT in terms of latency and UE battery life, with an analytical framework (assuming no mobility and ideal TBS for Msg3). The authors in [7] conclude that both modes offer similar levels of quality of service, with a slightly better performance of UP EDT. Both modes of EDT improve the battery life more than 20% at extreme coverage level compared to Release 13 optimization. Although the evaluations assume an ideal uplink grant, or TBS, for Msg3, the authors indicate that EDT performance would be reduced if Msg3 grant differs from the data size. That is, if the TBS received is less than the data to transmit, EDT cannot be used. On the other hand, if the TBS received is too big, padding will have to be used and therefore, the efficiency of EDT will be reduced. Similar results are also provided by [8], where another analysis of EDT performance is conducted. The contents of Msg3 and Msg4 are analyzed and compared between both modes, UP and CP. For UL EDT (Msg3), header overhead of CP solution is larger (5 bytes) than in UP solution, whereas for DL EDT (Msg4), header overhead of CP solution is lower (12 bytes) than in UP solution. Furthermore, evaluations of latency and battery life for EDT are performed, where both modes obtain more or less the same quality of experience, with more than 10 years of battery life for extended coverage (MCL = 164 dB), which fulfills 5G requirement. This study obtains a battery life gain when using EDT that can range between 20% and 40% at extreme coverage level, compared to Release 13 optimization, depending on the traffic model used. We could conclude then that, in terms of network performance, UP solution is slightly more efficient than CP in the ideal case; i.e., adequate data size and without mobility, but when mobility occurs, UP might be not as efficient as CP, since context fetching must be performed, which could further cause additional errors, increasing network signaling. Additionally, taking into

TABLE 6. Summary of the security analysis of UP and CP EDT.

	CP	UP
Traceability	Traceable	Session Unlinkability
Replay messages	Intercepted/delayed and loss of reaction	Intercepted/delayed
Packet injections	Yes-low Degradation	Yes-medium/high Degradation-Denial
De-synchronization attacks	No	Yes-low
Known-plaintext attack	No	Yes-very low

account the 5G service categories, the use of the CP solution for mMTC devices may be beneficial when the number of simultaneous device requests is low (that is, NAS level congestion control is not necessary), since more resources will be available for the UP. These resources may be used by URLLC and eMBB service categories, which use the UP for data transmission.

As regarding the efficiency, the question of whether choosing one or another optimization remains open. Therefore, another important aspect to take into account is security. Table 6 collects the results of the security analysis carried out in the previous section. In order to prevent alarmist readings, a subjective evaluation of the severity of the attacks deemed feasible is provided, taking into account the goal, the effort and resources required. Packet injection vulnerability in the UP mode has been evaluated as medium/high since it could have uncontrollable effects if they are eventually (by chance) accepted by the system.

VII. CONCLUSION

In this paper, an overview of 5G security has been done, presenting the mechanisms for securing a newly established connection. Then, EDT has been also described in depth, specifying the steps to transmit a message using the two different approaches (UP and CP). The security issues of EDT have then been discussed, describing how each of the two approaches deals with problems such as replay attacks and packet injection and the vulnerabilities caused by the simplification in the protocols done by EDT. Another vulnerability is identified in the lack of any integrity protection in the MIB and SIB messages. Finally, an assessment of UP vs CP mode is done. Prior studies conclude that the efficiency of UP and CP are highly dependent on aspects such as network load and mobility of the users. Security adds more factors into the equation, which have been analyzed showing that in some cases UP may pose serious vulnerabilities to packet injection.

From the analysis described in this paper, several recommendations on EDT can be derived:

- 5G-AKA protocol has the LFM vulnerability with active adversaries. While this does not contradict SG5, it may still be a liability in some situations. Future releases of 3GPP should correct this vulnerability by protecting the authentication challenge against replay attacks (Section II-F).
- UEs where battery lifetime is critical should implement the indication of AS/NAS RAI in Msg3, indicating to the network their expectations of further messages; so that the network maintains them in RRC_IDLE mode (Section V).
- In future 3GPP releases, MIB and SIB messages should be signed, to prevent broadcasts from fake ng-eNBs with wrong parameters (Section V-A).
- To prevent traceability attacks on CP EDT, more frequent 5G-GUTI reassignations should be done. An equilibrium must be found between the increased security and the energy expense, so this can be an important line of future research (Section V-B).
- Also regarding CP EDT, replay attacks can be done if a certain message is prevented of reaching on time its destination in the uplink. Sending fresh values should add some protection such that the destination can detect that a packet is too old at a given point in time (Section V-B).
- Since UP EDT is vulnerable to packet injection, it is very important that critical applications have some kind of integrity check at application layer (Section V-C).
- De-synchronization attacks: this problem has no known solution. Research on making the protocol more robust to such attacks should be done (Section V-C).
- MT-EDT is subject to forgery of paging procedures. To solve this, future releases of 3GPP may include some kind of signing of the paging channel (Section V-D).
- On the question on whether to use CP or UP optimization, it is still open. UP is more efficient in ideal scenarios with no mobility, whereas CP saves resources in the network if the number of simultaneous transmissions is low. Regarding security, in CP replay attacks can cause a lack of reaction from the service, which may be critical in some cases; and, while UP does not have this vulnerability, it is more vulnerable to packet injection. Therefore, it is of utmost importance that these vulnerabilities are patched. Packet injection can be prevented at higher layers, therefore it may be recommendable in systems where a lack of response can be hazardous (Section VI).

REFERENCES

[1] *3rd Generation Partnership Project (3GPP)*, Accessed: Nov. 2021. [Online]. Available: https://www.3gpp.org/
[2] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2nd Quart., 2020.
[3] *Security Architecture and Procedures for 5G System*, document TS 33.501, Version 16.8.0, 3GPP, Sep. 2021.
[4] K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2019.
[5] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*, document TS 36.300, Version 15.2.0, 3GPP, Jun. 2018.

[6] J. Thota and A. Aijaz, "On performance evaluation of random access enhancements for 5G uRLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–7, doi: 10.1109/WCNC.2019.8885815.

[7] A. Hoglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.

[8] *Evaluation for Early Data Transmissions*, document R2-1713058, TSG-RAN WG2 #100, 3GPP, Nov. 2017.

[9] O. Liberg, J. Bergman, A. Hölund, T. Khan, G. A. Medina-Acosta, H. Rydén, A. Ratilainen, D. Sandberg, Y. Sui, T. Tirronen, and Y. P. E. Wang, "Narrowband Internet of Things 5G performance," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5, doi: 10.1109/VTCFall.2019.8891588.

[10] F. J. Dian and R. Vahidnia, "A simplistic view on latency of random access in cellular Internet of Things," in *Proc. 11th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. (IEMCON)*, Nov. 2020, pp. 0391–0395.

[11] R. Barbau, V. Deslandes, G. Jakllari, and A.-L. Beylot, "An analytical model for evaluating the interplay between capacity and energy efficiency in NB-IoT," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2021, pp. 1–9, doi: 10.1109/ICCCN52240.2021.9522178.

[12] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2*, document TS 36.300, Version 16.6.0, 3GPP, Jun. 2021.

[13] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol Specification*, document TS 36.331, Version 16.6.0, 3GPP, Sep. 2021.

[14] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document TS 36.213, Version 16.6.0, 3GPP, Jun. 2021.

[15] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, document TS 36.321, V16.6.0, 3GPP, Sep. 2021.

[16] *Procedures for the 5G System (5GS)*, document TS 23.502, Version 16.10.0, 3GPP, Sep. 2021.

[17] D. Dolev and A. C. Yao, "On the security of public key protocols," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 2, pp. 198–208, Mar. 1983.

[18] C. Yu, S. Chen, F. Wang, and Z. Wei, "Improving 4G/5G air interface security: A survey of existing attacks on different LTE layers," *Comput. Netw.*, vol. 201, Dec. 2021, Art. no. 108532, doi: 10.1016/j.comnet.2021.108532.

[19] J. Arkko, V. Lehtovirta, and P. Eronen, *Improved Extensible Authentication Protocol Method for 3rd Generation Authentication and Key Agreement (EAP-AKA)*, document IETF RFC 5448, May 2009. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5448

[20] H. Krawczyk, M. Bellare, and R. Canetti, *HMAC: Keyed-Hashing for Message Authentication*, document IETF RFC 2104, Feb. 1997. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc2104

[21] *3G Security; Specification of the MILENAGE Algorithm Set: An Example Algorithm Set for the 3GPP Authentication and Key Generation Functions F1, F1*, F2, F3, F4, F5 and F5*; Document 5: Summary and Results of Design and Evaluation*, document TR 35.909, Version 16.0.0, 3GPP, Jul. 2020.

[22] *Generic Authentication Architecture (GAA); Generic Bootstrapping Architecture (GBA)*, document TS 33.220, Version 16.4.0, 3GPP, Jun. 2021.

[23] *Information Technology—Security techniques—Hash-functions—Part 3: Dedicated Hash-Functions*, document ISO/IEC 10118-3:2004, Mar. 2004.

[24] A. Shaik, R. Borgaonkar, J.-P. Seifert, N. Asokan, and V. Niemi, "Practical attacks against privacy and availability in 4G/LTE," in *Proc. 23rd Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, Feb. 2016, pp. 1–16, doi: 10.14722/ndss.2016.23236.

[25] D. Fox, "Der IMSI-catcher," *Datenschutz und Datensicherheit*, vol. 26, no. 4, pp. 212–215, Apr. 2002.

[26] *Numbering, Addressing and Identification*, document TS 23.003, Version 17.3.0, 3GPP, Sep. 2021.

[27] J. Munilla, A. Hassan, and M. Burmester, "5G-compliant authentication protocol for RFID," *Electronics*, vol. 9, no. 11, p. 1951, Nov. 2020.

[28] D. Basin, J. Dreier, L. Hirschi, S. Radomirovic, R. Sasse, and V. Stettler, "A formal analysis of 5G authentication," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 1383–1396.

[29] *NR; Medium Access Control (MAC) Protocol Specification*, document TS 38.321, V16.6.0, 3GPP, Sep. 2021.

[30] A. Koutsos, "The 5G-AKA authentication protocol privacy," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Jun. 2019, pp. 464–479.

[31] H. Khan and K. M. Martin, "A survey of subscription privacy on the 5G radio interface—The past, present and future," *J. Inf. Secur. Appl.*, vol. 53, Aug. 2020, Art. no. 102537, doi: 10.1016/j.jisa.2020.102537.

[32] M. Arapinis, L. Mancini, E. Ritter, M. Ryan, N. Golde, K. Redon, and R. Borgaonkar, "New privacy issues in mobile telephony: Fix and verification," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, Oct. 2012, pp. 205–216.

[33] M. Arapinis, L. I. Mancini, E. Ritter, and M. D. Ryan, "Analysis of privacy in mobile telephony systems," *Int. J. Inf. Secur.*, vol. 16, no. 5, pp. 491–523, Oct. 2017.

[34] *Service Requirements for Machine-Type Communications (MTC); Stage 1*, document TS 22.368, Version 13.2.0, 3GPP, Dec. 2016.

[35] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document TR 38.913, Version 16.0.0, 3GPP, Jul. 2020.

[36] *Coexistence Between NB-IoT NR*, document TR 37.824, Version 16.0.0, 3GPP, Jun. 2020.

[37] *Coexistence Between LTE-MTC NR*, document TR 37.823, V16.0.0, 3GPP, Jun. 2020.

[38] 5G Americas. (2020). *The 5G Evolution: 3GPP Releases 16–17*. [Online]. Available: https://www.5gamericas.org/5g-evolution-3gpp-Releases-16-17/

[39] *New SID Support Reduced Capability NR Devices*, document RP-193238, TSG RAN Meeting #86, 3GPP, Dec. 2019.

[40] *LTE-M Deployment Guide to Basic Feature Set Requirements*, GSMA, London, U.K., Jun. 2019. [Online]. Available: https://www.gsma.com/iot/resources/ltem-deployment-guide-v3/

[41] *Release 14 Description; Summary Rel-14 Work Items*, document TR 21.914, Version 14.0.0, 3GPP, May 2018.

[42] *Release 15 Description; Summary Rel-15 Work Items*, document TR 21.915, Version 15.0.0, 3GPP, Sep. 2019.

[43] *Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer; General Description*, document TS 36.201, Version 13.2.0, 3GPP, Jun. 2016.

[44] M. Kanj, V. Savaux, and M. Le Guen, "A tutorial on NB-IoT physical layer design," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2408–2446, 4th Quart., 2020.

[45] *Architecture Enhancements to Facilitate Communications With Packet Data Networks and Applications*, document TS 23.682, Version 16.10.0, 3GPP, Sep. 2021.

[46] *Non-Access-Stratum (NAS) Protocol for Evolved Packet System (EPS); Stage 3*, document TS 24.301, Version 16.8.0, 3GPP, Mar. 2021.

[47] *NB-IoT Deployment Guide to Basic Feature Set Requirements*, GSMA, London, U.K., Jun. 2019. [Online]. Available: https://www.gsma.com/iot/resources/nbiot-deployment-guide-v3/

[48] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, document TS 36.321, Version 14.9.0, 3GPP, Jan. 2019.

[49] *General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access*, document TS 23.401, V17.2.0, 3GPP, Sep. 2021.

[50] *System Architecture for the 5G System (5GS); Stage 2*, document TS 23.501, Version 17.2.0, 3GPP, Sep. 2021.

[51] *Email Discussion on Support of MT-EDT for 5GC*, document, Pre SA2#138E e-meeting Email Discussion, 3GPP, 2020. Accessed: Sep. 4, 2022. [Online]. Available: https://www.3gpp.org/ftp/tsg_sa/WG2_Arch/TSGS2_138e_Electronic/Inbox/CCs/Moderated_Email_Discussion/SA2%23138E_Email_Discussion_MT-EDT_OpenIssues_r0.doc

[52] S. R. Hussain, M. Echeverria, A. Singla, O. Chowdhury, and E. Bertino, "Insecure connection bootstrapping in cellular networks: The root of all evil," in *Proc. 12th Conf. Secur. Privacy Wireless Mobile Netw.*, May 2019, pp. 1–11, doi: 10.1145/3317549.3323402.

[53] J. Munilla, M. Burmester, and R. Barco, "An enhanced symmetric-key based 5G-AKA protocol," *Comput. Netw.*, vol. 198, Oct. 2021, Art. no. 108373, doi: 10.1016/j.comnet.2021.108373.

[54] A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*. Boca Raton, FL, USA: CRC Press, 1996.

[55] 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 Netw.*, vol. 31, no. 6, pp. 8–15, Nov./Dec. 2017.

DAVID SEGURA received the B.Sc. degree in telematics engineering and the M.Sc. degree in telematics and telecommunication networks from the University of Malaga, Spain, in 2019 and 2020, respectively, where he is currently pursuing the Ph.D. degree in cellular communications. In 2019, he started to work as a Researcher with the Department of Communication Engineering, University of Malaga.

JORGE MUNILLA is currently working as an Associate Professor with the Department of Engineering, University of Málaga, Spain. He has been a Guest Researcher with the IAIK Krypto Group, University of Graz, Austria, in 2006, and a Visiting Faculty Member with the Center for Security and Assurance in IT (C-SAIT), Florida State University, USA, in 2009, 2011, and 2015, the Centre for Computer and Information Security Research, University of Wollongong, Australia, in 2012 and 2014, and the Security Research Group, University of Kent, U.K., in 2019. He has coauthored several book chapters and more than 50 papers in high impact international journals and conferences. His research interests include resilient cyber-physical systems, security in 5G and pervasive/ubiquitous systems (RFID and the IoT), and the application of big data and machine learning techniques.

