FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

5G SA Private Networks Design

Daniel Girão Pereira

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Professor Manuel Pereira Ricardo

July 31, 2023

Abstract

Ever since the introduction of the first generation of mobile communications in the 1980s, there is an increasing evolution of mobile communications in terms of access capabilities and functionalities. Research and development in this area during the last decades gave us 2G, 3G, 4G, 5G. 5G is the last generation of mobile communications already under deployment and commercially available. One of the biggest selling points of this technology is the provisioning of higher data rates, lower latencies and a higher device density, than prior generations. Mobile technologies rely on communications and computing infrastructure whose architecture is defined by standard organizations.

5G technology was also envisaged to support industrial use cases, that are critical for daily operations and integration with other subsystems. 5G ensures that these subsystems can communicate with each other with low latency, high reliability and with sizable amount of security. 5G Private Networks will provide these characteristics to guarantee that an industrial use case is efficient and has an high productivity.

The main goal of this dissertation is to identify 5G topologies that can support the requirements of relevant 5G industrial use cases. In particular, the main achievement of this dissertation was to assess a specific 5G SA Private Network for an Automated Guided Vehicles (AGV) use case, and conclude how did it support the use case, and suggest changes to the network topology of this network.

Resumo

Desde do início da comunicações móveis que tem havido uma evolução exponencial da mesma em termos de funcionalidade, capacidade e acessibilidade. Durante estas últimas décadas, a investigação e o desenvolvimento produziram a 2ª, 3ª, 4ª, 5ª, e nos próximos anos, também a 6ª geração de comunicações móveis será uma realidade. A 5ª geração ou o 5G é atualmente a tecnologia mais recente que está pronta para o mercado. Existe um enorme interesse por esta geração de comunicações móveis devido ao aumento que ela proporciona em termos de débitos de dados, diminuição de latências, e aumento de densidade de dispositivos, isto comparando com gerações anteriores. Esta tecnologia tem na sua génese uma arquitetura e infraestrutura base que é normalizada por entidades globais.

Um dos principais alvos de utilização do 5G é a indústria mas não descurando os possíveis futuros requisitos de um utilizador comum. O 5G entra no mercado como um possível catalizador de diversas operações industriais. O seu propósito é suportar a comunicação entre diversos sistemas, e garantir que essa comunicação é feita de forma resiliente, entre outras características. As redes privadas 5G vão permitir fornecer uma maior eficiência e produtividade às diversas operações feitas nas diferentes indústrias.

O principal objetivo desta tese é identificar um conjunto de topologias e arquiteturas de rede 5G que suportem os requisitos de aplicações industriais. A principal contribuição desta dissertação foi a análise de uma rede privada 5G relativamente ao suporte de um caso de uso específico de "Automated Guided Vehicles" (AGV). São retiradas conclusões sobre o a adequabilidade da topologia de rede escolhida ao caso de uso selecionado. E, em consequência dos resultados obtidos são sugeridas alterações à topologia usada para esta rede 5G.

Acknowledgements

I would like to thank my supervisors, Professor Manuel Ricardo and Engineer Ricardo Santos, for all of the support and commitment during this dissertation. Furthermore, to all the people in Deloitte gTEE team, I want to extend my deepest gratitude for the opportunity and all the knowledge and resources that you gave me to do this work, and I am sure that I will use this knowledge for the rest of my journey as engineer. A special shoutout to Válter Dias, Ericson Bittencourt, Daniel Rodrigues, André Laranjeira, and João Pinto for the support you gave me during the integration of the gTEE's Testbed.

I would also like to thank my friends and colleagues with whom I shared the journey through this last year at FEUP. To Queirós, Rodrigo, Catarina and Gabriella, thank you for keeping it fun.

To my friends and family, I would like to thank for the huge support you gave me throughout my life, and for believing in me.

Daniel Pereira

"Never trust a computer you can't throw out a window."

Steve Wozniak

Contents

List of Figures

List of Tables

Abbreviations

3GPP	Third-Generation Partnership Project
3G	Third Generation
4G	Forth Generation
5G	Fifth Generation
5GC	5G Core
5G CN	5G Core Network
ACP	Airspan Control Platform
AGV	Automated-Guided Vehicles
AMF	Access and Mobility Management Function
API	Application Programming Interface
ARQ	Automatic Repeat Request
ARP	Address Resolution Protocol
AUSF	Authentication Server Function
AWS	Amazon Web Services
BBU	Band Base Unit
BIOS	Basic Input/Output System
C-MTC	Critical Machine-Type Communications
CN	Core Network
COTS	Commercial of-the-shelf
CP	Control-Plane
CPRI	Common Public Radio Interface
CU	Central Unit
DHCP	Dynamic Host Configuration Protocol
DL	Downlink
DN	Data Network
DU	Distributed Unit
eCPRI	evolved Common Public Radio Interface
eMBB	Enhanced Mobile Broadband
eNB	eNodeB
EPC	Evolved Packet Core
F1AP	F1 Application Protocol
F1-c	F1 Control-Plane
F1-u	F1 User-Plane
gNB	gNodeB
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GTP	GPRS Tunnelling Protocol
GTP-U	GTP User-Plane
IEEE	Institute Of Electrical and Electronics Engineers
IP	Internet Protocol
IMT	Internet Mobile Telecommunications
IQ	In-phase and Quadrature
ITU-T	International Telecommunication Union - Telecommunication

ITU-R	International Telecommunication Union - Radiocommunications
k8s	Kubernetes
KPI	Key Performance Indicator
MAC	Medium Access Control
MEC	Multi-access Edge Computing
MIMO	Multiple-Input Multiple-Output
mMTC	Massive Machine Type Communications
MNO	Mobile Network Operator
NAS	Non-Access Stratum
NAT	Network Address Translation
NEF	Network Exposure Function
NGAP	NG Application Protocol
NG-c	NG Control-Plane
NG-u	NG User-Plane
NR	New Radio
NRF	Network Repository Function
OAM	Orchestration and Management
O-CU	Open-CU
O-DU	Open-DU
O-RU	Open-RU
OS	Operating System
PCF	Policy Control Function
PDU	Protocol Data Unit
PDCP	Packet Data Convergence Protocol
PHY	Physical
PLMN	Public Land Mobile Network
PRB	Physical Resource Block
PTP	Precision Time Protocol
QAM	Quadrature Amplitude Modulation
QFI	Qos Flow Identifier
QoS	Quality-of-Service
RAN	Radio Access Network
RAT	Radio A ccess Technology
RF	Radio Frequency
RIC	RAN Intelligent Controllers
RLC	Radio-Link Control
RRC	Radio Resource Control
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RU	Radio Unit
SA	Standalone
SCTP	Stream Control Transmission Protocol
SDAP	Service Data Application Protocol
SINR	Signal to Interference plus Noise Ratio
SMF	Session Management Function
SQL	Structure Query Language
SSH	Secure Shell
syncE	Synchronization Ethernet
TLS	Transport Layer Security
UDM	Unified Data Management
UDP	User Datagram Protocol
UDR	Unified Data Repository

UE	User Equipment
UL	Uplink
UP	User-Plane
UPF	User-Plane Function
uRLLC	Ultra-Reliable Low Latency Communications
URSP	Universal Software Radio Peripherals
VLAN	Virtual LAN
VM	Virtual Machine
VPN	Virtual Private Network
XnAP	Xn Application Protocol
Xn-c	Xn Control-Plane
Xn-u	Xn User-Plane

Chapter 1

Introduction

1.1 Context

The fifth generation (5G) of mobile communications began to be discussed around 2012 and it is being deployed currently. 5G aims to respond to the demands of the end-users which have specific communications requirements for data rates, latencies, coverage, and device density. These requirements were derived from a set of use cases, grouped into 3 main types: Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (uRLLC), Massive Machine Type Communications (mMTC).

These use cases require capabilities from the network that the prior generation of mobile communications (4G) did not provide, except for the eMBB use case, in a certain measure. Furthermore, 5G introduces New Radio (NR) access that provides new developments in the radio air interface such as improved Multiple-Input Multiple-Output (MIMO) capabilities by enabling more antennas to be used in a base station, interference mitigation techniques, and increased available bandwidth. NR brings a flexible scheme for managing time slots and frequency sub-carriers which enables the device and base stations to achieve the required quality of service.

Mobile networks have an architecture which is defined by standards. 5G, in particular, is formed by a Radio Access Network (RAN) part and a Core Network (CN) part. Each network segment is composed of standard functional blocks that are normalized by several organizations including Third-Generation Partnership Project (3GPP). Currently, most of the 5G functional blocks are virtual, can run on ordinary computers, and be deployed anywhere.

1.2 Problem and Motivation

Nowadays, several industries are developing new services and applications such as Augmented Reality, Industry Automation, and Automated Guided Vehicles (AGV). These applications demand specific requirements in terms of data traffic, service availability, and resilience of the network. 5G aims to provide the access network infrastructure for these applications. Moreover, at the beginning of 5G, industries presented performance demands from the network which were associated to the three use cases mentioned in the previous section. Currently, there are standards for 5G in which architecture, functions, interfaces, and other performance requirements will be the baseline for the deployment of Private Networks that will support the industries' digital transformation to increase business revenue and operational efficiency targets.

In this current phase, it is essential that 5G topologies can be created to provide the quality of service demanded by each use case. It is imperative that tests can be made for evaluating different topologies of networks and functions and draw benchmarks of each topology to cross reference with the requirements defined by the standards organizations.

1.3 Objectives

The main goal of this dissertation is to achieve the end-to-end solution architecture of 5G functional blocks, in the specific domain of Private Standalone 5G networks, that can satisfy the performance requirements of 5G use cases. These deployments will be tested, simulated and recommendations will be provided for an easy selection of the best architecture for a given set of use cases. This work will follow these streams:

- Identify technical requirements and aggregate use cases;
- Set the key performance indicators (KPIs) for each representative usage scenario;
- Run test scenarios in a 5G network, in order to evaluate the performance requirements of the use cases;
- Establish the end-to-end architecture to meet the requirements of the use cases.

By developing this work, we aim to identify a path for defining efficient end-to-end Private Standalone architectures. We also intend to create standards to select the best-profited architecture considering a specific industry use case.

We do not have time in this thesis to evaluate a large number of use scenarios and the focus will be in finding the most adequate 5G topology for an AGV use case by assessing the use case's performance via testing in a 5G laboratory testbed, which has a pre-established 5G architecture/topology. To address other use scenarios, we will extrapolate from this testing different conclusions that can help in the design of 5G networks for other use cases.

1.4 Document Structure

The rest of this document is structured in eight chapters. Chapter 2 describes the background and fundamental aspects of 5G. Chapter 3 presents and discusses articles exposing solutions related to our challenges. Chapter 4 focuses in the architecture, specification and implementation that was done to finalize the 5G testbed that was used to test the AGV use case. Chapter 5 presents the methodology behind the different test scenarios that were made to evaluate the 5G architecture, and also explains how the use case was implemented. Chapter 6 presents the results that were achieved following the testing phase. Chapter 7 evaluates the 5G testbed as the correct (or not) 5G deployment for the AGV use case and others. Finally, Chapter 8 presents conclusions of the work, and also addresses key-points that may be explored in the context of future work.

Chapter 2

5G Fundamental Aspects

This chapter covers the topics considered relevant to understand the design of 5G. It is divided into the following sections:

- 5G Architecture - the main entities that composed the 5G architecture, and the protocol stack used for radio access;
- Open-RAN (O-RAN) - description of the latest RAN architecture;
- 5G use cases - description of the use cases defined by the 5G standards organizations. Exposition of the main KPIs that are used to define the performance of 5G use cases. Exposition of the minimum performance requirements by the standards organization;
- Interfaces - description of the interfaces that can be used in a 5G end-to-end architecture between different functional entities.

2.1 5G Architecture

The overall system architecture for 5G can be divided into two main parts, the RAN and the CN. The RAN is responsible for radio-related functions. The CN is responsible for non-radio access and control functions such as authentication and setup of end-to-end communications. Having these functions in the CN and not in the RAN is beneficial since it allows several RANs to be served by the same CN. The RAN connects several user equipments (UEs) and devices to the CN and it is the main element of the 5G infrastructure. It is formed by the gNodeB (gNB), which is a logical node that is responsible for all radio-related functions. Since this dissertation is focused on Standalone (SA) 5G networks, it will focus exclusively on the gNB and the 5G CN.

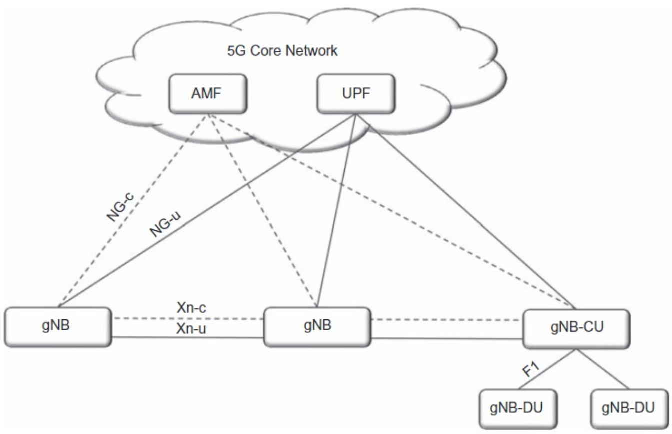


Figure 2.1: Overall 5G Architecture [1]

2.1.1 5G RAN - Radio Access Network

The RAN is represented by the logical node gNB, which is a logical node and not a physical implementation. The gNB can be split into two entities, a central unit (CU) and one or more distributed units (DUs).

The RAN is responsible for the radio-related functions. Firstly, for establishing a connection between a UE and the Core, and secondly for transporting data between the UE and the Core, or between the UE and another UE.

The gNB serves UEs through user-plane (UP) and control-plane (CP) protocols. Each gNB has implemented the RAN protocol stack which can be divided into these two planes, the CP, and UP. Both planes share the following protocols: Packet Data Convergence Protocol (PDCP), Radio-Link Control (RLC), Medium Access Control (MAC), and Physical Layer (PHY). In terms of specific protocols for each plane, the UP has the Service Data Application Protocol (SDAP), and the CP includes the Radio Resource Control (RRC) protocol, as shown in Fig. 2.2.

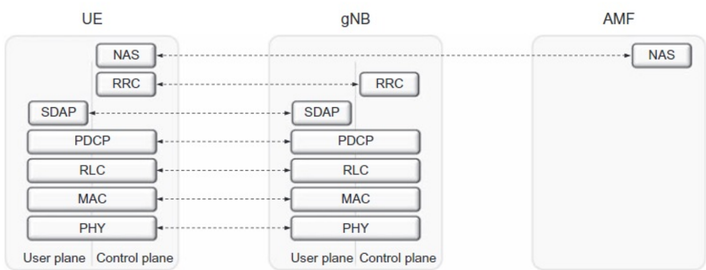


Figure 2.2: Radio Protocol Stack [1]

The Non-Access Stratum (NAS) functionality is responsible for the establishment of communication between the UE and the Access and Mobility Management Function (AMF) of the CN.

The RRC protocol is responsible for handling the RAN-related CP procedures. These can include the broadcasting of system information to the devices, so they can communicate with a cell, managing the connection, including the setup of bearers and mobility, and also handling the device capabilities [1].

The SDAP protocol is responsible for the mapping between a Quality-of-Service (QoS) flow from the 5G Core (5GC) and a data radio bearer. It is also responsible for marking the QoS flow identifier (QFI) in both uplink (UL) and downlink (DL) packets [15].

The handling of QoS control is done in the 5G CN, and as so, the RAN does not have knowledge of this service. The RAN receives from the CN a QoS flow having specific requirements, that come in the QFI, such as priority, packet error ratio, and other information. With this knowledge, the SDAP protocol allocates a radio bearer to properly serve this flow of data [1].

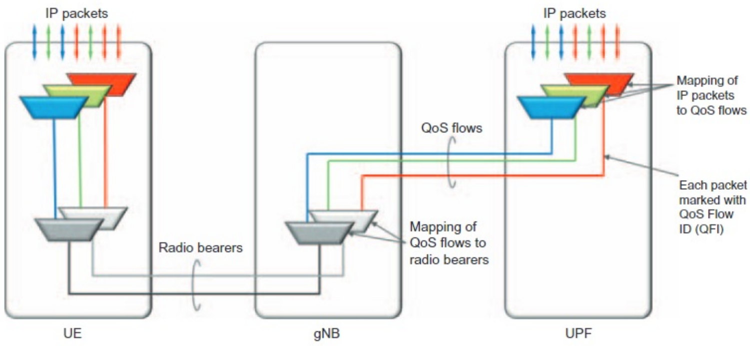


Figure 2.3: SDAP's QoS handling [1]

The PDCP protocol is responsible for internet protocol (IP) header compression, ciphering, integrity protection and verification, retransmissions, duplicate removal, and in-sequence delivery [15]. The PDCP protocol performs IP header compression to reduce the number of bits being transported over the radio interface. The ciphering is done to guarantee that no one can eavesdrop, and integrity protection ensures that the data comes from a reliable source. The PDCP can also perform retransmissions, duplicate removal, and in-sequence delivery. The UE will use the PDCP to handle retransmissions to the gNB of packets not yet delivered. In this case, a gNB may receive duplicates, and the PDCP entity in the gNB will remove these duplicates. In terms of in-sequence delivery, the PDCP can perform reordering to ensure in-sequence delivery to higher-layer protocols.

RLC protocol is responsible for segmentation and retransmission handling. It is responsible for segmenting the payload received by the PDCP into suitably sized RLC frames and reassembling the segments when delivering again to the PDCP protocol. It will handle the retransmission of packets that arrived with errors, different from PDCP. In this way, it guarantees error-free delivery of data to higher layers. RLC can be configured in one of three modes. The transparent mode does not add any headers. The unacknowledge mode supports segmentation and duplicate detection. The acknowledge mode supports, in addition, the retransmission of erroneous packets [1].

The MAC protocol is responsible for handling logical-channel multiplexing, hybrid-Automatic Repeat Request (ARQ) retransmission, and scheduling (control the assignment of time-frequency resources to users) [1].

The RLC protocol sends its frames to the MAC protocol. These frames are sent through different logical channels. The type of information that is present in the RLC frame will dictate the type of logical channel to be used. The logical channels are mainly of two types: control and traffic. Control channels are characterized by having control information, and the traffic channel is characterized by transporting user data. The type of logical channel will be signalized in the MAC header.

The frame encapsulated by the MAC protocol is sent through a transport channel to the Physical (PHY) layer. A transport channel is defined by how and with what characteristics the information is transmitted over the radio interface [1]. The data in a transport channel is organized in transport blocks. The MAC protocol is responsible for the mapping of logical channels into transport channels.

The PHY layer is responsible for coding, modulation, multi-antenna processing, and mapping of the signal to the appropriate physical time-frequency resources [1]. The physical layer will analyze the transport blocks in the transport channel to see how to transmit the data over the radio interface (transport block size, modulation, coding scheme, antenna mapping). The PHY layer will map the transport channels to physical channels. A physical channel corresponds to the set of time-frequency resources used for the transmission of a particular transport channel. Each transport channel is mapped into a physical channel.

2.1.2 5G Core Network

The 5GC is responsible for giving the end-user devices reliable and secure access to its services and the 5G network. The Core is responsible for a wide variety of functions in the 5G network, such as connectivity and mobility management, authentication and authorization, subscriber data management, policy management, and many more [16].

The 5GC was developed upon the existing 4G Core but with enhancements. It has a service-based architecture, which is formed by services and functionalities, rather than nodes. In this way, the services and functionalities are virtualized and can run on commercial-off-the-shelf (COTS) servers.

As in the gNB, so the CN supports the CP/UP split. With this split, the Core has the flexibility to scale both planes in different manners. So that scaling the CP does not influence the scaling of the UP.

The main user-plane function is the User-Plane Function (UPF) which acts as a gateway between the RAN and external networks. It is through the UPF that the user data will be forwarded to the correct destination. The CP main services provided by the CN are the AMF and the Session Management Function (SMF). Apart from AMF and SMF, the CP has more services.

The CN can be implemented in a single physical node or distributed across multiple nodes. It can be deployed on-site, on a cloud platform, or part on-site and part on a cloud platform (hybrid approach).

Fig. 2.4 presents a high-level 5G Core architecture.

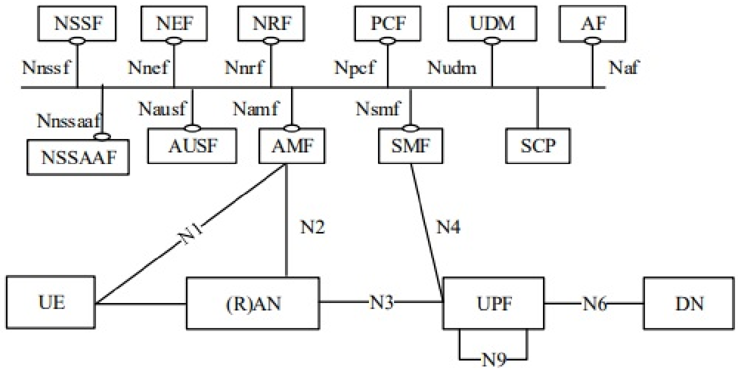


Figure 2.4: 5G Core Architecture [2]

The AMF network function is connected to the UE, via the N1 interface (NAS interface), and to the RAN (control-plane (CP)), via the N2 interface. The AMF function will work as the middle-man communicator between a UE and the Core CP, and the same for the RAN. The AMF function is in charge of registration management, connection management, and mobility management.

In terms of registration management, the AMF is in charge of registering a device in the network and establishing the user context in the network. According to [2], some services are only accessible through registration. Also, the initial registration involves the execution of user authentication and access authorization based on subscription profiles in the Unified Data Management (UDM).

Connection management is important for establishing and releasing the NAS signaling connection between the UE and the AMF over the N1 interface. This connection also includes the signaling connection between the UE and the RAN.

In terms of mobility management, the AMF is in charge of handling mobility restrictions, radio resources, and the UE mobility event notification. The AMF can restrict mobility handling or service access of a UE [2], by applying mobility restrictions. An example of mobility restrictions can be prohibiting a UE to access the network via a specific radio access technology (RAT), or prohibiting a UE from communicating with the network in a specific area. When the AMF is managing connection requests from users, these restrictions need to be assessed. Furthermore, for radio resource management, the AMF can control the frequency bands and RATs that the UEs are using to communicate with the network [2]. The UE mobility event notification is a subscription-based service provided by the AMF, that allows other network function services (SMF or the other services we are going to see) to know a specific UE location.

