5G CONNI

Private 5G Networks for Connected Industries

Deliverable D4.1

Initial specification and implementation of the building blocks

Co-funded by the Horizon 2020 programme
of the European Union in collaboration with Taiwan

Date of Delivery:	31.03.2021		
Project Start Date:	01.10.2019	**Duration:**	36 Months

Document Information

Project Number:	861459
Project Name:	Private 5G Networks for Connected Industries

Document Number:	D4.1
Document Title:	Initial specification and implementation of the building blocks
Editor:	Mickael Maman (CEA)
Authors:	Mickael Maman (CEA) Ngoc-Lam Dinh (CEA) Cheng-Yi Chien (Chunghwa Telecom, CHT) Jiun-Cheng Huang (Chunghwa Telecom, CHT) Yueh-Feng Li (Chunghwa Telecom, CHT) Ling-Chih Kao (Chunghwa Telecom, CHT) Mattia Merluzzi (SAP) Sergio Barbarossa (SAP) Stefania Sardellitti (SAP) Daniele Munaretto (ATH) Daniele Ronzani (ATH) Marco Centenaro (ATH) Nicola di Pietro (ATH) Jack Shi-Jie Luo (ITRI) Shuo-Peng Liang (ITRI) CC Weng (ANI) Tzu-Ya Wang (III)
Dissemination Level:	Public
Contractual Date of Delivery:	31.03.2021
Work Package	WP 4
File Name:	861459-5G CONNI-D4.1-Initial specification and implementation of the building blocks-v1.0.docx

Revision History

Version	Date	Comment
0.1	25.09.2020	First ToC
1.0	31.03.2021	Version for submission

Executive Summary

WP4 (Technical Enablers for Industrial Applications) covers Mobile Edge Computing (MEC) cloud development, industrial application technical development, radio network technical development, and core network technical development for industrial field. The main goal of this work package is to ensure industrial use cases can be implemented on private 5G networks successfully for industrial requirements, including high data rates (eMBB) and low latency (URLLC).

D4.1 provides initial specification and implementation of private 5G networks building blocks. This deliverable will be extended in D4.2. These innovative components will fuel the lab integration reported in D5.1.

During WP4, activities are carried out on both the European and the Taiwanese side of the consortium with partners either mirroring or complementing competences across the two regions. Section 2 aims to harmonize the research and development of innovative new technologies required for the target use cases. Section 3 provides an update of the state of the art for the four building blocks. Sections 4-7 detail the different contributions according to the building blocks. For the Radio Network Technical Enablers (Task 4.1), Alpha Network has developed a 5G RAN system composed of a CPE and a gNB, and CEA has designed an orchestrator and showed simulation results using NS-3 for investigating how to enable deterministic URLLC services. In Task 4.2 on Core Network Technical Enablers, III and Athonet have developed a 5G core network. The Mobile Edge Cloud Enablers have been developed by Athonet based on the hybrid architecture in smart industries and by Chunghwa Telecom with a SA MEC prototype. In Industrial Application Enablers (Task 4.4), ITRI has implemented three use cases prototype applications. The data collection and process diagnosis methods have been investigated and were demonstrated in a prototype implementation in one of the production cells targeted for the 5G CONNI demo. 3D models for these machines have been created for AR applications. The prototype of a cloud CNC controller has been developed and tested in a distributed network environment. Moreover, one of the objectives of this task is to rethink the network in a holistic manner by jointly optimizing all enabling technologies. Thus, SAP has developed an algorithm on dynamic resource allocation for wireless edge machine learning exploring energy-latency-reliability trade-offs.

Table of Contents

Executive Summary 4
Table of Contents 5
List of Figures 6
List of Tables 8
List of Acronyms 9
1 Introduction 12
1.1 Scope 12
1.2 Structure 12
2 Preliminary considerations 13
2.1 Recall of the 5G CONNI architecture 13
2.2 Design considerations 13
2.3 Modelling Methodologies 19
3 Background and State of the Art 25
3.1 Radio Network Technical Enablers 25
3.2 Core Network Technical Enablers 26
3.3 Mobile Edge Cloud Enablers 27
3.4 Industrial Application Enablers 29
4 Radio Network Technical Enablers 32
4.1 5G RAN system composed of CPE and gNB 32
4.2 Deterministic URLLC protocols 35
4.3 Conclusions 40
5 Core Network Technical Enablers 41
5.1 NFV-like lightweight orchestration framework for the core network 41
5.2 5G Core prototype 44
5.3 Conclusions 45
6 Mobile Edge Cloud Enablers 46
6.1 MEC based on 5G-CONNI's hybrid architecture 46
6.2 MEC 5G SA 47
6.3 Conclusions 50
7 Industrial Application Enablers 51
7.1 Three use cases prototype applications 51
7.2 Joint optimization of enabling technologies 54
7.3 Conclusions 59
8 Conclusions 60
References 62

List of Figures

Figure 1: Hybrid model. UEs can connect to the private CN by accessing from a private RAN or a public one. The enterprise's CN may be placed in a private datacenter or a central public cloud. 14
Figure 2 : NG-RAN Overall architecture [TS38300]......15
Figure 3 : NG-RAN architecture for gNB and 5GC [TS30401]......15
Figure 4 : Function split between NG-RAN and 5GC [TS38300]......16
Figure 5 : Protocol Stack for NG-C and NG-U [TS38300] 17
Figure 6 : Functions break down in RU for fronthaul Interface [ORANCUS] 17
Figure 7: Evaluation parameters for indoor-office scenarios [TR38901]......19
Figure 8: Evaluation parameters for InF [TR38901] 20
Figure 9: 3GPP step wise procedure for channel realizations [TR38901] 21
Figure 10: 3GPP LOS probability [TR38901] 21
Figure 11: 3GPP Pathloss model [TR38901] 22
Figure 12: 3GPP Channel model parameters for Indoor [TR38901]......22
Figure 13: 3GPP Channel model parameters for InF [TR38901] 23
Figure 14: A two-state Markov chain representing ON/OFF traffic model [Marvi2019] 24
Figure 15: Classification of the mechanisms as a function of the diversity and the network layer. 25
Figure 16: The Service-based Architecture in 5G core 27
Figure 17: ETSI MEC reference architecture 28
Figure 18: CPE architecture 32
Figure 19: 5G NR RU architecture......33
Figure 20: 5G NR CDU architecture 33
Figure 21: The proposed 8 possible CU DU split options......34
Figure 22: Framework to design a global orchestration 35
Figure 23: Classification of KPIs according to the network layers. 36
Figure 24: Distribution of E2E latency across the network layers......38
Figure 25: E2E Latency and Packet Error Rate 38
Figure 26: E2E latency distribution of MCS17 with or without HARQ 39
Figure 27: CDF of E2E Latency of MCS17 with or without HARQ 39
Figure 28: CDF of E2E Latency for MCS5, MCS17 and AMC......40
Figure 29: Overview of NFV frameworks and interfaces 41
Figure 30: Content of SOL002 request. 42
Figure 31: Confirmation of successful instantiation and configuration of the VNF EPC. 43
Figure 32: VNF instantiation, configuration and deletion time. 44
Figure 33: III-5GC Architecture......44
Figure 34: Integration of the MEC host into the 5G network architecture. 46
Figure 35: The architecture of bump-in-the-wire MEC cloud 48
Figure 36: The architecture of MEC 5G SA version 48
Figure 37: EcoreClould Platform Web UI and the MEE VNF deployment 49
Figure 38: MEC 5G SA prototype integrated with III's 5G core network and Alpha's 5G base station 50
Figure 39: Test result of MEC 5G SA prototype integrated with AR Process Diagnostics application 50
Figure 40: Prototype for the Using Augmented/Virtual Reality for Process Diagnosis use case 52