EMIL J. KHATIB (Member, IEEE) received the Ph.D. degree in machine learning, big data analytics and knowledge acquisition applied to the troubleshooting in cellular networks, in 2017. He is currently a Postdoctoral Juan de la Cierva Fellow with the University of Málaga. He has participated in several national and international projects related to Industry 4.0 projects. He is working on the topic of security and localization in industrial scenarios.

RAQUEL BARCO is currently a Full Professor in telecommunication engineering at the University of Malaga. Before joining the University, she worked at Telefonica, Madrid, Spain and the European Space Agency (ESA), Darmstadt, Germany. As a Researcher, she is specialized in mobile communication networks and smart-cities, having led projects funded by several million euros. She has published more than 100 papers in high impact journals and conferences and authored five patents. She received several research awards.

Part III

Achievements

Chapter 7

Conclusions

This chapter provides a summary of the research conducted during this thesis. For this purpose, this chapter is divided into three sections. Section 7.1 provides a review of the objectives pursued in this thesis, highlighting the main contributions of each of them. Section 7.2 suggests some lines of future work related to the research carried out. Finally, Section 7.3 shows a list of publications as well as other activities related to this thesis.

7.1 Contributions

This thesis aims at assessing and improving the performance of mobile networks in the Industry 4.0 paradigm. To this end, a set of challenges in the industrial scenario have been identified and required objectives to solve these challenges have been defined. A total of six objectives have been established through this work, which are distributed as follows. Obj. 1 and 2 are related to the performance assessment of the network in an indoor industrial environment. Obj. 3 and 4 refer to the development of optimization algorithms to improve network performance. Finally, Obj. 5 and 6 aim to cover CIoT signaling optimizations, first assessing the impact of these optimizations in the network and then providing a security analysis of the latest optimization proposed by the 3GPP. The contributions related to each of these objectives are presented below:

Obj. 1. To study the impact of 5G numerologies on the latency for critical services.

- An analysis on the impact of the different 5G numerology configurations on

users' latency has been performed. In this analysis, a more detailed study than in those found in the state of the art has been performed. The 5G numerologies have been evaluated under two different channel conditions (LOS and NLOS) and with different packet sizes for an AGV use case.

- The study has been carried out in a 5G simulated environment. The results showed that the numerology selection is not trivial, and an intermediate value is more suitable under NLOS conditions. This study opens the way to algorithms that can be used to dynamically adjust the numerology configuration in the network for a better performance.

Obj. 2. To assess and compare network scalability with different technologies in an industrial environment.

- Following this objective, an empirical assessment has been carried out with different number of devices, packet sizes and scenarios to evaluate the network performance in terms of latency and packet loss. In particular, the technologies selected to compare their performance have been 5G, Wi-Fi 6, and the use of multi-connectivity between both with a PD approach.
- More specifically, measurement campaigns were performed with commercial equipment in the "5G Smart Production Lab" at Aalborg University (Denmark), which consists of a small-scale indoor industrial factory environment composed of two halls and a wide range of industrial manufacturing and production equipment.
- As a result of the measurement campaigns, it has been demonstrated that the 5G technology provides lower tails in the latency distribution and is more reliable than Wi-Fi 6. On the other hand, the multi-connectivity solution demonstrated a significant reduction in the latency tails and a zero packet loss, being very effective to fulfil the use cases with very restrictive latency and reliability requirements.

Obj. 3. To propose a mechanism to enhance reliability for critical services.

- To meet this objective, an algorithm based on ML to dynamically activate PD in an industrial environment has been designed. This algorithm relies on network metrics such as the Signal to Interference plus Noise Ratio (SINR), the modulation index, and the HARQ feedback to predict the latency. The

output of the predictor is used for the PD decision in the downlink direction, based on a latency threshold.

- The latency predictor was trained with the RF algorithm, and the performance of the proposed algorithm has been validated under different tests conducted in a 5G simulated environment. Under this simulator, the DC feature with PD approach was implemented, as described in Appendix A.
- The performance of the algorithm has been compared with other approaches in the state of the art, thus demonstrating that the proposed solution is able to achieve better results and minimize resource wastage in the network.

Obj. 4. To evaluate the network performance in a distribution center.

- The contributions corresponding to this objective are, firstly, the design and implementation of an open-source simulator based on the ns-3 framework and the 5G-LENA module which includes new features to support the assessment of the network in a distribution center. In particular, features such as a distribution center scenario, the activities involved there, the resource allocation per slice, and the industrial channel and propagation loss model have been implemented.
- Secondly, once the above features were implemented, two NS strategies using the 5G network have been evaluated under different logistics activities. These NS strategies consist in the use of a static slice with a balanced division of network resources, and the use of a slice that dynamically adjust its size depending on the activity taken place.
- Finally, the QoS performance provided by these NS strategies across various traffic profiles have been evaluated via simulations, and the results demonstrated that a dynamic slice improves the QoS especially under high traffic load, whereas the static slice performs well when the traffic load is low.

Obj. 5. To study the impact of CIoT signaling optimizations in the network.

- In relation to this objective, an analysis on the impact of the different CIoT signaling optimizations proposed by the 3GPP on user's latency has been performed. Specifically, in this study the NB-IoT technology has been used for the evaluation of these optimizations via the CP.

- The study has been carried out with commercial equipment from Amarisoft, with which several measurement campaigns were performed under different coverage levels and packet sizes.
- From the results of these measurements, it has been demonstrated that EDT, unlike Release 13 optimization, fulfils the 3GPP latency requirement for infrequent small data transmissions under extreme coverage level.

Obj. 6. To analyse the security of 5G EDT optimization for CIoT.

- A comprehensive study of the EDT optimization in CIoT has been provided as a final contribution. In this study, the EDT feature has been described in detail for both supported operation modes, CP and UP.
- Moreover, a security analysis of this feature has been provided, extracting the main vulnerabilities found in each of its operation modes. Specifically, vulnerabilities such as packet injection, replay attacks and the injection of fake values to disable EDT have been found.
- Finally, after the exhaustive security analysis, a set of recommendations for researchers and manufacturers have been provided, which include solutions to patch these vulnerabilities in future 3GPP releases, and which operation mode is more suitable to use.

7.2 Future work

Possible lines of research that might continue the work in this thesis are the following:

- Regarding the 5G numerologies assessment, this thesis has conducted a study that aims to serve as a guide for a better understanding of which numerology configuration is more suitable in an industrial environment for critical services. One of the main lines to be addressed in this context would be the study of other mechanisms such as preemptive scheduling and resource reservation to complement the numerology selection. Another line would be the developing of algorithms to dynamically select the proper numerology according to radio conditions.
- In this thesis, a study of the network scalability with different technologies have been conducted in an indoor industrial scenario, making it easier for the manufacturing sector to opt for one technology or another based on the network

performance and their business use case. A possible future line to extend this work would be the integration of spectrum interference on the study, to better provide and compare the performance of the network with and without interference. Another line would be the enhancement of the multi-RAT tool to take into account network metrics, thereby enabling the PD only when necessary. This would result in a more efficient usage of network resources.

- The proposed algorithm based on ML to dynamically duplicate packets has been evaluated and has demonstrated its effectiveness in reducing resource wastage while improving the reliability. Some aspects that have not been covered in this thesis and could be the subject of future study are the implementation of this algorithm in a commercial network, the selection of the proper SN, the required model update frequency due to network or environment changes, and the study of using Federated Learning (FL) for the creation and enrichment of the model.

- The performance of the network in a distribution center has been evaluated in this thesis with a customized open-source simulator. One of the main lines to be addressed in this context would be the study of linking wireless performance with the production performance. This would further demostrate the impact of network optimizations on the vertical scenario.

- Regarding the optimization of CIoT transmissions, this thesis has covered CIoT optimizations up to Release 16, the latest one corresponding to the EDT feature. A similar mechanism to EDT has been proposed in native 5G, namely two-step RACH. A possible future line to extend this work would be the comparison of these two features from an analytical and a performance point of view. Moreover, a new type of device has been proposed by the 3GPP in the Release 17, namely NR RedCap. These devices are focused on use cases such as wearables, video surveillance and industrial IoT. A future line would be to consider RedCap devices to further study the performance and the energy saving in the network.

- Finally, in terms of network security, multiple future lines of research can be launched. First, with the arrival of Open-RAN networks, where NFs are virtualized and standard open interfaces between virtualized network elements are used, new security challenges that need to be analyzed and solved arise. Secondly, the massive use of AI algorithms expected in future networks posses new threats into the models such as data poisoning, model evasion or model inversion that need

to be analysed, and defense mechanisms need to be developed and deployed in the network.

7.3 Publications and projects

The following subsections present the publications and activities related to this thesis.

7.3.1 Journals

Publication arising from this thesis

The work carried out in this thesis has resulted in four papers published in high impact journals plus one in the process of revision, listed as follows.

[I] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "5G Numerologies Assessment for URLLC in Industrial Communications," *Sensors*, vol. 21, no. 7, p. 2489, Apr. 2021.

[II] D. Segura, E.J. Khatib, and R. Barco, "Dynamic Packet Duplication for Industrial URLLC," *Sensors*, vol. 22, no. 2, p. 587, Jan. 2022.

[III] D. Segura, J. Munilla, E.J. Khatib, and R. Barco, "5G Early Data Transmission (Rel-16): Security Review and Open Issues," *IEEE Access*, vol. 10, pp. 93289–93308, Sep. 2022.

[IV] D. Segura, E.J. Khatib, and R. Barco, "Evaluation of Mobile Network Slicing in a Logistics Distribution Center," *IEEE Transactions on Network and Service Management*, Under review, 2024.

[V] D. Segura, S.B. Damsgaard, A. Kabaci, P. Mogensen, E.J. Khatib, and R. Barco, "An Empirical Study of 5G, Wi-Fi 6, and Multi-Connectivity Scalability in an Indoor Industrial Scenario," *IEEE Access*, vol. 12, pp. 74406-74416, May. 2024.

7.3.2 Conferences and Workshops

Conferences arising from this thesis

Several works have also been presented at national and international conferences, as shown below.

[VI] D. Segura, E.J. Khatib, and R. Barco, "Evaluación de numerologías 5G para URLLC," in *XXXV Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2020)*, Málaga (España), Sept. 2020.

[VII] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "Evaluación de los modos de conexión para NB-IoT," in *XXXVI Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2021)*, Vigo (España), Sept. 2021.

[VIII] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "NB-IoT latency evaluation with real measurements," in *2022 IEEE Workshop on Complexity in Engineering (COMPENG)*, Florence (Italy), Jul. 2022.

[IX] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "Evaluación de la latencia de NB-IoT con medidas reales," in *XXXVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2022)*, Málaga (España), Sept. 2022.

[X] D. Segura, S.B. Damsgaard, P. Mogensen, E.J. Khatib, and R. Barco, "Comparativa empírica del rendimiento de 5G y Wi-Fi en un escenario industrial de interior," in *XXXVIII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2023)*, Cáceres (Spain), Sep. 2023.

[XI] D. Segura, H.Q. Luo-Chen, C. Baena, E.J. Khatib, S. Fortes, and R. Barco, "Testbed para la evaluación de los ataques de envenenamiento y evasión en un servicio E2E," in *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

Conferences related to this thesis

[XII] J. Llanes, E.J. Khatib, D. Segura, and R. Barco, "Seguridad en B5G/6G," in *XXXVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2022)*, Málaga (Spain), Sep. 2022.

[**XIII**] S.B. Damsgaard, D. Segura, M.F. Andersen, S.A. Markussen, S. Barbera, I. Rodríguez, and P. Mogensen, "Commercial 5G NPN and PN Deployment Options for Industrial Manufacturing: An Empirical Study of Performance and Complexity Tradeoffs," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Toronto (Canada), Sep. 2023.

[**XIV**] H.Q. Luo-Chen, D. Segura, C. Baena, E.J. Khatib, and R. Barco, "Detection of anomalous samples based on automatic thresholds," in *2024 IEEE Workshop on Complexity in Engineering (COMPENG)*, Florence (Italy), Jul. 2024.

[**XV**] H.Q. Luo-Chen, D. Segura, C. Baena, E.J. Khatib, S. Fortes, and R. Barco, "Alteración de datos E2E: impacto de un ataque de envenenamiento y evasión en una red celular," in *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

[**XVI**] C.S. Álvarez-Merino, D. Segura, C. Baena, E.J. Khatib, and R. Barco, "Infraestructura para la monitorización del consumo energético en redes b5G/6G," in *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

[**XVII**] E.J. Khatib, D. Segura, A. Tarrías, and R. Barco, "Estudio del ataque de cadena de suministro sobre XZ utils y sus consecuencias en telecomunicaciones," in *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

7.3.3 Related projects

This thesis has contributed to the following projects:

- National projects:
  - EDEL4.0: Seguridad y fiabilidad en las comunicaciones 5G/IoT para la Industria 4.0. Número de proyecto UMA18-FEDERJA-172, receiving funds from Junta de Andalucia and European Comission, within the framework of "Proyectos de I+D+i en el marco del Programa Operativo FEDER Andalucia 2014-2020".

– PENTA: Provisión de servicios PPDR a través de Nuevas Tecnologías de Acceso radio. Número de proyecto PY18-4647, receiving funds from Junta de Andalucía and European Comission, within the framework of "Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI 2020)".

– MAORI: Massive AI for the OpenRadIo b5G/6G network. Project number TSI-063000-2021-72, receiving funds from Ministerio de Asuntos Económicos y Transformación Digital and European Union - NextGenerationEU within the framework "Recuperación, Transformación, y Resiliencia".

7.3.4 Research stay

This thesis involved a five-month stay in Aalborg (Denmark), collaborating with Aalborg University on several measurement campaigns with 5G, Wi-Fi 6, and multi-connectivity in an industrial environment. The stay took place between February 2023 and June 2023 and was supervised by Preben E. Mogensen.

Appendices

Appendix A

Assessment tools and testbeds

A.1 5G LENA ns-3 simulator

NS-3 is an open-source network simulation software with discrete events, and it is composed of different modules developed in C++, each of them providing a particular function [108]. Among the functionalities implemented with this simulator is the 5G-LENA module [109], which is oriented to 5G NSA networks.

The 5G-LENA module is focused on the new 5G NR specifications of the 3GPP standard, particularly on the Release 15. This module adds some functionalities adapted to 5G in the PHY and MAC layers. It supports both the millimeter FR (FR2) and the non-millimeter FR (FR1), as well as beamforming, modulation schemes and error recovery (HARQ) implementations, among others. The most important features of this module are listed:

- Support of 5G numerologies. It is possible to select the numerologies from 0 to 4, which corresponds to Release 15 specification.

- Multiplexing of different Bandwidth Parts (BWPs). Allows to divide the total bandwidth into different parts, with different configurations (e.g., different numerology). This is useful for multiplexing different services with different requirements.

- Scheduler based on OFDMA and Time Division Multiple Access (TDMA), with the typical scheduling algorithms such as round robin, proportional fair, etc.

- Propagation and channel model based on the 3GPP 38.901 standard [110]. Implements the propagation and channel models for rural, urban and office scenarios.

A.1.1 Author's contribution

Throughout the development of this thesis, many features have been included using as a baseline the NS-3 simulator and the 5G-LENA module, with the aim to assess network performance in an industrial environment for those works carried out with simulations. This resulted in an open-source simulator with the code available on Github [111]. The different enhancements and features developed are described below.