The SMF function is responsible for session management, IP address allocation and management, and traffic steering at the UPF. It also has a dynamic host configuration protocol (DHCP) and address resolution protocol (ARP) function and is the termination of interfaces towards policy control functions. Session management includes session establishment, modification, and release, including tunnel maintenance between UPF and RAN node [2]. The SMF provides a protocol data unit (PDU) connectivity service, which supports the exchange of PDUs between a UE and a Data Network (DN). This is supported by PDU sessions that are established upon UE request [2]. So a PDU session is the establishment of connectivity between a UE and a DN, for the exchange of user data. The PDU session will be deployed between the UE, the RAN node, the UPF, and the DN.

Each PDU session supports a single PDU session type [2]. It can only support the exchange of a singular type of PDUs. They are defined by the version of IP that they support: IPv4, IPv6, and IPv4v6. The IP address that is allocated to the UE can depend on several details. It can depend on the IP stack capabilities of the UE, the DN's IP configuration, or the type of PDU session that the UE specifies in its request. This IP address allocation will be done via DHCP messages exchanged between the UE and the SMF service (v4 or v6 depending on the PDU session type).

In terms of ARP functionality, the SMF can respond to ARP requests and or IPv6 Neighbour Solicitation requests.

The SMF will also be responsible for traffic steering at the UPF to route traffic to the proper destination. When creating the PDU sessions it will give the destination IP address to the UPF. In this case, the SMF will work as the CP of a router, it will configure the routing table that the UPF will use to forward the user data to the proper destination. The routing table created by the SMF service can be influenced by any policy created by a network service in the Core, that is sent to the SMF. This is one of the reasons why it is said that the SMF is the termination of interfaces toward policy control functions.

The UPF is the main network function service of the CN in the UP. The UPF is the gateway between the RAN and external DNs. The UPF behaves as an anchor point for different RANs. It is also responsible for traffic routing and forwarding. Other important functions of the UPF include packet inspection, QoS handling, and ARP proxy [2].

The Policy Control Function (PCF) supports a unified policy framework to manage network behavior. It provides policies for the CP functions (SMF, for example) to enforce them. The Network Exposure Function (NEF) is responsible for exposing other network functions' capabilities and events to third-party applications and Edge Computing. These applications will interact with the Core Network and provide services through the NEF. The NEF will retrieve information about other network functions through the Network Repository Function (NRF), which maintains each network function profile with its supported services. All procedures involving user authentication, and access authorization, will be handled by the Unified Data Management (UDM) network function, which is responsible for the generation of authentication credentials, user identification handling, and access authorization based on subscription data. The UDM will retrieve and store subscription data in the Unified Data Repository (UDR) network function. The Authentication

Server Function (AUSF) will be responsible for giving access to the network. In conclusion, the main control network functions (AMF and SMF) need to be in constant communication with the other network functions of the Core Network, if they want to provide their services.

2.2 O-RAN

The most recent Radio-Access Network developed was the Open-RAN (O-RAN), standardized by the O-RAN Alliance organization in 2018.

According to [3], O-RAN deploys a solution that is based on disaggregated, virtualized, and software-based components, connected through open and standardized interfaces, which are fundamental for having interoperability between different vendors' equipment (white-box development). In this way, O-RAN abandons all-in-one solutions (monolithic units, black-box development) that implement each and every layer of the radio protocol stack in a static manner. With this open solution, we can configure the RAN with different control policies and goals in terms of requirements, and split the radio functionality through different locations and hardware platforms.

O-RAN splits the gNB into three nodes which are the open-CU (O-CU), the open-DU (O-DU), and the open-RU (O-RU). These nodes are interconnected standard interfaces (fronthaul, midhaul) that will be described in the following section. Furthermore, they are also connected to RAN Intelligent Controllers (RICs) that perform management and control of the RAN.

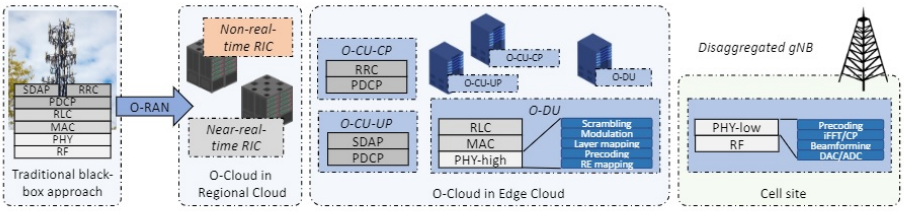


Figure 2.5: Evolution from traditional RAN to the O-RAN approach [3]

As was already mentioned, the RAN splits the gNB into different functional units. The gNB is split into an O-CU, O-DU, and an O-RU. Another important detail is that the O-CU can be further split. It is possible to have two logical components of the O-CU, one for the CP and another for the UP. The split used for decoupling the radio protocol stack between the O-RU and O-DU is the 7.2x split, standardized by the O-RAN Alliance [6]. The PHY functions are distributed between the O-RU and the O-DU. The O-DU will also implement the MAC and RLC protocols. The O-CU will implement the PDCP protocol and the SDAP and RRC protocol in their respective plane.

O-RAN brings programmable components that can run optimization routines with closed-loop control and orchestrate the RAN. Both RICs (near-real-time and non-real-time) process data (key performance measurements) to determine and apply control policies and actions on the RAN.

O-RAN enables virtualization, so it is possible to decouple the software from the hardware. Radio-related functions can be deployed in commercial-off-the-shelf (COTS) servers. Furthermore, by encouraging virtualization and distributing the RAN functionalities between different hardware, it will demand the creation of open standardized interfaces to connect the different components of the O-RAN architecture.

2.3 5G Use Cases

The International Telecommunication Union - Radiocommunications (ITU-R), in its Internet Mobile Telecommunications (IMT)-2020 recommendation [4], defined three main cases that aggregate the majority of 5G usage scenarios:

- Enhanced Mobile Broadband (eMBB): Continues the work in Mobile Broadband Services made in prior generations (3G and 4G). It is an important use case since the demand for these types of services is increasing as more people have access to technology and more demanding applications are being created. This type of use case imposes high data rates and capacity in a 5G network to keep up with large volumes of traffic.

- Ultra-Reliable Low Latency Communications (uRLLC): This use case is related to both human and machine-centric communications (critical machine type communications - C-MTC). It is characterized by very low latencies and extremely high reliability and availability.

- Massive Machine-Type Communications (mMTC): This use case is a pure machine-centric use case. It is characterized by a very large number of connected devices (high connection density) that typically transmit small amounts of data that are not delay-sensitive. These devices tend to be low-cost and powered by batteries.

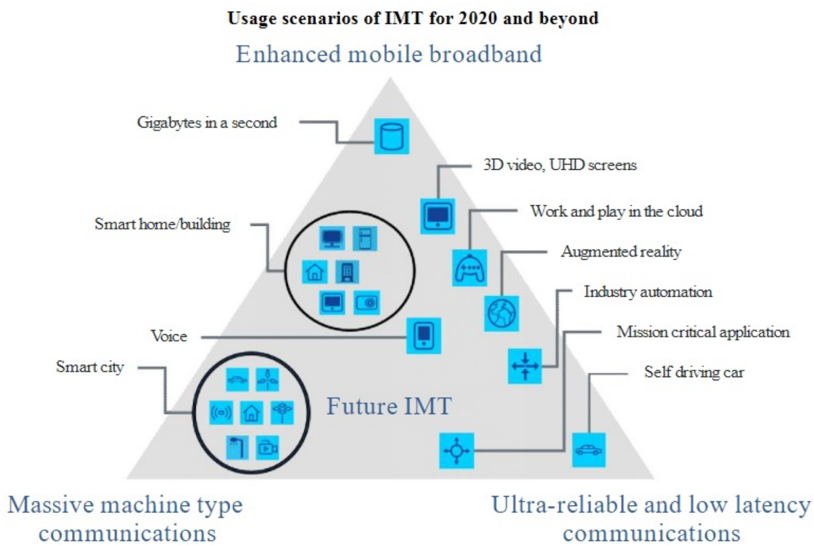


Figure 2.6: 5G Use Cases [4]

This categorization of 5G in three types of use cases is not static. It is possible that a use case will need a low latency but does not demand critical reliability requirements.

2.3.1 IMT-2020 Required Capabilities

In the IMT-2020 recommendation, it is also defined a set of capabilities that are used to evaluate the infrastructure being used to support 5G. In reference to [4], these capabilities are the following:

- Peak Data rate: the maximum achievable data rate under ideal conditions by a user/device;
- User Experienced Data Rate: the data rate that can be achieved over a large coverage area for a majority of users;
- Spectrum Efficiency: average data throughput per Hz of the spectrum and per cell (radio equipment);
- Area Traffic Capacity: total traffic throughput served per geographic area;
- Network Energy Efficiency: defined by the energy consumption of the radio access network;
- Latency: defined as the contribution by the radio access network to the time from when the source sends a packet to when the destination receives it;
- Mobility: is in the context of key capabilities only defined as mobile speed;
- Connection Density: total numbers of connected and/or accessible devices per unit area.

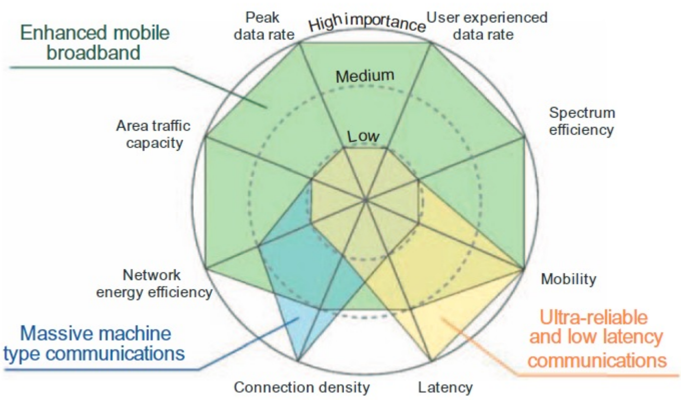


Figure 2.7: Key Capabilities per Use Case [1]

Other capabilities defined by the IMT-2020 are spectrum and bandwidth flexibility, reliability, resilience, security and privacy, and operational lifetime.

2.3.2 IMT-2020 Performance Requirements

Based on the use cases and capabilities described above, the ITU-R developed a set of minimum performance requirements for the 5G infrastructure. The performance requirements are described in [17]. These are shown, summarized, in Fig. 2.8.

Parameter	Minimum Technical Performance Requirement
Peak data rate	Downlink: 20 Gbit/s Uplink: 10 Gbit/s
Peak spectral efficiency	Downlink: 30 bit/s/Hz Uplink: 10 bit/s/Hz
User-experienced data rate	Downlink: 100 Mbit/s Uplink: 50 Mbit/s
Fifth percentile user spectral efficiency	3 × IMT-Advanced
Average spectral efficiency	3 × IMT-Advanced
Area traffic capacity	10 Mbit/s/m^2 (indoor hotspot for eMBB)
User plane latency	4 ms for eMBB 1 ms for URLLC
Control plane latency	20 ms
Connection density	1,000,000 devices per km^2
Energy efficiency	Related to two aspects for eMBB: a. Efficient data transmission in a loaded case b. Low energy consumption when there is no data The technology shall have the capability to support a high sleep ratio and long sleep duration
Reliability	$1-10^{-5}$ success probability of transmitting a layer 2 PDU (Protocol Data Unit) of 32 bytes within 1 ms, at coverage edge in Urban Macro for URLLC
Mobility	Normalized traffic channel data rates defined for 10, 30, and 120 km/h at ~1.5 × IMT-Advanced numbers Requirement for high-speed vehicular defined for 500 km/h (compared to 350 km/h for IMT-Advanced)
Mobility interruption time	0 ms
Bandwidth	At least 100 MHz and up to 1 GHz in higher-frequency bands. Scalable bandwidth shall be supported

Figure 2.8: Minimum performance requirements, taken from [1]

2.3.3 3GPP Performance Requirements for Usage Scenarios

In [5], performance requirements are defined for scenarios of high data rate and traffic density (eMBB use case), and also for scenarios of low latency and high reliability (uRLLC use case). Several scenarios require the support of very high data rates or traffic densities of the 5G system. Others require the support of very low latency and very high communications service availability.

For the first case, we have scenarios like urban macro areas, which are a wide-area scenario in an urban area, rural macro, which are a wide-area scenario in a rural area, indoor hotspots, for offices, homes, and residential deployments, broadband access in a crowd, for very dense crowds which causes a very high connection density and also a high data rate in the uplink, dense urban, broadcast-like services, high-speed train, high-speed vehicle, and airplane connectivity.

Table 7.1-1 Performance requirements for high data rate and traffic density scenarios.

	Scenario	Experienced data rate (DL)	Experienced data rate (UL)	Area traffic capacity (DL)	Area traffic capacity (UL)	Overall user density	Activity factor	UE speed	Coverage
1	Urban macro	50 Mbps	25 Mbps	100 Gbps/km^2 (note 4)	50 Gbps/km^2 (note 4)	10 000/km^2	20%	Pedestrians and users in vehicles (up to 120 km/h	Full network (note 1)
2	Rural macro	50 Mbps	25 Mbps	1 Gbps/km^2 (note 4)	500 Mbps/km^2 (note 4)	100/km^2	20%	Pedestrians and users in vehicles (up to 120 km/h	Full network (note 1)
3	Indoor hotspot	1 Gbps	500 Mbps	15 Tbps/km^2	2 Tbps/km^2	250 000/km^2	note 2	Pedestrians	Office and residential (note 2) (note 3)
4	Broadband access in a crowd	25 Mbps	50 Mbps	[3,75] Tbps/km^2	[7,5] Tbps/km^2	[500 000]/km^2	30%	Pedestrians	Confined area
5	Dense urban	300 Mbps	50 Mbps	750 Gbps/km^2 (note 4)	125 Gbps/km^2 (note 4)	25 000/km^2	10%	Pedestrians and users in vehicles (up to 60 km/h)	Downtown (note 1)
6	Broadcast-like services	Maximum 200 Mbps (per TV channel)	N/A or modest (e.g., 500 kbps per user)	N/A	N/A	[15] TV channels of [20 Mbps] on one carrier	N/A	Stationary users, pedestrians and users in vehicles (up to 500 km/h)	Full network (note 1)
7	High-speed train	50 Mbps	25 Mbps	15 Gbps/train	7,5 Gbps/train	1 000/train	30%	Users in trains (up to 500 km/h)	Along railways (note 1)
8	High-speed vehicle	50 Mbps	25 Mbps	[100] Gbps/km^2	[50] Gbps/km^2	4 000/km^2	50%	Users in vehicles (up to 250 km/h)	Along roads (note 1)
9	Airplanes connectivity	15 Mbps	7,5 Mbps	1,2 Gbps/plane	600 Mbps/plane	400/plane	20%	Users in airplanes (up to 1 000 km/h)	(note 1)

NOTE 1: For users in vehicles, the UE can be connected to the network directly, or via an on-board moving base station.
NOTE 2: A certain traffic mix is assumed; only some users use services that require the highest data rates [2].
NOTE 3: For interactive audio and video services, for example, virtual meetings, the required two-way end-to-end latency (UL and DL) is 2-4 ms while the corresponding experienced data rate needs to be up to 8K 3D video [300 Mbps] in uplink and downlink.
NOTE 4: These values are derived based on overall user density. Detailed information can be found in [10].
NOTE 5: All the values in this table are targeted values and not strict requirements.

Figure 2.9: Table with performance requirements for high data rate and traffic density scenarios, taken from [5]

For the second case, we have scenarios like discrete automation, which demands high reliability and availability, and are characterized by systems that are deployed in geographically limited areas, and have limited access to public network resources. Process automation is similar to discrete automation in terms of requirements (except latency-wise), but its focus is on automation for reactive flows (refineries and water distribution). These scenarios are usually served by private networks. Another scenario is automation for electricity distribution, which has similar requirements to process automation, also served by a private network, but it is immersed in the public space. The last scenario is intelligent transport systems which are composed of automation solutions for the infrastructure supporting street-based traffic. Similar to the last scenario, the network will be also immersed in the public space.

Table 7.2.2-1 Examples of performance requirements for low-latency and high-reliability scenarios.

Scenario	Max. allowed end-to-end latency (note 2)	Survival time	Communication service availability (note 3)	Reliability (note 3)	User experienced data rate	Payload size (note 4)	Traffic density (note 5)	Connection density (note 6)	Service area dimension (note 7)
Discrete automation	10 ms	0 ms	99,99%	99,99%	10 Mbps	Small to big	1 Tbps/km^2	100 000/km^2	1000 x 1000 x 30 m
Process automation – remote control	60 ms	100 ms	99,9999%	99,999%	1 Mbps up to 100 Mbps	Small to big	100 Gbps/km^2	1 000/km^2	300 x 300 x 50 m
Process automation – monitoring	60 ms	100 ms	99,9%	99,9%	1 Mbps	Small	10 Gbps/km^2	10 000/km^2	300 x 300 x 50
Electricity distribution – medium voltage	40 ms	25 ms	99,9%	99,9%	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	100 km along power line
Electricity distribution – high voltage (note 1)	5 ms	10 ms	99,9999%	99,999%	10 Mbps	Small	100 Gbps/km^2	1 000/km^2 (note 8)	200 km along power line
Intelligent transport systems – infrastructure backhaul	30 ms	100 ms	99,9999%	99,999%	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	2 km along a road

NOTE 1: Currently realised via wired communication lines.
NOTE 2: This is the maximum end-to-end latency allowed for the 5G system to deliver the service in the case the end-to-end latency is completely allocated to the 5G system from the UE to the Interface to Data Network.
NOTE 3: Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets may take place in order to satisfy the reliability requirement.
NOTE 4: Small: payload typically ≤ 256 bytes
NOTE 5: Based on the assumption that all connected applications within the service volume require the user experienced data rate.
NOTE 6: Under the assumption of 100% 5G penetration.
NOTE 7: Estimates of maximum dimensions; the last figure is the vertical dimension.
NOTE 8: In dense urban areas.
NOTE 9: All the values in this table are example values and not strict requirements. Deployment configurations should be taken into account when considering service offerings that meet the targets.

Figure 2.10: Table with performance requirements for low-latency and high-reliability scenarios, taken from [5]

2.4 5G Interfaces - Fronthaul, Midhaul, Backhaul

Based on the O-RAN architecture, we know that there are three main entities in the RAN, the O-CU, the O-DU, and the O-RU. The next step is to know how do they communicate with each other, what interfaces do they use, and how does the data reach the Core. Some of these interfaces are standardized by the O-RAN Alliance because they only appeared by the necessity created from the O-RAN architecture. The rest was standardized by the 3GPP.

2.4.1 Fronthaul Interface

The fronthaul interface is the interface that connects the O-RU to the O-DU. This interface was standardized by O-RAN Alliance [3]. This interface is used to connect an O-DU to one or more O-RUs, inside the same gNB. The O-RAN fronthaul presents four functionalities/planes, which are the control, the user, the synchronization, and the management-plane.

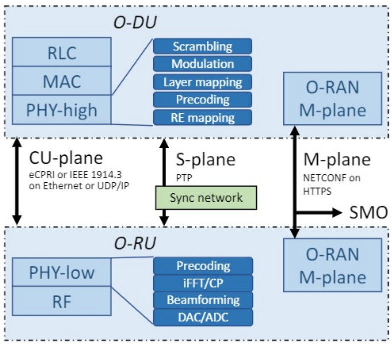


Figure 2.11: O-RAN Fronthaul Interface [3]

The CP is used for the transport of PHY control messages. It transports commands from the High-PHY in the O-DU to the Low-PHY in the O-RU. The UP is in charge of transporting the in-phase and quadrature (I/Q) data between the O-RU and the O-DU. Both planes are represented by the same protocol structure to transport control/user data.

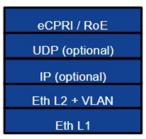


Figure 2.12: Control and User-plane Protocol Structure [6]

Both control and user-plane messages are encapsulated and transported in evolved Common Public Radio Interface (eCPRI) or IEEE 1914.3/Radio over Ethernet headers as payload. Before these protocols, and in the prior mobile communications, we had the Common Public Radio interface (CPRI), a circuit-based (point-to point) protocol running over fibre, which would bring inflexibility in terms of RAN functions splitting. eCPRI and IEEE 1914.3 will be transported over

Ethernet layer 2 (packet-based protocol which allows a more efficient transport of data), and in comparison to CPRI, they provide flexibility in RAN functional splitting, allowing to support different network requirements (bandwidth, latency, etc.) in the fronthaul, which is essential in 5G. In the case of O-RAN, UDP and IP encapsulation is not used because of bandwidth and stringent latency requirements [6].

The synchronization-plane is used to control time, frequency, and phase synchronization between the clocks of the O-DU and the O-RUs. This is a key element if we want a functioning frequency-slotted system distributed across multiple radio units. To synchronize the different O-RUs with the O-DU, the O-RAN Alliance has defined the Precision Time Protocol (PTP or IEEE 1588) as the protocol responsible for the synchronization of distributed clocks in a network.

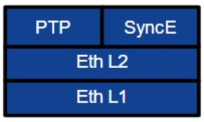


Figure 2.13: Synchronization-plane Protocol Structure [6]

The frequency and time synchronization of O-DUs and O-RUs via Ethernet layer 2 uses Synchronization Ethernet (syncE) and PTP, respectively. The PTP protocol can be encapsulated in UDP and IP but it does not provide the synchronization performance needed [6].

The management-plane enables the initialization and management of the connection between the O-DU and the O-RU, and the configuration of the O-RU. This management-plane can connect the O-RU to an O-RU controller [7]. This controller can be the O-DU or an external entity to the RAN.

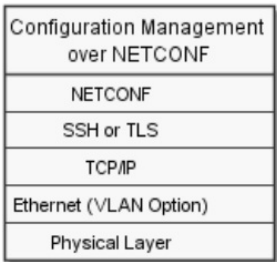


Figure 2.14: Management-plane Protocol Structure [7]

The transport network layer that is used to transport management-plane messages is composed of IP, Secure Shell (SSH), or Transport Layer Security (TLS) protocols [7]. The management-plane messages will be organized in YANG data modules and sent via NETCONF.

2.4.2 Midhaul Interface

The midhaul interface is the interface that connects the central unit to the distributed unit. This interface was standardized by the 3GPP organization [8]. The midhaul interface was denominated as the F1 interface by the 3GPP. Since it is possible to have the disaggregation of the central unit in two units the CU for the CP and the CU for the UP, 3GPP defined an F1 interface for each plane. The interface for the CP supports the exchange of signaling information, and the interface for the UP supports user data transmission.