Figure 41: The implementation architecture of the Process Diagnostics by CNC and Sensing Data Collection...52
Figure 42: The tool condition monitoring system used in "Process Diagnostics by CNC and Sensing Data Collection" use case...53
Figure 43: The augmented reality app developed for the "Using Augmented/Virtual Reality for Process Diagnosis" use case...53
Figure 44: Demonstration of the "Using Augmented/Virtual Reality for Process Diagnosis" use case...54
Figure 45: Architecture of the cloud CNC....54
Figure 46: Average E2E delay vs. average energy consumption...56
Figure 47: Reliability function of the E2E delay...57
Figure 48: Energy-delay-accuracy trade-off...58

List of Tables

Table 1 : Function and non-functional requirements of use cases on the European side......14
Table 2 : MEC functions support in 5G CONNI project......18
Table 3 : Non-functional requirements of use cases on the Taiwan side......18
Table 4 : End-to-end test of 5G NR gNodeB and CPE equipment......34
Table 5 : Simulation parameter of URLLC simulations using NS-3......37
Table 6: Virtualization Host characteristics......42
Table 7: Software and tools version.......43
Table 8 : Simulation parameters of dynamic resource allocation evaluation......55

List of Acronyms

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5GC	5G Core
5G CONNI	5G for Connected Industries
ACIA	Alliance for Connected Industries and Automation
AF	Application Function
AMF	Access and Mobility Management Function
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
AS	Angular Spread
ATH	Athonet
AUSF	Authentication Server Function
CBR	Constant Bit Rate
CBRS	Citizens Broadband Radio Service
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CHT	ChungHwa Telecom
CN	Core Network
CNC	Computerized Numerical Control
COVID-19	Corona Virus Disease 2019
CP	Control Plane
CPE	Customer Premises Equipment
CU	Central Unit
DFE	Digital Front End
DMTD	Dual-Mixer Time Difference
D-RAN	Distributed RAN, Distributed Radio Access Network
DS	Delay Spread
DU	Data Unit
DoA	Description of Action
E2E	End-to-End
ECC	ECoreCloud
eMBB	Enhanced Mobile Broadband
EML	Edge Machine Learning
EPC	Evolved Packet Core
ENI	Experimental Networked Intelligence
eSIM	Embedded SIM, Embedded Subscriber Identity Module
ETSI	European Telecommunications Standards Institute
FIB	Forwarding Information Base
FoF	Factories of the Future
FR-x	Functional Requirement x
FR1	3GPP Frequency Range 1

FR2	3GPP Frequency Range 2
gNB	Gigabit Node B, 5G Base Station
GTP	GPRS Tunneling Protocol
HARQ	Hybrid ARQ, Hybrid Automatic Repeat Request
HHI	Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute
IEEE	Institute of Electrical and Electronics Engineers
III	Institute for Information Industry
IIoT	Industrial Internet of Things
IMTC	Intelligent Machine Tool Center
InF	Indoor Factory
ITRI	Industrial Technology Research Institute
I-UPF	Intermediate User Plane Function
KPI	Key Performance Indicator
LOS	Line of Sight
MANO	Management and Orchestration
MAC	Medium Access Control
MEC	Mobile Edge Cloud / Multi-access Edge Computing
MEE	Mobile Edge Enabler
MEH	Mobile Edge Cloud Host
MIMO	Multiple Input, Multiple Output
MNO	Mobile Network Operator
MSD	Mean Squared Deviation
MQTT	Message Queuing Telemetry Transport
MVNO	Mobile Virtual Network Operator
NGMN	Next Generation Mobile Network
NFV	Network Function Virtualization
NLOS	Non Line of Sight
NMSE	Normalized Mean Square Error
NPN	Non-Public Network
NR	5G New Radio
NS3	Network Simulator Version 3
NSA	5G Non Stand Alone
OAM	Operation, Administration and Maintenance
PCF	Policy Charging Function
PDCP	Packet Data Convergence Protocol
PHY	Physical Layer
PLL	Phase Locked Loop
PPS	Pulse Per Second
QoS	Quality of Service
RAN	Radio Access Network
RIB	Routing Information Base
RLC	Radio Link Control
RRC	Radio Resource Control

RU	Radio Unit
SA	(5G) Stand Alone / Services and System Aspects
SAP	University La Sapienza
SBA	Service Based Architecture
SBI	Service Based Interface
SDN	Software Defined Networking
SF	Shadow Fading
SLA	Service Level Agreement
SMF	Session Management Function
SPS	Semi Persistent Scheduling
TR	Technical Report
TS	Technical Specification
UC	Use Case
UDM	Unified Data Management
URLLC	Ultra-Reliable Low Latency Communication
UP	User Plane
UPF	User Plane Function
VIM	Virtual Infrastructure Manager
VNF	Virtualized Network Function
VNFM	Virtualized Network Function Manager
VR	Virtual Reality
VUCA	Virtual Uniform Circular Antenna Array
WP	Work Package

1 Introduction

WP4 (Technical Enablers for Industrial Applications) covers Mobile Edge Computing (MEC) cloud development, industrial application technical development, radio network technical development, and core network technical development for industrial field. The main goal of this work package is to ensure industrial use cases can be implemented on private 5G networks successfully for industrial requirements, including high data rates (eMBB) and low latency (URLLC).

1.1 Scope

D4.1 provides the initial specification and implementation of private 5G networks building blocks. This deliverable will be extended in D4.2. These innovative components will fuel the lab integration reported in D5.1.

1.2 Structure

Section 2 aims to harmonize the research and development of innovative new technologies required for the target use cases. Section 3 provides an update of the state of the art for the four building blocks. Sections 4-7 details the different contributions according to the building blocks: Radio Network Technical Enablers from Task 4.1, Core Network Technical Enablers from Task 4.2, Mobile Edge Cloud Enablers from Task T4.3 and Industrial Application Enablers from Task 4.4

2 Preliminary considerations

2.1 Recall of the 5G CONNI architecture

In D2.1, the envisioned private network architectures are described and divided into two categories:

1. Private networks deployed as an isolated and standalone network.
2. Private networks deployed in conjunction with a public network.

The first category refers to the fully private network described in Section 3.1 of D2.1.

The second category includes three options, according to the interaction and infrastructure sharing with the public network, and they are described from Section 3.2 to 3.4 in D2.1, which refer to the MVNO, Hybrid and MNO's private network. In these scenarios, the overall network is a combination of public and private networks. The public network refers to the MNO's network that offers services to general public, whereas the private network refers to the non-public network (NPN) that provides services to the organization.

On 5G CONNI's European side, the preferred network architecture is the *hybrid model* defined in Section 3.3 of D2.1. We recall that, in this architecture, the CN can be split into a centralized Control Center (typically containing the 5G control-plane elements) interacting with the local RAN and devices through locally deployed Edge Nodes (containing user-plane elements (UPF) and the MEC platform). Such a platform allows the deployment and management of several distributed private networks, each anchored by an Edge Node.