1. Industrial channel model and propagation loss

The 5G-LENA module does not provide the industrial scenario. Therefore, the first feature added to the simulator was the inclusion of the industrial channel and propagation loss model defined in the Release 16 of the 3GPP standard [110]. The Indoor Factory (InF) scenario focuses on factory halls of varying sizes and with varying levels of density of clutter, e.g., machinery, assembly lines, storage shelves, etc. In particular, four different industrial scenarios are defined based on the clutter density and base station height:

- InF with Sparse clutter and Low base station height (InF-SL). It is characterized by having a factory ceiling height of between 5-25 meters composed of large machinery with regular metal surfaces (e.g., several mixed production areas with open spaces and storage/commissioning areas). The typical size of these machinery is 10 meters and the base station is located at a height below the average height of the machinery.

- InF with Dense clutter and Low base station height (InF-DL). It is characterized by having a factory ceiling height of between 5-15 meters and, composed of small and medium-sized machinery, and metal objects with irregular structure (e.g., assembly and production lines surrounded by small mixed machinery). The typical size of these machinery is 2 meters and the base station is located at a height below the average height of the machinery.

- InF with Sparse clutter and High base station height (InF-SH). Same scenario as InF-SL, but with the base station height being above the height of the machinery.

- InF with Dense clutter and High base station height (InF-DH). Same scenario as InF-DL, but with the base station height being above the height of the machinery.

With the aim of providing maximum flexibility to perform simulations and allowing to recreate different scenarios, all InF scenarios were implemented and Table A.1 shows the different parameters that can be modified in the simulator.

Scenario	Parameter	Value	Description
All	**InFTotalSurface**	Any	Represents the total area of the factory in squared meters
All	**InFVolume**	Any	Represents the total surface of the factory in cubic meters
InF-SL, InF-SH	**LowClutterDensity**	0 to 0.39	Percentage of area occupied by clutter
InF-DL, InF-DH	**HighClutterDensity**	0.4 to 1	Percentage of area occupied by clutter
InF-SH, InF-DH	**ClutterHeight**	0 to 10	Clutter height in meters

Table A.1: Configurable simulator parameters according to the industrial scenario.

2. RRC Idle state

In 5G-LENA, by default, when starting a simulation, all UEs move from RRC Idle state to RRC Connected state, performing the RA procedure and establishing a DRB, even when there are UEs that are not going to transmit any data. This causes radio resources to be consumed, which could be used by other UEs that are transmitting data. In order to make the simulations more realistic, the possibility of establishing the connection later (just when the UE has data to send) was implemented in the simulator and, in addition, the inclusion of the transition from RRC Connected to RRC Idle state after an inactivity period of the UE.

To achieve this goal, the code associated to the RRC layer have been modified, both in the UE and base station, together with the NAS layer and part of the network core. Figure A.1 shows the different states through which the UE transits and the modifications made are highlighted in red:

1. When the RRC Idle connection mode is activated, the UE remains in the IDLE_CAMPED_NORMALLY state, without sending the preamble.

2. The UE switches to CONNECTED_NORMALLY state when alerted from the NAS layer that the UE has data to transmit. When this occurs, the RA preamble is sent, in addition to performing the steps from IDLE_CAMPED_NORMALLY state to CONNECTED_NORMALLY state, where, once the RRC connection is established, the packet shall be sent over the corresponding DRB.

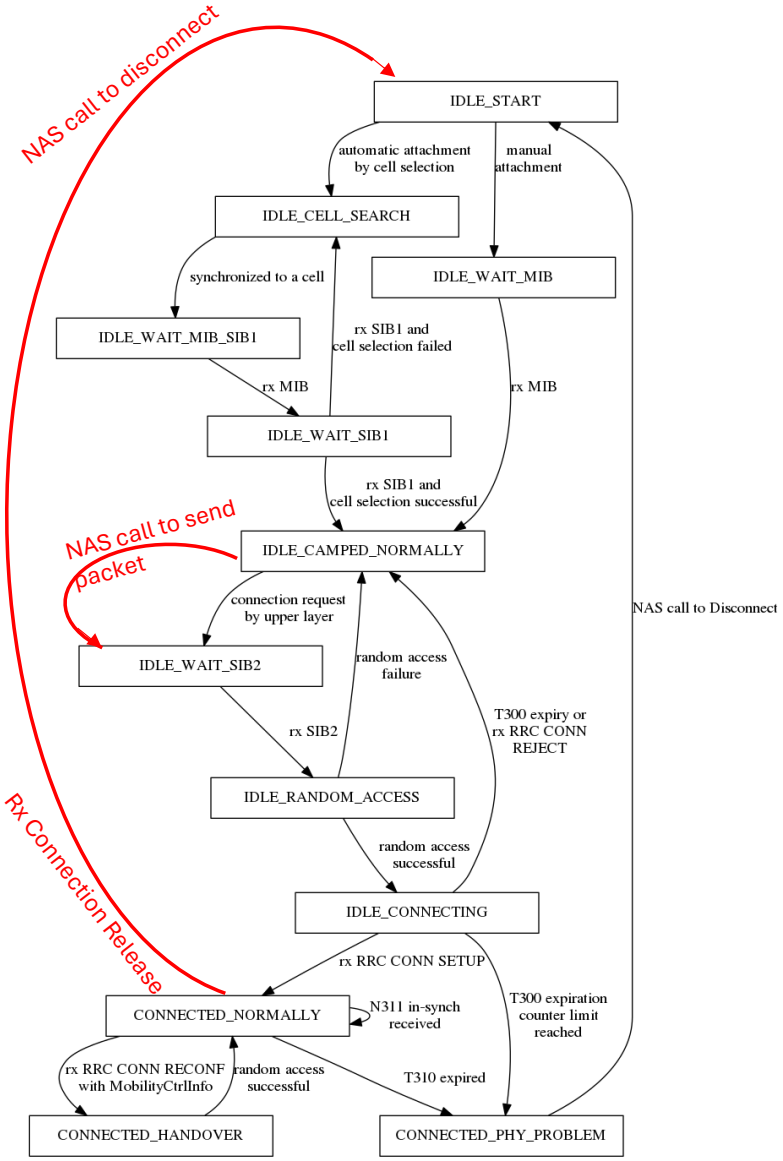


Figure A.1: RRC layer state machine at the UE.

3. The NAS layer accumulates user packets in a buffer while the RRC connection is being established.

4. When the UE does not transmit any data during a time period, it receives a signaling message from the network (RRC Connection Release), which causes the UE context in the network core, the SRBs (SRB0, SRB1) and DRBs to be removed. It also forces the UE to go to IDLE_START state and to remain in the IDLE_CAMPED_NORMALLY state.

5. In case of reconnection from the UE, the same process will be repeated from step 2.

Figure A.2 shows the different states that the UE context stored in the base station goes through each time a connection establishment request is received, with the implemented modifications highlighted in red. In this case, only the UE inactivity detection has been implemented, which causes its context to be removed. In general, the steps to be followed are:

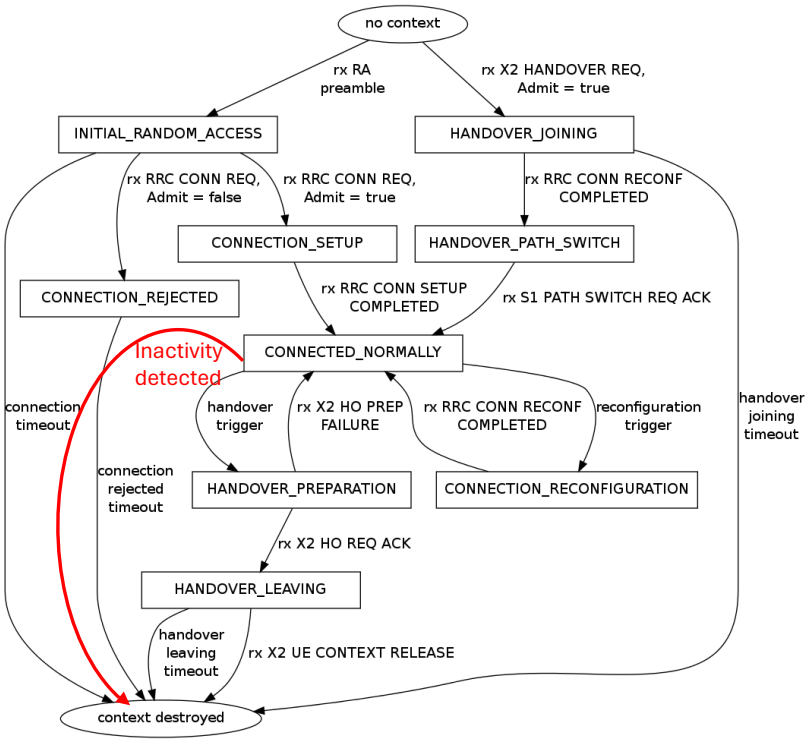


Figure A.2: RRC layer state machine at the base station for each UE.

1. When a packet is received in the CONNECTED_NORMALLY state (this state is reached after a successful establishment of the RRC connection), an inactivity timer is started and its value is configurable.

2. In case the network continues receiving packets from the UE, this timer value is reset. Otherwise, when the inactivity timer expires, the UE context is removed.

3. Before proceeding to remove the UE context, the network sends an RRC Connection Release message to the UE, which moves the UE from RRC Connected to RRC Idle state.

4. Finally, the base station communicates with the network core and all bearers associated with that UE are deleted.

Table A.2 shows the different configurable parameters to activate RRC Idle mode and to set the inactivity timer.

Parameter	Value	Description
UseIdealConnection	**Boolean** (True, False)	Indicates whether the mode used is the original simulator mode (True) or the one implemented with RRC Idle mode (False)
InactivityTimer	**TimeValue** (Value in seconds)	Indicates the inactivity time configured in the gNB to switch a UE to RRC Idle state when the timer expires

Table A.2: Configurable parameters to activate RRC Idle mode.

3. Dual Connectivity and Packet Duplication feature

The DC feature has been implemented in the simulator. This feature allows the UE to be connected with two base stations simultaneously. In order to support this feature in the simulator, two 5G network devices have been inserted in the same node, each of them associated and connected to the corresponding gNB and with their respective band configuration. On the other hand, the possibility of connecting a UE with two different technologies simultaneously, in this case LTE and 5G, has also been included. The implementation is equivalent to MR-DC, allowing to test different DC configurations, such as LTE as MN and 5G as SN, or both nodes being 5G.

Once DC has been inserted in a node, it is essential to implement an application that is capable of managing two sockets (each of them connected to the target application), otherwise, the UE will only send data over the primary socket and one of the links will be completely unusable. The implementation details of this application are described in the next subsection. On the other hand, the DC solution with the PD approach for the downlink has been included with the 5G technology, being used in a journal article of this thesis ([**II**]) corresponding to Chapter 5. The operation and implementation of this approach is described below.

1. Upon arrival of a packet for a UE at the MN from the network core, it is processed and a PDCP header is added, as indicated in the standard. Thus, both packets will contain the same identifier and the duplication can be detected at the receiver.

2. Once the PDCP header has been processed and added, the duplicated packet is sent thought the Xn interface to the SN. MN and SN are always the same throughout the simulation and the necessary functions have been added so that in the simulation configuration file the pairing (i.e., who acts as MN and who as

SN, as well as activating the Xn interface between them) is carried out prior to the start of the simulation. These parameters are described in Table A.3.

3. Both packets are processed independently by the lower layers (RLC, MAC, and PHY) at each node. Upon arrival at the UE, the first packet received is sent to the upper layers and, when the duplicate packet is detected (based on the sequence number), it is discarded and not sent to the upper layers.

4. Finally, the packet reaches the application layer, where it is processed and the corresponding traces are obtained.

Function	Parameter	Description
SetMasterGnb	**Boolean** (True, False)	Indicates whether the gNB acts as a master node in the case of DC
SetSecondaryGnb	**Boolean** (True, False)	Indicates whether the gNB acts as a secondary node in the case of DC
ActivatePacketDuplication	**sourceRnti**: UE RNTI in master node; **sourceCellId**: master cell identifier; **targetCellId**: secondary cell identifier; **targetRnti**: UE RNTI in secondary node	The master node activates PD for the indicated RNTI
DeActivatePacketDuplication	**sourceRnti**: UE RNTI in the master node	The master node deactivates PD for the indicated RNTI

Table A.3: Simulator functions implemented at RRC layer for the establishment of DC and PD.

4. Application module for DC

Two new modules have been included in the simulator, which are responsible for providing a User Datagram Protocol (UDP) uplink application with two different sockets, each one connected with the corresponding interface when DC feature is used. These modules are namely DualSocket and DualSocket5G.

This application allows the user data to be sent to the network, either via one interface or the other, or via both network interfaces. In this case, the operation of both modules is equivalent, with the only difference being the type of connection of the network interface (LTE-5G in DualSocket module and 5G-5G in the DualSocket5G module).

Figure A.3 depicts a visual scheme of the application. When a new packet arrives, an algorithm with a set of functions determines which network interface is selected.

In this case, if a number of conditions are met, the selected interface will be the best between the primary and secondary ones. Alternatively, the packet could be sent duplicated on both interfaces. Once the decision has been made, the packet will be sent to the lower layers.

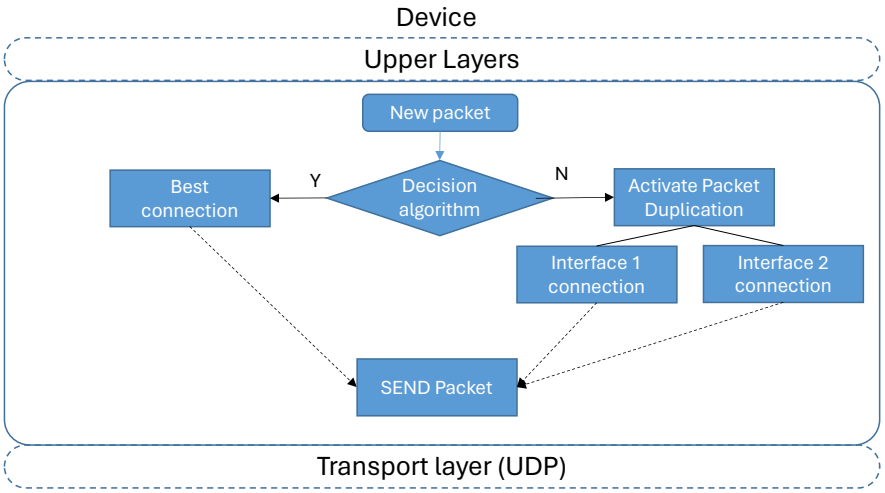


Figure A.3: Visual scheme of the UDP application.

On the other hand, Figure A.4 shows the operation of the decision algorithm to send the packet through one interface or another. The algorithm uses ML techniques to predict whether it is necessary to activate PD or, on the contrary, to choose the best connection. This algorithm has a number of inputs that are updated based on how packets arrive on that interface from the network (e.g., SINR, modulation used, delay obtained, etc.). In addition, feedback is provided on how the packet arrived at its destination (End-to-End (E2E) feedback).

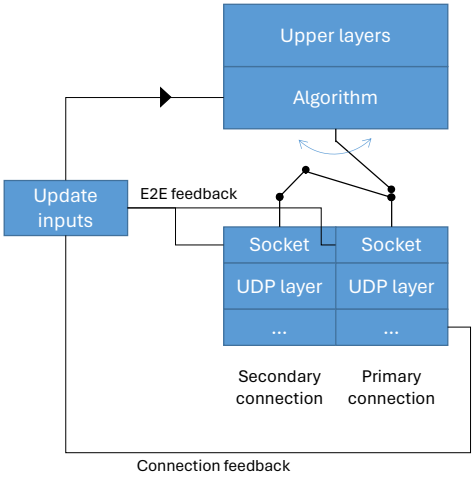


Figure A.4: Visual scheme of the decision algorithm.