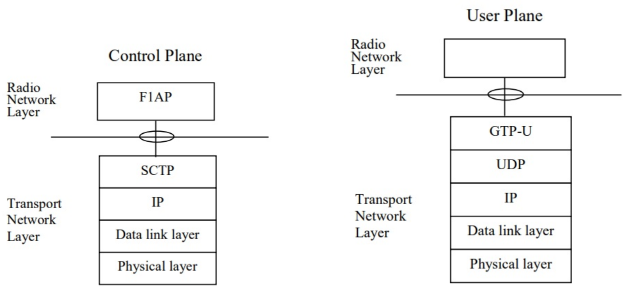


Figure 2.15: F1 Interface protocol structure [8]

The F1-c interface is based on IP transport, comprising the Stream Control Transmission Protocol (SCTP) on top of IP for reliable transport of data. The signaling service is provided by the F1 application protocol (F1AP) [8]. The F1AP consists of elementary procedures, and each procedure is a different type of interaction between the central unit and the distributed unit, that can lead to different control and management actions [18].

The F1-u interface is also based on IP transport, comprising the User Datagram Protocol (UDP) and GPRS Tunnelling Protocol (GTP-U) protocols on top of IP. The GTP-U is a tunneling protocol that is used to encapsulate and transport PDUs. Each tunnel is identified in each node with an identification number, an IP address, and a UDP port number. It is mandatory to have a tunnel if we want to forward data between two GTP-U entities [19]. This way, the GTP-U protocol is going to be used to forward data between both the CU and the DU.

2.4.3 Backhaul Interface

The backhaul interface will be the interface far-off of the cell site, and it will be the interface that will connect the RAN with the CN. In this section, we will also discuss the interface that connects a gNB to another gNB. Nonetheless, both interfaces are standardized by 3GPP. These interfaces are represented in Fig. 2.1.

The NG interface is used to connect a gNB to the 5G CN. The NG interface is split into two interfaces, the NG-c for the CP connection between the CU-CP and the AMF service, also known as the N2 interface, and the NG-u for the UP connection between the CU-UP and the UPF service, also known as N3 interface.

According to [9], the NG-c interface supports procedures to establish, maintain and release PDU sessions between the RAN and CN, procedures to perform intra-RAT handover, the separation of each UE on the protocol level for user-specific signaling management, the transfer of NAS signaling messages between a UE and the AMF service (N1 interface represented in Fig. 2.4), and mechanisms for resource reservation for packet data streams.

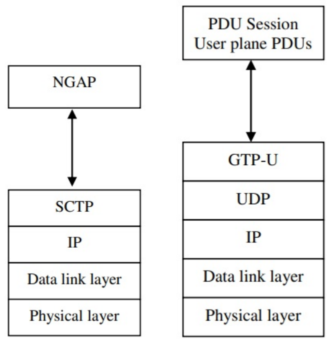


Figure 2.16: NG-c and NG-u interface protocol structure [9]

In terms of protocol structure, this interface is based on IP transport, and for reliable transport, the SCTP protocol is added on top of the IP. This interface is very similar to the F1-c interface, in terms of the transport of packets and also, so too has, this interface, a signaling protocol to transport different signaling messages. The NG application protocol (NGAP) is very similar to the F1AP protocol. The protocols just differ in the interface where they are used and in the type of procedures that they have.

According to [9], the NG-u interface provides non-guaranteed delivery of user data between the CU-UP and the UPF. In terms of protocol structure, this interface is similar to the F1-u interface, so it uses the GTP-U protocol to transport the user-plane PDUs.

Moving to the Xn interface, this interface is used to connect gNBs to each other, more accurately, the CU to another gNB. In this interface, there is also a split between the CP and the UP.

According to [20], the Xn-c interface provides the provision of reliable transfer of Xn application protocol (XnAP) messages, networking and routing functions, redundancy in the signaling network, and flow and congestion control.

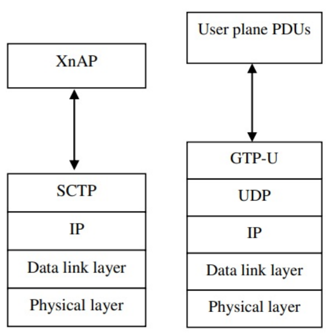


Figure 2.17: Xn-c and Xn-u interface protocol structure [10]

The protocol structure that we see in Figure 2.17 is very similar to CP interfaces that we see for the F1 and NG interfaces. The transport of packets is done over IP and it also used SCTP to guarantee a reliable transfer of packets. The XnAP is very similar to the F1AP and the NGAP protocols, and just differs in the elementary procedures that it offers.

The Xn-u interface provides non-guaranteed delivery of user data between two gNBs, and also uses the GTP-U protocol to transport the PDUs.

The network reach for each interface [11], fronthaul, midhaul, and backhaul is respectively 1-20, 20-40, and 300 km.

Chapter 3

5G Private Network Design

In the last chapter, we presented the 5G infrastructure and its main entities. We also showed how O-RAN is organized, and what interfaces are used to connect these entities. Furthermore, we presented the performance requirements for specific 5G scenarios proposed by 3GPP.

With this information, in this chapter, we will focus on the study and analysis of the state-of-art of topics that need to be clarified so that a good basis is made to begin our work.

The main topics of focus in this chapter are the following:

- the split of the radio protocol stack between the RAN entities (CU, DU, and RU);
- the different private network topologies;
- the deployment scenario for a uRLLC use case, more specifically for an automated-guided vehicle use case.

3.1 5G RAN Protocol Architecture - Split Functionality

As described in Chapter 2, the radio protocol stack provides many functionalities and the RAN architecture can be composed with different entities. So, it is important to understand how the RAN protocols should be splitted between these entities.

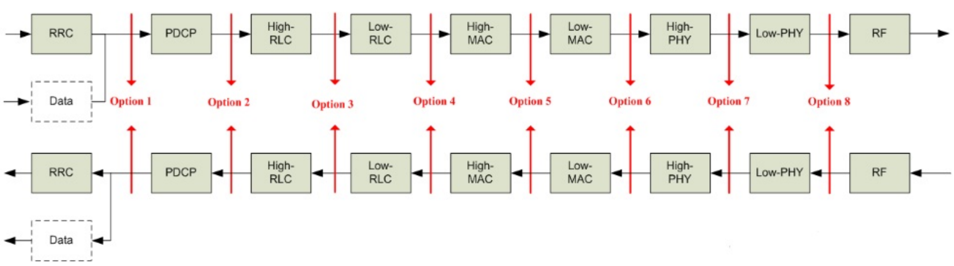


Figure 3.1: 3GGP Split Options [11]

Figure 3.1 shows eight options for functional splitting presented by the 3GPP organization [21]. Initially, these splits were defined as a separation between a CU and a DU, not considering the disaggregation presented by the O-RAN architecture. The splits were characterized as high or low-layer splits.

A high-layer split is less demanding in terms of transport network performance but has fewer coordination gains since functionalities are decentralized. Its counterpart, the low-layer split, is more demanding in terms of bandwidth, latency (transport network requirements), but it needs less investment in terms of deployment at the cell site.

As shown in [11], the 3GPP has defined split-option 2 as its standardized high-layer split option, since it has already a standardized interface, the F1 interface. However, the 3GPP postpone the decision for the low-layer split option, having split-option 6 and 7 as the main candidates.

The split-option 2 implements RRC, SDAP, and PDCP protocols in the CU, and the rest of the radio protocols in the DU. This option has two variants, one with the separation of control and user-plane in the CU, and the other not. The separation between CP from the UP adds security and resiliency. Furthermore, latency and bandwidth requirements in the F1 interface are relatively relaxed, since the real-time functions are located below the PDCP layer [22], and this way most of the raw data is processed in the radio unit. Since the PDCP layer is in the CU and PDCP supports in-sequence delivery, with this split, we will require a re-sequencing buffer in both CU and DU, which will add latency in the transport of data through the F1 interface. The coordination of mobility and handover procedures are made in the CU because it has the RRC functionalities, however, the coordination of scheduling will go to the DU, since it has the MAC layer. This can be a disadvantage because we limited the potential of coordination of scheduling between DUs [22]. This will be true for a single-layer split. Still, with the O-RAN architecture, the O-DU will coordinate the scheduling between different O-RUs, and this disadvantage disappears.

As referred above, for the low-layer split there are several options being explored. There are two main contenders, split-option 6, being proposed by the Small Cell Forum [11], and split-option 7, which has three variants, 7.1, 7.2, and 7.3. The O-RAN alliance has proposed the fronthaul of its architecture to be based in split 7.2 [3]. The split-option 6 implements a MAC/PHY split between the CU and DU. The split-option 7 implements an intra-PHY split that has three different ways of distributing PHY functionalities (sub-options 7.1, 7.2, and 7.3).

According to [22], the split-option 6 presents a more relaxed transport network between the CU and DU, in terms of bandwidth and latency requirements in comparison to split-option 7, but the DU is more complex in terms of functionalities that need to be implemented at the cell site. Another disadvantage of the split-option 6 is the reduction in benefits achievable by the centralization of radio functionalities since PHY functions represent 80% of baseband functionality and Data-Link functions only represent 20%, and this split-option implements all PHY functions in the radio unit.

With the split-option 7, we get a good balance between the RU complexity and interface requirements. Split-option 7 has three variants, however, the best option is 7.2. The split-option 7.3 is viewed by 3GPP as a downlink option only and provides the same bitrate as option 7.2 [22].

Furthermore, split-option 7.2 presents better bandwidth and latency requirements in the transport network in comparison to split-option 7.1, since it has a variable bitrate. This is true because the resource element mapping (PHY functionality) is located in the radio unit and not in the centralized unit, detecting unused subcarriers from the radio frequency (RF) link and this is not the case for the sub-option 7.1 [22].

As explained, split-option 2 implements all the real-time functions below the PDCP layer, which are deployed in the radio unit. So, most processing is done in the radio unit. Fewer data needs to be transported through the transport network to the destination, which brings down latency. For this reason, this high-layer split is regarded as an optimal split for the uRLLC case. In [23], the authors conclude by their results that the processing time of data to be sent to an UE is much lower in split-option 2 than in split-option 7.2. This delay in processing the data through the radio protocol stack is the most impactful overall.

According to [24], eMBB use cases require high gNB coordination, and a low-layer split, as split-option 7.2, can provide better coordination and resource sharing in comparison to a high-layer split.

The O-RAN Alliance supports two-tier functional splits, with its architecture that is based on disaggregation. The O-CU will implement the SDAP, RRC, and PDCP protocols. The O-DU will implement the RLC, MAC, and High-PHY protocols. The O-RU will implement the Low-PHY layer. It is possible to see that O-RAN implements the split-option 2 in its midhaul and the split-option 7.2x in its fronthaul. With this dual-split implemented, we introduce now the concept of RAN deployment scenarios, mentioned in [11]. The RAN deployment scenarios defined are the following:

- Independent RU, DU, and CU locations: there are fronthaul, midhaul, and backhaul networks;
- Co-located CU and DU: the CU and the DU are located together. There is no midhaul;
- RU and DU integration: the RU and the DU are deployed close to each other (for example, in the same building). Since they are very close, the fronthaul network is ignored, and there is only a midhaull, and backhaul network;
- RU, DU, and CU integration: all-in-one solution. There is only a backhaul network.

This way, by choosing the RAN deployment scenario, we can shape our architecture to support a specific use case, by manipulating the location of the DU. When we want to support a use case more related to the uRLLC type, we place the DU closer to the RU. If we want to support an eMBB use case, we try to centralize the DU and the CU.

3.2 5G Private Network Topologies

A private network or non-public network provides 5G network services to a defined user organization. It is deployed on the organization's defined premises (specific area) [25]. The reasons for implementing this kind of network are the need for support of high QoS requirements, security and privacy, and creating isolated networks as a form of protection against malfunctions of external networks.

Private networks can be splitted into two categories. The first one is private networks that are deployed as isolated and standalone networks, and the second is when the private networks are made in a collaborative perspective with Mobile Network Operators (MNOs) networks. In the following paragraphs, a description of the different deployments of private networks is made according to [25] and [26].

In the isolated deployment, the private network is implemented as a standalone network. All network functions are located inside the logical perimeter of the master and the owner entity defined premises. This means that the CN is placed inside this logical perimeter. Besides the CN, the gNB and possible Multi-access Edge Computing (MEC) units are also inside the logical perimeter. Because of this, the private network is physically separated from the public network, and it is an independent system [26], where there is no sharing of the operator network's resources.

This deployment ensures privacy and security since the private network has no contact with the MNO network because it is physically separated. The management of data is done within the organization's premises. Furthermore, this deployment provides ultra-low latency communications, because of the distance between the devices, the gNB, the UPF, the MEC, and other application services. Another great advantage is the fact that any malfunction that can occur in the public/operator's network will not influence the operation of the private network. The only disadvantage is the deployment cost of an end-to-end private network. Although it is an isolated deployment, it is still possible to create an optional connection to public network services.

The next deployments to be described are deployments that are implemented in conjunction with the MNO network. The first one is characterized by sharing the radio access network, the same gNB. The gNB is located within the organization and it is shared between the private and public networks. This gNB will not be physically separated, instead, it will be logically separated. The CN will remain segregated from the public network and will be deployed within the organization's premises. The data traffic will be delivered to the corresponding UPF. Public data will go to the public UPF and private data will be forwarded to the organization's UPF. Since the UPF, MEC units, and other application servers are within the enterprise, this deployment provides ultra-low latency communications.

The second deployment is characterized by the sharing of the RAN and also the Core's CP with a MNO network. In this topology, only the UPF, MEC units, and application servers are built in the enterprise and are physically separated from the public network. Other entities are logically separated. From this deployment, the only thing that changes in relation to the last deployment is that subscription and operation information will be stored in the public infrastructure. This deployment can still provide ultra-low delay communications because of the distance between the UPF and the UEs (all are within the organization's premises).

The third and last deployment will implement a private network hosted on the MNO network. Every entity of the network is logically separated, and only the gNB is deployed within the organization's premises. The CN exists in the public network. Since the UPF is located in the public network, the traffic needs to travel to the UPF and come back to communicate possible replies from MEC units or applications servers. Network latency in this deployment can have a huge impact on the performance of the use cases, depending on the distance between the organization's UEs and the location of the public UPF.

These deployments are focused on the owner of the network, and if the private network needs the support of MNOs. This is not the baseline to cover in this thesis, but with this information, which is considered factual by many articles, we can extrapolate some knowledge in terms of the impact of having a physical, hybrid, or cloud CN.

From these deployments, we can conclude that the placement of the UPF is very important for ultra-low delay communications. What we can conclude from these articles is that the UP of the CN needs to be at the site or near it to be possible to provide ultra-low delay communications. Looking from this perspective, we can conclude that a good deployment for the uRLLC case needs to have the UPF near the cell site. This perspective is also explored in [27], where it is explored different placements of the UPF for services with a specific latency requirement, in a multi-tier architecture. And what is shown in this report is that when the latency requirement becomes more and more stringent, the UPF and the RAN need to be closer and closer to the cell site.

3.3 5G Use Case - Automated Guided Vehicles Use (Industry 4.0)

The goal of 5G is to enhance communications used in current applications and support new up-coming services that until now were impossible to implement. The fourth industrial revolution, also known as Industry 4.0, is the next stage of industrial manufacturing, which is focused on the digitization of all industrial processes. In a factory of the future, each phase of production needs to be digital and interconnected with each other. The objective is to increase the efficiency of all industrial processes.

Automated Guided Vehicles (AGV) are one of the most relevant use cases that are being worked on in 5G. AGVs have been enhanced through the years, and one of the main issues has been the access type communication between AGVs and the main control entity.

An AGV is a self-drivable robot, which is often used in industrial applications to transport materials/products through a factory. It is composed of a processing module that captures sensor data, processes it, and uses it for self-navigation and to deliver it to the control station. The control station exists to ensure the coordination between different AGVs. The AGV also needs to integrate a 5G compatible device to connect to the network.

Before 5G, the existing communication technologies, like WIFI or wired technologies presented several limitations. Wired technologies can be complex in terms of installation and tend to be costly. Also, in most industries, scalability is key to adapting to new ways of working, and this may drive the placement of workstations. A WIFI network can also not be the best solution, since it is possible to have different subsystems with several devices connected that are communicating internally with each other and with the respective control entities, and the amount of data that is circulating can create interference and delay between the different subsystems. Some 5G network topologies and architectures were created to support this type of use case. The following paragraphs will present some of these solutions.

3.3.1 The 5G MEC Application in Smart Manufacturing

In [12], the authors introduce a pre-commercial 5G MEC deployment, which integrates a MEC unit in a 5G SA network. This MEC unit will run the AGV management platform, which communicates with AGV units. In terms of deployment scenario, the RAN and the UP part of the Core are deployed inside the factory. The rest of the Core is deployed outside the factory, in the middle of the factory's city.

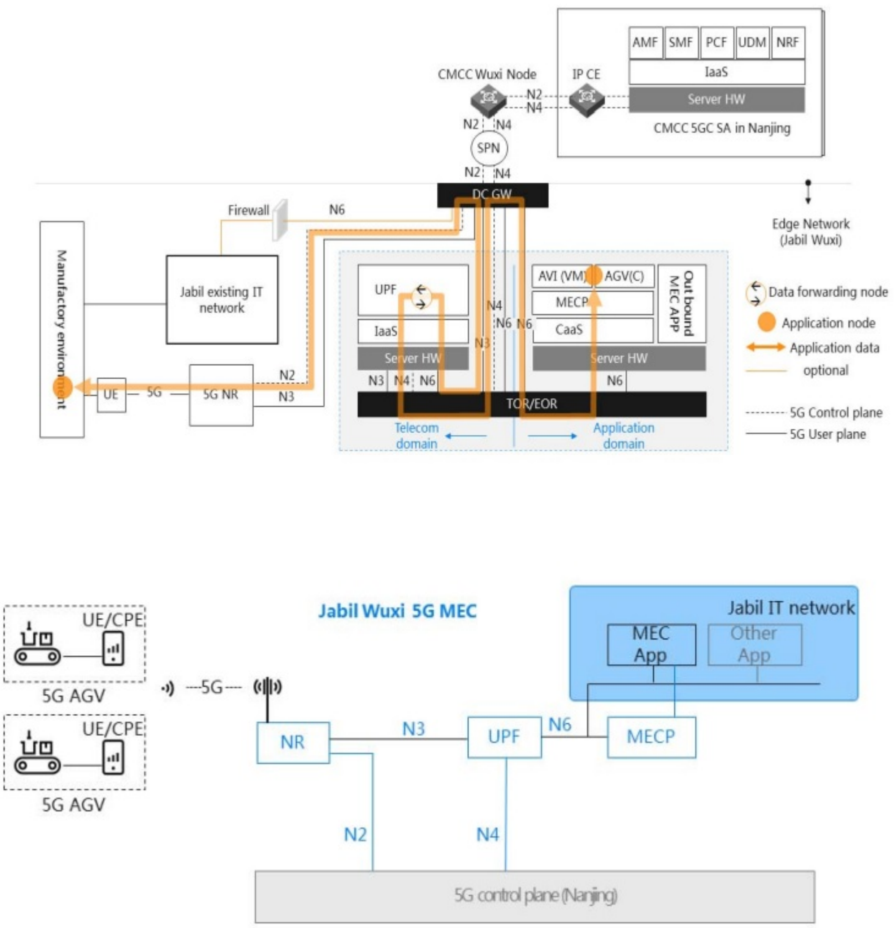


Figure 3.2: Deployment Architecture and Deployment Diagram [12]

The RAN has several radio units (quantity not specified) that are connected via Common Public Radio Interface (CPRI) to a Band Base Unit (BBU). This means that the split-option 8 was used to implement the Radio protocol stack. In this regard, the first conclusion that we can make is that if more functions were implemented in the radio unit, very small latencies could be achieved (in reference to Section 3.1). The UPF is deployed in a server located inside the factory, and the MEC unit will be deployed in another server, also located inside the factory. The placement of the UP in the factory was made to provide lower latency, security integrity, and privacy (as it was seen in the previous Section 3.2).

Fig. 3.2 presents the deployment scenario implemented for this use case. As shown in the figure, all user data will circulate inside the premises of the factory, because of privacy requirements. In terms of bands, a band in the 2.6 GHz is used. In terms of results, the authors return the measured average round-trip time of 10 ms, with a jitter of 2 ms approximately, but did not specify the number of AGVs that were deployed.

3.3.2 Open Source 5G-NSA Network for Industry 4.0 Applications

In [13], an implementation of a 5G private non-standalone (NSA) network is presented, that is used as the communication network between an AGV and a remote control station. In this use case scenario, the AGV is remotely controlled and monitored by an operator in a control station. Both entities are UEs of the network. The main purpose of the network is to transport the commands (steering, velocity) given by the operator in the control station to the AGV, and to transport the image being captured by the AGV to the control station. The main purpose of the network is to establish a connection between the AGV and the control station so that they can communicate with each other.

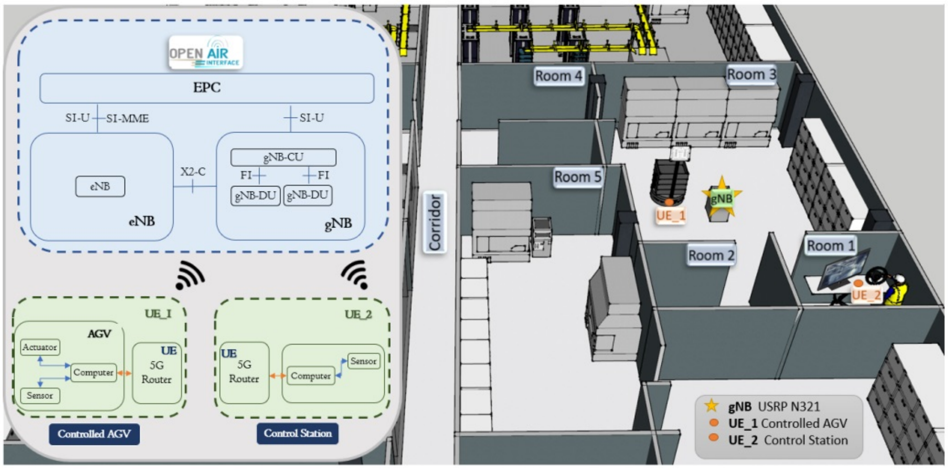


Figure 3.3: Deployment Architecture and Placement of the 5G NSA network in the real scenario evaluated [13]

The deployment configuration consists of one server for the Evolved Packet Core (EPC) (4G Core), and two universal software radio peripherals (URSP) that are configured as a gNB and an eNodeB (eNB). The RAN and the EPC are deployed on the same server. The deployment is installed in a factory, which has five rooms. The control station is located in one room, and the gNB is located in another room. The tests are only made in three rooms, for the other two rooms an extrapolation in terms of results is made. The three rooms are the one with the gNB, the one with the control station, and another room where the AGV has physical access to.