On 5G CONNI's Taiwanese side, the target deploying network architecture is based on section 3.1 in D2.1, which is a fully private network. A fully private ownership expects that the enterprise owns almost every dimension, that is, spectrum, RAN, MEC, CN, and applications. For the target use cases, a fully private 5GC (including AMF, SMF, UPF, etc.) will be deployed on the enterprise datacenter; meanwhile, the MEC(bump-in-wire), RAN and CPE will be deployed on the factory site to achieve the higher efficacy base on the local data flow and storage processing.

2.2 Design considerations

On 5G CONNI's European side, the final objective architecture is the hybrid model and will be implemented in two steps. The first step will be performed as follows:

- a fully on-site deployment – see Section 3.1 of D2.1 – at the BOSCH premises to correctly configure the environment at the local factory;
- interoperability tests of the 5GC between the mobile core network vendor (ATH) and HHI, which owns the 5G RAN infrastructure.

As an outcome of these preliminary tests, the best split of control-plane functionalities between the local factory and the Enterprise's datacenter/central cloud for the use cases will be found, and the hybrid model will be finally implemented.

The hybrid model can be seen as a combination of the Fully Private and MVNO models. As shown in Figure 1, the enterprise hosts a local private RAN and MEC platform, which are connected to a private CN, also owned by the enterprise. However, radio access of the enterprise's UEs can also take place by roaming through public MNO's RAN, which forwards control and management traffic to the private CN. The Edge Node is located inside the enterprise firewall and keeps traffic and user data local to meet low latency, data security, and edge computing requirements.

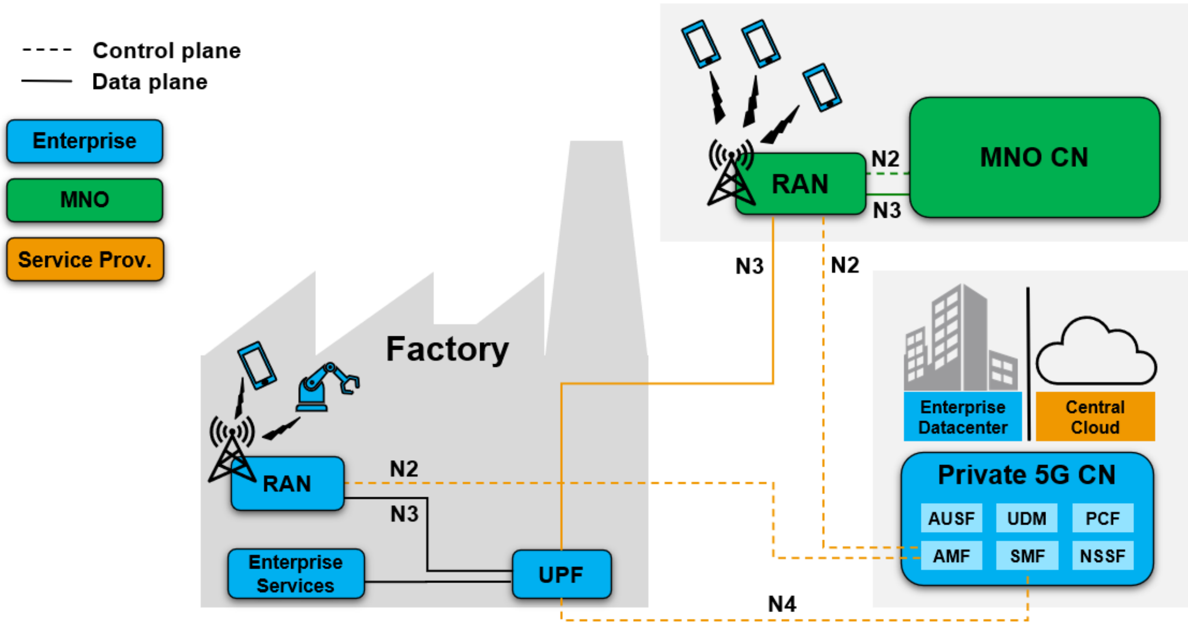

Figure 1: Hybrid model. UEs can connect to the private CN by accessing from a private RAN or a public one. The enterprise's CN may be placed in a private datacenter or a central public cloud.

The envisioned hybrid architecture will be configured so to meet the requirements of the UC-3, which were set in D1.1 and are hereafter reported for the reader's convenience:

Table 1 : Function and non-functional requirements of use cases on the European side

Use Case	Functional requirements	Non-functional requirements
UC-3 Robot platform with Edge Intelligence and Control	• *Primary*: FR-3 (end-to-end QoS), FR-4 (network capability exposure), FR-5 (priority, QoS and policy control), FR-6 (time synchronization), FR-8 (context-aware network), FR-9 (real-time end-to-end QoS monitoring), FR-13 (edge computing) • *Secondary*: FR-1 (mobility management), FR-10 (5G-LAN), FR-11 (proximity services), FR-12 (secure remote access)	• Service bitrate: 228 kbps – 1.6 Mbps for control traffic; 100s Mbps for video traffic • Communication area: 100 m x 100 m x 15 m • Connectivity density: 1 to few tens per shop floor • Area traffic capacity: 100s Mbps per 100 m² • UE speed < 2 km/h • End-to-end latency: 1 ms – 7 ms • Transfer interval: 5 ms – 20 ms • Transmission time: 1.4 ms – 7 ms • Survival time: 20 ms • Message size: 200 bytes • Video latency < N times transfer interval

On 5G CONNI's Taiwanese side, the use case requirements of use case 1 and use case 2 are defined in D1.1. In order to realize two use cases at the ITRI IMTC site, the design considerations of radio network, core network, MEC, and industrial applications are described in the following.

- Radio network
  In order to fulfil the interoperability of networking components from different vendors, the design consideration of RAN architecture is based the industrial specification, e.g. the 3GPP specification. For the 5G RAN architecture, 3GPP TS38.300 [TS38300] and TS38.401 [TS38401] are taken into consideration. Figure 2 shows the NG-RAN overall architecture. There composed gNB, ng-eNB and AMF/UPF network components. In this 5G-CONNI project, only gNB and NG interfaces are addressed.

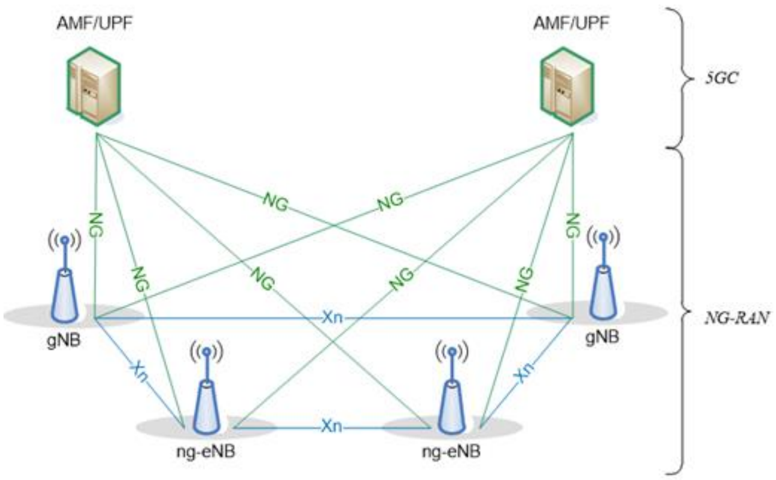

Figure 2 : NG-RAN Overall architecture [TS38300]

Figure 3 shows the gNB architecture. A gNB can be composed of a gNB-CU and one or several gNB-DU. F1 interface in between of gNB-CU and gNB-DU. gNB-DU is connecting to 5GC through NG interface.