5. A distribution center scenario

The enhancements made to the simulator also includes a realistic representation of a distribution center scenario, including the logistics activities that take place there. The details of this implementation are described in Chapter 5, corresponding to a journal article of this thesis ([**IV**]).

A.2 Random access simulator for cellular devices

As part of the journal publication [**IV**] of this thesis, a RA simulator for cellular devices has been developed, resulting in an open-source simulator with the code available on Github [120].

This simulator has been implemented on Python and enables the evaluation of the performance of the contention-based RA channel for 5G cellular networks. The implementation is based on the 3GPP standard [121–123] and the parameters that can be modified are described in Table A.4.

Parameter	Type	Description
PRACH Configuration Index	**Integer** ([0, 1, 2, 3, 4, 5])	Defines the periodicity of the RA slots. The periodicity ranges between a maximum of one RA slot per subframe to a minimum of one RA slot every two frames
Number of available preambles	**Integer**	Corresponds to the number of preambles reserved for the contention-based procedure
preambleTransMax	**Integer**	Maximum number of preamble attempts for a device before declaring RA failure
RAR Window Size	**Integer**	Time window to monitor the RA response
Backoff Indicator	**Integer**	Random backoff that is used by the UEs to wait a time when a preamble collision occurs before retrying a new access attempt. This backoff is intended to disperse the access attempts and thus, reduce the probability of preamble collision

Table A.4: Simulator parameters.

The simulator extracts the following metrics from the simulations:

- Blocking probability: probability that a device reaches the maximum number of transmission attempts (*preambleTransMax*) and is unable to complete an access

process.

- Average number of preamble retransmissions: measure the average number of preamble retransmissions required to have a success access.
- Access delay: time elapsed between the transmission of the first preamble and the reception of the Random Access Response (RAR) by the device. Only for devices that do not reach the maximum number of transmission attempts.

A.3 AAU 5G Smart Production Lab

The AAU 5G Smart Production Lab consists of a small-scale industrial factory environment of approximately 1250 m^2 and a wide range of industrial manufacturing and production equipment from different vendors, such as robotics arms, welding machines, production lines, AMRs, etc. The lab is equipped with multiple networks from different wireless technologies, such as private deployments of LTE, 5G NR, and Wi-Fi 6; and dedicated operator-managed network slices of LTE and 5G NR. The lab also contains a dedicated positioning system based on Ultra-Wide Band (UWB).

As part of the measurement campaigns performed in publication **[V]** of this thesis, a testbed was created to perform latency measurements with 5G SA and Wi-Fi 6 technologies. The testbed was composed of an Intel NUC [124], equipped with an Intel M2 Wi-Fi 6 AX200 card, and running Arch Linux; and with a 5G modem (Simcom SIM8202G-M2 [125]) connected to the NUC through a M2 to USB3 adapter. Figure A.5 illustrates the equipment used and the data path for each technology.

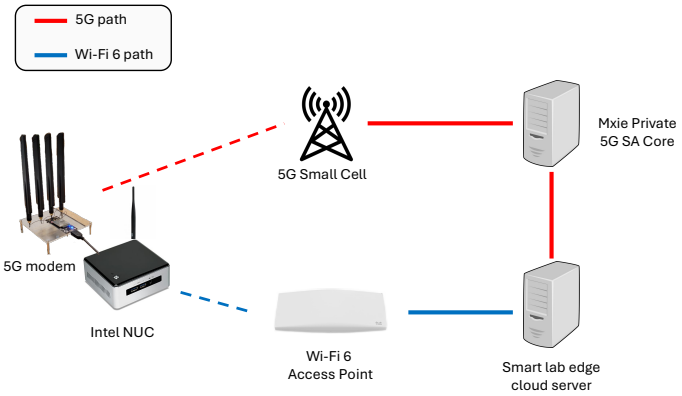


Figure A.5: Testbed in the AAU 5G Smart Production Lab.

For mobility measurements, a MiR200 AMR [126] was used, with the aforementioned equipment placed on top of the AMR. The AMR allowed to perform different

reproducible mobility tests within the AAU 5G Smart Production Lab, which guarantees a consistency on the measurements.

For the latency assessment, the Linux ping tool was used on the NUC, in which the interface for data transmission was indicated via command line. Python scripts were developed to automate the process of configuring devices and interfaces, launching multiple measurements to obtain statistical data, controlling the AMR robot path and collecting the data from the logs.

A.3.1 Mpconn tool

For the multi-connectivity measurements performed in the journal article [**V**] of this thesis, a tool developed at Aalborg University was used, namely mpconn [127]. This tool duplicates the packets at Layer 3 and sends them over IP in Layer 4 (UDP) packets through 5G and Wi-Fi technologies. An overview of the functionality of this tool is provided in Figure A.6, where a ping request packet is sent from a NUC to a server.

First, a virtual tunnel IP address is created in the NUC and in the server, where an instance of mpconn is running. This virtual tunnel IP address is used to communicate the mpconn instance running in the NUC with the mpconn instance running in the server. Then, for each packet sent by the NUC, a custom UDP packet adding a sequence number is created, and the packet is duplicated and sent via both interfaces. At the receiver side (server), the first packet received from the client is decapsulated, while the duplicated packet received is discarded based on the sequence number. In the example illustrated in Figure A.6, the packet duplication process for the ping reply will be the same but inverted.

A.4 Testbed for the evaluation of CIoT optimizations

The measurement campaigns in publications [**VIII**] and [**IX**] of this thesis were performed using a testbed with Amarisoft equipment. Specifically, the AMARI Callbox Classic and AMARI UE Simbox solutions from Amarisoft were used, and Figure A.7 illustrates these solutions and the different configurations that can be used.

Both devices have a completely software-based network implementation, where different network elements are deployed in a virtualized way. In the case of the Callbox,

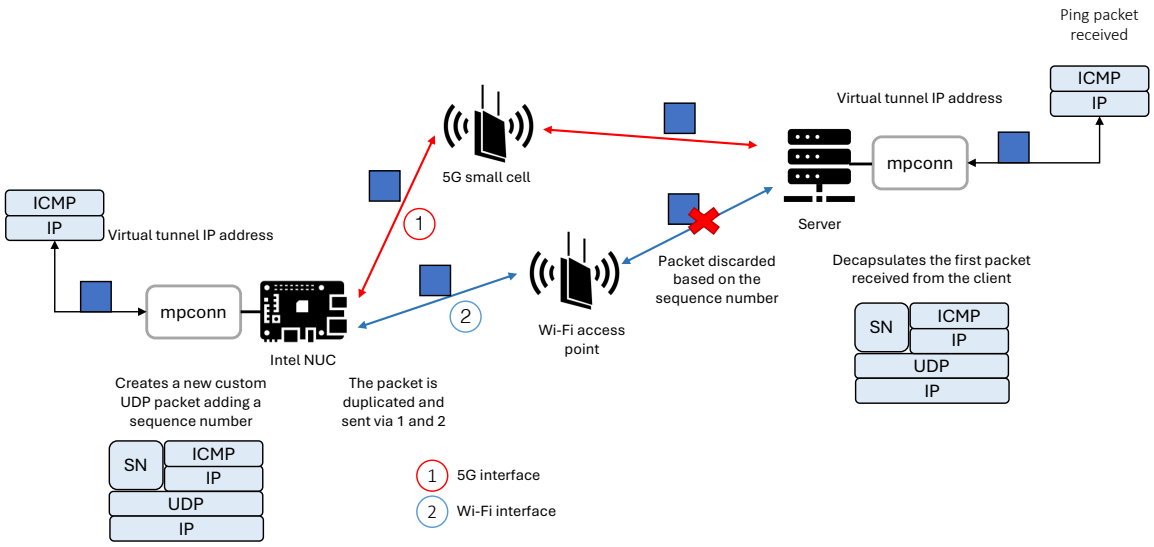


Figure A.6: An example of mpconn tool when transmitting a ping request.

the type of base station and the core network elements are virtualized. On the other hand, in the UE Simbox a UE terminal and its components are virtualized. The virtualization on these devices allows the configuration of different networks in the same physical device, thereby adding more flexibility. With respect to the Callbox, it allows the implementation of many LTE/NR network elements, such as the MME/AMF, as well as a large number of protocols and interfaces of these networks, thereby creating a virtual core. Similarly, the software allows to create different number of instances of eNB/ng-eNB/gNB, through which it is allowed to manage the Software Defined Radio (SDR) card of the device. All of this is implemented on a PC running on top of the Linux operating system.

Same as the Callbox, the UE Simbox allows a software implementation of a virtualized UE, where the different network elements of the UE are implemented along with its protocol layers. In this case, the UE Simbox allows the configuration of LTE, NB-IoT/LTE-M and NR devices. The entire implementation of both devices is based on the 3GPP standard with support up to Release 17.

Under this testbed, first, the configuration files for testing the CIoT optimizations were created. This involves configuring the scripts to define the network elements in both (Crowdcell and UE Simbox), the antennas, the spectrum and bandwidth, and the applications that run on top of the UE with the Amarisoft script format. In this case, the technology was configured as NB-IoT with support of EDT and the base station was configured as a ng-eNB connected to a 5GC. Furthermore, the Linux ping tool was implemented on top of the UE to evaluate the latency performance.

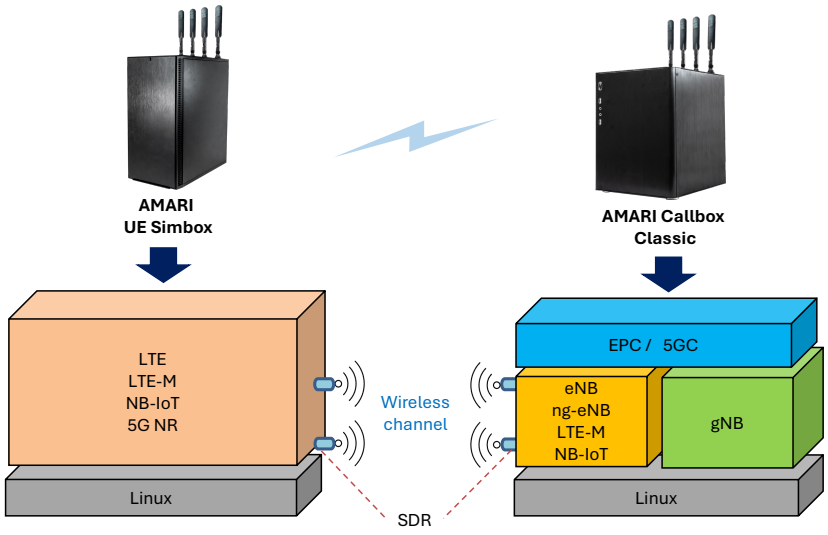


Figure A.7: Testbed with Amarisoft equipment.

Finally, a Python script was developed to automate the process of launching multiple measurements to obtain statistical data. In particular, the script was in charge of changing network conditions such as the cell gain, launching the tests and collecting the data from the logs. A diagram of the testbed for the latency evaluation of CIoT optimizations is depicted in Figure A.8.

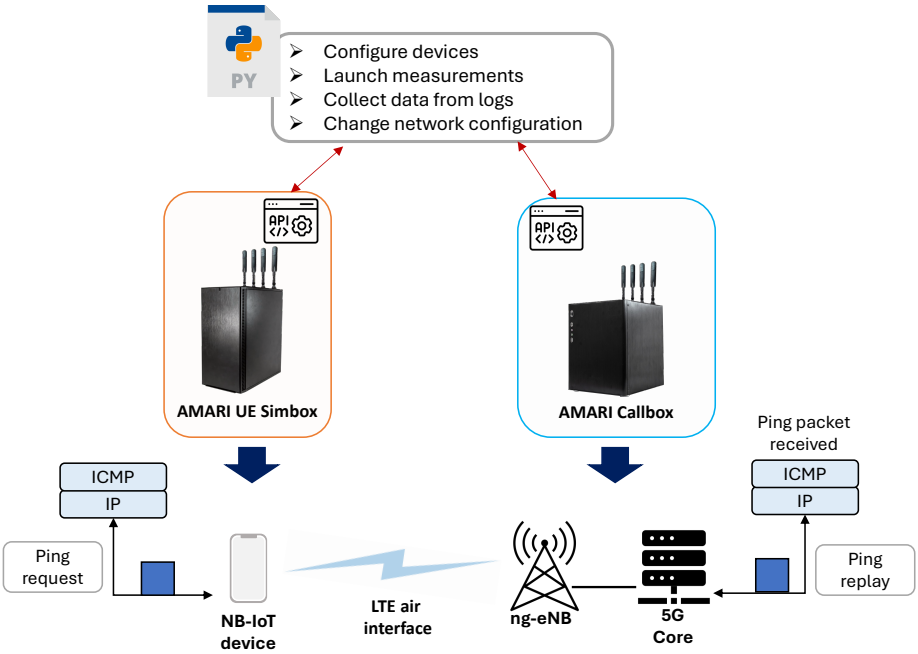


Figure A.8: Diagram of the testbed for the latency evaluation of CIoT optimizations.

A.5 Testbed for the evaluation of poisoning and evasion attacks in an E2E service

As part of publications [**XI**], [**XIV**] and [**XV**] of this thesis, a testbed for the evaluation of poisoning and evasion attacks in an E2E service has been used. More specifically, the testbed allows the extraction of metrics from the network and from the service, thus allowing the generation of datasets with radio and service parameters in situations with and without attack. The testbed has been implemented under a 5G network and the E2E service provided is the download of video on demand from the Youtube platform.

The physical architecture of the testbed is composed of different blocks, as depicted in Figure A.9. The testbed is partly inherited from the one proposed by the authors of [128], but with slight differences. First, a new block to include background traffic has been introduced, thereby providing a more realistic network scenario. Furthermore, a new block for the generation of attack samples is also included, thus allowing the generation of samples with altered values. These new blocks are marked in Figure A.9. Each of the different blocks that compose the testbed are detailed below.

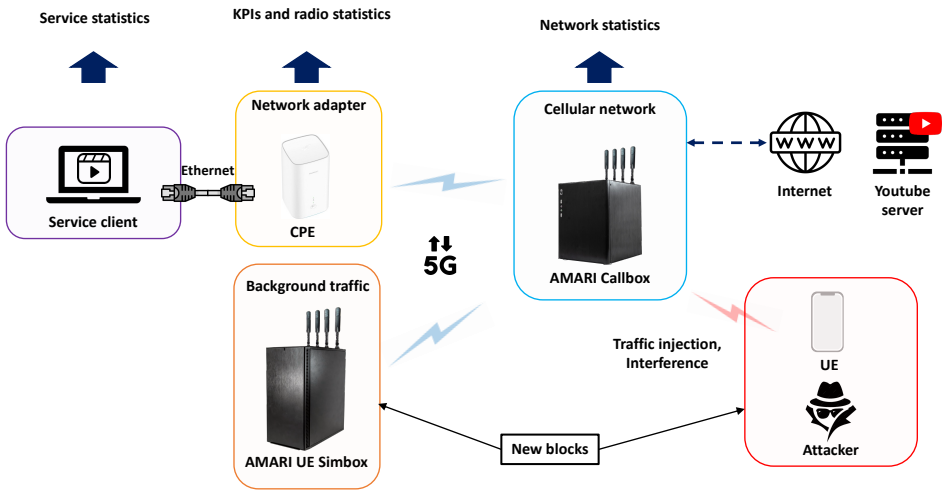


Figure A.9: An overview of the physical architecture of the testbed.

- **Service client**: the E2E service used is a video on demand service from the Youtube platform that is executed in a laptop. The video download is automated with different Python scripts that make use of the Selenium web driver. This block allows the collection of service metrics such as the buffer health, the initial time or freeze events.