In terms of measurements, it is assessed the coverage, which is comprised of measurements of the Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), and Signal to Interference plus Noise Ratio (SINR) of the five rooms. The authors conclude that the frequency being used (3.6 GHz) limits indoor propagation and the range is not as good as what they desired. Also, they established that values of 10 or more dB for the SINR can guarantee a good performance of the system. Coverage gaps can be justified by walls and obstacles, that need to be taken into account in the indoor network design.

Furthermore, it is also assessed the difficulty of AGV driven operability due to high latency, and it is recommended that the control of the AGV must be executed over a network whose latency does not exceed 30 ms.

3.3.3 Deployment and Evaluation of an Industry 4.0 Use Case over 5G

In [14], the authors present and validate a solution for the remote control of AGVs using a real 5G deployment. To validate their deployment and to assess its performance, a group of service KPIs, in correlation with 5G network KPIs, are analyzed. The service KPIs being analyzed are the guide error, which is the deviation of the AGV from the correct path, and the current consumption, which is the instantaneous current consumed by the AGV. The 5G network KPIs that are manipulated are end-to-end latency and packet loss. The two service KPIs are related since large and frequent guide errors will trigger the control station to transmit different control actions, to correct the trajectory of the AGV, which in turn will result in an increase in energy consumption by the AGV to apply these actions (apply different angular velocities). Furthermore, the 5G network KPIs can also influence the service KPIs. A large delay or packet loss can result in bigger guide errors, which increases energy consumption.

Figure 3.4: 5G Network Deployment Architecture [14]

For the deployment scenario, the authors implement a 5G NSA network, with two radios, one for 4G communications and another for 5G communications. Both are connected to an EPC, which in turn is connected to a data center where the control station is located.

The AGV is composed of sensors that measured critical variables (guide error, current consumption, battery status, and velocity). This data is sent to the control station. The AGV also has actuators, which are comprised of motors and wheels, to perform the movement imposed by the control station. The AGV will have a 4G and 5G router to connect to the network.

For the test scenario, the AGV needs to go through a specific path that begins in the AGV's charging station and ends in the same position. The authors of this article evaluate initially the performance of the AGV using 5G and 4G radios to access the network. As expected the performance of the AGV with 5G is much better, because of the lower latencies that the 5G radio can provide compared to the 4G radio. With these lower latencies, the guide errors are much smaller, which consequently decreases also the energy consumption of the AGV.

Another detail that is analyzed in this paper is the impact of the control station placement on the performance of the AGV. The placement of the control station affects the communication delay between the AGV and the control station. To evaluate this detail, it is added incrementally end-to-end delay to the network, and it is measured the service KPIs. It was concluded that a delay of 200 ms can be tolerated by the AGV, although it is suggested to stay in lower values.

It is also analyzed the impact of radio conditions, by manipulation of the packet loss in the network. The same packet loss is applied in both downlink and uplink directions. It is concluded that a packet loss of over 30 percent results in a bad performance by the AGV.

3.3.4 Main Conclusions

To conclude this section, we can take important details from each and every article. In the first one, we can see a real deployment scenario for these use cases, particularly the choice of the core implementation, being a hybrid core. In the second article, it is important to notice the KPIs that were evaluated, in terms of coverage and UE-to-UE latency. And in the last paper, it is important to take with us the service KPIs, which can be helpful to evaluate a network in relation to the performance of the application. And also, the network KPIs that were manipulated in order to assess the performance of the application. In this case, the packet loss and the end-to-end latency.

Chapter 4

5G Network Architecture

The focus of this chapter will be in the deployment of the 5G Private Network architecture used in this thesis. This chapter will be structured in the following sections:

- Design and Specification - description of the RAN and Core Architecture. Description of the infrastructure that supports both entities. Furthermore, we will also mention the system requirements and performance specifications;
- Implementation - description of what needed to be done, in terms of implementation, to complete the deployment of the 5G Network.

4.1 Design and Specification - Architecture

We will describe the deployment architecture for both the RAN and Core solutions that were implemented in this 5G Private Network Deployment. The main goal is to explain the infrastructure that needs to be implemented for this 5G deployment.

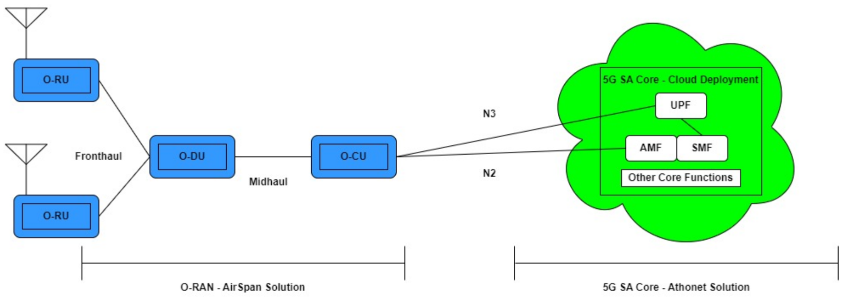


Figure 4.1: High Level Architecture

Figure 4.1 represents our 5G deployment, which is characterized by using the AirSpan O-RAN solution, and the Athonet 5G SA Core solution.

The RAN solution was fully deployed on premises, and follows the O-RAN architecture that was explained in Section 2.2, except that it does not have RICs. This RAN solution is composed by two O-RUs, one O-DU, and one O-CU. The split 7.2x and the split 2 are deployed for the fronthaul and midhaul connections. The O-DU and the O-CU are deployed through the AirSpan vRAN solution, which are containers configured directly in COTS servers or Virtual Machines (VM). For both O-RUs, the AirSpan AirVelocity 2700 was used as the radio unit. For the Core, its implementation was done in a hyperscaler, in this case Amazon Web Services (AWS). The solution was implemented in a single instance, and it was the 5G SA Core Athonet platform.

4.1.1 RAN Architecture

As explained, we are implementing an O-RAN architecture with least three entities, that will run all radio protocol stack functions as we saw in Section 2.1.1, with a particular split of functionality, as we saw in Section 3.1. To ensure the proper behaviour of this solution, we need to guarantee that the infrastructure can connect all of these entities and also provide all standardized interfaces by the O-RAN Alliance, as we saw in Section 2.4.

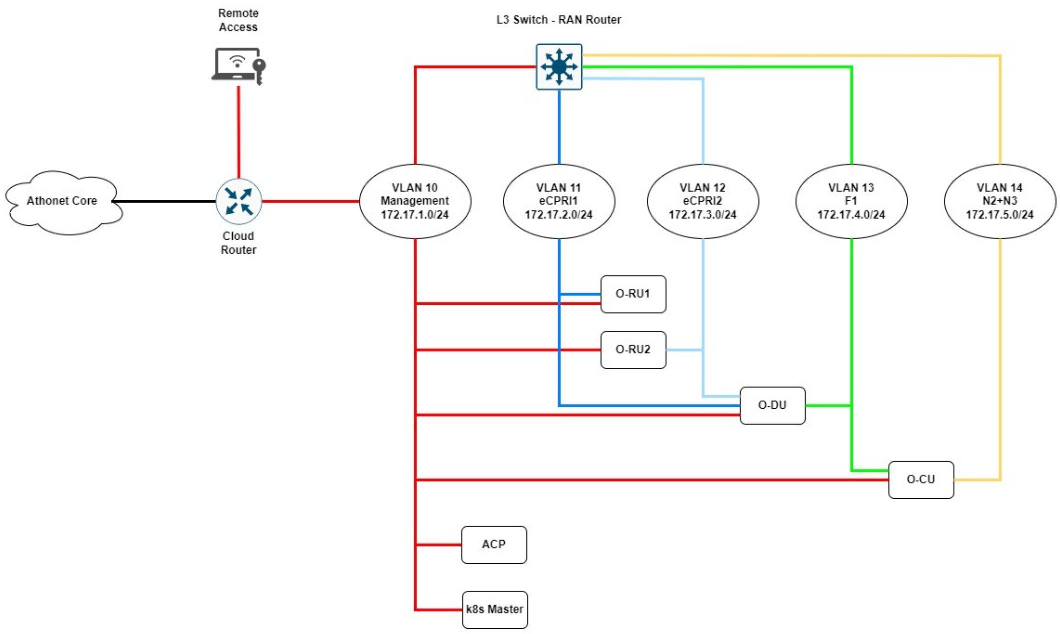


Figure 4.2: O-RAN Logical Diagram

Figure 4.2 expresses the Logical Diagram that needs to be followed to ensure that the O-RAN requirements are fulfilled, and also how the AirSpan solution was implemented in this 5G Private Network.

As we can see in this figure, most of the interfaces discussed in Section 2.4 are present and each interface has its own sub-network. This deployment includes two routers, one is responsible for the connection between the gNB and the Core, and the other is responsible for the organization of the RAN.

Beside the regular O-RAN entities, we also have two entities that are mandatory for this AirSpan solution: AirSpan Control Platform (ACP) and the kubernetes (k8s) master. The ACP is responsible for the orchestration and management of the RAN. For example, it provides provisioning in terms Core connection, setting up the N2 and N3 interfaces that will connect to the Core, or in terms of radio configurations, setting up the frequency band in which the radio units will operate. It works as an API to facilitate the provision of the RAN. The k8s master is responsible for the deployment of the vRAN containers in the O-DU and O-CU. This deployment uses k8s and its way of operating will be described later when describing the implementation of this 5G network.

In terms of sub-networks, as we can see in Figure 4.2, most of the interfaces mentioned in Section 2.4 are represented in this diagram. There is a sub-network for the eCPRI connection between the O-RU and the O-DU. In this case, two subnetworks, each one for each O-RU in our deployment. This connection is comprised of the control and user-plane of the physical layer connection between the O-RU and the O-DU. There is also a subnetwork for the F1 connection between the O-DU and the O-CU. Furthermore, there is a sub-network for the connection between the Core and the gNB, for the support of the N2 and N3 connection. Lastly, a management subnetwork was also created, which was responsible for the management plane of each O-RAN entity, which provides the basic function of accessing each entity in the RAN, k8s communication for deploying containers, and communication between ACP and O-RAN entities.

An important detail that needs to be mentioned is that the synchronization-plane in the fronthaul does not exist in accordance to the O-RAN Alliance requisites, in this AirSpan Solution. The PTP synchronization is done through separate connections to the RAN router which acts as the grandmaster clock. So the O-DU and both O-RUs are synchronized through a connection to the router. Another important detail is the F1-c connection which is done through the management interface between the O-CU and the O-DU.

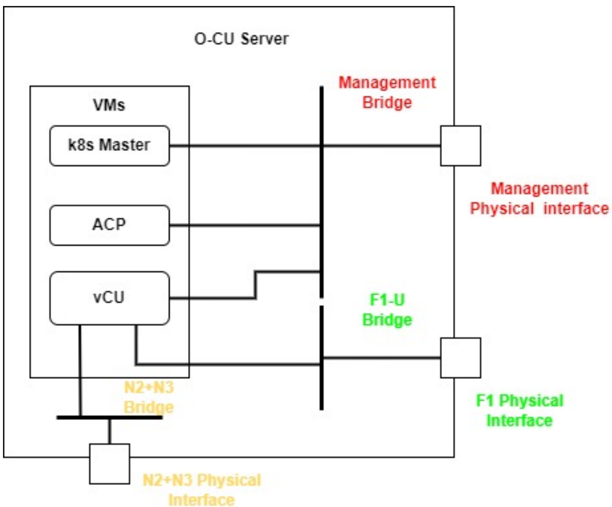


Figure 4.3: O-CU Server Organization

Figure 4.3 represents the organization of the server that houses the O-CU, the ACP and the k8s Master VMs. As we can see, all of these VMs require a management interface. The O-CU has also two interfaces for the F1 connection and a N2 and N3 interface for the connection between the gNB and the Core. All of the VMs' interfaces need to be linked with virtualized bridges to connect to the server's physical interfaces.

For the O-DU, its environment is implemented over a bare-metal server. It has four interfaces. One for management purposes, which also has the F1-c connection with the O-CU, as prior mentioned. In terms of O-RAN interfaces, it has one for the F1-u connection with the O-CU, another trunked eCPRI interface for both O-RUs, and an interface solely for the PTP synchronization.

The O-RUs possess two main interfaces, one for management and another for the eCPRI connection for the control and user-plane. Through the management interface, the O-RUs receive also the PTP packets, from the the router that organizes the RAN.

4.1.2 Core Architecture

The Athonet 5G SA Core solution was implemented in a cloud environment, specifically in the Amazon Web Services (AWS) Cloud. This solution was implemented in a single AWS instance.

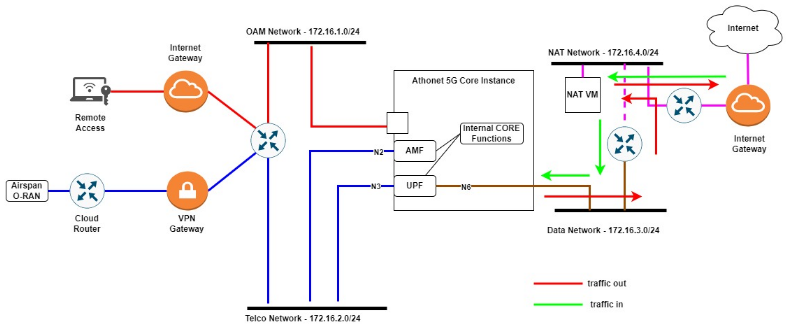


Figure 4.4: Athonet Solution Logical Diagram

This Core architecture requires three interfaces, one for management purposes, another for the N2 and N3 connection with the gNB, and another for the N6 interface that will connect the UPF to the Data Networks. In our case, we have one extra interface to ensure that we have one interface for the AMF (N2) and another for the UPF (N3). Since this solution is deployed in a single instance, the UPF is implemented with the rest of the Core functions, which means that we could have one single interface that would direct the traffic coming from the gNB to the proper Core function. In terms of networking, we have one management sub-network, one sub-network called Telco for the AMF and UPF interfaces, and two sub-networks for the Data Networks (Data Network and the NAT Network).

As we can see in the Figure 4.4, there is a Network Address Translation (NAT) VM, and as its name suggests, it is responsible for translating all UEs' IP traffic that leaves the 5G Core instance through the N6 interface. This procedure is mandatory since the IP address that can be allocated to an UE does not integrate any sub-network in the AWS deployment, and if the AWS cloud receives any packet with an IP address outside the existing sub-networks in its environment, it discards packets. Furthermore, in Figure 4.4, there are also two gateways, the Internet and Virtual Private Network gateways. The Internet gateway provides Internet access to any instance that has a route to that gateway. The Virtual Private Network gateway is responsible for providing the connection between the Core and the RAN.

All the traffic flows arriving in the UPF from the UEs leave through the N6 interface and are forwarded to the NAT VM. All the traffic going to the UEs have to be forwarded to the NAT VM before going to the N6 interface.

4.2 Design and Specification - Requirements and Performance

In this section, we will describe the equipment used for the deployment of the RAN, and also some specifications of said equipment. We will also describe some requirements for the deployment on a Cloud environment. Some performance specifications will be also shared.

4.2.1 RAN System

In terms of the equipment used for the RAN deployment, Table 4.1 presents a high-level overview of what was used to implement the AirSpan solution.

Table 4.1: Required Equipment for RAN Deployment

Equipment	Description	Comments
AirVelocity 2700	Airspan Indoor Radio Unit	Operating in the n77 band (3800-3900 MHz)
COTS Server	O-DU Server	Only the O-DU runs in this server
COTS Server	O-CU Server	ACP, vCU and k8s Master VMs run in this server through a hypervisor
Router with Switch Capabilities	RAN Router	Router that organizes the RAN network, and supports all traffic that goes through the O-RAN
Router	Cloud Router	Router that connects the RAN to the Core
GPS Antenna	Outdoor GPS Antenna	For PTP synchronization, and included with the RAN Router
SFP+ Cable	25 Gbps SFP+ Cable	For RAN Connections
SFP+ Cable	10 Gbps SFP+ Cable	For RAN Connections
RJ45 Cable	Copper Cable (Ethernet)	For RAN Connections
SFP+ Fiber Cable	10 Gbps Single Mode Single Fiber	Cables used for connecting the O-RUs to the RAN router

In terms of radio specifications, Figure 4.5 shows the radio pattern of the O-RUs. Table 4.2 presents some specifications about the AirVelocity 2700 radio unit.

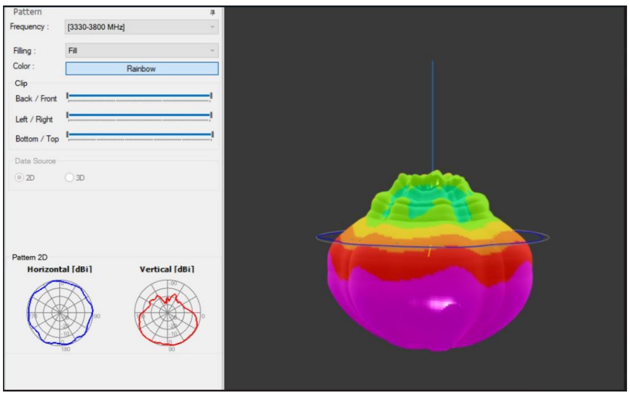


Figure 4.5: Radio Pattern of the AirVelocity 2700 RU

Table 4.2: AirSpan Velocity 2700 Specifications

Main Specifications	
Tx Power	4x 250 mW (4x 24 dBm)
Total Tx Power	1 W (30 dBm)
Antenna Gain	6 dBi
Total EIRP	36 dBm
vRAN Split	7.2a
MIMO Layers DL/UL	2 DL / 1 UL (future software upgrade for 4 DL / 4 UL)
BW Support	20 / 40 / 100 MHz
DL Modulation Schemes	QPSK / 16 QAM / 64 QAM / 256 QAM
UL Modulation Schemes	QPSK / 16 QAM / 64 QAM
Operating Band	3800-3900 MHz
Available Frequency Band	3700-4000 MHz
Antenna Configuration	Integrated / Pointed to the floor / 160º radio pattern
Fronthaul	1x 10 Gbit/s Copper & 1x SFP+
Power Consumption	40 W

In terms of performance, Table 4.3 presents the performance and scale specifications of these AirSpan 5G O-RAN solution.

Table 4.3: AirSpan Solution - System Scale and Performance

Description	**Performance**
Max Number of 5G Devices per O-RU	32
Max Number of O-RUs per O-DU	3
Max Number of O-DUs per O-CU	2
Max Number of O-CU-UP per O-CU-CP	1
Max DL Throughput per O-RU (100 MHz)	630 Mbit/s
Max UL Throughput per O-RU (100 MHz)	64,3 Mbit/s

4.2.2 5G Core System

Since the Athonet 5G SA Core was deployed in a cloud environment, some AWS instance-wise requirements need to be ensured. The Core instance needs at least a *t3.xlarge*, characterized by having 4 vCPUs, 16 GiB of memory. The storage size value selected was 41 GiB.

As we saw in Figure 4.4, for this solution we need to create three virtual routers, one for granting access to the Core from the gNB, another to connect the Data Network sub-network to the NAT VM, and another to connect the NAT VM to the Core and to the Internet. The NAT VM just needs to be a Linux based instance. In our deployment, we used for this VM an *t3.small* instance, which is characterized for having 2 vCPUs, 2 GiB of memory, and we chose a storage size of 20 GiB.

4.3 Implementation

The purpose of this section is to describe the main steps of implementation and deployment of this 5G testbed network. It will focus on the criterias, mandatory conditions, and also will highlight some issues and limitations.

4.3.1 RAN - On Site Installation and Assembling

In terms of physical assembling and wiring, the following figure represents the connections used to deploy the RAN, from a physical perspective. This 5G installation was made for an indoor scenario, and for a building with four floors.

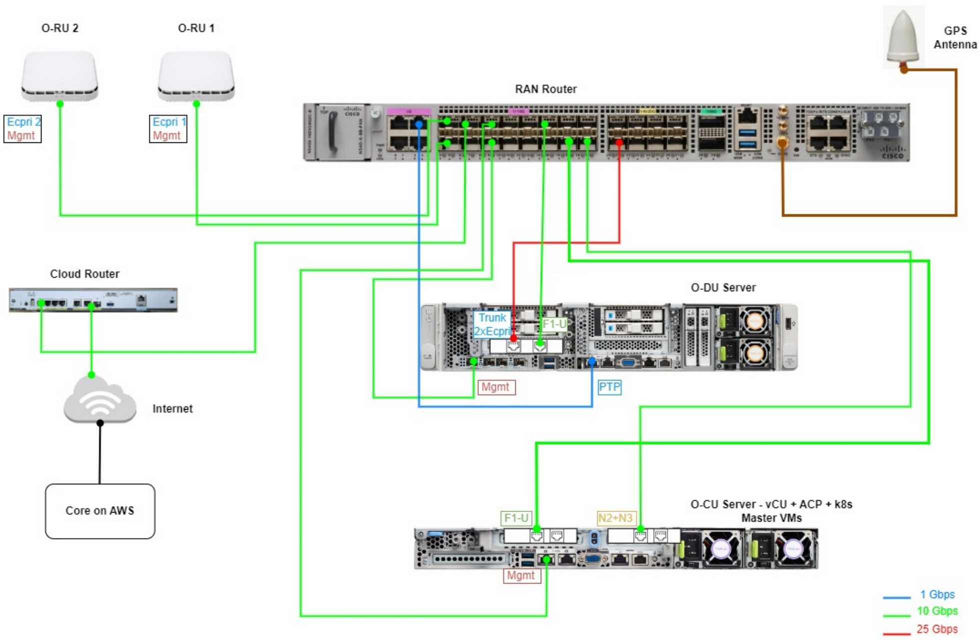


Figure 4.6: RAN Physical Diagram

Some important detail that need to be mentioned are:

- The distance of each O-RU to the O-DU is 40 and 60 m. One O-RU is installed in the second floor and the other in the third. Both are mounted in the ceiling;
- The Global Positioning System (GPS) antenna has a distance of 65 m to the RAN router, and it is mounted outdoor, in the rooftop;
- Besides the O-RUs and the GPS antenna, all the routers and COTS servers are installed in a Datacenter in the building, in the third floor;
- The connection of the O-RUs to the RAN router is made through single fiber in single mode, so it is mandatory to use fibre transceivers and receivers with different wavelengths for each endpoint of the connection;

- Since each O-RU has a single fiber, both eCPRI and management connections go through that fiber. Because of this, the O-RU has two sub-interfaces, and in the RAN router, we have two sub-interfaces in the corresponding ports that are connected to the O-RU;
- The management connection of the O-RUs also includes the synchronization plane, and it is through this connection that the PTP packets reach the O-RU;
- To ensure the PTP synchornization, the RAN router has connected to itself a GPS antenna to provide time synchronicity to the O-RUs and the O-DU;
- As mentioned, the F1-c connection between the O-DU and the O-CU is done through their management connection.