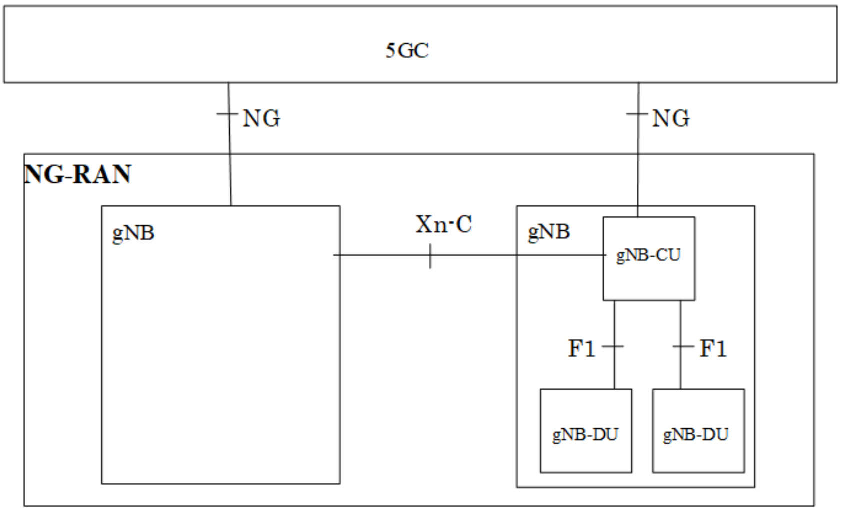

Figure 3 : NG-RAN architecture for gNB and 5GC [TS30401]

Figure 4 shows the functions in gNB and AMF/UPF respectively. In addition to functions of Radio Resource Management: Radio Bearer Control, Radio Admission Control, Connection Mobility Control, Dynamic allocation of resources to UEs in both uplink and downlink (scheduling), gNB should host other functions such as:

- IP header compression, encryption and integrity protection of data;
- Selection of an AMF at UE attachment when no routing to an AMF can be determined from the information provided by the UE;
- Routing of User Plane data towards UPF(s);
- Routing of Control Plane information towards AMF;
- Connection setup and release;
- Scheduling and transmission of paging messages;
- Scheduling and transmission of system broadcast information (originated from the AMF or OAM);
- Measurement and measurement reporting configuration for mobility and scheduling;
- Transport level packet marking in the uplink;
- Session Management;
- Support of Network Slicing;
- QoS Flow management and mapping to data radio bearers;
- Support of UEs in RRC_INACTIVE state;
- Distribution function for NAS messages;

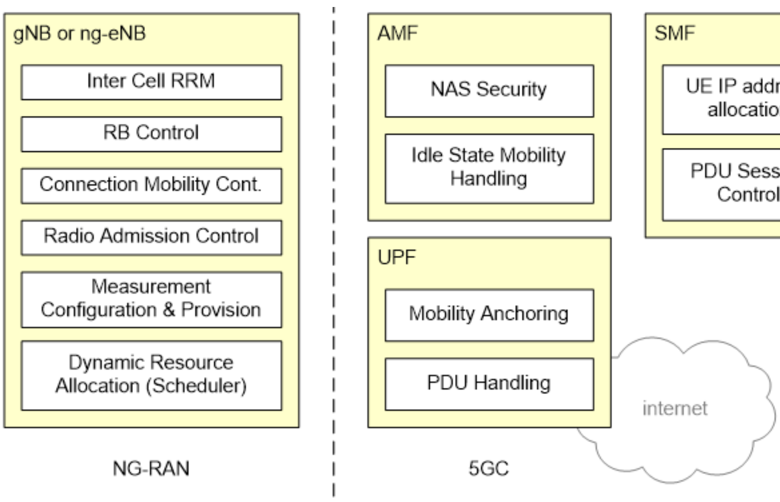

Figure 4 : Function split between NG-RAN and 5GC [TS38300]

AMF and UPF which are components in 5GC are designed for connecting NG control plane interface (NG-C) and NG user plane interface (NG-U) to gNB, should host functions shown in Figure 4 , more details are refer to [TS38300].
To breakdown the high level design for NG-C and NG-U, Figure 5 breakdowns the protocol stack. For NG-C plane, NG-AP layer is used to connecting to AMF and GTP layer is used to connect to UPF.
The high level consideration of NG-RAN is focusing on the inter-operability of network components from multi-vendors. The implementation of NG-RAN, especially the NG interfaces will based on 3GPP industrial specifications.

As for the fronthaul interface which connects the RU and DU, the O-RAN specification will be taken into consideration [ORANCUS].

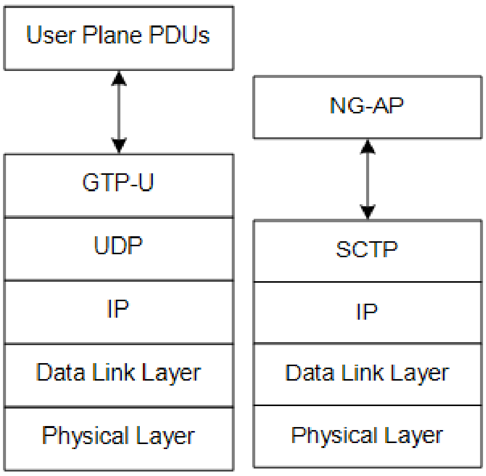

Figure 5 : Protocol Stack for NG-C and NG-U [TS38300]

According to [ORANCUS], functional split option 7-2 and RU category A is proposed for the implementation. Details please refer to section 4. Figure 6 shows the downlink low level PHY functions for category A O-RU. Different from category B, the function of RE mapping is in DU side. Functions of iFFT and CP are in the O-RU side.

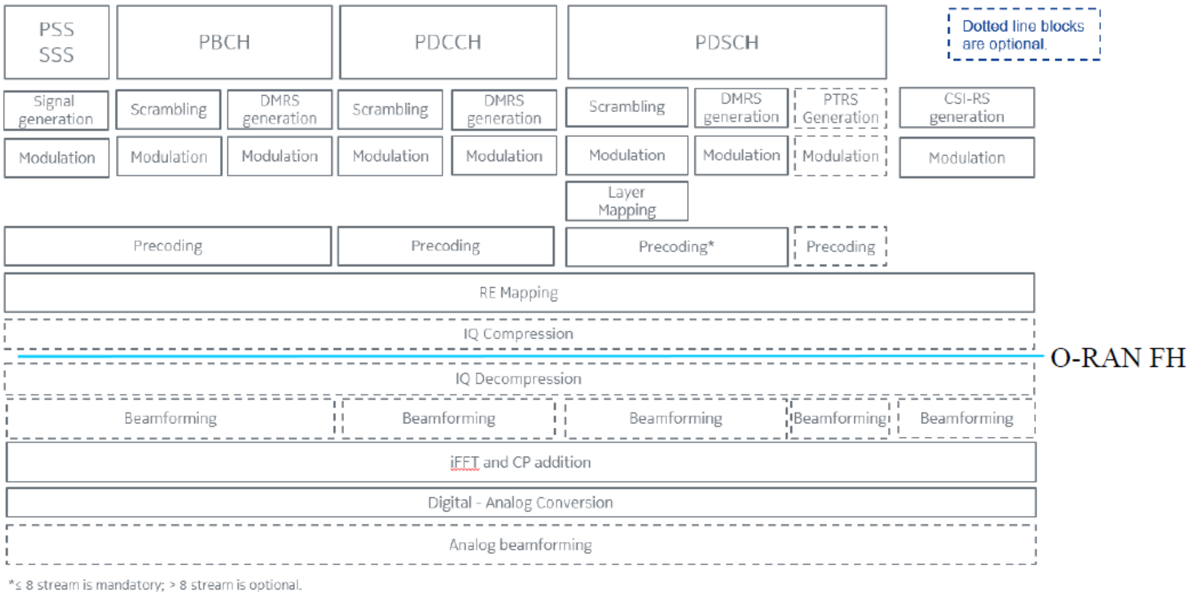

Figure 6 : Functions break down in RU for fronthaul Interface [ORANCUS]