- **Network adapter**: this block provides network connectivity to the client. For that purpose, the Huawei CPE PRO 2 has been used as a network adapter, which provides connectivity to the client via Ethernet connection, and backhaul connection with the cellular network.

- **Cellular network**: this block provides cellular connectivity to the users. It is composed of an AMARI Callbox Classic equipment, which creates a virtualized RAN and core network, providing 5G cellular service and internet access to the users.

- **Background traffic**: this new block adds background traffic in the network, thus generating a realistic scenario for the collection of network and service metrics. The equipment used in this block is the AMARI UE Simbox, that allows the emulation of multiple users (up to 64) connected to the AMARI Callbox. Each emulated user is able to run independent traffic, such as video content services, File Transfer Protocol (FTP), ping, etc.

- **Attacker**: this new block allows the collection of metrics in attack situation. In this block, the adversary can be considered as a legitimate user in the network, with its own identity and connected to the cellular network, injecting traffic into the network; or as an external adversary performing an attack to the legitimate network, such as an interference attack. The aim of this block is to alter network and service metrics collected.

Appendix B

Summary (Spanish)

B.1 Introducción

B.1.1 Motivación

La llegada de la cuarta revolución industrial o Industria 4.0 [1] marca un cambio en la fabricación y el sector industrial. El término Industria 4.0 fue usado por primera vez en 2011 en el encargo que el Gobierno alemán hizo a la *Industry-Science Research Alliance* para la consolidación del liderazgo de la industria alemana [2]. Posteriormente, esta iniciativa se extendió al resto de la Unión Europea y en la actualidad, la Industria 4.0 hace referencia a la interconexión de máquinas y sistemas dentro de los centros de producción, así como entre estos y el mundo exterior. Esta revolución digital está transformando las fábricas en fábricas inteligentes, donde la digitalización es clave. En una fábrica conectada, los sensores, el almacenamiento en la nube y el análisis de datos en tiempo real se utilizan para optimizar los procesos de producción. Un aspecto central de esta revolución es la necesidad de que los procesos de producción y distribución sean robustos, eficientes y más flexibles. Para alcanzar estas necesidades, existen diferentes tecnologías facilitadoras que están en el núcleo de la Industria 4.0:

- Sistemas ciberfísicos (*Cyber-Physical Systems*, CPS) [3, 4]. Integran la capacidad de computación y de red en un proceso físico. Las tecnologías CPS permiten el desarrollo de las fábricas inteligentes, donde las máquinas y los equipos están interconectados, lo que permite su monitorización, control y optimización en tiempo real.

- Internet de las cosas (*Internet of Things*, IoT) [5]. IoT es una red de objetos físicos a los que se han incorporado sensores, *software* y otras tecnologías que les permiten conectarse e intercambiar datos. En la Industria 4.0, el uso de IoT facilita el flujo continuo de información a través de las líneas de producción, mejorando la visibilidad operativa y la toma de decisiones.

- Inteligencia Artificial (IA) [6]. Los algoritmos de IA analizan grandes cantidades de datos generados por los CPS y los dispositivos IoT. Esta tecnología permite el mantenimiento predictivo, el control de calidad y la adaptación de los procesos de fabricación, reduciendo el tiempo de inactividad y mejorando la calidad del producto.

- Computación en la nube [7]. La computación en la nube desempeña un papel importante en la Industria 4.0 al proporcionar la infraestructura y la plataforma para almacenar, procesar y analizar las grandes cantidades de datos generados por los dispositivos IoT y otros sensores en el proceso de fabricación. Además, la computación en la nube puede proporcionar la potencia de cálculo necesaria para ejecutar algoritmos de IA.

- Realidad Aumentada (RA) [8]. La aplicación de la tecnología de RA puede mejorar una serie de procesos, como la formación, el mantenimiento y el diseño. Al superponer información digital al mundo físico, la tecnología de RA puede proporcionar a los trabajadores datos e instrucciones en tiempo real, facilitando así flujos de trabajo más eficientes y eficaces.

- Robótica [9, 10]. Los robots y los sistemas de automatización en la Industria 4.0 son más inteligentes, flexibles y colaborativos. Estos sistemas pueden realizar tareas complejas junto a los trabajadores, aumentando la productividad y la seguridad en los entornos de fabricación.

- Análisis de grandes datos [11, 12]. La recopilación y el análisis de grandes conjuntos de datos permiten mejorar las previsiones, mejoras en la eficiencia y descubrir nuevos conocimientos. La toma de decisiones basada en datos se encuentra en el núcleo de la Industria 4.0, impulsando prácticas de fabricación más ágiles y con mayor capacidad de respuesta.

Aunque el concepto de Industria 4.0 se centra principalmente en la industria de producción, las tecnologías y principios mencionados también se aplican en distintos sectores industriales, como la logística, la sanidad, la agricultura o la energía.

Las redes industriales tradicionales se basan principalmente en conexiones cableadas y tecnologías inalámbricas heredadas. Algunas de las conexiones cableadas que se han utilizado son ProfiNET [13], EtherCAT [14] y el conjunto de protocolos de redes sensibles al tiempo (*Time Sensitive Networks*, TSN) [15]. En el campo de las tecnologías inalámbricas, las principales tecnologías utilizadas son las basadas en la familia IEEE 802.11, comúnmente denominada Wi-Fi, pero también soluciones personalizadas para fábricas basadas en IEEE 802.15.1 y 802.15.4, como *Wireless Interface to Sensors and Actuators* (WISA) y WirelessHART [16]. Sin embargo, estas redes a menudo se quedan cortas en términos de escalabilidad, flexibilidad y capacidad de respuesta en tiempo real que requieren las aplicaciones industriales modernas [17]. La naturaleza dinámica de las fábricas inteligentes, los sistemas autónomos y las cadenas de suministro complejas requieren una infraestructura de comunicación que pueda soportar sin problemas un gran número de dispositivos conectados, facilitar el intercambio de datos en tiempo real y garantizar altos niveles de seguridad y fiabilidad.

Las redes celulares, con su adopción generalizada, fiabilidad demostrada y evolución continua, se encuentran en una posición única para satisfacer estas necesidades, ofreciendo una tecnología fundamental para impulsar la Industria 4.0. Las redes celulares, especialmente con la llegada de la quinta generación (5G) de redes móviles y las próximas tecnologías 6G [18], ofrecen capacidades sin precedentes que se alinean perfectamente con las demandas de la Industria 4.0. Entre ellas se incluye el soporte de casos de uso relacionados con comunicaciones críticas, que se conocen como comunicaciones ultra fiables de baja latencia (*Ultra-Reliable Low Latency Communications*, URLLC), el uso masivo de dispositivos de tipo máquina, también conocido como comunicaciones masivas de tipo máquina (*massive Machine-Type Communications*, mMTC), y los servicios mejorados de banda ancha (*enhanced Mobile Broadband*, eMBB). La capacidad de proporcionar una comunicación determinista, la compatibilidad con un número masivo de dispositivos IoT y un alto caudal de datos son habilitadores críticos para aplicaciones como el mantenimiento predictivo, la monitorización remota y la robótica autónoma. Además, la naturaleza modular y escalable de las redes celulares permite despliegues a medida en diversos entornos industriales, desde plantas de fabricación a gran escala hasta instalaciones remotas y aisladas. Esta flexibilidad admite la creación de redes privadas [19] dedicadas a necesidades industriales específicas, garantizando que se cumplan eficazmente los requisitos únicos de los distintos sectores.

El impulso mundial hacia la sostenibilidad y la eficiencia en las operaciones industriales [20] subraya aún más la importancia de aprovechar las redes de comunicación

avanzadas. Al permitir una gestión más eficiente de los recursos, reducir el tiempo de inactividad mediante el mantenimiento predictivo y facilitar la perfecta integración de las fuentes de energía renovables, las redes celulares [21] contribuyen significativamente a los objetivos de sostenibilidad de las industrias modernas.

Dado que la adopción e implantación de la tecnología celular se está llevando a cabo de manera progresiva en las industrias, especialmente la tecnología 5G [22], es necesario estudiar su aplicabilidad, evaluando el rendimiento de la red a través de los diferentes servicios y casos de uso involucrados en la fábrica inteligente.

B.1.2 Objetivos

El objetivo principal de esta tesis es evaluar y mejorar el rendimiento de la red celular en un entorno industrial de interior. Para ello, en esta tesis se abordan diferentes técnicas y optimizaciones de la red. En primer lugar, se realizan tareas relacionadas con el estudio de la latencia de servicios críticos y la escalabilidad en la red. En segundo lugar, se desarrollan diferentes herramientas para evaluar el rendimiento de la red en un entorno industrial y mejorar la fiabilidad de los servicios críticos mediante el uso de la solución de multiconectividad. En tercer lugar, se desarrollan y evalúan algoritmos de optimización con los siguientes propósitos: mejorar la fiabilidad de los servicios críticos sin desperdiciar recursos, y mejorar la calidad de servicio (*Quality of Service*, QoS) de los diferentes perfiles de tráfico involucrados en una fábrica. Por último, se ha evaluado el rendimiento de las distintas optimizaciones propuestas por el *Third Generation Partnership Project* (3GPP) para dispositivos IoT celulares (*Cellular IoT*, CIoT), incluyendo también un análisis de seguridad de la última optimización. Específicamente, las líneas de investigación abordadas en esta tesis se pueden resumir en los siguientes objetivos:

Obj. 1. Estudiar el impacto de las numerologías 5G en la latencia de los servicios críticos.

El objetivo de este estudio consiste en analizar el comportamiento de las diferentes configuraciones de numerología en la latencia percibida por los usuarios bajo diferentes condiciones de canal y tamaños de paquete. En este sentido, este estudio pretende sentar las bases para futuras optimizaciones en la reducción de la latencia, ya que una numerología adecuada puede seleccionarse en función de las condiciones radio experimentadas.

Obj. 2. Evaluar y comparar la escalabilidad de la red con diferentes tecnologías en un entorno industrial.

El propósito de este objetivo es el de evaluar y comparar empíricamente el rendimiento de la red respecto a la latencia y las pérdidas de paquetes con distintas tecnologías en un escenario industrial de interior. En concreto, la evaluación debe tener en cuenta diferentes tamaños de paquete y escenarios con distintos número de dispositivos transmitiendo datos. Como resultado, este estudio debería proporcionar una visión clara sobre qué tecnología se adapta mejor al sector industrial.

Obj. 3. Proponer un mecanismo para mejorar la fiabilidad de los servicios críticos.

Este objetivo se refiere al diseño y desarrollo de un algoritmo que cumpla los requisitos de fiabilidad de los servicios críticos. Así, el algoritmo propuesto debería ser capaz de adaptar y controlar dinámicamente la activación de la duplicación de paquetes para evitar el malgasto de los recursos en la red.

Obj. 4. Evaluar el rendimiento de la red en un centro de distribución.

La finalidad de este objetivo es la de realizar una evaluación de la red 5G en un escenario correspondiente a un centro de distribución, teniendo en cuenta los diferentes perfiles de tráfico implicados en este escenario. En concreto, en este trabajo se debería comparar la QoS de estos perfiles de tráfico bajo diferentes actividades logísticas con diferentes enfoques de *Network Slicing* (NS).

Obj. 5. Estudiar el impacto de las optimizaciones de señalización para CIoT en la red.

El objetivo de este estudio es el de analizar el comportamiento de las diferentes optimizaciones de señalización de CIoT en la latencia percibida por el usuario cuando transmite de forma infrecuente pequeños datos en la red.

Obj. 6. Analizar la seguridad de EDT en 5G para CIoT.

Este objetivo se relaciona con el Obj. 5 y se refiere a un análisis en profundidad de la seguridad de la optimización de la transmisión temprana de datos (*Early Data Transmission*, EDT), describiendo en detalle sus modos de operación y analizando las principales vulnerabilidades asociadas a esta optimización. Como resultado, un conjunto de recomendaciones para los proveedores debería ser derivado del análisis de seguridad.

B.2 Descripción de los resultados

En esta sección se presentan los resultados derivados de esta tesis. Estos trabajos abordan los retos identificados y los objetivos definidos en la Sección 1.2. La Figura B.1 ilustra la relación entre los retos, los objetivos y los resultados obtenidos. En la figura, cada publicación se representa como un bloque individual, indicando el capítulo de la tesis en el cual se incluye.

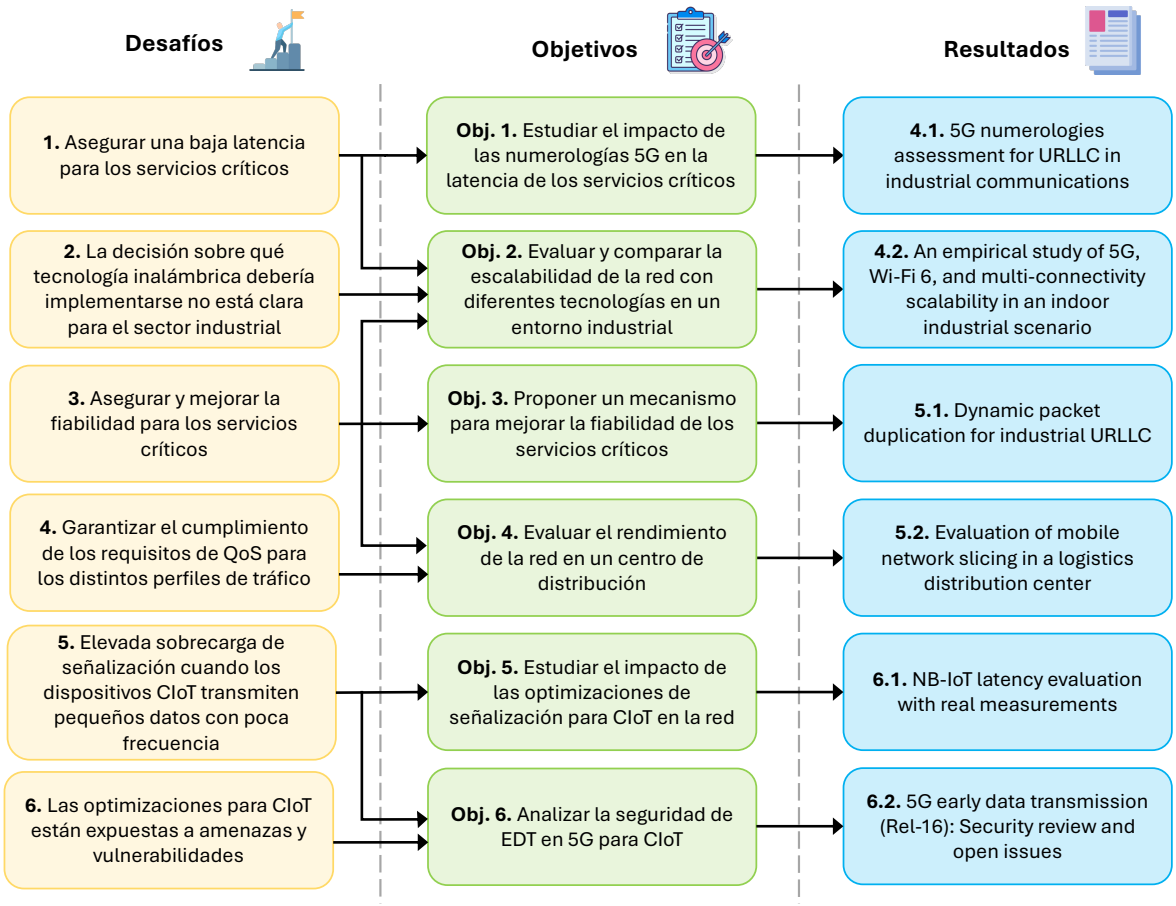


Figura B.1: Desafíos, objetivos y publicaciones.