4.3.2 RAN - Networking Configuration

In this subsection, some details about both routers will be shared, focusing on configurations that needed to be done for the proper functioning of the 5G Network. As mentioned, the cloud router has the purpose of establishing the connection to the AWS cloud where the 5G Core was deployed. The RAN router, which includes switch functionalities, has the purpose of organizing the RAN.

For the organization of the RAN, the following configuration was made in the RAN router:

- **Bridge Configuration** - created a bridge domain for each sub-network shown in Figure 4.2, and associated the proper router's ports (its sub-interfaces) to the different bridges;
- **Port Configuration** - in this architecture, ports can be in access or trunk mode. For trunked ports, the traffic needs to be tagged with the correct VLAN tag. Access mode ports use untagged encapsulation. An important detail about this AirSpan Solution is that the management connection between the O-RUs and the RAN router needs to be configured in Native VLAN mode. This meaning that the correpsonding sub-interface of the router's port need to encapsulate the traffic in untagged mode but it continues to be part of the management bridge group. This is required because the PTP packets go through this management connection;
- **PTP Configuration** - this configuration is mandatory to ensure PTP synchronization in the O-RUs and in the O-DU. One aspect important to mention is that the way the PTP synchronization is implemented in this architecture is not completely aligned with the O-RAN Alliance norms, because the router provides the clock to the O-RUs and not the O-DU. This meaning that both O-RUs and the O-DU are slaves. Some important details that need to be configured are:
  - The type of clock the router is - grandmaster;
  - The type of ITU-T Telecom PTP profile used - this is specified by the AirSpan solution;
  - The time-source being used by the grandmaster clock - GPS;

  - The transport of PTP packets is made over Ethernet as discussed in Section 2.4;
  - The sending of PTP packets is made in multicast mode;
  - The router ports connected to the O-RUs and the O-DU need to be configured as master-only ports.

- **Global Navigation Satellite System (GNSS) Configuration** - configured to do its survey by gathering satellites from GPS, Galileo and Glonass to improve accuracy.

For the cloud router, and having the main goal of connecting the RAN to the Core, the main configuration made in this router was to create an IPsec tunnel between this router and the AWS cloud. All the traffic that arrives from the Core will be forwarded to the RAN router to reach the O-CU UP.

4.3.3 RAN - Servers Provision (O-DU and O-CU)

The following paragraphs describe the various steps that needed to be made to provision the COTS servers before deploying AirSpan vRAN solution. The focus will be in the more important details that needed to be implemented to support this AirSpan solution.

4.3.3.1 Kubernetes - vRAN deployment

The deployment of the vRAN, meaning all the radio functions hosted in the O-DU and the O-CU, are deployed using k8s. The vRAN solution of AirSpan is inserted in a k8s cluster composed by 3 nodes, as we can see in the following figure.

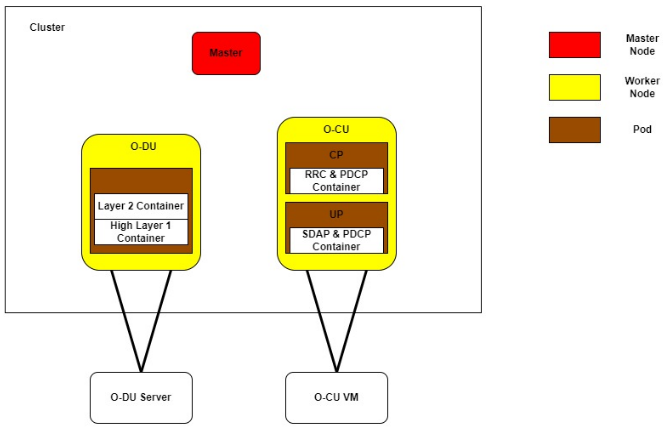


Figure 4.7: k8s Deployment Diagram

In this deployment, we have a master node that controls the k8s environment, mainly each worker node. We have two worker nodes, the O-DU and the O-CU. Both master and worker nodes can be hosted in VMs, or bare metal servers. In our case, the k8s master and the O-CU node are hosted in VMs, while the O-DU is hosted in a bare metal server.

The k8s master represents the control-plane, while the worker nodes represent the data-plane, where the applications runs. In this case, the k8s master will deploy the different radio functions in the proper worker node. In k8s, applications run in containers, but also the container needs to be encapsulated in a pod. A pod is the unit of deployment in k8s technology. The AirSpan solution possesses 3 pods. One pod is deployed in the O-DU worker node and possesses the layer 2 and layer 1 functionalities running in separate containers. The other 2 pods are deployed in the O-CU worker node, having one pod for the O-CU CP functionalities running in one container, and the other pod with the O-CU UP functionalities, running in a single container. To operate the containers, the hosts that house the worker nodes need to have a container runtime. In AirSpan case, they use Docker. The deployment of the pods in each worker node is done using HELM, which grouped all configuration files in a single package, deploying not only the pods but also running the containers within the pods according to the configurations of the YAML file. This file has information about the different networks showned in Figure 4.2, for example.

4.3.3.2 O-CU Server Provision

The O-CU server, as seen in Figure 4.3, hosts the k8s master, the O-CU, and the ACP. Each entity is deployed in a VM. To create the different VMs, we used kVM, as our hypervisor. This hypervisor needs to be installed over a Linux-based OS, for example CentOS Linux. In terms of VM deployment, each VM was deployed following the AirSpan requirements (in terms of OS installation, and network interfaces), which are presented in their manuals.

For the O-CU provision, we created a VM with the CentOS image, and in its installation followed the partitions required by AirSpan. After having the VM running, we downloaded and installed an AirSpan package with all system dependencies (for example Docker) to deploy the AirSpan CU. The following step in the provision of the O-CU VM was to open via *firewall-cmd* command a group of ports for k8s management. The ports that are opened are set by AirSpan. The last step of the O-CU VM provision was to join the O-CU Worker node to the k8s cluster.

Before joining the O-CU worker node in the k8s cluster, we first needed to deploy the k8s cluster in the k8s master VM, which requires a series of steps to be done in the k8s master VM (which also has a CentOS Linux OS). The whole process of deploying a k8s cluster can be found online in multiple websites [28].

4.3.3.3 O-DU Server Provision

The O-DU server requires a different set of steps to provision it, mainly because the O-DU was deployed in a bare metal server, and there are some details like PTP that needed to be configured in this entity. In terms of provisioning in the O-DU, these were the main steps done:

- **Basic Input/Output System (BIOS) Configuration** - change specific values in the BIOS setup of the server, following the guidelines given by AirSpan;

- **OS Installation** - installation of a CentOS Linux image, given by AirSpan, following their guidelines in terms of partitions and base environment installation;

- **Dependecies Installation** - download and install all dependencies required to deployed the AirSpan O-DU;
- **O-DU Cores Isolation** - need to define the right set of cores to be isolated for the AirSpan application;
- **Server CPU Frequency** - need to update the CPU frequency to ensure the proper operation of the O-DU node and the accelerator card used in the COTS server;
- **Disabling Online Sync Sources** - disable all time sources to ensure that the main time is synced to PTP;
- **PTP Synchronization** - configured the PTP synchronization (*ptp4l* and *phc2sys*) to a specific physical interface in the server and as a slave port;
- **O-DU security** - open a group of ports using *firewall-cmd*, for k8s management, as done in the O-CU and k8s master VMs;
- **Dual Socket Servers** - ensure that the network interface card and the accelarator card for the physical-layer functions in the O-DU are working with the same isolated cores;
- **Enabling SR-IOV** - following AirSpan's guidelines, and if our deployment has multiples O-RUs, the SR-IOV needs to be enable for multiples O-RUs to communicate with a single network interface card in the O-DU server. In this case, the eCPRI interfaces in the O-DU were virtualized;
- **Kubernetes** - join O-DU worker node to k8s cluster.

4.3.3.4 ACP Installation

The ACP, which is the AirSpan platform for orchestration and management of the O-RAN, was deployed in a VM in the O-CU server, as already described. The setup of VM was done following the AirSpan guidelines for VM requiremnts, in terms of vCPUS and RAM memory. This VM runs an Ubuntu-Linux OS, with 50 GB for OS partition and 350 GB for Data Partition. After this initial VM setup, the following task were done:

- Structured Query Language (SQL) Server Installation - this server is needed as a database for storing all data about the O-RAN, including states of different O-RAN entities, and performance counters/KPIs;
- ACP installation - installation of AirSpan software to manage the O-RAN. This software provides a Web API, for easy access to all management tools;
- DHCP Server Installation - this DHCP server is needed to allocate management IP addresses to the O-RUs (AirVelocity 2700 RUs). In this case, the IP address allocation was done by attaching a static IP address to the corresponding O-RUs identified by their MAC addresses.

4.3.4 RAN - O-RAN Deployment

After provision both servers, setup the k8s cluster, and install AirSpan O-RAN management platform (ACP), we deployed, commissioned and provisioned the AirSpan vRAN (O-RAN Deployment). While following the AirSpan guidelines, these next actions were taken to ensure a sucessful deployment of the vRAN:

- **Load the Docker Images** - load the Docker images to the worker nodes. With these images, we can run the containers that will perform the RAN functionality of the O-DU and the O-CU. In this case we have 4 Docker images, 2 for the O-DU and 2 for the O-CU. As already mentioned, the O-DU worker node has one pod running two containers, one for running the High-Physical Layer and the other for running the Layer 2 (MAC and RLC). The O-CU worker node has two pods running, each one with one container. One runs O-CU CP functions (RRC and PDCP) and the other O-CU UP functions (SDAP and PDCP);

- **Load HELM packages** - load the O-CU and O-DU HELM packages to the k8s master. These HELM packages were created by AirSpan to launch their pods more efficiently. They contain configuration files that ease the deployment of the vRAN;

- **Modify HELM Files** - each HELM package contains YAML files for each pod. These files need to be configure with the proper values in terms of gNB identification, IP addresses of the interfaces, Destination Network and Gateways for midhaul and backhaul connections, and VLAN tags as seen in Figure 4.2. These were configurations made for the O-CU in relation to the F1 and Core connection. For the O-DU, the configuration was more focused on the O-RUs, in terms of number of O-RUs connected to the O-DU, the MAC addresses (source and destination) for each eCPRI connection, and the VLAN tags being used for each eCPRI connection, which considers the Figure 4.2. In the O-DU we also configured the F1-u interface, in terms of source IP, destination network and gateway, and VLAN tag;

- **Deployment of the vRAN** - after the configuration of the HELM files we used a HELM client to launch the pods, effectively launching the AirSpan vRAN. To ensure the pods were running we checked its status via *kubectl* commands.

After the deployment of the O-RAN, the next phase was the provision of the O-RAN in the ACP, which is comprised in these main steps (all done in the AirSpan platform and following their guidelines):

- **MNO Configuration** - The first thing that needed to be provisioned was the Mobile Country Code (MCC) and Mobile Network Code (MNC). In our case this is a Laboratory Testbed and is identified with the code 001 and 01, respectively. We also needed to provision the slice being used in terms of Slice Service Type and Slice Descriptor.

- **Node Discovery** - In this step, we created the gNB, giving its name, MNO information and identification number. Even more relevant, we checked if all O-RAN entities had a clear connection to the ACP through the management plane, more precisely, through the O1 interface, which is responsible for the orchestration and management of all O-RAN entities [3]. After that, the gNB was registered in the ACP as a viable node;

- **gNB Components Configuration** - in this step we needed to configure several characteristics of the gNB, such as:
  - types and number of slices being used;
  - in the O-CU CP, we needed to provision the endpoint of the backhaul connection, which was resumed to define the IP address used by the AMF. We also needed to create 2 Cells, one for each O-RU, where we configured the cell number and local cell ID;
  - in the O-CU UP, we needed to provision the E1 connection, which is the connection between the O-CU CP and O-CU UP [3];
  - in the O-DU, we needed to provision both F1-c and F1-u endpoint connection, but most importantly, we needed to configure the O-DU cell properties for each cell. It is in this configuration that we setup the type of band, the bandwidth, the central frequency, the start frequency being used by each cell.

After this configuration, we needed to send these information to the gNB. The ACP does this through an action called re-provision. After a sucessful re-provision, the last step of this deployment was to re-deploy all pods in the cluster, in order to apply the configurations that were made in the ACP.

To conclude the deployment of the O-RAN, we needed to check a few details, including:

- in the O-CU CP, both E1 and NG connections were up and running;
- in the O-DU, both cells were running and also the High-Physical layer functionality was running;
- in both O-RUs, we needed to ensure that the PTP status was locked, and that the RF power of channel 1 and 2 should be different from zero.

4.3.5 Core - Athonet API Configuration

The 5G SA Core was deployed in a Cloud service. In this case, the Athonet solution was deployed in a AWS instance, as already described in Subsection 4.2.2. The Athonet software provides an web API to configure the 5G SA Core through a web browser. In this subsection, we are going to describe the configuration done to get the Core up and running:

- **Network Configuration** - create the interfaces for the N2, N3 and N6 interface. For each interface associate its IP adress, and routes. In the case of the N2 and N3 interfaces, we needed to establish a route for the gNB, and in the case of the N6 interface we needed to provide a default gateway. We also created a tunnel for the user-plane that is used in the PDU sessions for the 5G UEs' data going through the UPF to the gNB;

- **AMF Configuration** - in the AMF Core function, we needed to configure its identity in terms of Network Name, PLMN ID, and also defined the exisiting slices for this PLMN. In our case, we are just working with one slice. In terms of networking, we associated the AMF N2 interface to the one created in the network configuration via IP address;

- **SMF Configuration** - in the SMF Core function, we needed to declare the Data Networks that the UPF will have access to. In our case, we only defined one, which is to access the Internet. Besides declaring the Data Network, we also needed to define the IP pool being used for the 5G UEs. The Data Network needed to be associated to the type of a slice that is going to be used by a specific UPF to access said Data Network. Lastly, the DSCP to QoS Identifier mapping needed to be configured. This configuration helps define the different parameters for QoS Priority in the UPF's QoS flows;

- **UPF Configuration** - in the UPF Core function, we needed to associate its N3 interface to the interface created in the network configuration via IP address. We needed to associate the user-plane tunnel created in network configuration with the Data Network, so that the 5G UEs can establish PDU sessions to the Data Network through that tunnel;

- **UDR Configuration** - in this Core function, we provisioned the SIM cards used by the 5G UEs. For each SIM card, we can define different usage profiles (Provisioned Data Profile). And in this profile, we can configure:

  - Downlink and uplink throughput for a specific UE;
  - Slice used by a specific UE;
  - Data Network the UE wants to communicate with;
  - QoS Identifier the UE wants to use - crucial for QoS Priority;
  - PDU session type allocated to a specific UE - in our case we configured for IPV4;

Chapter 5

Testing - Scenarios and Methodology

In this chapter, our focus will be on describing the use case that was used to assess our 5G deployment, and also explain the various test scenarios that were performed and its motivations.

5.1 Automated Guided Vehciles (AGV) - Use Case

Our thesis will be focused on the AGV use case, which is linked to the uRLLC use cases. In Section 3.3, we saw different 5G deployments to support this use case. Furthermore, these articles gave an approach to begin our work in terms of deployment scenarios, and the relevant KPIs that needed to be evaluated to assess the performance of the use case.

The AGV scenario that we intend to work with is similar to the one described in Subsection 3.3.2, where we have an AGV that comunicates and is managed by the controller through the 5G Network, meaning that both entities will be UEs in our network. In terms of data flows, the AGV will be uploading a video stream, sending sensor information (UL), and receiving control commands from the controller (DL). At the same time, the controller will watch the AGV's video stream, and control the AGV.

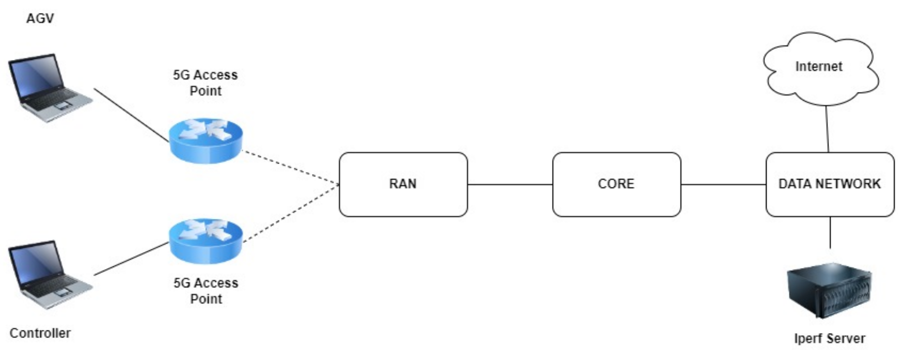


Figure 5.1: Solution Architecture

Due to time constraints, it was not possible to get a 5G vehicle device and we had to emulate an environment that recreates the AGV behaviour and also induces the same stress that a real AGV would induce in a 5G network. As we can see in Figure 5.1, we used two laptops connected via Ethernet cable to different 5G Access Points, directly attached to the 5G Private Network, one for AGV and the other for the controller. They are connected via Ethernet cable to remove Wifi connection limitations (interference for example) between the laptops and the 5G Access Point. The 5G Access Points being used are from the AirSpan product line (AirSpot 9621).

To emulate the data flows that were explained before, in the case of the AGV, we used Iperf to generate traffic between the AGV's laptop and a server (NAT VM) residing in the cloud, on both downlink and uplink directions. For the Controller, we choose to run an online multiplayer game called Counter-Strike Global Offensive, that requires low latency and low percentage of lost packets. For the video stream, we used the Twitch application, where the AGV will have the role of streamer (streams its webcam image) and the controller the role of viewer. In this way, we can ensure that all data flows and 3GPP AGV use case requirements are being implemented by this emulation.

5.2 Test Scenarios - Motivation and Realization

In this section, we will describe the different scenarios that were created to assess the end-to-end 5G network. These scenarios were established to emulate different conditions that could be present in an AGV use case. The test scenarios will be described in terms of motivation and realization.

5.2.1 Default

The default scenario is the basic scenario that emulates one single AGV communicating with the controller. This scenario will provide the anchor to compare with other test scenarios.

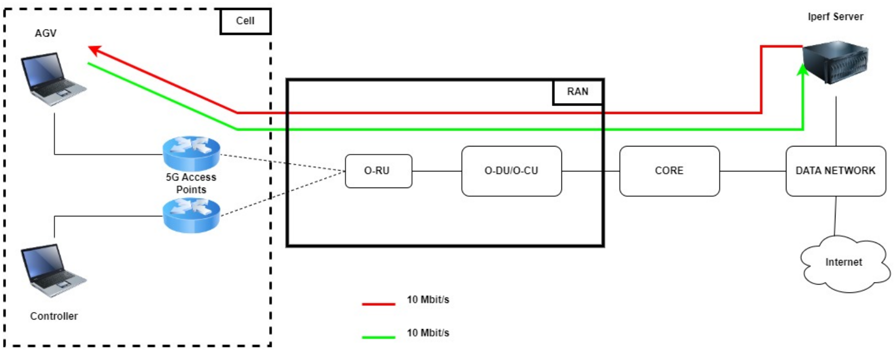


Figure 5.2: Default Scenario

For this scenario's setup, the AGV generates traffic from itself to the NAT VM, which will be our Iperf Server. In this way we have an uplink data flow from the AGV to the endpoint of the Core, so it goes through the entirety of the 5G Network, including the 5G Core user-plane. In the downlink the same situation will happen, the data flow is ensured by generating traffic from the NAT VM to the AGV. This traffic is symmetric meaning that the same quantity of data going in the uplink direction will also go in the downlink direction. As mentioned above, the tool used to generate this traffic was Iperf. In this case, we generate UDP traffic with a rate of 10 Mbit/s and with packet size of 1300 Bytes, in both directions. One important detail to mention about the AGV is the reason for using a rate of 10 Mbit/s to generate traffic. This value is used because the AGV use case is grouped in the Discrete Automation use cases described by 3GPP and mentioned in Section 2.3.3. Discrete automation is a use case scenario for industry applications, mainly controlling equipment/robots. As we can see in Figure 2.10, the experience user data rate is 10 Mbit/s.

```
ubuntu@ip-172-16-4-20:~$ iperf3 -u -c 10.100.0.2 -b 10M  -t 240 -l 1300
```

Figure 5.3: Iperf Command in NAT VM

For the video stream, as mentioned, we used the Twitch application, called Twitch Studio. In terms of specifications of the video streamed by the AGV, the video stream can be characterized by having a frame rate of 30 frames/s, an uplink rate of 1,5 Mbit/s, and video quality with 420p resolution.

For the controller, in terms of video stream, it has to open a web browser and access twitch.tv slash the username of the streamer (AGV). In terms of the emulation of controlling the AGV, the controller runs the online multiplayer game Counter Strike, which as an uplink rate of 2 to 3 Mbps.

5.2.2 Uplink Congestion

This test scenario has the purpose of evaluating the capacity of one O-RU, meaning that from this scenario we want to find out the uplink congestion threshold but, more importantly, how many AGVs can one cell support in this 5G deployment. To evaluate this scenario, we maintain most of the conditions of the default scenario, but we increment the rate of the traffic being generated in both uplink and downlink directions. The values used were 30, 40, 50 and 52 Mbit/s. We have chosen these values so as to have multiples of 10, and this value 52 Mbit/s was chosen because, as we are going to see in the next chapter, it was identified as a true uplink limit.

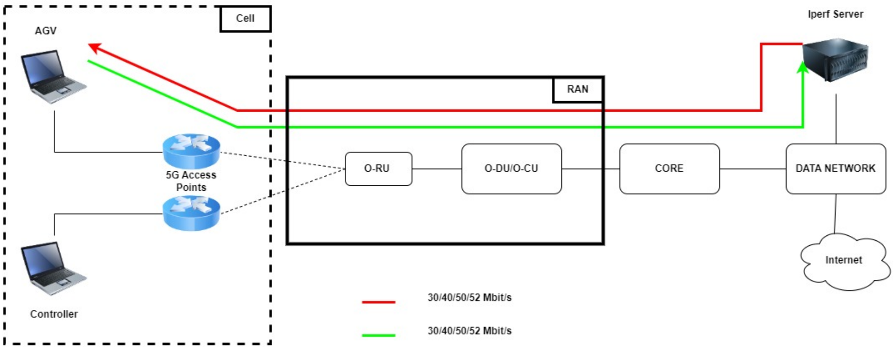


Figure 5.4: Uplink Congestion Scenario

5.2.3 Coverage Simulation through Modulation Control

In this test scenario and due to device static conditions, we try to simulate the impact of coverage conditions to conclude the performance constraints in the performance of the AGV. To simulate these conditions we can only change the downlink modulation in the ACP. One important detail that needs to be mentioned is that ACP only allows to manipulate the downlink modulation, and this modulation can only change in best case logic, this meaning that we can choose the best case modulation scheme achieved, but not lock the modulation scheme. In the AirSpan case, we can enable or disable 256-QAM downlink modulation per Cell, which means that when this option is enabled, in the best case scenario we can get 256-QAM downlink modulation, and when this option is disabled, we can get in best case scenario 64-QAM downlink modulation through the radio channel.