The design of O-RU support 4x4 MIMO and 4 steams. The radio characteristic will comply with 3GPP TS38.104 [TS38104] specification.

- Core network

The fourth industrial revolution is being conceived around a smarter, fully connected, and automated paradigm for future factories and industries. It will combine smart objects and digital systems capable of autonomously exchanging information, triggering actions, and controlling each other independently. The innovation technology like 5G connectivity, IoT technologies, Fog/Edge/Cloud solutions, Big Data crunchers, and cyber-security arise

as key components to realize this paradigm shift. The goal is to create flexible and resource-efficient production systems in manufacturing and to integrate industry, business, and internal processes through computer-based systems.
To support this goal, the 5G core network plays a crucial role in industrial *private* networks. 5G networks have been conceived to support the needs of a hyper-connected society, demanding simultaneously very high data rate access, very low latency, and wider coverage for an increasing number of almost permanently connected devices.The 5G core network will accommodate simultaneously a mix of different service typologies with very distinct needs on top of the same physical infrastructure. Apart from that, 5G strongly leverages the introduction of three new technological paradigms on provider's core networks: network virtualization, network slicing and network programmability. The virtualization of network functions enables the separation of the execution of a given network function from the specific physical device. There is flexibility at the time of deploying and scaling network functions, allowing at the same time the commoditization of the devices on top of which those functions run. Aspects such as guaranteed latency, high throughput, isolation, etc., are made available for the verticals consuming the network, permitting sophisticated services to gracefully co-exist on top of the same infrastructure. 5GC manages the end-to-end configuration and deployment in accordance with services' traffic characteristics and service level agreements (SLA).

- MEC

MEC provides functions according to Functional Requirements defined in D1.1, as shown in Table 2. MEC must provide the traffic local breakout features according to FR-13. Due to Mobility Management Support (FR-1), there is more than one gNB the IMTRI IMTC site, so MEC has to support handover. MEC also provides the priority traffic handling feature because there are two use cases at IMTRI IMTC site and some traffic needs higher priority processing.

Table 2 : MEC functions support in 5G CONNI project

Functions	Functional Requirements
Support handover	Mobility Management Support (FR-1)
Meet application requirements	End-to-end QoS Support (FR-3)
Priority traffic handling according to applications	Priority, QoS and Policy Control Support (FR-6)
Report MEC status to monitoring platform	Real-time end-to-end QoS Monitoring Support (FR-9)
Remote access	Secure Remote Access Support (FR-12)
Traffic local breakout	Edge Computing Support (FR-13)

MEC provides KPIs including 10G throughput, latency less than 1 ms and UE less than 20 UEs, that could achieve the non-functional requirements of use cases on the Taiwan side as shown in Table 3.

Table 3 : Non-functional requirements of use cases on the Taiwan side

Use Case	Non-functional requirements
UC-1 Process Diagnostics by CNC and Sensing Data Collection	• per machine is approximately 208 Mbps • 11 stations
UC- 2 Using Augmented/Virtual Reality for Process Diagnosis	• *service bit rate up to 1 Gbps* • *positioning service latency <15 ms* • *motion-to-photon latency<20ms* • *end-to-end latency<=10ms*

- Industrial applications

  Design of the industrial application should consider the actual development progress of the 5G system of project members. As currently commercialized or emerging 5G system components supports only R15, which covers only the eMBB characteristics of the 5G network. Therefore, applications developed for data collection and AR/VR were designed without strict low latency and high reliability requirements.

  Moreover, the selected use case should be designed to utilize or extend the capability of the IoT infrastructure to minimize the development effort and to evaluate the benefit of factory IoT based on 5G network.

2.3 Modelling Methodologies

2.3.1 Scenario of interest

According to 3GPP technical report TR38.901 [TR38901], the scenarios of interest are:

- Indoor: This scenario is intended to capture various typical indoor deployment scenarios, including office environments, and shopping malls. The typical office environment is comprised of open cubicle areas, walled offices, open areas, corridors etc. The BSs are mounted at a height of 2-3 m either on the ceilings or walls. The shopping malls are often 1-5 stories high and may include an open area (or "atrium") shared by several floors. The BSs are mounted at a height of approximately 3 m on the walls or ceilings of the corridors and shops.
  - Example: [Tx height: 2-3m, Rx height: 1.5m, area: 500 square meters]

Parameters		Indoor - office open office	Indoor - office mixed office
Layout	Room size (WxLxH)	120mx50mx3m	
	ISD	20m	
BS antenna height h_{BS}		3 m (ceiling)	
UT location	LOS/NLOS	LOS and NLOS	
	Height h_{UT}	1 m	
UT mobility (horizontal plane only)		3 km/h	
Min. BS - UT distance (2D)		0	
UT distribution (horizontal)		Uniform	

Figure 7: Evaluation parameters for indoor-office scenarios [TR38901]

- Indoor industrial scenarios: the indoor factory (InF) scenario focuses on factory halls of varying sizes and with varying levels of density of "clutter", e.g. machinery, assembly lines, storage shelves, etc

Parameters		InF				
		InF-SL (sparse clutter, low BS)	InF-DL (dense clutter, low BS)	InF-SH (sparse clutter, high BS)	InF-DH (dense clutter, high BS)	InF-HH (high Tx, high Rx)
Layout	Room size	Rectangular: 20-160000 m^2				
	Ceiling height	5-25 m	5-15 m	5-25 m	5-15 m	5-25 m
	Effective clutter height h_c	< Ceiling height, 0-10 m				
	External wall and ceiling type	Concrete or metal walls and ceiling with metal-coated windows				
Clutter type		Big machineries composed of regular metallic surfaces. For example: several mixed production areas with open spaces and storage/commissioning areas	Small to medium metallic machinery and objects with irregular structure. For example: assembly and production lines surrounded by mixed small-sized machineries.	Big machineries composed of regular metallic surfaces. For example: several mixed production areas with open spaces and storage/commissioning areas	Small to medium metallic machinery and objects with irregular structure. For example: assembly and production lines surrounded by mixed small-sized machineries.	Any
Typical clutter size, $d_{clutter}$		10 m	2 m	10 m	2 m	Any
Clutter density r (percentage of surface area occupied by clutter)		Low clutter density (<40%)	High clutter density (≥40%)	Low clutter density (<40%)	High clutter density (≥40%)	Any
BS antenna height h_{BS}		Clutter-embedded, i.e. the BS antenna height is below the average clutter height		Above clutter		Above clutter
UT location	LOS/NLOS	LOS and NLOS				100% LOS
	Height h_{UT}	Clutter-embedded				Above clutter

Figure 8: Evaluation parameters for InF [TR38901]

2.3.2 Indoor Channel Models for industrial environments

In 5G-CONNI studies, we also consider 3GPP channel models from 3GPP technical report TR38.901 [TR38901]. The radio channel realizations are obtained by a step wise procedure.