B.2.1 Evaluación de las numerologías 5G para URLLC en comunicaciones industriales

La llegada de la red 5G ha facilitado la introducción de características novedosas, permitiendo el desarrollo de nuevos casos de usos y servicios. Una de estas características es la numerología, que permite un proceso de asignación de recursos más rápido debido al uso de *slots* de tiempo más cortos. Esta característica es de particular importancia

para servicios con restricción de latencia, como los empleados en la operación de los vehículos guiados automatizados (*Automated Guided Vehicles*, AGVs), ya que permite una reducción de la latencia.

Sin embargo, en los escenarios industriales, el principal desafío proviene de la presencia de muros de hormigón y de las grandes estructuras y máquinas metálicas, lo que da lugar a interferencia y propagación multicamino. Como consecuencia, seleccionar una numerología apropiada es una tarea desafiante, y esta debe adaptarse a las condiciones de radio experimentadas.

Por ello, el primer artículo presentado en el Capítulo 4 se enfoca en la evaluación del impacto de la numerología en el retardo experimentado en el enlace radio para un servicio de control remoto (comunicación de AGVs), cubriendo con ello el Obj. 1 de esta tesis. Específicamente, este estudio abarca la evaluación con distintos tamaños de paquete y condiciones de canal en un entorno de industria simulado, con un foco especial en la identificación y el análisis de los valores anómalos.

Los resultados demuestran que la premisa de que una alta numerología tiende a un menor retardo no siempre se cumple, particularmente en condiciones de no línea de visión directa (*Non-Line-of-Sight*, NLOS). En estos casos, una numerología intermedia es más adecuada para este tipo de servicio.

B.2.2 Estudio empírico de la escalabilidad de 5G, Wi-Fi 6 y multiconectividad en un escenario industrial de interior

El sector industrial está adoptando la Industria 4.0 para mejorar la flexibilidad y reducir los costes de instalación mediante el uso de la conectividad inalámbrica. Sin embargo, persiste la pregunta sobre qué tecnología inalámbrica debería implementarse en la fábrica para cumplir con los requisitos de las aplicaciones de próxima generación, como los robots móviles autónomos (*Autonomous Mobile Robots*, AMRs). Mientras que la tecnología Wi-Fi es la más prevalente y fácil de desplegar, la red 5G ha sido diseñada para soportar las necesidades del sector industrial. Por lo tanto, es importante comparar ambas tecnologías desde el punto de vista del rendimiento, especialmente bajo diferentes condiciones de carga y con diferentes número de dispositivos. El uso de la multiconectividad con diferentes tecnologías de acceso radio también se considera un habilitador clave para satisfacer los requisitos de las aplicaciones en tiempo real más

críticas.

Por lo tanto, el segundo artículo presentado en el Capítulo 4 se centra en la evaluación empírica y comparación de la escalabilidad de la red 5G, Wi-Fi 6 y multiconnectividad desde el punto de vista de la latencia y las pérdidas de paquetes, cubriendo con ello el Obj. 2 de esta tesis. Este trabajo se ha llevado a cabo en el laboratorio "5G Smart Production Lab" en Aalborg (Dinamarca), donde se han realizado diferentes campañas de medidas para diversos escenarios (estático y movilidad) y tamaños de paquete.

Los resultados obtenidos muestras en general latencias bajas con Wi-Fi, pero largas colas en la distribución de la latencia, con unas pérdidas de paquetes mayores en comparación con 5G. Por otro lado, la latencia de 5G es muy consistente con colas acotadas y obteniendo una baja pérdida de paquetes. En términos de escalabilidad, 5G escala mejor que Wi-Fi, viéndose esta última muy afectada por el número de dispositivos transmitiendo datos. Finalmente, la solución de multiconectividad mostró una mayor fiabilidad y menores latencias en todos los casos evaluados.

B.2.3 Duplicación de paquetes dinámica para URLLC industrial

Este trabajo sigue la línea comenzada con la primera publicación del Capítulo 4. Esto es, una vez se selecciona una numerología apropiada para reducir la latencia, el segundo paso consiste en mejorar la fiabilidad de las comunicaciones críticas. Una de las formas de mejorar la fiabilidad de estas comunicaciones es mediante el uso de la multiconectividad, particularmente con el enfoque de duplicación de paquetes. No obstante, esta solución lleva consigo un aumento de la redundancia en la red, lo que puede llevar a un uso inapropiado de los recursos de red.

Por ello, para reducir el malgasto de los recursos de red, el primer artículo del Capítulo 5 propone un algoritmo de duplicación de paquetes dinámico basado en aprendizaje automático (*Machine Learning*, ML), que determina cuando la duplicación de los paquetes es requerida en una transmisión específica de datos para enviar un mensaje crítico de manera exitosa (Obj. 3). En concreto, un estimador de latencia basado en bosque aleatorio (*Random Forest*, RF) fue entrenado y evaluado, el cual decide cuando realizar la duplicación basándose en un umbral de latencia. La metodología presentada fue evaluada en un simulador 5G y el rendimiento de la red fue comparado con distintos

enfoques: no duplicar los paquetes y, una duplicación estática de los paquetes.

Los resultados de la evaluación demostraron que el algoritmo dinámico de duplicación de paquetes propuesto reduce en un 81% el número de paquetes duplicados enviados mientras que mantiene el mismo nivel de latencia (esto es, la latencia obtenida se encuentra por debajo del umbral) que la técnica de duplicación estática. Esta reducción en el número de paquetes duplicados resulta en un uso más eficiente de los recursos de la red.

B.2.4 Evaluación de *Network Slicing* de la red móvil en un centro de distribución logística

El segundo artículo incluido en el Capítulo 5 aborda el problema de optimizar los recursos de la red para los diferentes perfiles de tráfico involucrados dentro de un centro de distribución de logística. En concreto, estos perfiles de tráfico corresponden a eMBB, URLLC y mMTC, con distintos requisitos de latencia, fiabilidad, *throughput*, etc.

Específicamente, este artículo primero introduce un novedoso simulador de código abierto desarrollado, basado en el *framework* de ns-3, con una representación realista de un escenario de centro de distribución, donde están presentes diferentes actividades logísticas. Las comunicaciones de estas actividades han sido modeladas y usadas para estimar el rendimiento de los diferentes perfiles de tráfico. Como resultado, el simulador desarrollado sirve como base para evaluar el rendimiento de la red 5G en un escenario de logística inteligente (Obj. 4).

En segundo lugar, bajo el simulador desarrollado, este trabajo evalúa y compara el rol de dos estrategias de NS en 5G para la logística inteligente: el uso de una *slice* estática con una división balanceada de los recursos de red y el uso de una *slice* dinámica que adapta los recursos basándose en la carga del tráfico, dependiendo de la actividad que se esté realizando. En concreto, este trabajo evalúa estas estrategias en términos de QoS para los diferentes perfiles de tráfico, resultando en las siguientes métricas: *throughput* para el tráfico eMBB, fiabilidad para el tráfico URLLC y el canal de acceso aleatorio (*Random Access*, RA) para el tráfico mMTC.

Los resultados obtenidos muestran que una *slice* dinámica realiza un uso más eficiente de los recursos radio, mejorando la QoS de los diferentes perfiles de tráfico, incluso cuando hay un pico de tráfico en un perfil específico. Esta mejora va desde el

6.48% a el 95.65%, dependiendo del perfil de tráfico específico y la métrica evaluada.

B.2.5 Evaluación de la latencia de NB-IoT con medidas reales

Diferentes optimizaciones han sido propuestas por el 3GPP para dispositivos CIoT con el objetivo de mejorar la vida de la batería y reducir la señalización con la red. Estas optimizaciones comenzaron con la llegada de la *Release* 13, donde la transmisión por el plano de control (*Control Plane*, CP) fue introducida. Esta optimización permite la transmisión de datos utilizando el CP en lugar del plano de usuario (*User Plane*, UP), evitando con ello el establecimiento de las portadoras radio de datos (*Data Radio Bearers*, DRBs) del UP.

Además, con la llegada de la *Release* 15, la optimización de EDT fue introducida para soportar las transmisiones infrecuentes de datos pequeños, soportando tanto el modo de transmisión por el plano de control CP como por el UP. Esta última optimización permite la transmisión de datos durante el procedimiento de RA, con una reducción significativa en la señalización entre el UE y la red, y sin la necesidad de realizar un cambio de estado de la capa *Radio Resource Control* (RRC). Esto es, el UE transmite los datos en el estado RRC *Idle*.

De este modo, el primer artículo del Capítulo 6 se centra en la evaluación y comparativa de las optimizaciones propuestas por el 3GPP para CIoT previamente mencionadas a través del CP en términos de rendimiento de latencia con la tecnología NB-IoT, cubriendo con ello el Obj. 5 de esta tesis. En concreto, en este trabajo se ha realizado una campaña de medidas con equipos de Amarisoft (AMARI Crowdcell y AMARI UE Simbox) bajo diferentes tamaños de paquetes y niveles de cobertura.

Los resultados evaluados mostraron bajas latencia para EDT, particularmente en el caso de paquetes pequeños, donde se utiliza un *transport block* reducido, siendo así más eficiente desde una perspectiva de la red. Además, se ha demostrado que EDT, al contrario que la optimización de la *Release* 13, logra el requisito de latencia definido por el 3GPP (10 segundos) bajo cobertura extrema.

B.2.6 EDT en 5G: revisión de seguridad y problemas abiertos

Esta sección presenta el segundo de los trabajos llevado a cabo en relación con el Capítulo 6 de esta tesis. En este caso, este trabajo extiende la línea comenzada con la primera publicación del Capítulo 6, ofreciendo una descripción detallada de la optimización de EDT junto a un análisis de la seguridad de este mecanismo. Por tanto, este trabajo cubre el Obj. 6 de esta tesis.

Como se ha mencionado anteriormente, la optimización de EDT fue introducida en la *Release* 15 para permitir la transmisión de datos durante el proceso de RA. Esta característica, destinado especialmente para transmisiones infrecuentes y con tamaños pequeños, trata de reducir la latencia y el consumo de los UEs. No obstante, a pesar de la importancia de esta novedad y el acuerdo general sobre su efectividad, existen pocos trabajos en la literatura que proporcionen información sobre su implementación y analicen las ventajas y desventajas de sus dos diferentes opciones de implementación (CP y UP).

Además, a pesar de que la seguridad es reconocida como un aspecto crucial para el correcto despliegue de esta tecnología, la literatura carece de una revisión de los problemas de seguridad y las características de este mecanismo. Como consecuencia de esta falta de trabajos y la complejidad de los protocolos de redes móviles, existe una división entre los expertos en seguridad y los investigadores de EDT, que impide el fácil desarrollo de esquemas de seguridad.

Para combatir esta importante brecha, este artículo ofrece un tutorial de EDT y su seguridad, analizando las principales vulnerabilidades y concluyendo con un conjunto de recomendaciones para investigadores y fabricantes. En concreto, debido a las simplificaciones en los protocolos llevado a cabo por EDT, se han encontrado vulnerabilidades como la inyección de paquetes, ataques por repetición y la inyección de valores falsos para deshabilitar EDT en la red.

B.3 Conclusiones

B.3.1 Contribuciones

Esta tesis tiene como objetivo evaluar y mejorar el rendimiento de las redes móviles en el paradigma de la Industria 4.0. Para ello, se han identificado un conjunto de desafíos en el entorno industrial y se han definido los objetivos necesarios para resolver estos desafíos. A lo largo de este trabajo se han establecido un total de seis objetivos, los cuales están distribuidos de la siguiente manera. Los Obj. 1 y 2 están relacionados con la evaluación del rendimiento de la red en un entorno industrial de interior. Los Obj. 3 y 4 se refieren al desarrollo de algoritmos de optimización para mejorar el rendimiento de la red. Finalmente, los Obj. 5 y 6 pretenden cubrir las optimizaciones de señalización de CIoT, primero evaluando el impacto de estas optimizaciones en la red y luego proporcionando un análisis de la seguridad de la última optimización propuesta por el 3GPP. A continuación se presentan las contribuciones relacionadas con cada uno de estos objetivos:

Obj. 1. Estudiar el impacto de las numerologías 5G en la latencia de los servicios críticos.

- Se ha realizado un análisis del impacto de las diferentes configuraciones de numerología 5G en la latencia de los usuarios. En este análisis, se ha llevado a cabo un estudio más detallado que los encontrados en el estado del arte. Se han evaluado las numerologías 5G bajo diferentes condiciones de canal (LOS y NLOS) y con diferentes tamaños de paquete para el caso de uso de un AGV.
- El estudio se ha llevado a cabo en un entorno 5G simulado. Los resultados mostraron que la selección de la numerología no es trivial, siendo un valor intermedio más adecuado bajo condiciones NLOS. Este estudio abre la puerta a algoritmos que puedan ser utilizados para dinámicamente ajustar la configuración de la numerología en la red para un mejor rendimiento.

Obj. 2. Evaluar y comparar la escalabilidad de la red con diferentes tecnologías en un entorno industrial.

- Siguiendo este objetivo, se ha llevado a cabo una evaluación empírica con diferente número de dispositivos, tamaños de paquete y escenarios para eval-

uar el rendimiento de la red respecto a la latencia y las pérdidas de paquetes. En concreto, para realizar esta comparativa de rendimiento, se han seleccionado las tecnologías 5G, Wi-Fi 6 y el uso de multiconectividad entre ambas con un enfoque de duplicación de paquetes.

- Específicamente, se han llevado a cabo campañas de medidas con equipo comercial en el laboratorio "5G Smart Production Lab" de la Universidad de Aalborg (Dinamarca), que consiste en un entorno de fábrica industrial de interior a pequeña escala compuesto por dos salas y una amplia gama de equipos de fabricación y producción industrial.
- Como resultado de las campañas de medida, se ha demostrado que la tecnología 5G proporciona menor latencia en las colas y es más fiable que Wi-Fi 6. Por otro lado, la solución de multiconectividad demostró una significativa reducción en las colas de la latencia y cero paquetes perdidos, siendo esta solución muy efectiva para lograr los casos de uso con requisitos de latencia y fiabilidad muy restrictivos.

Obj. 3. Proponer un mecanismo para mejorar la fiabilidad de los servicios críticos.

- Para cumplir este objetivo, se ha diseñado un algoritmo basado en ML para activar dinámicamente la duplicación de paquetes en un entorno industrial. Este algoritmo se basa en métricas de red como la relación señal/interferencia más ruido (SINR), el índice de modulación y la retroalimentación *Hybrid Automatic Repeat reQuest* (HARQ) para predecir la latencia. La salida del predictor se utiliza para la decisión de duplicación de paquetes en el enlace descendente, basándose en un umbral de latencia.
- El predictor de latencia se entrenó con el algoritmo de RF y el rendimiento del algoritmo propuesto se validó mediante diferentes pruebas realizadas en un entorno simulado 5G. En este simulador, se implementó la función de conectividad dual (*Dual Connectivity*, DC) con un enfoque de duplicación de paquetes, tal y como se describe en el Apéndice A.
- Se ha comparado el rendimiento del algoritmo con otros enfoques en el estado del arte, demostrando así que la solución propuesta es capaz de obtener mejores resultados y de minimizar el desperdicio de recursos en la red.

Obj. 4. Evaluar el rendimiento de la red en un centro de distribución.