5.2.4 Different Cells

It is possible that, on a factory floor, the AGVs and the controller's operation center are not located in the same space. The distribution and possible UL congestion from a large number of AGVs in the same place cannot affect the controller.

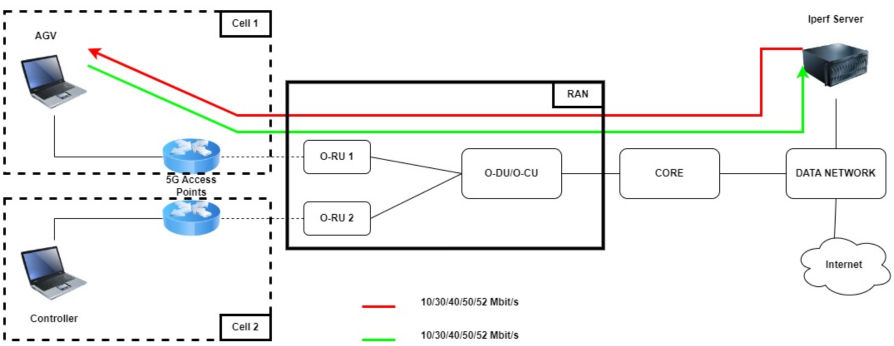


Figure 5.5: Different Cells Scenario

To implement this scenario we just need to attach each 5G Access Point to separate O-RUs, as we can see in Figure 5.5. Also, with this disaggregation, we can distinguish the impact of the controller and the AGV in the network, in terms of traffic load.

5.3 Test Methodology

To initiate an attempt of a test scenario, the AGV begins its stream, and then we ensure that the controller has access to the AGV's video stream, and can see the AGV's image. After this, the controller initiates the online multiplayer game. Once both of these services are stabilized, we can start to generate, in the Iperf server (NAT VM), the downlink traffic that reaches the AGV, and in the AGV, the uplink traffic that reaches the Iperf Server. Several attempts were done and each one is characterized by generating traffic for 4 minutes 5 times, meaning that each attempt will take up to 20 minutes. Every 4 minutes, the test KPIs will be retrieved for a post processing and analysis of the network and the use case behaviour.

The different cells scenario guarantees that the AGV generates continuous traffic, since the AGV is on a different floor in relation to the controller, due to the physical position of the O-RUs. The downlink traffic continues to be generated for 4 minutes, and 5 times.

To summarize, the uplink congestion, the coverage simulation and the different cells scenarios, all have their anchor in the default scenario. But, all of them are characterized by having attempts with different traffic rates (10, 30, 40, 50 and 52 Mbit/s) in the AGV data flow with the Iperf server. For the coverage simulation scenario, we used in the best conditions a 64-QAM modulation in the downlink, and not 256-QAM. Lastly, for the different cells scenario, the AGV and the controller are attached to different cells.

Chapter 6

Results

The focus of this chapter is to showcase the results gathered from the attempts made in each scenario described in Chapter 5. Before presenting the results, we need to describe the KPIs used to assess the performance of the use case and the 5G Network. After that, the chapter will be organized by test scenario, showcasing the results of each one. These results will help assessing the 5G deployment used, in the next chapter.

6.1 5G Network's Limitations

Before starting to describe the KPIs and present the results from the test scenarios, we will address a limitation found in the final phase of deploying the 5G Network. As mentioned in Chapter 4, this 5G architecture has its Core deployed in a Cloud service. In this particular case, the datacenter used from this Cloud service is located in Ireland. Since the connection between the RAN network to the Core network is done through the Internet with no dedicated path, the latency of the path can reveil itself as an issue.

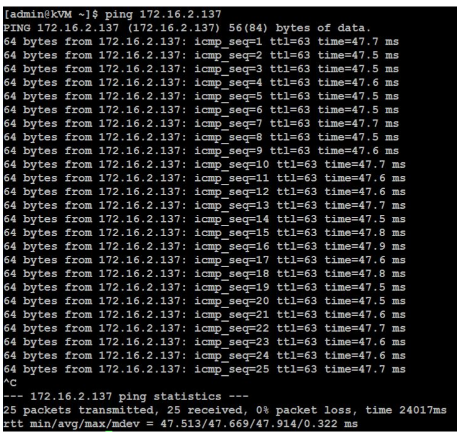


Figure 6.1: Latency brought by this Core on Cloud Deployment

In this case and from doing a simple test using the *ping* command, we registered an average of 48 ms of round-trip time between the endpoint of the gNB closer to the Core (O-CU) and the endpoint of the Core closer to the gNB (AMF/UPF).

To reduce this latency, there are three options, and throughout each one, the implementation and operation cost tends to increase. The options are:

- Change the Core to a closer datacenter, keeping the implementation of the Core full on Cloud;
- Implement a hybrid deployment, with the UPF on-premises and keep the CP on Cloud;
- Implement a physical deployment, with all Core functions on-premises;

In our case, and during this dissertation, none of these options were possible. The deployment of the Core on cloud was not possible in a location closer to the RAN, because the AWS account being used is placed in Ireland. The other two options were not possible due to the lack of capacity in the existin on-premises servers to install the UPF or the whole Core on-premises.

Nevertheless, the first and second option should be the ones to explore in this 5G deployment, since we have a end-to-end latency of approximately 25 ms, and in terms of latency we could mitigated to 15 or even 10 ms with another datacenter location or below the 10 ms mark with an UPF on-premises, as we saw in Subsection 3.3.1.

With previous information, this 5G deployment cannot support the latency requirement of an AGV use case, according to Figure 2.10 (Discrete Automation usage scenario), but this architecture with a 5G Core implemented on Cloud, in terms of latency, could be a possibility if the location of the datacenter ensures a low latency path to the RAN. The hybrid deployment of the Core can be the follow up solution, if the latency is a key requirement. In my opinion, a Core fully on-premises should not be deployed only for latency reasons. Topics like security and end-to-end closed environments are more relevant for choosing a Core on-premises.

Since latency is a huge bottleneck in this architecture, we will not analyze this KPI for the following results presented in this chapter.

6.2 KPIs - Description

The KPIs considered are the following:

- **Throughput** - Iperf's measure of throughput based on the amount of data that a network transports. Its units are Mbit/s;
- **Packet Loss** - Iperf's measure of packet loss based on the amount of packets received at the endpoint versus the total of packets sent to the same endpoint;
- **Skipped Frames** - Number of dropped video frames related to network constrains - related to the video stream (Twitch platform);

- **PRB Usage** - Extracted from ACP KPI's dashboard, and according to Airspan, this counter provides the percentual usage of PRBs on the downlink or uplink directions;
- **Modulation Scheme Usage** - This measure represents the percentage of the modulation scheme used to transport data through the radio interface channel, between the O-RU and the UE. ACP provides the number of transport blocks that were used for each modulation scheme, during a certain period of time.

In order to organize the data, tables and graphics were created for a better understanding of the AGV use case performance. Some tables include some abbreviations:

- **Service Crash** - this represents the exhaustion of the uplink radio resources, which crashes the applications running in the AGV and controller;
- **SF** - Stream Freeze - this means that some of the video stream statistics could not be retrieved because the stream was frozen, in the moment of gathering the data;
- **N/A** - No Measure - this happened in a specific attempt where the data was retrieved only 4 times, since the AGV stopped generating uplink traffic. The AGV's uplink traffic started way before the start of the attempt, so the attempt needed to be ended.

As we present the results, the expression that the AGV and the controller lost their connection to the network will be highlighted. For the scenarios where both are attached to the same cell, this will mean that the congestion driven by the AGV's traffic load, clashes the controller and the AGV to not be able to send data from their running services (game and video stream), which makes the services alert about a network connectivity problem. This is the main cause for the term lost connection to the network to be used. Also, when this phenomenon happened throughout the testing phase, we would considered that the attempt was finished, but what was seen is that they could restore their connections. For the last scenario, where the AGV and the controller were attached to different cells, we are going to see that the same happened to the AGV when the traffic load reaches the uplink congestion threshold, but in this case, we let the AGV "*reconnect*" to the network.

The tables that come next will present the value obtained for a specific KPI, in the end of each reading, in each attempt for each test scenario.

6.3 Default Scenario

Table 6.1: Throughput UL - Default Scenario (10 Mbit/s)

Throughput UL (Mbit/s) - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	9,9	9,89	9,91	9,89	9,9	9,898
2	9,91	9,92	9,9	9,87	9,88	9,896
3	9,9	9,92	9,92	9,93	9,91	9,916

Table 6.1 exhibits the uplink throughput registered in the Default Scenario.

Table 6.2: Packet Loss UL - Default Scenario (10 Mbit/s)

Packet Loss UL (%) - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0,93	1,1	0,9	1	0,92	0,97
2	0,85	0,79	0,92	1,2	1,2	0,992
3	0,94	0,81	0,81	0,69	0,81	0,812

In Table 6.2, we registered values of packet loss in the uplink direction that round the 0,9%, which means that the packet loss will be higher than what is required from this type of use case, for reliability.

Table 6.3: Skipped Frames - Default Scenario (10 Mbit/s)

Skipped Frames - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	1	1	0	4	0	1,2
2	0	0	0	4	1	1
3	4	2	1	0	1	1,6

Table 6.3 present the skipped frames registered in the video stream created by the AGV. As we will see from other results that will be presented, these values are insignificant in terms of performance of the stream.

Table 6.4: Default Scenario (10 Mbit/s)

Default Scenario - 10 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	9,87	9,93	9,903
Packet Loss UL	0,69	1,2	0,925
Skipped Frames	0	4	1,27

Table 6.4 shows the range each KPI reaches for the Default Scenario, and also the average value of each KPI for this scenario. From this table, we can understand that the packet loss registered for this scenario is already very high for this type of use case.

Table 6.5: PRB Usage (ACP Counter) - Default Scenario (10 Mbit/s)

Attempts	UL PRB Usage (%)	DL PRB Usage (%)
1	19	3
2	19	3
3	19	3
Average	19	3

Table 6.5 present the usage of the available radio physical resource blocks for this scenario.

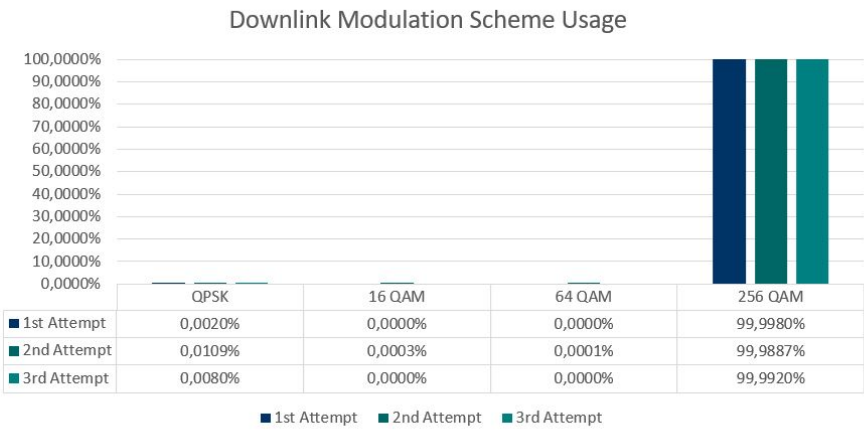


Figure 6.2: Modulation Scheme Usage (Downlink) - Default Scenario (10 Mbit/s)

Figure 6.2 illustrates that the preferable modulation scheme used for downlink communication was 256-QAM. All the other modulation schemes can be ignored, since its usage is low in comparison to 256-QAM.

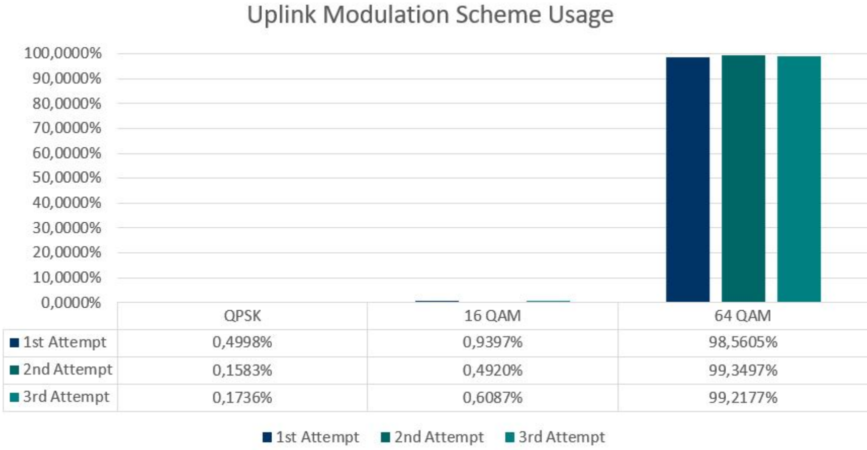


Figure 6.3: Modulation Scheme Usage (Uplink) - Default Scenario (10 Mbit/s)

As mentioned before, the best modulation scheme that can be used in the uplink channel is 64-QAM. From Figure 6.3, we can identify that in normal conditions the modulation scheme used is 64-QAM.

6.4 Uplink Congestion

For this scenario, we will present the results for both set of attempts where the traffic generated by the AGV was 50 and 52 Mbit/s, respectively.

Table 6.6: Throughput UL - Uplink Congestion Scenario (50 Mbit/s)

Throughput UL (Mbit/s) - 50 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	49,9	49,9	49,9	49,9	49,9	49,9
2	47,9	**Service Crash**				47,9
3	49,6	49,9	45,5	**Service Crash**		48,33
4	49,9	49,8	49,8	49,9	49,9	49,86

From Table 6.6, we can notice that in the second attempt the AGV and the controller lost connection to the network, in the end of the first reading. We also identify the same behaviour in the third attempt, where the controller and the AGV lost connection to the network in the end of the third reading. For both attempts, the cause for this disconnection is the congestion provoked by the traffic generated by the AGV.

Table 6.7: Throughput UL - Uplink Congestion Scenario (52 Mbit/s)

Throughput UL (Mbit/s) - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	48,7	45,7	48,5	48,1	51	48,4
2	47,9	**Service Crash**				47,9
3	49,6	51,5	51,5	51,4	49,5	50,7

Table 6.7 shows a similar behaviour to Table 6.6. The second attempt presents a loss of connection by the AGV and the controller in the end of the first reading. We can see that the throughput is below the desired throughput of 52 Mbit/s, which was not the case for the 50 Mbit/s scenario above, and this will show in the percentage of packet loss.

Table 6.8: Packet Loss UL - Uplink Congestion Scenario (50 Mbit/s)

Packet Loss UL (%) - 50 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0,22	0,21	0,22	0,22	0,21	0,216
2	4,5	**Service Crash**				4,5
3	0,85	0,19	8,9	**Service Crash**		3,31
4	0,21	0,25	0,42	0,19	0,22	0,258

As we can notice, from Table 6.8, the congestion of the uplink in the second attempt cause a packet loss of 4,5%, which impact the AGV and the controller to lose their connectivity to the network. The third attempt had a similar event in the end of the third reading, as we can see by the 8,9% of lost packets, which congested the uplink channel and cause the AGV and the controller to lose connection to the network.

Table 6.9: Packet Loss UL - Uplink Congestion Scenario (52 Mbit/s)

Packet Loss (%) - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	6	12	6,6	7,5	1,8	6,78
2	7,9	**Service Crash**				7,9
3	4,5	0,69	0,91	1,2	4,6	2,38

In comparison to Table 6.8, we can see in Table 6.9 that the packet loss increased, and Table 6.11 will show that this increase impacted the quality of the stream. Furthermore, in the second attempt for generated traffic at 52 Mbit/s, we can see that this packet loss made the AGV and the controller to lost connection to the network in the end of the first reading, due to congestion in the uplink.

Table 6.10: Skipped Frames - Uplink Congestion Scenario (50 Mbit/s)

Skipped Frames - 50 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0	0	0	0	0	0
2	**Service Crash**					
3	0	0	**Service Crash**			0
4	0	0	1	0	0	0,2

Table 6.10 shows us that at 50 Mbit/s of generated traffic, there was not a great impact in the video stream. Which makes sense since the overall packet loss registered was less than 1%.

Table 6.11: Skipped Frames - Uplink Congestion Scenario (52 Mbit/s)

Skipped Frames - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	31	**SF**	47	42	20	35
2	7	**Service Crash**				7
3	39	14	18	22	**Service Crash**	23,25

For Table 6.11, the same said above does not apply, since the number of skipped frames overall was very high, due to the increase in packet loss. In terms of performance, this was shown in the stream, by the distortion and re-buffering of the video.

Table 6.12: Uplink Congestion Scenario (50 Mbit/s)

Uplink Congestion Scenario - 50 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	45,5	49,9	48,99
Packet Loss UL	0,19	8,9	2,07
Skipped Frames	0	1	0,07

From Table 6.12, we can conclude that the mark of the 50 Mbit/s represents the threshold for the uplink congestion, since the packet loss increased exponentially versus what was saw in the default scenario. Also, some attempts were made for the scenario where the traffic generated was around 30 and 40 Mbit/s, and it was registered an average packet loss of 0,76% and 0,31%, respectively. In these attempts the use case ran smoothly.

Table 6.13: Uplink Congestion Scenario (52 Mbit/s)

Uplink Congestion Scenario - 52 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	45,7	51,5	49
Packet Loss UL	0,69	12	5,69
Skipped Frames	7	47	21,75

With Table 6.13, we can conclude that increasing the uplink traffic above 50 Mbit/s will force an uplink congestion that will result in a severe increase of packet loss.

Table 6.14: PRB Usage (ACP Counter) - Uplink Congestion Scenario (50 Mbit/s)

Attempts	**UL PRB Usage (%)**	**DL PRB Usage (%)**
1	68	13
2	60	12
3	63,5	12
4	69	13
Average	65,13	12,5

From Table 6.14, we can identify that the increase in traffic load results in an increase of PRBs used by the use case versus what was seen in the default scenario.

Table 6.15: PRB Usage (ACP Counter) - Uplink Congestion Scenario (52 Mbit/s)

Attempts	**UL PRB Usage (%)**	**DL PRB Usage (%)**
1	74,33	12,33
2	73	13
3	77,5	13
Average	74,94	12,78

For a scenario where the AGV or AGVs are generating a traffic with 52 Mbit/s, the PRB usage is similar to the one seen in the scenario with 50 Mbit/s. There was a small increase in the uplink PRB usage.

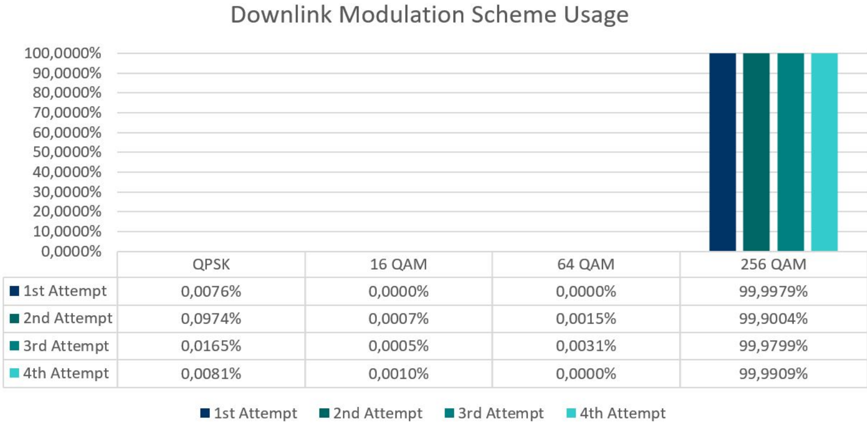


Figure 6.4: Modulation Scheme Usage - Uplink Congestion Scenario (52 Mbit/s)

In terms of downlink modulation schemes, we can see a similar behaviour from what was saw in Figure 6.2, where the scheme used predominately was 256-QAM.

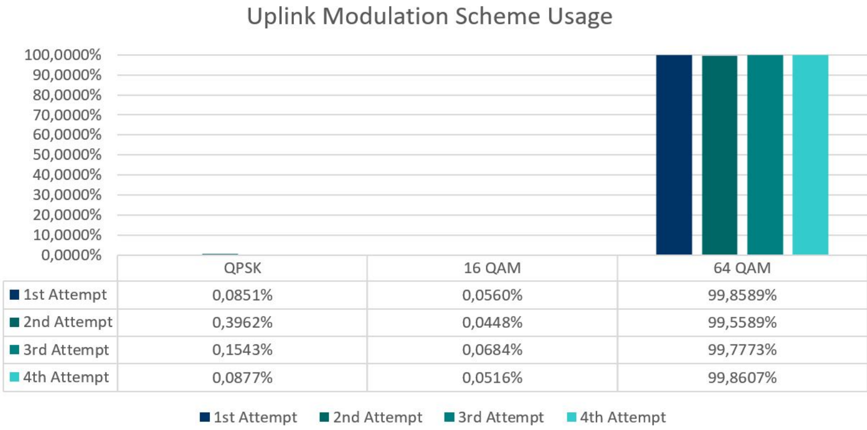


Figure 6.5: Modulation Scheme Usage - Uplink Congestion Scenario (50 Mbit/s)

From the figure above, we can identify that even in a congestion scenario the uplink modulation scheme most used was 64-QAM.

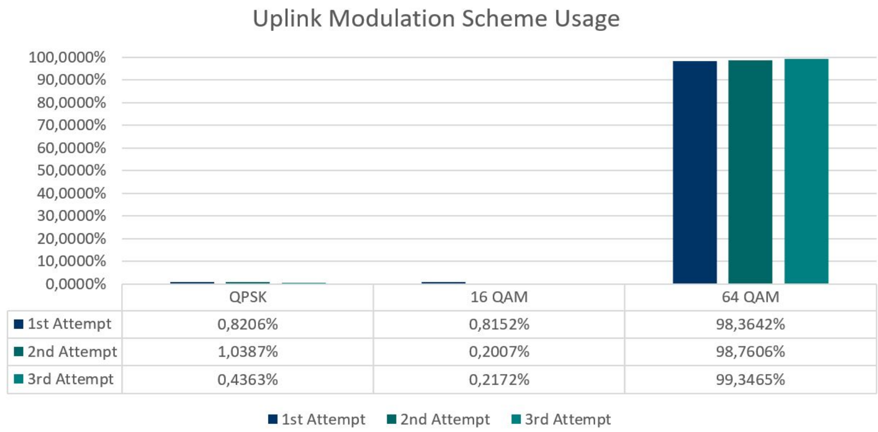


Figure 6.6: Modulation Scheme Usage - Uplink Congestion Scenario (52 Mbit/s)

Even in the worst scenario that was experimented, we do not see a significant impact in the uplink modulation scheme. The chosen scheme continues to be 64-QAM.