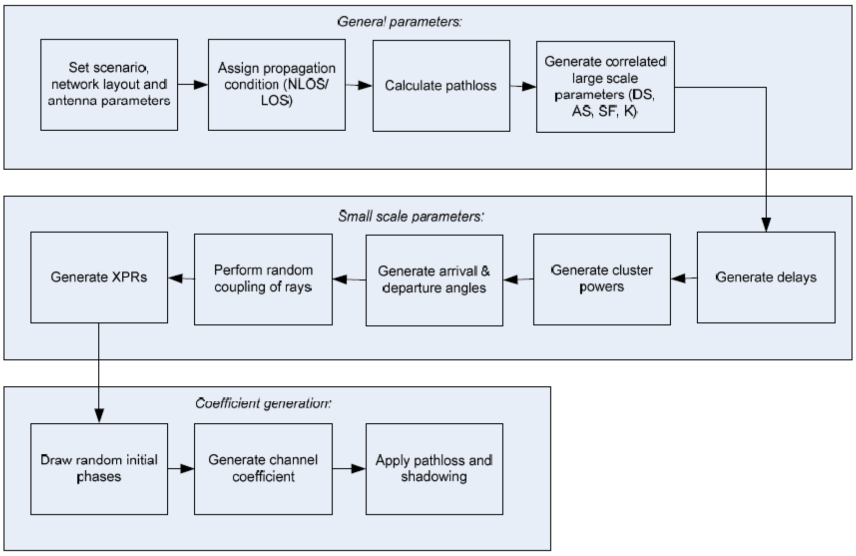

Figure 9: 3GPP step wise procedure for channel realizations [TR38901]

The first step is to set environment, network layout, and antenna array parameters. In 5G-CONNI, we are considering InH-office or InF.

The second step is to assign propagation condition (LOS/NLOS).

Scenario	LOS probability (distance is in meters)
Indoor - Mixed office	$\mathrm{Pr}_{\mathrm{LOS}} = \begin{cases} 1 & , d_{\text{2D-in}} \leq 1.2\text{m} \\ \exp\left(-\frac{d_{\text{2D-in}} - 1.2}{4.7}\right) & , 1.2\text{m} < d_{\text{2D-in}} < 6.5\text{m} \\ \exp\left(-\frac{d_{\text{2D-in}} - 6.5}{32.6}\right) \cdot 0.32 & , 6.5\text{m} \leq d_{\text{2D-in}} \end{cases}$
Indoor - Open office	$\mathrm{Pr}_{\mathrm{LOS}} = \begin{cases} 1 & , d_{\text{2D-in}} \leq 5\text{m} \\ \exp\left(-\frac{d_{\text{2D-in}} - 5}{70.8}\right) & , 5\text{m} < d_{\text{2D-in}} \leq 49\text{m} \\ \exp\left(-\frac{d_{\text{2D-in}} - 49}{211.7}\right) \cdot 0.54 & , 49\text{m} < d_{\text{2D-in}} \end{cases}$
InF-SL InF-SH InF-DL InF-DH	$\mathrm{Pr}_{\mathrm{LOS,subsce}}(d_{\mathrm{2D}}) = \exp\left(-\frac{d_{\mathrm{2D}}}{k_{\mathrm{subsce}}}\right)$ where $k_{subsce} = \begin{cases} -\frac{d_{clutter}}{\ln(1-r)} & \text{for InF-SL and InF-DL} \\ -\frac{d_{clutter}}{\ln(1-r)} \cdot \frac{h_{BS} - h_{UT}}{h_c - h_{UT}} & \text{for InF-SH and InF-DH} \end{cases}$ The parameters $d_{clutter}$, r, and h_c are defined in Table 7.2-4
InF-HH	$\mathrm{Pr}_{\mathrm{LOS}} = 1$

Figure 10: 3GPP LOS probability [TR38901]

The third step is to calculate the pathloss.

Scenario	LOS/NLOS	Pathloss [dB], f_c is in GHz and d is in meters, see note 6	Shadow fading std [dB]	Applicability range, antenna height default values
InH - Office	LOS	$PL_{\text{InH-LOS}} = 32.4 + 17.3\log_{10}(d_{3D}) + 20\log_{10}(f_c)$	$\sigma_{SF} = 3$	$1\text{m} \le d_{3D} \le 150\text{m}$
	NLOS	$PL_{\text{InH-NLOS}} = \max(PL_{\text{InH-LOS}}, PL'_{\text{InH-NLOS}})$ $PL'_{\text{InH-NLOS}} = 38.3\log_{10}(d_{3D}) + 17.30 + 24.9\log_{10}(f_c)$	$\sigma_{SF} = 8.03$	$1\text{m} \le d_{3D} \le 150\text{m}$
		Optional $PL'_{\text{InH-NLOS}} = 32.4 + 20\log_{10}(f_c) + 31.9\log_{10}(d_{3D})$	$\sigma_{SF} = 8.29$	$1\text{m} \le d_{3D} \le 150\text{m}$
InF	LOS	$PL_{LOS} = 31.84 + 21.50\log_{10}(d_{3D}) + 19.00\log_{10}(f_c)$	$\sigma_{SF} = 4.$	$1 \le d_{3D} \le 600\,m$
	NLOS	InF-SL: $PL = 33 + 25.5\log_{10}(d_{3D}) + 20\log_{10}(f_c)$ $PL_{NLOS} = \max(PL, PL_{LOS})$	$\sigma_{SF} = 5.7$	
		InF-DL: $PL = 18.6 + 35.7\log_{10}(d_{3D}) + 20\log_{10}(f_c)$ $PL_{NLOS} = \max(PL, PL_{LOS}, PL_{InF-SL})$	$\sigma_{SF} = 7.2$	
		InF-SH: $PL = 32.4 + 23.0\log_{10}(d_{3D}) + 20\log_{10}(f_c)$ $PL_{NLOS} = \max(PL, PL_{LOS})$	$\sigma_{SF} = 5.9$	
		InF-DH: $PL = 33.63 + 21.9\log_{10}(d_{3D}) + 20\log_{10}(f_c)$ $PL_{NLOS} = \max(PL, PL_{LOS})$	$\sigma_{SF} = 4.0$	

Figure 11: 3GPP Pathloss model [TR38901]

The forth step is to generate large scale parameters, e.g. delay spread (DS), angular spreads (ASA, ASD, ZSA, ZSD), Ricean K factor (K) and shadow fading (SF).