- Las contribuciones correspondientes a este objetivo son, en primer lugar, el diseño e implementación de un simulador de código abierto basado en el *framework* de ns-3 y el módulo 5G-LENA, que incluye nuevas características para soportar la evaluación de la red en un centro de distribución. En concreto, se han implementado características como el escenario de un centro de distribución, las actividades involucradas allí, la asignación de recursos por *slice*, y el modelo de propagación y canal industrial.
- En segundo lugar, cuando las características anteriores fueron implementadas, se han evaluado dos estrategias de NS utilizando la red 5G. Estas estrategias consisten en el uso de una *slice* estática con una división de los recursos de red balanceada, y el uso de una *slice* que dinámicamente ajusta su tamaño dependiendo de la actividad que se esté llevando a cabo.
- Finalmente, se ha evaluado mediante simulaciones el rendimiento de QoS proporcionado por esas estrategias de NS sobre distintos perfiles de tráfico, y los resultados han demostrado que una *slice* dinámica mejora la QoS especialmente con alta carga de tráfico, mientras que la *slice* estática rinde bien cuando la carga de tráfico es baja.

Obj. 5. Estudiar el impacto de las optimizaciones de señalización para CIoT en la red.

- En relación con este objetivo, se ha llevado a cabo un análisis del impacto en la latencia de los usuarios de las diferentes optimizaciones de señalización en CIoT propuestas por el 3GPP. Concretamente, en este estudio se ha utilizado la tecnología NB-IoT para la evaluación de las distintas optimizaciones por el CP.
- El estudio se ha llevado a cabo con equipo comercial de Amarisoft, con el cual se han realizado varias campañas de medidas bajo diferentes niveles de cobertura y tamaños de paquete.
- A partir de los resultados de estas mediciones, se ha demostrado que EDT, a diferencia de la optimización de la *Release* 13, cumple con el requisito de latencia definido por el 3GPP para transmisiones de datos pequeños y poco frecuentes bajo un nivel de cobertura extremo.

Obj. 6. Analizar la seguridad de EDT en 5G para CIoT.

- Como contribución final se ha llevado a cabo un estudio de la optimización de EDT en CIoT. En este estudio, la optimización de EDT se ha descrito en detalle para sus dos modos de operación soportados, el CP y el UP.
- Además, se ha proporcionado un análisis de seguridad de esta optimización, extrayendo las principales vulnerabilidades encontradas en cada uno de sus modos de operación. Concretamente, se han encontrado vulnerabilidades como la inyección de paquetes, los ataques por repetición o la inyección de valores falsos para deshabilitar EDT.
- Finalmente, tras un análisis de seguridad exhaustivo, se ha proporcionado un conjunto de recomendaciones para investigadores y fabricantes, que incluyen soluciones para remediar estas vulnerabilidades en futuras versiones del estándar 3GPP, y que modo de operación es más recomendable de utilizar.

B.3.2 Publicaciones

Revistas

Publicaciones derivadas de esta tesis

El trabajo realizado en esta tesis ha dado lugar a cuatro artículos publicados en revistas de alto impacto más otra en proceso de revisión, que se enumeran a continuación.

[I] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "5G Numerologies Assessment for URLLC in Industrial Communications," *Sensors*, vol. 21, no. 7, p. 2489, Abr. 2021.

[II] D. Segura, E.J. Khatib, and R. Barco, "Dynamic Packet Duplication for Industrial URLLC," *Sensors*, vol. 22, no. 2, p. 587, Ene. 2022.

[III] D. Segura, J. Munilla, E.J. Khatib, and R. Barco, "5G Early Data Transmission (Rel-16): Security Review and Open Issues," *IEEE Access*, vol. 10, pp. 93289–93308, Sep. 2022.

[IV] D. Segura, E.J. Khatib, and R. Barco, "Evaluation of Mobile Network Slicing in a Logistics Distribution Center," *IEEE Transactions on Network and Service Management*, Bajo revisión, 2024.

[V] D. Segura, S.B. Damsgaard, A. Kabaci, P. Mogensen, E.J. Khatib, and R. Barco, "An Empirical Study of 5G, Wi-Fi 6, and Multi-Connectivity Scalability in an Indoor Industrial Scenario," *IEEE Access*, vol. 12, pp. 74406-74416, May. 2024.

Conferencias

Conferencias derivadas de esta tesis

También se han presentado varios trabajos en congresos nacionales e internacionales, como se muestra a continuación.

[VI] D. Segura, E.J. Khatib, and R. Barco, "Evaluación de numerologías 5G para URLLC," en *XXXV Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2020)*, Málaga (España), Sept. 2020.

[VII] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "Evaluación de los modos de conexión para NB-IoT," en *XXXVI Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2021)*, Vigo (España), Sept. 2021.

[VIII] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "NB-IoT latency evaluation with real measurements," en *2022 IEEE Workshop on Complexity in Engineering (COMPENG)*, Florencia (Italia), Jul. 2022.

[IX] D. Segura, E.J. Khatib, J. Munilla, and R. Barco, "Evaluación de la latencia de NB-IoT con medidas reales," en *XXXVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2022)*, Málaga (España), Sept. 2022.

[X] D. Segura, S.B. Damsgaard, P. Mogensen, E.J. Khatib, and R. Barco, "Comparativa empírica del rendimiento de 5G y Wi-Fi en un escenario industrial de interior," en *XXXVIII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2023)*, Cáceres (España), Sept. 2023.

[XI] D. Segura, H.Q. Luo-Chen, C. Baena, E.J. Khatib, S. Fortes, and R. Barco, "Testbed para la evaluación de los ataques de envenenamiento y evasión en un servicio E2E," en *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

Conferencias relacionadas con esta tesis

[**XII**] J. Llanes, E.J. Khatib, D. Segura, and R. Barco, "Seguridad en B5G/6G," in *XXXVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2022)*, Málaga (España), Sept. 2022.

[**XIII**] S.B. Damsgaard, D. Segura, M.F. Andersen, S.A. Markussen, S. Barbera, I. Rodríguez, and P. Mogensen, "Commercial 5G NPN and PN Deployment Options for Industrial Manufacturing: An Empirical Study of Performance and Complexity Tradeoffs," en *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Toronto (Canada), Sept. 2023.

[**XIV**] H.Q. Luo-Chen, D. Segura, C. Baena, E.J. Khatib, and R. Barco, "Detection of anomalous samples based on automatic thresholds," en *2024 IEEE Workshop on Complexity in Engineering (COMPENG)*, Florencia (Italia), Jul. 2024.

[**XV**] H.Q. Luo-Chen, D. Segura, C. Baena, E.J. Khatib, S. Fortes, and R. Barco, "Alteración de datos E2E: impacto de un ataque de envenenamiento y evasión en una red celular," en *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

[**XVI**] C.S. Álvarez-Merino, D. Segura, C. Baena, E.J. Khatib, and R. Barco, "Infraestructura para la monitorización del consumo energético en redes b5G/6G," en *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

[**XVII**] E.J. Khatib, D. Segura, A. Tarrías, and R. Barco, "Estudio del ataque de cadena de suministro sobre XZ utils y sus consecuencias en telecomunicaciones," en *XXXIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2024)*, Cuenca (España), Sept. 2024.

B.3.3 Proyectos relacionados

Esta tesis ha contribuido a los siguientes proyectos:

- Proyectos nacionales:
  - EDEL4.0: Seguridad y fiabilidad en las comunicaciones 5G/IoT para la Industria 4.0. Número de proyecto UMA18-FEDERJA-172, recibiendo fondos de la Junta de Andalucia y la Comisión Europea, perteneciente a la convocatoria "Proyectos de I+D+i en el marco del Programa Operativo FEDER Andalucia 2014-2020".
  - PENTA: Provisión de servicios PPDR a través de Nuevas Tecnologías de Acceso radio. Número de proyecto PY18-4647, recibiendo fondos de la Junta de Andalucía y la Comisión Europea, perteneciente a la convocatoria del "Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI 2020)".
  - MAORI: Massive AI for the OpenRadIo b5G/6G network. Número de proyecto TSI-063000-2021-72, recibiendo fondos del Ministerio de Asuntos Económicos y Transformación Digital y la Unión Europea - NextGenerationEU dentro del marco de "Recuperación, Transformación, y Resiliencia".

B.3.4 Estancia de investigación

Como parte de esta tesis se ha realizado una estancia de investigación de cinco meses en Aalborg (Dinamarca), colaborando con la Universidad de Aalborg en la realización de varias campañas de medidas con 5G, Wi-Fi 6 y multiconectividad en un entorno industrial. La estancia tuvo lugar entre febrero de 2023 y junio de 2023, y fue supervisada por Preben E. Mogensen.

Bibliography

[1] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Business & information systems engineering*, vol. 6, pp. 239–242, 2014.

[2] European Commission, "Digital Transformation Monitor. Germany: Industrie 4.0," 2017, (accessed June 2024). [Online]. Available: https://de.sistematica.it/docs/379/Germay_Industrie_4.0.pdf

[3] D. G. Pivoto, L. F. de Almeida, R. da Rosa Righi, J. J. Rodrigues, A. B. Lugli, and A. M. Alberti, "Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review," *Journal of manufacturing systems*, vol. 58, pp. 176–192, 2021.

[4] S. J. Oks, M. Jalowski, M. Lechner, S. Mirschberger, M. Merklein, B. Vogel-Heuser, and K. M. Möslein, "Cyber-physical systems in the context of Industry 4.0: A review, categorization and outlook," *Information Systems Frontiers*, pp. 1–42, 2022.

[5] M. Soori, B. Arezoo, and R. Dastres, "Internet of things for smart factories in Industry 4.0, a review," *Internet of Things and Cyber-Physical Systems*, 2023.

[6] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022.

[7] A. Sunyaev and A. Sunyaev, "Cloud computing," *Internet computing: Principles of distributed systems and emerging internet-based technologies*, pp. 195–236, 2020.

[8] P. Fraga-Lamas, T. M. Fernández-Caramés, O. Blanco-Novoa, and M. A. Vilar-Montesinos, "A review on industrial augmented reality systems for the Industry 4.0 shipyard," *IEEE Access*, vol. 6, pp. 13 358–13 375, 2018.

[9] J. Wen, L. He, and F. Zhu, "Swarm robotics control and communications: Imminent challenges for next generation smart logistics," *IEEE Communications Magazine*, vol. 56, no. 7, pp. 102–107, 2018.

[10] R. Goel and P. Gupta, "Robotics and Industry 4.0," *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*, pp. 157–169, 2020.

[11] J. Lee, H.-A. Kao, and S. Yang, "Service innovation and smart analytics for Industry 4.0 and big data environment," *Procedia cirp*, vol. 16, pp. 3–8, 2014.

[12] M. Khan, X. Wu, X. Xu, and W. Dou, "Big data challenges and opportunities in the hype of Industry 4.0," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[13] R. Pigan and M. Metter, *Automating with PROFINET: Industrial communication based on Industrial Ethernet.* John Wiley & Sons, 2008.

[14] D. Orfanus, R. Indergaard, G. Prytz, and T. Wien, "Ethercat-based platform for distributed control in high-performance industrial applications," in *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*, 2013, pp. 1–8.

[15] F. Zezulka, P. Marcon, Z. Bradac, J. Arm, T. Benesl, and I. Vesely, "Communication systems for Industry 4.0 and the IIoT," *IFAC-PapersOnLine*, vol. 51, no. 6, pp. 150–155, 2018.

[16] V. K. Huang, Z. Pang, C.-J. A. Chen, and K. F. Tsang, "New trends in the practical deployment of industrial wireless: From noncritical to critical use cases," *IEEE Industrial Electronics Magazine*, vol. 12, no. 2, pp. 50–58, 2018.

[17] M. Alabadi, A. Habbal, and X. Wei, "Industrial internet of things: Requirements, architecture, challenges, and future research directions," *IEEE Access*, vol. 10, pp. 66 374–66 400, 2022.

[18] L. Qiao, Y. Li, D. Chen, S. Serikawa, M. Guizani, and Z. Lv, "A survey on 5G/6G, AI, and robotics," *Computers and Electrical Engineering*, vol. 95, p. 107372, 2021.

[19] J. Ordonez-Lucena, J. F. Chavarria, L. M. Contreras, and A. Pastor, "The use of 5G Non-Public Networks to support Industry 4.0 scenarios," in *2019 IEEE*

Conference on Standards for Communications and Networking (CSCN), 2019, pp. 1–7.

[20] European Commission, "Industry and the Green Deal," (accessed June 2024). [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/industry-and-green-deal_en

[21] M. Attaran, "The impact of 5G on the evolution of intelligent automation and industry digitization," *Journal of ambient intelligence and humanized computing*, vol. 14, no. 5, pp. 5977–5993, 2023.

[22] A. Mahmood, S. F. Abedin, T. Sauter, M. Gidlund, and K. Landernäs, "Factory 5G: A review of industry-centric features and deployment options," *IEEE Industrial Electronics Magazine*, vol. 16, no. 2, pp. 24–34, 2022.

[23] M. Cheffena, "Propagation channel characteristics of industrial wireless sensor networks [wireless corner]," *IEEE Antennas and Propagation Magazine*, vol. 58, no. 1, pp. 66–73, 2016.

[24] E. A. Oyekanlu, A. C. Smith, W. P. Thomas, G. Mulroy, D. Hitesh, M. Ramsey, D. J. Kuhn, J. D. Mcghinnis, S. C. Buonavita, N. A. Looper *et al.*, "A review of recent advances in automated guided vehicle technologies: Integration challenges and research areas for 5G-based smart manufacturing applications," *IEEE Access*, vol. 8, pp. 202 312–202 353, 2020.

[25] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *2018 15th international symposium on wireless communication systems (ISWCS)*, 2018, pp. 1–6.

[26] M. Darabi, V. Jamali, L. Lampe, and R. Schober, "Hybrid puncturing and superposition scheme for joint scheduling of URLLC and eMBB traffic," *IEEE Communications Letters*, vol. 26, no. 5, pp. 1081–1085, 2022.

[27] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, 2018.

[28] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," in *2018 IEEE Symposium on Computers and Communications (ISCC)*, 2018, pp. 00 136–00 141.

[29] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1080–1093, 2020.

[30] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[31] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovács, and T. K. Madsen, "System level analysis of uplink grant-free transmission for URLLC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–6.

[32] C. Wang, Y. Chen, Y. Wu, and L. Zhang, "Performance evaluation of grant-free transmission for uplink URLLC services," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–6.

[33] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, 2019, pp. 607–612.

[34] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 741–755, 2020.

[35] A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. Da Silva, "Waveform and numerology to support 5G services and requirements," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 90–98, 2016.

[36] J. Flores de Valgas, J. F. Monserrat, and H. Arslan, "Flexible numerology in 5G NR: Interference quantification and proper selection depending on the scenario," *Mobile Information Systems*, vol. 2021, no. 1, p. 6651326, 2021.

[37] A. Hossain and N. Ansari, "5G multi-band numerology-based TDD RAN slicing for throughput and latency sensitive services," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1263–1274, 2023.

[38] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G new radio numerologies and their impact on the end-to-end latency," in *2018 IEEE 23rd Inter-*

national Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2018, pp. 1–6.

[39] S. Senk, S. A. W. Itting, J. Gabriel, C. Lehmann, T. Hoeschele, F. H. P. Fitzek, and M. Reisslein, "5G NSA and SA campus network testbeds for evaluating industrial automation," in *European Wireless 2021; 26th European Wireless Conference*, 2021, pp. 1–8.

[40] J. Rischke, P. Sossalla, S. Itting, F. H. P. Fitzek, and M. Reisslein, "5G campus networks: A first measurement study," *IEEE Access*, vol. 9, pp. 121 786–121 803, 2021.