6.5 Coverage Simulation through Modulation

For this scenario, we will present different results gathered for the group of attempts where the AGV generated 10 and 52 Mbit/s of traffic.

Table 6.16: Throughput UL - Coverage Simulation Scenario (10 Mbit/s)

Throughput UL (Mbit/s) - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	9,93	9,93	9,93	9,93	9,93	9,93
2	9,91	9,91	9,92	9,92	9,91	9,914
3	9,92	9,93	9,92	9,92	9,92	9,922
4	9,91	9,92	9,92	9,92	9,93	9,922

In comparison to Table 6.1, the values presented in the table above are very similar.

Table 6.17: Throughput UL - Coverage Simulation Scenario (52 Mbit/s)

Throughput UL (Mbit/s) - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	48,5	48	48,8	47,7	47,6	48,12
2	49,3	49,7	49,7	49,8	49,6	49,62
3	48,1	**Service Crash**				48,1
4	49,7	48,7	49,3	49,8	**Service Crash**	49,375

Table 6.17 presents a variability in its values. It is possible to see that, in the third attempt, the AGV and the controller lost connection to the network in the end of the first reading. Also, in the forth attempt, the same happen but in the middle of the fifth reading.

Table 6.18: Packet Loss UL - Coverage Simulation Scenario (10 Mbit/s)

Packet Loss UL (%) - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0,61	0,62	0,64	0,67	0,67	0,642
2	0,81	0,81	0,78	0,76	0,81	0,794
3	0,73	0,69	0,76	0,76	0,76	0,74
4	0,83	0,72	0,74	0,63	0,7	0,724

Table 6.18 present similar values of packet loss to those seen in the default scenario.

Table 6.19: Packet Loss UL - Coverage Simulation Scenario (52 Mbit/s)

Packet Loss UL (%) - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	2,9	3,9	1,7	4,2	4,5	3,44
2	1,1	0,4	0,46	0,33	0,74	0,606
3	3,7	**Service Crash**				3,7
4	0,45	2,5	1,3	0,42	**Service Crash**	1,17

From Table 6.19, we can identify in the third attempt, the network crash that was witnessed occured because of the uplink congestion that can be seen by the packet loss of 3,7%. For the forth attempt, and during the fifth reading, the throughput in the uplink direction collapsed to 0 Mbit/s, provoking a crash on the applications being run by the controller and the AGV.

Table 6.20: Skipped Frames - Coverage Simulation Scenario (10 Mbit/s)

Skipped Frames - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	1	1	0	1	3	1,2
4	0	0	0	0	0	0

From Table 6.18, there was not any lack of performance from the video stream in this scenario, which has the AGV producing a traffic load of 10 Mbit/s, similar to the default scenario.

Table 6.21: Skipped Frames - Coverage Simulation Scenario (52 Mbit/s)

Skipped Frames - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	51	20	15	43	26	31
2	80	17	8	6	41	30,4
3	**Service Crash**					
4	7	56	37	8	**Service Crash**	27

From Table 6.21, we can see that the uplink congestion caused the video stream to disrupt the its motion, because of the number of skipped frames per reading, just like what we saw in Table 6.11.

Table 6.22: Coverage Simulation Scenario (10 Mbit/s)

Coverage Simulation Scenario - 10 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	9,91	9,93	9,922
Packet Loss UL	0,61	0,83	0,725
Skipped Frames	0	3	0,3

In comparison to Table 6.4, the table above presents similar values. The packet loss values are smaller but this difference is not significant.

Table 6.23: Coverage Simulation Scenario (52 Mbit/s)

Coverage Simulation Scenario - 52 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	47,6	49,8	48,8
Packet Loss UL	0,33	4,5	2,23
Skipped Frames	6	80	29,47

A similar comment can be provided about Table 6.23 in comparison to Table 6.12. The values that we are seeing are very similar, which ensures that the uplink congestion threshold starts around 50 Mbit/s.

Table 6.24: ACP's PRB Usage - Coverage Simulation Scenario (10 Mbit/s)

Attempts	UL PRB Usage (%)	DL PRB Usage (%)
1	17	4
2	17,5	4
3	17,67	4
4	18	4
Average	17,54	4

In terms of PRB usage, Table 6.24 presents values similar to the ones of the default scenario.

Table 6.25: ACP's PRB Usage - Coverage Simulation Scenario (52 Mbit/s)

Attempts	UL PRB Usage (%)	DL PRB Usage (%)
1	75,5	17
2	78	17
3	74	17
4	75	17,5
Average	75,63	17,13

For the attempts with a traffic load of 52 Mbit/s, the same behaviour that was seen in the uplink congestion scenario is seen in this one too.

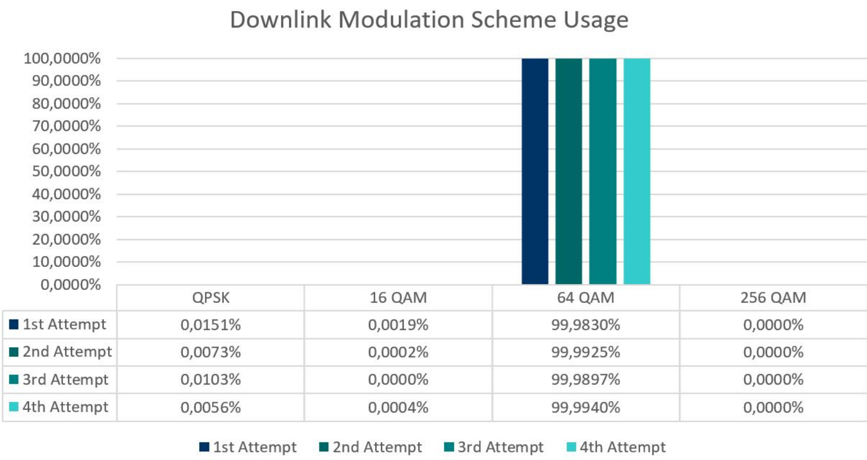


Figure 6.7: Modulation Scheme Usage - Coverage Simulation Scenario (10 Mbit/s)

In terms of downlink modulation scheme, what we can notice in both Figure 6.7 and Figure 6.8 is that the scheme mostly used was the best one available. In this case, it was the 64-QAM modulation. We can conclude that forcing the downlink modulation to be in the best case scenario of 64-QAM, and not of 256-QAM, did not affect the perfomance of the AGV.

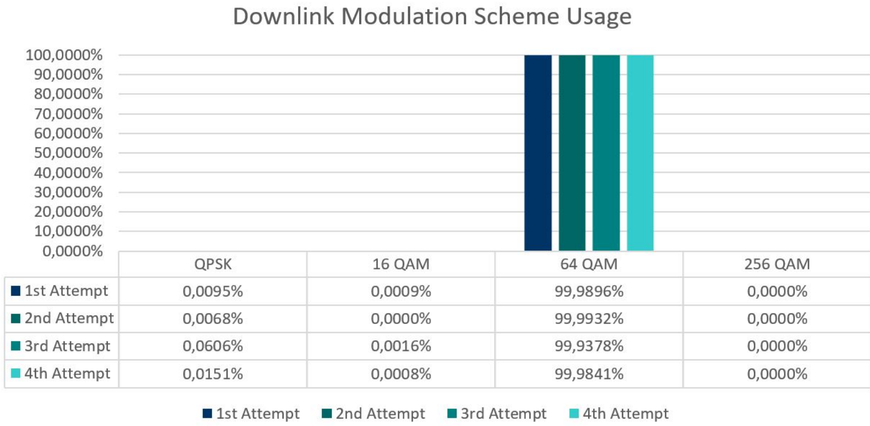


Figure 6.8: Modulation Scheme Usage - Coverage Simulation Scenario (52 Mbit/s)

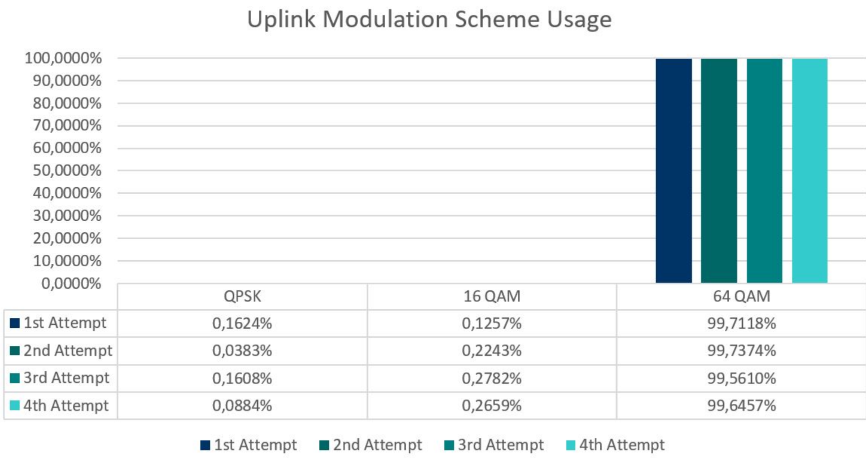


Figure 6.9: Modulation Scheme Usage - Coverage Simulation Scenario (10 Mbit/s)

In terms of used uplink modulation schemes, we can notice that we have similar results to the ones that we saw in the prior scenarios. Being that the modulation scheme used predominately is 64-QAM, even in situations of congestion.

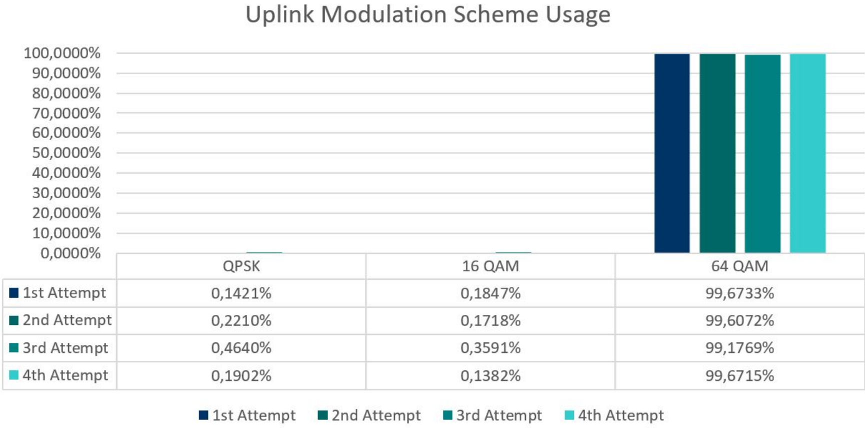


Figure 6.10: Modulation Scheme Usage - Coverage Simulation Scenario (52 Mbit/s)

From this Coverage Simulation scenario, we did not see a different outcome from the default or uplink congestion scenario. The downlink modulation did not affect the network or the performance of the use case.

6.6 Different Cells

For this testing scenario, we will present the different set of attempts where the traffic load created by the AGV was of 10, 30 and 52 Mbit/s. An important detail that needs to be mentioned is that the AGV and the controller are attached to different cells. The cell where the AGV is attached is called Cell 1, and the cell where the controller is attached is called Cell 2. We chose to present this scenario for a traffic load of 10, 30 and 52 Mbit/s so that it is possible to understand the evolution of the network resources usage by the use case. Since, we are using a Cell for each entity of the use case, we can identify the individual impact and load of the controller and AGV separately.

Table 6.26: Throughput UL - Different Cells Scenario (10 Mbit/s)

Throughput UL (Mbit/s) - 10 Mbit/s	
Attempts\Readings	**1/2/3/4/5**
1	9,89
2	9,93
3	9,94
4	9,93

For a traffic load of 10 Mbit/s, Table 6.26 presents similar behaviour in the uplink channel as that seen in the prior scenarios that had the same traffic load.

Table 6.27: Throughput UL - Different Cells Scenario (30 Mbit/s)

Throughput UL (Mbit/s) - 30 Mbit/s	
Attempts\Readings	**1/2/3/4/5**
1	29,9
2	29,9
3	29,9

From Table 6.27, we can conclude that the traffic load generated by the AGV at 30 Mbit/s did not cause any disruption to 5G network, which is logical since in terms of traffic load we are at half point of what is the uplink congestion threshold.

Table 6.28: Throughput UL - Different Cells Scenario (52 Mbit/s)

Throughput UL (Mbit/s) - 52 Mbit/s	
Attempts\Readings	**1/2/3/4/5**
1	50,5
2	50,6
3	50,7

As we can see from Table 6.28, in all 3 attempts it was recorded an estimated throughput relatively smaller than the desired (52 Mbit/s). This congestion in the uplink, caused by the traffic generated by the AGV, will cause a bad performance of the video stream, since this congestion exhaust the AGV connection to the network. Since the controller is not attached to the same cell, it will continue to be connected to the network.

Table 6.29: Packet Loss UL - Different Cells Scenario (10 Mbit/s)

Packet Loss UL (%) - 10 Mbit/s	
Attempts\Readings	**1/2/3/4/5**
1	0,7
2	0,66
3	0,64
4	0,66

In terms of packet loss, Table 6.29 presents similar values to the ones seen in prior scenarios with the same traffic load generated by the AGV.

Table 6.30: Packet Loss UL - Different Cells Scenario (30 Mbit/s)

Packet Loss UL (%) - 30 Mbit/s	
Attempts\Readings	**1/2/3/4/5**
1	0,28
2	0,23
3	0,38

Table 6.30 shows that at half of the uplink congestion threshold, the packet loss is minimal.

Table 6.31: Packet Loss UL - Different Cells Scenario (52 Mbit/s)

Packet Loss UL (%) - 52 Mbit/s	
Attempts\Readings	**1/2/3/4/5**
1	2,8
2	2,8
3	2,5

The values in Table 6.31 can reflect the congestion that was observed when registering this scenario for these conditions. This packet loss in the uplink direction force the data plane to lose a significant amount of video stream packets sent by the AGV, which reflected in the constant re-buffering events by the controller in its stream player. For the controller, the video stream was freezed, and its stream player tried to re-buffer its video buffer. In this case, we can also say that the AGV lost its connection to the network several times, since this loss of packets cause its stream service to alert to network connectivity problems. However, in lower periods of congestion, the AGV would reconnect to the network. At the same time the services/applications running in the controller did not have any issue, since the controller is attached to another cell.

Table 6.32: Skipped Frames - Different Cells Scenario (10 Mbit/s)

Skipped Frames - 10 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0

For this scenario, where the traffic load of the AGV was 10 Mbit/s, by looking at Table 6.32, we can infer that the video stream ran smoothly in all attempts, which it did, since there were no skipped frames.

Table 6.33: Skipped Frames - Different Cells Scenario (30 Mbit/s)

Skipped Frames - 30 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	0	0	0	0	0	0
2	0	0	0	2	0	0,4
3	0	0	0	2	0	0,4

From Table 6.33, we can conclude the same that was said for Table 6.32.

Table 6.34: Skipped Frames - Different Cells Scenario (52 Mbit/s)

Skipped Frames - 52 Mbit/s						
Attempts\Readings	**1**	**2**	**3**	**4**	**5**	**Average**
1	94	107	64	109	**N/A**	93,5
2	35	29	85	31	127	61,4
3	108	98	78	48	0	66,4

As we can notice, from Table 6.34, the congestion in the uplink provoked a number of frames to be lost, and cause the video stream to have a poor performance in the controller. This high number of skipped frames, caused in the controller, a high number of re-buffering events.

Table 6.35: Different Cells Scenario (10 Mbit/s)

Different Cells Scenario - 10 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	9,89	9,94	9,923
Packet Loss UL	0,64	0,7	0,665
Skipped Frames	0	0	0

Table 6.35 presents similar values to the ones seen in other scenarios with the same AGV's traffic load (10 Mbit/s).

Table 6.36: Different Cells Scenario (30 Mbit/s)

Different Cells Scenario - 30 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	29,9	29,9	29,9
Packet Loss UL	0,23	0,38	0,297
Skipped Frames	0	2	0,267

Table 6.36 shows that at half of what is said to be the traffic load that causes congestion, the use case runs smoothly.

Table 6.37: Different Cells Scenario (52 Mbit/s)

Different Cells Scenario - 52 Mbit/s			
KPIs	**Min**	**Max**	**Average**
Throughput UL	50,5	50,7	50,6
Packet Loss UL	2,5	2,8	2,7
Skipped Frames	0	127	81,43

Table 6.37 demonstrates us again that increasing the traffic load above the uplink congestion threshold (50 Mbit/s) causes a lack of performance from the AGV, since its connection becomes even more unreliable.

Table 6.38: ACP's PRB Usage - Different Cells Scenario (10 Mbit/s)

	Cell 1 - AGV		**Cell 2 - Controller**	
Attempts	**UL PRB Usage (%)**	**DL PRB Usage (%)**	**UL PRB Usage (%)**	**DL PRB Usage (%)**
1	17,7	4	5	7,67
2	20	4	5	6,5
3	19,5	4	5	6
4	19,5	4	5	6
Average	19,17	4	5	6,54

Table 6.38 and the following tables present the impact that each UE, in this case the controller and the AGV, has in radio resources.

Table 6.39: ACP's PRB Usage - Different Cells Scenario (30 Mbit/s)

	Cell1 1 - AGV		Cell 2 - Controller	
Attempts	**UL PRB Usage (%)**	**DL PRB Usage (%)**	**UL PRB Usage (%)**	**DL PRB Usage (%)**
1	51,5	12	5	6
2	51,5	11	5	5
3	51,5	11	5	5,5
Average	51,5	11,33	5	5,5

In reference to Table 6.39, we can identify that the AGV uses up to half of the PRB resources in the uplink direction in a scenario where it generates a traffic load of 30 Mbit/s. In terms of downlink usage, we can conclude the the usage goes a bit over the 10% mark.

Table 6.40: ACP's PRB Usage - Different Cells Scenario (52 Mbit/s)

	Cell 1 - AGV		Cell 2 - Controller	
Attempts	**UL PRB Usage (%)**	**DL PRB Usage (%)**	**UL PRB Usage (%)**	**DL PRB Usage (%)**
1	97,5	20	6	3,5
2	97,5	10	6,5	3,5
3	94,67	13,67	6,33	5
Average	96,56	14,57	6,28	4

From Table 6.40, we can identify that in the second and third attempt the dowlink PRB usage had a lower percentage in comparison to the other attempts. From this observation, we can understand that the uplink congestion can influence the downlink connection estability.

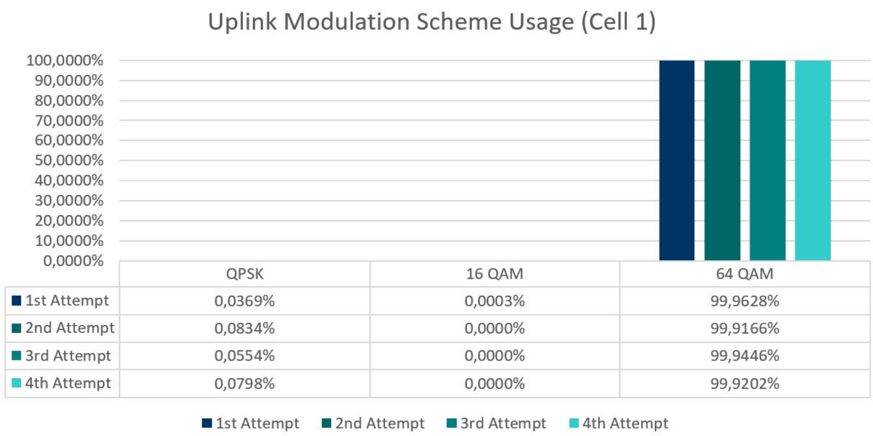
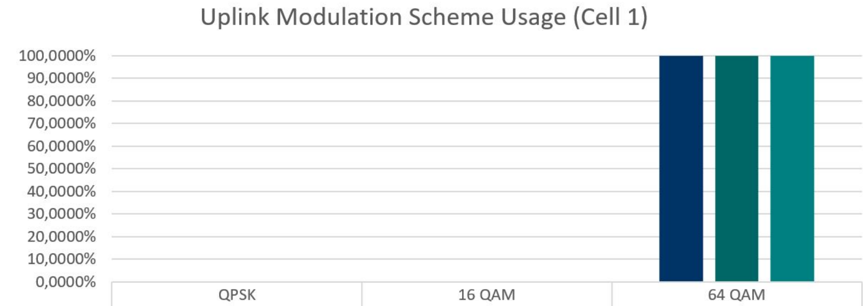


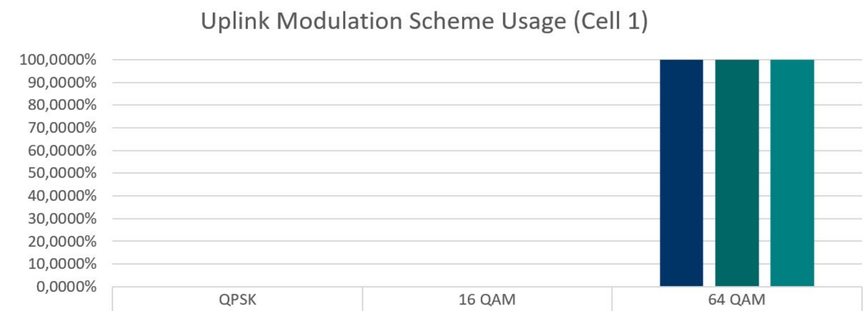
Figure 6.11: Modulation Scheme Usage - Different Cells Scenario (10 Mbit/s)



	QPSK	16 QAM	64 QAM
1st Attempt	0,0439%	0,0000%	99,9561%
2nd Attempt	0,0949%	0,0000%	99,9051%
3rd Attempt	0,0520%	0,0000%	99,9480%

1st Attempt 2nd Attempt 3rd Attempt

Figure 6.12: Modulation Scheme Usage - Different Cells Scenario (30 Mbit/s)



	QPSK	16 QAM	64 QAM
1st Attempt	0,0487%	0,0022%	99,9491%
2nd Attempt	0,0804%	0,0011%	99,9184%
3rd Attempt	0,0735%	0,0098%	99,9167%

1st Attempt 2nd Attempt 3rd Attempt

Figure 6.13: Modulation Scheme Usage - Different Cells Scenario (52 Mbit/s)

The three figures above show that the most used uplink modulation scheme was 64-QAM. Even though we had congestion in the uplink direction in the 52 Mbit/s scenario, this did not affect the modulation scheme used in the radio channel. The 64-QAM modulation scheme continued to be used predominately.

Chapter 7

Assessment of the 5G testbed Architecture

The main goal for this chapter is to assess the 5G deployment used, and try to take it apart and identify the properties that need to be changed. We assess this architecture for the AGV use case, but we will also extrapolate and assess the performance of this architecture when supporting other Industrial use cases, with key requirements in latency and reliability.