Scenarios		Indoor-Office LOS	Indoor-Office NLOS
Delay spread (DS) lgDS=log10(DS/1s)	μ_{lgDS}	-0.01 log10(1+f_c) - 7.692	-0.28 log10(1+f_c) - 7.173
	σ_{lgDS}	0.18	0.10 log10(1+f_c) + 0.055
AOD spread (ASD) lgASD=log10(ASD/1°)	μ_{lgASD}	1.60	1.62
	σ_{lgASD}	0.18	0.25
AOA spread (ASA) lgASA=log10(ASA/1°)	μ_{lgASA}	-0.19 log10(1+f_c) + 1.781	-0.11 log10(1+f_c) + 1.863
	σ_{lgASA}	0.12 log10(1+f_c) + 0.119	0.12 log10(1+f_c) + 0.059
ZOA spread (ZSA) lgZSA=log10(ZSA/1°)	μ_{lgZSA}	-0.26 log10(1+f_c) + 1.44	-0.15 log10(1+f_c) + 1.387
	σ_{lgZSA}	-0.04 log10(1+f_c) + 0.264	-0.09 log10(1+f_c) + 0.746
Shadow fading (SF) [dB]	σ_{SF}	See Table 7.4.1-1	
K-factor (K) [dB]	μ_K	7	N/A
	σ_K	4	N/A
Cross-Correlations	ASD vs DS	0.6	0.4
	ASA vs DS	0.8	0
	ASA vs SF	–0.5	–0.4
	ASD vs SF	–0.4	0
	DS vs SF	–0.8	–0.5
	ASD vs ASA	0.4	0
	ASD vs K	0	N/A
	ASA vs K	0	N/A
	DS vs K	-0.5	N/A
	SF vs K	0.5	N/A
Cross-Correlations [1)]	ZSD vs SF	0.2	0
	ZSA vs SF	0.3	0
	ZSD vs K	0	N/A
	ZSA vs K	0.1	N/A
	ZSD vs DS	0.1	-0.27
	ZSA vs DS	0.2	-0.06
	ZSD vs ASD	0.5	0.35
	ZSA vs ASD	0	0.23
	ZSD vs ASA	0	-0.08
	ZSA vs ASA	0.5	0.43
	ZSD vs ZSA	0	0.42
Delay scaling parameter r_τ		3.6	3
XPR [dB]	μ_{XPR}	11	10
	σ_{XPR}	4	4
Number of clusters N		15	19
Number of rays per cluster M		20	20
Cluster DS (c_{DS}) in [ns]		N/A	N/A
Cluster ASD (c_{ASD}) in [deg]		5	5
Cluster ASA (c_{ASA}) in [deg]		8	11
Cluster ZSA (c_{ZSA}) in [deg]		9	9
Per cluster shadowing std ζ [dB]		6	3
Correlation distance in the horizontal plane [m]	DS	8	5
	ASD	7	3
	ASA	5	3
	SF	10	6
	K	4	N/A
	ZSA	4	4
	ZSD	4	4

Figure 12: 3GPP Channel model parameters for Indoor [TR38901]

Scenarios		InF LOS	InF NLOS
Delay spread (DS) lgDS=log10(DS/1s) [4)]	μ_{lgDS}	log10(26(V/S)+14)-9.35	log10(30(V/S)+32)-9.44
	σ_{lgDS}	0.15	0.19
AOD spread (ASD) lgASD=log10(ASD/1°)	μ_{lgASD}	1.56	1.57
	σ_{lgASD}	0.25	0.2
AOA spread (ASA) lgASA=log10(ASA/1°)	μ_{lgASA}	-0.18*log10(1+fc) + 1.78	1.72
	σ_{lgASA}	0.12*log10(1+fc) + 0.2	0.3
ZOA spread (ZSA) lgZSA=log10(ZSA/1°)	μ_{lgZSA}	-0.2*log10(1+fc) + 1.5	-0.13*log10(1+fc) + 1.45
	σ_{lgZSA}	0.35	0.45
Shadow fading (SF) [dB]	σ_{SF}	Specified as part of path loss models	
K-factor (K) [dB]	μ_K	7	N/A
	σ_K	8	N/A
Cross-Correlations	*ASD* vs *DS*	0	0
	ASA vs *DS*	0	0
	ASA vs *SF*	0	0
	ASD vs *SF*	0	0
	DS vs *SF*	0	0
	ASD vs *ASA*	0	0
	ASD vs *K*	-0.5	N/A
	ASA vs *K*	0	N/A
	DS vs *K*	-0.7	N/A
	SF vs *K*	0	N/A
Cross-Correlations [1)]	*ZSD* vs *SF*	0	0
	ZSA vs *SF*	0	0
	ZSD vs *K*	0	N/A
	ZSA vs *K*	0	N/A
	ZSD vs *DS*	0	0
	ZSA vs *DS*	0	0
	ZSD vs *ASD*	0	0
	ZSA vs *ASD*	0	0
	ZSD vs *ASA*	0	0
	ZSA vs *ASA*	0	0
	ZSD vs *ZSA*	0	0
Delay scaling parameter r_τ		2.7	3
XPR [dB]	μ_{XPR}	12	11
	σ_{XPR}	6	6
Number of clusters N		25	25
Number of rays per cluster M		20	20
Cluster *DS* (c_{DS}) in [ns]		N/A	N/A
Cluster *ASD* (c_{ASD}) in [deg]		5	5
Cluster *ASA* (c_{ASA}) in [deg]		8	8
Cluster *ZSA* (c_{ZSA}) in [deg]		9	9
Per cluster shadowing std ζ [dB]		4	3
Correlation distance in the horizontal plane [m] [5)]	*DS*	10	10
	ASD	10	10
	ASA	10	10
	SF	10	10
	K	10	N/A
	ZSA	10	10
	ZSD	10	10

Figure 13: 3GPP Channel model parameters for InF [TR38901]

The last steps are about the small scale parameters: generate cluster delays, cluster powers, arrival/departure angles…

In the simulations for computation offloading presented in section 7.2, 3GPP path loss and shadowing models are taken into account, with a statistical model for fast fading due to the dynamic nature of the proposed scenarios.

Within Work Package 3, in Task 3.1, channel measurement campaigns are foreseen to be conducted at Bosch facilities to build channel models for different factory environments, both at FR1 and FR2 (i.e. sub-6 GHz and above 6 GHz). A channel sounding system has been developed by HHI and is ready to be used for the measurements, which have been affected by the COVID-19 pandemic due to the limited access to Bosch facilities. Deliverable 3.1 (M22) will report the results of the channel measurement campaign. Therefore, these results could be exploited and incorporated in next deliverables of WP4, to harmonize simulation results with the channel models extracted in WP3.

2.3.3 Traffic models

In deterministic URLLC evaluations, the ON/OFF traffic models are preferred for modelling of Internet Protocol (IP) traffic [Aijaz2012]. By exploiting ON/OFF traffic models, several additional effects such as high variability or infinite variance can be created which results in an aggregate traffic that exhibits the Joseph effect (self-similar or long range dependent). The model are more realistic and provide the capability of application-specific or mixed traffic analysis. [Marvi2019]

The design and development of an ON-OFF traffic model relies on an accurate description of traffic entities from link level to application level. The analysis of the structure of IP traffic is performed predominantly using ON-OFF models. The ON-OFF model uses only 2 states, namely ON & OFF. The time spent between ON & OFF states, commonly referred to as the transition time, is expected to follow an exponential distribution. More details on the model can be found in [Marvi2019].