[41] S. B. Damsgaard, D. Segura, M. F. Andersen, S. Aaberg Markussen, S. Barbera, I. Rodríguez, and P. Mogensen, "Commercial 5G NPN and PN deployment options for industrial manufacturing: An empirical study of performance and complexity tradeoffs," in *IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–7.

[42] I. Rodriguez, R. S. Mogensen, A. Fink, T. Raunholt, S. Markussen, P. H. Christensen, G. Berardinelli, P. Mogensen, C. Schou, and O. Madsen, "An experimental framework for 5G wireless system integration into industry 4.0 applications," *Energies*, vol. 14, no. 15, p. 4444, 2021.

[43] J. Ansari *et al.*, "Performance of 5G trials for industrial automation," *Electronics*, vol. 11, no. 3, p. 412, 2022.

[44] A. Fink, R. S. Mogensen, I. Rodriguez, T. Kolding, A. Karstensena, and G. Pocovi, "Empirical performance evaluation of enterprise Wi-Fi for IIoT applications requiring mobility," in *European Wireless 2021; 26th European Wireless Conference*, 2021, pp. 1–8.

[45] V. Sathya, L. Zhang, and M. Yavuz, "A comparative measurement study of commercial WLAN and 5G LAN systems," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, 2022, pp. 1–7.

[46] V. Sathya, L. Zhang, M. Goyal, and M. Yavuz, "Warehouse deployment: A comparative measurement study of commercial Wi-Fi and CBRS systems," in *2023 International Conference on Computing, Networking and Communications (ICNC)*, 2023, pp. 242–248.

[47] A. Emami, H. Frank, W. He, A. Bravalheri, A.-C. Nicolaescu, H. Li, H. Falaki, S. Yan, R. Nejabati, and D. Simeonidou, "Multi - RAT enhanced private wireless networks with intent-based network management automation," in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1789–1794.

[48] S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard, and D. Chandramouli, "5G multi-RAT multi-connectivity architecture," in *2016 IEEE International Conference on Communications Workshops (ICC)*, 2016, pp. 180–186.

[49] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen, and G. Berardinelli, "Reliability oriented dual connectivity for URLLC services in 5G New Radio," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2018, pp. 1–6.

[50] M. Agiwal, H. Kwon, S. Park, and H. Jin, "A survey on 4G-5G dual connectivity: Road to 5G implementation," *IEEE Access*, vol. 9, pp. 16 193–16 210, 2021.

[51] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis," *IEEE Network*, vol. 32, no. 2, pp. 32–40, 2018.

[52] A. Aijaz, "Packet duplication in dual connectivity enabled 5G wireless networks: Overview and challenges," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 20–28, 2019.

[53] E. J. Khatib, D. A. Wassie, G. Berardinelli, I. Rodriguez, and P. Mogensen, "Multi-connectivity for ultra-reliable communication in industrial scenarios," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, 2019, pp. 1–6.

[54] A. Alicke, J. Rachor, and A. Seyfert, "Supply chain 4.0–the next-generation digital supply chain, mckinsey & company," *Supply Chain Management June*, 2016.

[55] Y. Ding, M. Jin, S. Li, and D. Feng, "Smart logistics based on the internet of things technology: An overview," *Int. J. Logist. Res. Appl.*, vol. 24, no. 4, pp. 323–345, Apr. 2021.

[56] Y. Song, F. R. Yu, L. Zhou, X. Yang, and Z. He, "Applications of the internet of things (IoT) in smart logistics: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4250–4274, Mar. 2020.

[57] Z. Yang, R. Wang, D. Wu, H. Wang, H. Song, and X. Ma, "Local trajectory privacy protection in 5G enabled industrial intelligent logistics," *IEEE Trans. Ind. Informat.*, vol. 18, no. 4, pp. 2868–2876, Apr. 2022.

[58] G. Li, "Development of cold chain logistics transportation system based on 5G network and internet of things system," *Microprocess. Microsyst.*, vol. 80, p. 103565, Feb. 2021.

[59] J. M. Marquez-Barja, S. Hadiwardoyo, B. Lannoo, W. Vandenberghe, E. Kenis, L. Deckers, M. C. Campodonico, K. dos Santos, R. Kusumakar, M. Klepper, and J. Vandenbossche, "Enhanced teleoperated transport and logistics: A 5G cross-border use case," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC) & 6G Summit*, Jun. 2021, pp. 229–234.

[60] E. J. Khatib and R. Barco, "Optimization of 5G networks for smart logistics," *Energies*, vol. 14, no. 6, p. 1758, Mar. 2021.

[61] J. Zhan, S. Dong, and W. Hu, "IoE-supported smart logistics network communication with optimization and security," *Sustain. Energy Technol. Assess.*, vol. 52, p. 102052, Aug. 2022.

[62] S. Iranmanesh, F. S. Abkenar, R. Raad, and A. Jamalipour, "Improving throughput of 5G cellular networks via 3D placement optimization of logistics drones," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1448–1460, Feb. 2021.

[63] M. Savic, M. Lukic, D. Danilovic, Z. Bodroski, D. Bajović, I. Mezei, D. Vukobratovic, S. Skrbic, and D. Jakovetić, "Deep learning anomaly detection for cellular IoT with applications in smart logistics," *IEEE Access*, vol. 9, pp. 59 406–59 419, 2021.

[64] J. Cheng, Y. Yang, X. Zou, and Y. Zuo, "5G in manufacturing: a literature review and future research," *The International Journal of Advanced Manufacturing Technology*, pp. 1–23, 2022.

[65] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, "URLLC and eMBB in 5G Industrial IoT: A survey," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1134–1163, 2022.

[66] S. Zhang, "An overview of network slicing for 5G," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.

[67] Y. Wu, H.-N. Dai, H. Wang, Z. Xiong, and S. Guo, "A survey of intelligent network slicing management for industrial IoT: Integrated approaches for smart transportation, smart energy, and smart factory," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1175–1211, 2022.

[68] T. Umagiliya, S. Wijethilaka, C. De Alwis, P. Porambage, and M. Liyanage, "Network slicing strategies for smart industry applications," in *2021 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2021, pp. 30–35.

[69] A. Hoglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.

[70] *Evaluation for early data transmissions*, TSG-RAN WG2 #100, document R2-1713058, 3GPP, Nov. 2017.

[71] O. Liberg, J. Bergman, A. Höglund, T. Khan, G. A. Medina-Acosta, H. Rydén, A. Ratilainen, D. Sandberg, Y. Sui, T. Tirronen, and Y. P. E. Wang, "Narrowband internet of things 5G performance," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, doi: 10.1109/VTCFall.2019.8891588.

[72] F. J. Dian and R. Vahidnia, "A simplistic view on latency of random access in cellular internet of things," in *Proc. 11th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. (IEMCON)*, Nov. 2020, pp. 0391–0395.

[73] R. Barbau, V. Deslandes, G. Jakllari, and A.-L. Beylot, "An analytical model for evaluating the interplay between capacity and energy efficiency in NB-IoT," in *Proc. Int. Conf. on Comput. Commun. and Netw. (ICCCN)*, Jul. 2021, doi: 10.1109/ICCCN52240.2021.9522178.

[74] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.5.0, 2023.

[75] 3GPP TR 36.913, "LTE; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced) (Release 10)," 3rd Generation Partnership Project, Tech. Rep. V10.0.0, 2011.

[76] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.9.0, 2023.

[77] 3GPP TS 36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.4.0, 2023.

[78] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2nd Quart., 2020.

[79] 3GPP TR 21.915, "Release 15 Description; Summary of Rel-15 Work Items (Release 15)," 3rd Generation Partnership Project, Tech. Rep. V15.0.0, 2019.

[80] 3GPP TS 38.300, "NR; NR and NG-RAN Overall Description; Stage 2 (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.5.0, 2023.

[81] 3GPP TS 23.501, "5G; System architecture for the 5G System (5GS) (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.10.0, 2023.

[82] 3GPP TS 38.104, "5G; NR; Base Station (BS) radio transmission and reception (Release 15)," 3rd Generation Partnership Project, Tech. Rep. V15.19.0, 2023.

[83] 3GPP TS 22.368, "Service Requirements for Machine-Type Communications (MTC); Stage 1 (Release 13)," 3rd Generation Partnership Project, Tech. Rep. V13.2.0, 2016.

[84] TSG RAN Meeting 86, "New SID Support Reduced Capability NR Devices," 3rd Generation Partnership Project, Tech. Rep. RP-193238, 2019.

[85] S. R. Borkar, "Long-term evolution for machines (LTE-M)," in *LPWAN technologies for IoT and M2M applications*. Elsevier, 2020, pp. 145–166.

[86] 3GPP TR 21.914, "Release 14 Description; Summary of Rel-14 Work Items (Release 14)," 3rd Generation Partnership Project, Tech. Rep. V14.0.0, 2018.

[87] 3GPP TR 21.917, "Release 17 Description; Summary of Rel-17 Work Items (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.0.1, 2023.

[88] G. Medina-Acosta, L. Zhang, J. Chen, K. Uesaka, Y. Wang, O. Lundqvist, and J. Bergman, "3GPP Release-17 physical layer enhancements for LTE-M and NB-IoT," *IEEE Communications Standards Magazine*, vol. 6, no. 4, pp. 80–86, 2022.

[89] M. Chen, Y. Miao, Y. Hao, and K. Hwang, "Narrow band internet of things," *IEEE Access*, vol. 5, pp. 20 557–20 577, 2017.

[90] M. Kanj, V. Savaux, and M. Le Guen, "A tutorial on NB-IoT physical layer design," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 4, pp. 2408–2446, 2020.

[91] 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 16)," 3rd Generation Partnership Project, Tech. Rep. V16.6.0, 2021.

[92] 3GPP TR 38.913, "5G; Study on scenarios and requirements for next generation access technologies (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.0.0, 2022.

[93] 3GPP TS 23.682, "Architecture enhancements to facilitate communications with packet data networks and applications (Release 16)," 3rd Generation Partnership Project, Tech. Rep. V16.10.0, 2021.

[94] 3GPP TS 24.301, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (Release 16)," 3rd Generation Partnership Project, Tech. Rep. V16.8.0, 2021.

[95] GSMA, "NB-IoT deployment guide to basic feature set requirements," Groupe Speciale Mobile Association (GSMA), Tech. Rep., 2019, (accessed June 2024). [Online]. Available: https://www.gsma.com/iot/wp-content/uploads/2019/07/201906-GSMA-NB-IoT-Deployment-Guide-v3.pdf

[96] ——, "LTE-M deployment guide to basic feature set requirements," Groupe Speciale Mobile Association (GSMA), Tech. Rep., 2019, (accessed June 2024). [Online]. Available: https://www.gsma.com/iot/wp-content/uploads/2019/08/201906-GSMA-LTE-M-Deployment-Guide-v3.pdf

[97] 3GPP TS 36.321, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 14)," 3rd Generation Partnership Project, Tech. Rep. V14.9.0, 2019.

[98] C. Rosa, K. Pedersen, H. Wang, P.-H. Michaelsen, S. Barbera, E. Malkamäki, T. Henttonen, and B. Sébire, "Dual connectivity for LTE small cell evolution: functionality and performance aspects," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 137–143, 2016.

[99] C. Pupiales, D. Laselva, Q. De Coninck, A. Jain, and I. Demirkol, "Multi-connectivity in mobile networks: Challenges and benefits," *IEEE Communications Magazine*, vol. 59, no. 11, pp. 116–122, 2021.

[100] 3GPP TS 36.323, "Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.2.0, 2023.

[101] 3GPP TS 33.501, "Security architecture and procedures for 5G System (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.11.1, 2023.

[102] J. Munilla, A. Hassan, and M. Burmester, "5G-compliant authentication protocol for RFID," *Electronics*, vol. 9, no. 11, p. 1951, 2020.

[103] J. Arkko, V. Lehtovirta, and P. Eronen, "Improved Extensible Authentication Protocol Method for 3rd Generation Authentication and Key Agreement (EAP-AKA')," IETF RFC 5448, May 2009, (accessed June 2024). [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5448

[104] B. Karakoc, N. Fürste, D. Rupprecht, and K. Kohls, "Never let me down again: Bidding-down attacks and mitigations in 5G and 4G," in *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WiSec '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 97–108. [Online]. Available: https://doi.org/10.1145/3558482.3581774

[105] 3GPP TS 38.213, "5G; NR; Physical layer procedures for control (Release 17)," 3rd Generation Partnership Project, Tech. Rep. V17.7.0, 2023.

[106] D. Dolev and A. Yao, "On the security of public key protocols," *IEEE Trans. Inf. Theory*, vol. 29, no. 2, pp. 198–208, Mar. 1983.

[107] C. Yu, S. Chen, F. Wang, and Z. Wei, "Improving 4G/5G air interface security: A survey of existing attacks on different LTE layers," *Computer Networks*, vol. 201, p. 108532, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128621004576

[108] "NS-3-A Discrete-Event Network Simulator for Internet Systems," https://www.nsnam.org/, (accessed June 2024).

[109] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simul. Model. Pract. Theory*, vol. 96, Nov. 2019, Art. no. 101933.

[110] *Study on Channel Model for Frequencies from 0.5 to 100 GHz*, document TR 38.901, V17.1.0, 3GPP, Jan. 2024.

[111] "5G-simulator: Extended 5G-simulator based on NS-3 and 5G-LENA," https://github.com/dsr96/5g-simulator, (accessed June 2024).

[112] "Amarisoft - AMARI Callbox Classic," https://www.amarisoft.com/test-and-measurement/device-testing/device-products/amari-callbox-classic, (accessed June 2024).

[113] "Amarisoft - AMARI UE Simbox," https://www.amarisoft.com/test-and-measurement/network-testing/network-products/amari-ue-simbox-e-series, (accessed June 2024).

[114] I. Rodriguez *et al.*, "5G swarm production: Advanced industrial manufacturing concepts enabled by wireless automation," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 48–54, 2021.

[115] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[116] G. Hackeling, *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.

[117] E. Bisong and E. Bisong, "Introduction to scikit-learn," *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pp. 215–229, 2019.

[118] W. McKinney *et al.*, "Pandas: a foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.

[119] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[120] "RA-simulator - A random-access channel simulator for cellular networks," https://github.com/dsr96/ra-simulator, (accessed June 2024).

[121] 3GPP TS 38.211, "NR; Physical channels and modulation," 3rd Generation Partnership Project, Tech. Rep. V16.10.0, 2022.

[122] 3GPP TS 38.321, "NR; Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project, Tech. Rep. V16.10.0, 2022.

[123] 3GPP TS 38.331, "NR; Radio Resource Control (RRC); Protocol specification," 3rd Generation Partnership Project, Tech. Rep. V16.10.0, 2022.

[124] Intel NUC Kit NUC5i3MYHE. (accessed June 2024). [Online]. Available: https://www.intel.co.uk/content/www/uk/en/products/sku/84860/intel-nuc-kit-nuc5i3myhe/specifications.html

[125] Simcom SIM8202G-M2. (accessed June 2024). [Online]. Available: https://www.simcom.com/product/SIM8202X_M2.html

[126] MiR200 Data Sheet. (accessed June 2024). [Online]. Available: https://www.ics-id.de/mir.html?file=files/ics-id.de/downloads/MIR200/Technische%20Daten%20MiR%20200%20%28EN%29.pdf

[127] Mpconn - The open source multi-path connectivity tool. (accessed June 2024). [Online]. Available: https://github.com/drblah/mpconn

[128] C. Baena, O. S. Peñaherrera-Pulla, L. Camacho, R. Barco, and S. Fortes, "Video streaming and cloud gaming services over 4G and 5G: A complete network and service metrics dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 154–160, 2023.