7.1 AGV - Use Case's Performance from each Test Scenario

In this chapter we draw conclusions about our architecture by analyzing the results and general behaviour of the AGV use case.

7.1.1 Default Scenario

As mentioned in Chapter 6, the default scenario is characterized by having both the controller and the AGV communicating to each other on the same cell. The AGV is generating 10 Mbit/s of traffic in both downlink and uplink directions.

As we can see in Section 6.3, the packet loss registered in the default scenario was high considering what is required by the 3GPP for Discrete Automation scenarios (figure 2.10). In this case, we are associating packet loss with the 3GPP's reliability definition. Reliability entails successful deliver of packets to a given targeted system [5]. The results show us an average packet loss ratio in terms of uplink traffic generated by the AGV of 0,92%, with a observed minimum value of 0,69% and maximum of 1,2%.

The reason for these values can be related to the path between the gNB and the Core, since it is made through the internet. It is possible that packets can be lost due to the fact that the path to the Core from the gNB goes through the internet. The solution for this limitation can be to establish a dedicated path to the Cloud, which is not a viable option in the private networks scope, meaning that the best plan of action is shifting the UPF to the end user location, having an hybrid deployment of the 5G Core. With this type of core implementation, the access of servers or the

Internet would be made on-premises, leaving only the Core control-plane functions deployed on Cloud, which in this case would be suitable since in a permanent regime this use case will not suffer changes in terms of connection/attachment communications. From the moment that the controller is attached to the network and the AGVs are set in term of number and attached to the network, the use case functions continuously without breaks. There would be some small latency, of about 25 ms (end-to-end latency) as was refered in the begining of the previous chapter, to start up or restart our network.

7.1.2 Uplink Congestion Scenario

In this scenario, the goal was to reach the uplink congestion level to predict the number of active AGVs on a radio cell. In Section 6.4, we can see that the congestion in the uplink begins when the traffic reaches 50 Mbit/s.

Looking at Table 6.8, the average packet loss of the uplink traffic at 50 Mbit/s was 2,07% with a maximum registered of 8,9%. Two of the four attempts of this scenario made with traffic being generated at 50 Mbit/s registered the exhaustion of the radio resources, meaning that both the AGV and the controller lost connection and could not receive or send data for a few seconds. This was seen by the crash of both applications running in the controller (the game and video stream) and a high percentage of packet loss in the uplink traffic of 4,5% and 8,9% in each attempt.

For the scenario where the AGV is generating 52 Mbit/s, we registered similar behaviour to the one seen at 50 Mbit/s, with the exception that in addition to the crashes, we also registered some poor performance of the video stream without the complete crash of the network. The two crashes of the network occurred because of uplink traffic congestion, with packet loss registered of 7,9% and 4,6%. The congestion in the uplink caused also the freezing of the video stream being watched by the controller. In one attempt, it was registered an average of 35 skipped frames, which in comparison to the default scenario is very high since the total average registered in the default scenario was 1,3 skipped frames. Furthermore, for both scenarios the average Uplink PRB Usage registered in the ACP was of 65 and 75%, respectively.

With this, it is possible to conclude that the uplink congestion begins at 50 Mbit/s of total traffic generated in the uplink. Knowing that each AGV generates traffic at 10 Mbit/s, according to Figure 2.10, we can conclude that each O-RU can support up to 5 AGVs, and for the case of 5 AGVs is possible to have some periods with no UL resources to ensure the demands of the use case.

7.1.3 Coverage Simulation through Modulation

The goal of this scenario was to try to simulate different coverage conditions to see the impact on the performance of the use case. From the results gathered for this scenario, the use case behave very similar to the Uplink Congestion scenario, which has its logic, since the changes made was to the downlink modulation and not the uplink, and this use case stresses more the uplink direction. The graphics shown in Section 6.3 and 6.4 are very similar to the ones presented in this scenario. As we can see, the modulation schemes usage doesn't suffer from this alteration, besides the Downlink 256-QAM scheme not being used, since the best scheme that can be used is the 64-QAM.

From this scenario, we could not gather conclusions about the impact of coverage in the performance of the use case. From section 6.5, we can also see the same behaviour registered in the Uplink Congestion scenario, in relation to having network crashes at 52 Mbit/s.

7.1.4 Different Cells

For this scenario we had the controller and the AGV attached to different O-RUs. The goal was to evaluate the behaviour of the network when the AGVs provoked congestion in the uplink, and also how that behaviour affects the controller and the AGV.

When the AGV only generates 10 Mbit/s of traffic, both in the uplink and downlink directions, we can conclude that we have the same impact on the network that we had in the default scenario. All KPIs shown in this scenario are very similar to the ones shown for the default scenario, except for the PRB Usage, since the traffic load of both the controller and the AGV are redistributed for both cells.

For the attempts where the AGV generates 52 Mbit/s of traffic, we registered uplink congestion in the AGV cell (Cell 1), in all 3 attempts (as we can see by the packet loss in Table 6.31). This uplink congestion impacted the video stream quality with several freezes and caused the controller to lose any sight of what the AGV was seeing. Although, the AGV was facing congestion in his cell, the controller kept having connection to the network, since the video game was running with no problems, and also we can see that the controller did not use more than 7% of its cell resources (Table 6.40). We can conclude that the uplink throughput limit for the O-RUs being used is around the 50 Mbit/s, and that the maximum number of AGVs in each cell can go up to five.

7.2 Possible Architectures for Other Industry Use Cases

In addition to the AGV use case, there are other applications that can take advantage from the 5G technology, mainly industry use cases. In this section, we describe two other use cases, and extrapolate a possible 5G architecture that can support them. This extrapolation will be based on the results and conclusions that we gathered from the AGV use case described in this document. As guidelines, we are going to use the requirements established by the 3GPP, as presented in Figure 2.10.

7.2.1 Process Automation - Refineries

In [5], 3GPP defines Process Automation as being the automation of reactive flows. In terms of communications, this usage scenario is characterized by having very stringent requirements for communications service availability. The 3GPP recommends this type of automation for deployments in restricted and isolated areas.

These usage scenarios are most expected in refineries and water distribution networks. In our case, we will be focusing on refineries. The use case we have in mind is what we see in Figure 2.10 as Process Automation - Monitoring. The path to produce oil refined products has very stages, steps and the whole process needs to be monitored to ensure efficiency, high quality and also safety. Each stage is characterized by specific conditions that need to be ensured to produce a high-quality product. KPIs like temperature, volume, or pressure need continuous monitoring and communication to an operation and alarmistic platform. These communication needs to have a low latency and low packet loss to provide real-time information about the stage of the processes being conducted in the refinery.

From Figure 2.10, we noticed that 3GPP establishes an end-to-end latency of 60 ms and a reliability of 99,9%, which means that end-to-end packet loss needs to be below 0,1%. From what we saw in Section 6.1, the end-to-end latency to our Core deployed in the cloud is approximately 25 ms, which from what we see in Figure 2.10 is below the 60 ms mark for this usage scenario. In terms of latency requirements, the Core conditions that our current architecture possesses are feasible for this use case. From what we saw in the results gathered from the default scenario and discussed in Subsection 7.1.1, in terms of Reliability, we cannot use this type of Core Deployment, since the packet loss registered is above the 0,1% established by 3GPP in Figure 2.10. To ensure the reliability proposed for these usage scenarios we need to at least deploy the UPF on refinery-premises. Furthermore, the monitoring and alarmistic application that processes all the data gathered from the sensors inside the refinery, would also need to be on-premises, deployed in a server. This would be also a plus in terms of security, since it would create and isolated data path which would be separated from the public network, and towards a Private Networks conception. In terms of traffic load, in Figure 2.10, the 3GPP defines the experienced user data rate as being 1 Mbit/s, which based on the radios used in our deployment architecture would support in terms of traffic around 50 sensors. However, the number of possible attached UEs to the radios is 32, so the number of sensors would be limited to this condition.

To conclude and in addition, it is possible to make a case to deploy the entire 5G Core on-premises to guarantee a closed end-to-end architecture, which would add safety and security.

7.2.2 Intelligent Transport Systems - Airports

In [5], the 3GPP defines Intelligent Transport Systems as automation solutions for street-based traffic support. The 5G networks for these usage scenarios will be immersed in the public space. Since, our focus is in isolated and restricted areas, we will morph this usage scenarios for a Airport use case, and also unlike the AGV and the Process Automation use case, this use case will include an outdoor use case. In terms of RAN and radio specifications, we cannot extrapolate a solution in this matter.

Airports are very complex infrastructures, that have an organized and well-thought logic. Many devices used in airports are high-definition cameras, air quality sensors, automated gates. All of these devices are connected to an Operation Center, that receives alerts and can issue commands. For example, throughout an airport's outdoor area, it is possible to install several cameras for a multitude of applications. It can be for vehicle detection. If a baggage vehicle, passengers' bus or a fuel truck are using the correct road. It can also be for detection of accidents, or non-desired animal intrusion. This information needs to be communicated to a Center of Operations.

By analyzing Figure 2.10, for this scenario, we can notice that the 3GPP defines the end-to-end latency as 30 ms and the reliability as 99,9999%, which means that the end-to-end packet loss needs to be 0,0001%. Once again, in terms of end-to-end latency, the current deployment of the Core can support this use case, since its end-to-end latency is below the 30 ms. In terms of reliability, the current Core deployment cannot support this required packet loss, so the best course of action is to deploy the UPF on-premises to ensure a minimal packet loss. The application that runs the Operation Center would also be deployed in a server on-premises, for the same reasons explained in the prior use case. Once again, the full deployment of the Core on-premises can be made to ensure a closed end-to-end architecture, which is isolated from public networks.

7.3 Core Deployment on Cloud

Throughout the last years, a discussion has been made about Cloud Infrastructure as an alternative for companies that for several reasons do not want to have their personal servers for housing their applications. To conclude this chapter, we want to expose and describe the costs of deploying a 5G Core in a Cloud Service, using our deployment as an example.

As already mentioned in Chapter 4, our 5G Core was deployed in the AWS Cloud services, with the described architecture. In AWS, there are two main sources of costs, the ones relative to infrastructure allocated to the costumers and the ones relative to traffic. In terms of infrastructure, we will have the cost of the instance created to support the Athonet 5G Core, the cost of the instance created to support the NAT VM and also the cost of the Site-to-Site VPN connection to establish a connection between the AWS and the on-premises network (this connection ensure the communication between the gNB and the Core). In terms of expenditure, we will have a cost per month of 1500 euros, 1400 euros and 65 euros, respectively. In terms of traffic, AWS does not charge any value for inbound traffic that comes from the Internet, only outbond traffic that goes to the Internet. As an example, we can consider the traffic generated for the AGV's default scenario, to better understand the cost of traffic per month. Assuming that the AGV and the controller receive 10 Mbit/s of data as described by the 3GPP requirements [5], this is equivalent to 414720 Gbytes of data in a month, which corresponds to a cost of 22000 euros per month. This cost is for an operation time of 24 hours in a day and 7 days a week. If we consider an operation time of 8 hours a day and 5 days a week, we produce 92160 Gbytes of data per month and it costs 7000 euros per month. We are just considering the traffic that is generated through the communication between the controller and the AGV, which is the traffic that goes from the Core to the gNB. However, if part of this traffic that reaches the UPF is destined to reach the internet, there will be an additional cost for this flow too. In conclusion, to run a 5G Core in a Cloud service, it would be necessary to spend around 300 thousand euros or 120 thousand euros per year, depending on the operation time of the use case. In addition, these prices were calculated with the AWS pricing calculator [29] and based on our deployed infrastructure that supports the 5G Core.

In contrast, a physical deployment of the Core would mean other types of costs. Special prices could be agreed between enterprises and the Cloud service to minimize the impact on the business return. If we use as reference the cost of the servers installed for the RAN network, we can estimate a cost around the 40 thousand euros mark to acquire a server that can support the 5G Core. Furthermore, the server is going to need maintenance and software updates which can cost around 5000 euros per year. Since our focus is in industrial use cases, some of these services would be classified as critical operations which means that an backup server with another provisioned Core would be necessary to ensure redundancy. This redundancy would cost an extra 40 thousand euros. Other costs should be considered in these calculation, for example ensuring a facility room with the minimal requirements to host both servers.

We can conclude that although performance requirements can have a significant importance in choosing the type of deployment of the Core, being physical, hybrid or on cloud, other aspects as costs also need to be analyze to decide the best solution for a specific 5G deployment.

Chapter 8

Conclusion & Future Work

8.1 Conclusion

The main goal of this dissertation was to develop and test different network topologies, in order to fulfill the requirements of different use cases. The proposed solution focused our study to a particular use case for a topology that we had access to.

In Chapter 2, we covered the main characteristics of 5G in a SA concept, in terms of architecture and performance requirements for different use cases. In Chapter 3, we described the current tendencies of 5G, regarding the functional splits being used to distribute the radio protocol architecture between RAN nodes, and also presented the advantages and disadvantages of different Core Network deployments. In Chapter 3 we showcased different AGV use cases deployed in a 5G network that support the emulation of the same use case.

In Chapter 4, we started by implementing and deploying our 5G SA Private Network. The enterprise architecture was already pre-established. The RAN was provided by AirSpan O-RAN implementation, and the 5G SA Core was determined by Athonet 5G Core Solution. The Core was deployed in a cloud based service hyperscaler because we did not have enough resources to deploy the Core in a physical server, close to the RAN. Despite having this limitation in terms of architecture and also being limited to the vendors' solutions that we used, this 5G private network presents itself as a real-life network that can be deployed to anyone that needs a 5G access network, meaning that with this solution, we can test 5G on a commercial deployment, and conclude about its limitations. This brings additional value for this dissertation by providing the opportunity of showcasing a commercial 5G network that can be deployed in a real-life use case. The implementation of this solution delayed the analysis of the use case and the network, which was the main goal of this thesis. Since both the RAN and the Core were closed environment solutions, their deployment brought some delay when it faced some obstacles.

In Chapter 5, we did not had access to an AGV due to availability, time reasons, and we did not have a proper space to use it. The solution for this problem was to emulate the AGV use case, by having two computers, one being the controller and the other the AGV, that stressed the 5G network as a AGV use case would. We created also different scenarios to analyze the performance and possible limitations of the network, to ensure the required performance of the use case.

In Chapter 6, we presented the results gathered from the different test scenarios performed. It is important to highlight that in the beginning of this chapter, an important limitation of the 5G network architecture being used is pointed out. Since our Core is deployed in a Cloud service, the latency between the gNB and Core is high, in comparison to the requirements for this type of use case. In terms of results, we highlight that there is a high packet loss ratio, that is also unacceptable for this type of use case, and that through some tests we conclude that the uplink maximum traffic load per radio unit is around 50 Mbit/s.

With these results and initial conclusions, in Chapter 6 we assessed the 5G network that was used to support the AGV use case in Chapter 7. The main conclusion taken from the results is that we need to change the deployment of the Core to ensure the low latency and high reliability requirements of the AGV use case. The best solution would be to deploy the user-plane of the 5G Core (UPF) near the RAN to ensure a lower and also a minimal packet loss between the RAN and the Core, just like mentioned in Chapter 3, when we discussed private networks. This change is crucial to achieve the reliability requirement, but, for the latency, it is possible to install the Core in a datacenter near the RAN and not need to create a local infrastructure to support a Core physical deployment. Another important aspect that was discussed in this chapter and needs to be considered was the cost of deploying a Core on cloud versus deploying a core on-premises. It is showed that besides network requirements, it is also important to take in account the cost of different Core deployments, and assessing the usage that the 5G network is going to have.

Currently, there are no mass 5G private commercial deployments. With this work, we hoped to make a contribution to accelerate the technology adaptation by the industries. One of the biggest obstacles that 5G is facing is the deployment cost of the end-to-end architectures, if it is worth it, due to the investment that needs to be made. With this work, we try to shed some light in a 5G topology and its advantages and disadvantages, so that a foundation can be created for others to pickup this topology and alter it to continue to do more performance testing, in order to achieve minimalist 5G architectures that can support the 5G use cases with the most reduced investment. In this way, it is possible to drive the organizations to invest in this technology, due to the enhancement in terms of communications efficiency that 5G provides compared to other technologies.

The dissertation's objectives were fulfilled. We identified the key requirements of a specific use case in a high-level usage scenario. Furthermore, we used performance KPIs to assess the performance of the use case in the 5G network, specifically the packet loss ratio of the generated traffic by the AGV, the modulation picked by the radio unit to communicate with the AGV and the controller, and also the frame loss in the video stream made by the AGV. By running different tests and assessing the results, we concluded that the current 5G topology cannot support this use case, and that the UPF needs to be deployed on-premises to reduce the packet loss and also the latency.

8.2 Future Work

Even though this thesis achieved the core of its objectives, there is a set of recommendations to improve this work and support the embrace of 5G O-RAN.

In terms of work that can be done with the same topology, it could be relevant to plan and dimension the radio coverage to better assess how to deploy a RAN and specifically its radio units distribution to ensure the best coverage scenario. We can also do a study in terms of different positioning of the RAN entities.

Furthermore, the Athonet Core presents QoS configuration, that with the right group of use cases could present as a challenge to project and design a QoS priority diagram for a group of use cases to achieve the best performance at the same time.

With extra resources, the deployment of a hybrid and physical Core could be interesting to compare the different performances of a use case. It would also bring extra value if the use case was not a emulation, and it was possible to have access to a real AGV.

References

[1] Dahlman, Parkvall, and Skold. *5G NR: The Next Generation Wireless Access Technology*. Elsevier, London, 2018.

[2] 3GPP. System architecture for the 5G System (5GS). TS 23.501, 3rd Generation Partnership Project (3GPP), September 2022.

[3] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. 2022. doi:https://doi.org/10.48550/arXiv.2202.01032.

[4] ITU-R. IMT Vision–Framework and overall objectives of the future development of IMT for 2020 and beyond. M. 2083-0, International Telecommunication Union-Radiocommunications, September 2015.

[5] 3GPP. Service requirements for next generation new services and markets. TS 22.261, 3rd Generation Partnership Project (3GPP), March 2019.

[6] O-RAN Alliance. O-RAN Working Group 4 (Open Fronthaul Interfaces WG) Control, User and Synchronization Plane Specification. O-RAN.WG4.CUS.0-v 09.00, International Telecommunication Union-Radiocommunications, 2022.

[7] O-RAN Alliance. O-RAN Working Group 4 (Open Fronthaul Interfaces WG) Management Plane Specification. O-RAN.WG4.MP.0-v 10.00, International Telecommunication Union-Radiocommunications, 2022.

[8] 3GPP. F1 general aspects and principles. TS 38.470, 3rd Generation Partnership Project (3GPP), July 2018.

[9] 3GPP. NG general aspects and principles. TS 38.410, 3rd Generation Partnership Project (3GPP), October 2021.

[10] 3GPP. Xn general aspects and principles. TS 38.420, 3rd Generation Partnership Project (3GPP), July 2020.

[11] ITU-T. Transport network support of IMT-2020/5G. Gstr-tn5g., International Telecommunication Union-Radiocommunications, February 2018.

[12] Ning Mu, Shulei Gong, Wanqing Sun, and Quan Gan. The 5G MEC Applications in Smart Manufacturing. In *2020 IEEE International Conference on Edge Computing (EDGE)*, 2020. doi:10.1109/EDGE50951.2020.00014.

[13] Elizabeth Palacios-Morocho, Pablo Picazo-Martínez, Saúl Inca, and Jose F. Monserrat. Open Source 5G-NSA Network for Industry 4.0 Applications. In *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2021. doi:10.1109/PIMRC50174.2021.9569481.

[14] Winnie Nakimuli, Jaime Garcia-Reinoso, J. Enrique Sierra-Garcia, Pablo Serrano, and Isaac Quintana Fernández. Deployment and Evaluation of an Industry 4.0 Use Case over 5G. *IEEE Communications Magazine*, 2021. doi:10.1109/MCOM.001.2001104.

[15] 3GPP. NR and NG-RAN Overall description. TS 38.300, 3rd Generation Partnership Project (3GPP), January 2021.

[16] Ericson. 5G CORE (5GC) - What is 5G Core? https://www.ericsson.com/en/core-network/5g-core (accessed Dec. 26,2022).

[17] ITU-R. Minimum requirements related to technical performance for IMT-2020 radio interface(s). M. 2410-0, International Telecommunication Union-Radiocommunications, November 2017.

[18] 3GPP. F1 Application Protocol (F1AP). TS 38.473, 3rd Generation Partnership Project (3GPP), January 2020.

[19] 3GPP. General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U). TS 29.281, 3rd Generation Partnership Project (3GPP), January 2020.

[20] 3GPP. Xn signalling transport . TS 38.422, 3rd Generation Partnership Project (3GPP), March 2020.

[21] 3GPP. Study on new radio access technology: Radio access architecture and interface. TS 38.801, 3rd Generation Partnership Project (3GPP), March 2017.

[22] Line M. P. Larsen, Aleksandra Checko, and Henrik L. Christiansen. A survey of the functional splits proposed for 5g mobile crosshaul networks. 2019. doi:10.1109/COMST.2018.2868805.

[23] Mountaser, G. and Condoluci, M. and Mahmoodi, T. and Dohler, M. and Mings, I. Cloud-RAN in Support of URLLC. 2017. doi:10.1109/GLOCOMW.2017.8269135.

[24] Behnam Ojaghi, Ferran Adelantado, Angelos Antonopoulos, and Christos Verikoukis. Impact of Network Densification on Joint Slicing and Functional Splitting in 5G. *IEEE Communications Magazine*, 2022. doi:10.1109/MCOM.001.2100680.

[25] 5G-ACIA. 5G Non-Public Networks for Industrial Scenarios 5G Alliance for Connected Industries and Automation 5G-ACIA White Paper, 2019.

[26] Miaowen Wen, Qiang Li, Kyeong Jin Kim, David López-Pérez, Octavia A. Dobre, H. Vincent Poor, Petar Popovski, and Theodoros A. Tsiftsis. Private 5G Networks: Concepts, Architectures, and Research Landscape. *IEEE Journal of Selected Topics in Signal Processing*, 2022. doi:10.1109/JSTSP.2021.3137669.

[27] Mike Gilson, Richard Mackenzie, Andy Sutton, and Jinri Huang. NGMN Overview on 5G RAN Functional Decomposition by NGMN Alliance, 2018.

[28] Vladimir Kaplarevic. How to Install a Kubernetes Cluster on CentOS 7. https://phoenixnap.com/kb/how-to-install-kubernetes-on-centos (accessed May 31,2023).

[29] AWS. AWS Pricing Calculator. https://calculator.aws/#/ (accessed Jun. 22,2023).