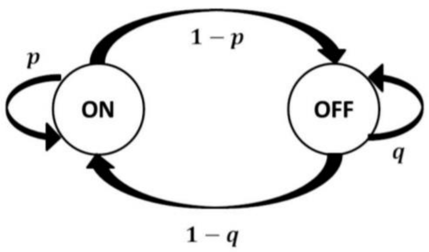

Figure 14: A two-state Markov chain representing ON/OFF traffic model [Marvi2019]

In NS-3, the traffic generator can be done using On/Off pattern which ON and OFF states alternate. The duration of each of these states is determined with the onTime and offTime random variables. During the “Off” state, no traffic is generated. During the “On” state, constant bit rate (CBR) traffic is generated. This CBR traffic is characterized by the specified “data rate” and “packet size”. When an On/Off application is stated, the first packet transmission occurs after a delay equal to packet size/ bit rate. When the off state is triggered during the packet transmission, the remaining time until when the next transmission is cached and will be used when the application is on again.

For example, packet size = 1000 bits, bit rate = 500 bits/sec. If the application is started at time 3 seconds, the first packet transmission will be scheduled for time 5 seconds (3 + 1000/500) and subsequent transmissions at 2 second intervals. If the above application were instead stopped at time 4 seconds, and restarted at time 5.5 seconds, then the first packet would be sent at time 6.5 seconds, because when it was stopped at 4 seconds, there was only 1 second remaining until the originally scheduled transmission, and this time remaining information is cached and used to schedule the next transmission upon restarting. [NS3APP]

In the simulations to evaluate computation offloading strategies (section 7.2), we will also use a simpler traffic model, in which data arrivals are generated from a Poisson distribution with mean arrival rate to be specified for each result. Poisson traffic is typical for evaluating performance of queueing systems.

3 Background and State of the Art

In this section, we present the state of the art of the four technical enablers: radio network, core network, mobile edge cloud and industrial applications

3.1 Radio Network Technical Enablers

In order to meet the deterministic URLLC requirements, several enhanced mechanisms can reduce latency and associated jitter, improve reliability, throughput, and availability of communication. These mechanisms can exploit 5-Dimension (5-D) diversity, based on: *(1)* Signal Processing, *(2)* Time, *(3)* Frequency, *(4)* Space and *(5)* Hardware. Besides, the end-to-end network enabling URLLC should consider the different involved layers of the network (i.e. not only RAN, but also MEC, Transport Network and Core Network). Figure 15 classifies the mechanisms proposed in the literature as a function of the 5-D diversity and the network layer it belongs to.

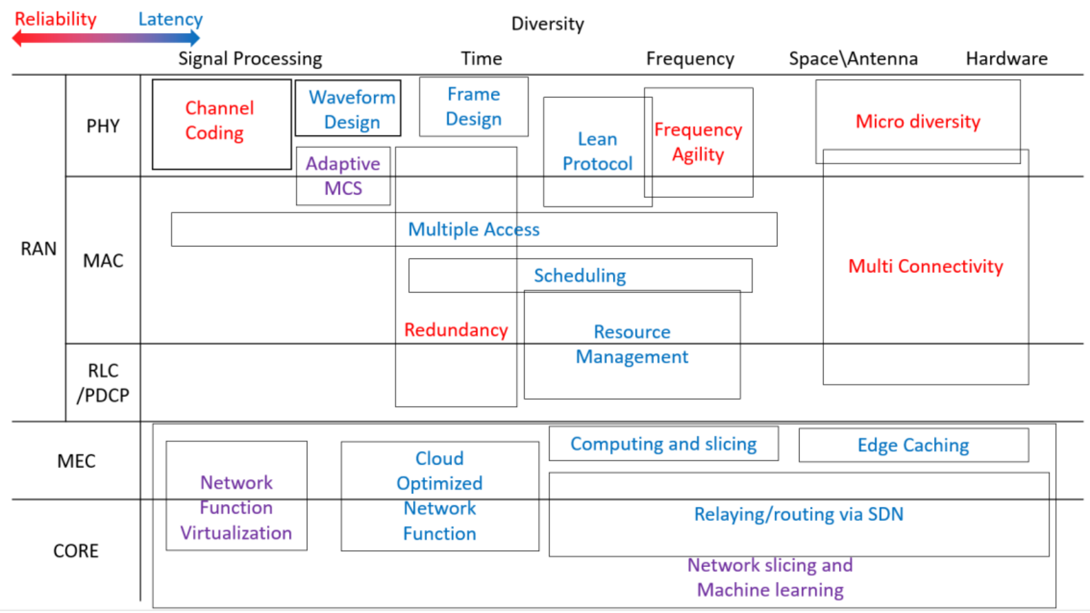

Figure 15: Classification of the mechanisms as a function of the diversity and the network layer.

Ultra reliable communication at RAN has been widely investigated for many years and can be divided into several mechanisms along 5-D (see Figure 15): channel coding in Signal Processing domain, redundancy in time and frequency domain, spatial diversity in space domain and multi-Radio Access Technology (multi-RAT) related to hardware's capability domain. The diversity in Signal Processing refers to the redundancy in terms of data processing, which includes channel coding and various modulation orders to achieve more robust transmission over the noisy channel. Small packets are expected to carry critical information that should be received with URLLC requirements. Thus, Liva et al. survey coding schemes for short blocks in [Liva9] and target acceptable performance in the short-block regime. Several efforts have been devoted to combine coding schemes and space diversity to further improve reliability as in [Chen18] which combines analog fountain codes with 10 antennas at the receiver to achieve 10^{-6} BLER. Besides, the antenna diversity, also known as space diversity, is considered as enabler for URLLC, since it provides more spatial degrees of freedom to improve the reliability of the system. Pocovi et al. depict in [Pocovi15] that joint microscopic MIMO schemes and macroscopic diversity are sufficient to guarantee the SINR outage performance. The joint spatial diversity is able to handle fast fading, and increases the robustness of the communication by reaching 10^{-5} SINR outage. Some approaches combine multiple mechanisms to enhance reliability. She et al. [She16] have proposed an uplink transmission optimization exploiting micro-diversity and frequency diversity infinite block-length channels that can reduce the total required bandwidth while guaranteeing high reliability.

Latency reduction techniques can be classified into several categories such as frame design, lean protocol, multiple access scheme and scheduling protocol, which relate to PHY and MAC layers at RAN. Frame structure design is an important aspect of 5G PHY and concerns a shorten version of the frame reducing latency communication without influencing reliability. Li et al. have demonstrated in [Li17] that low latency with guaranteed reliability design is achieved by using HARQ with Channel State Information (CSI) turnaround time and shorter Transmission Time Interval (TTI). Another flexible frame structure design is described in [Pedersen16] by Pedersen et al. They propose a multiplexing scheme in frequency-resource grid that dynamically adjusts the TTI size with user requirements to fulfill round trip interval requirement for URLLC. Moreover, they suggest an in-resource control channel design cooperating with data frame to exploit flexible bandwidth allocation, beamforming gain and control overhead. Besides, Ashraf et al. [Ashraf15] have proposed a fast control scheme, in which the control channel is reconfigured to support lower latency by reducing TTI by a factor in the order of five. This technique is then associated with antenna diversity to support a BLER of 10^{-9} and 0.2 ms TTI. Another method to offer high levels of reliability with latency reduction is to use multiple communication technologies ([Qadir15], [Nielsen8]). This method can be complementary to signal processing /time/frequency/space diversity to increase E2E reliability by exploiting interface diversity gain or path diversity gain. Furthermore, the dual connectivity concept, in which users simultaneously connect to primary and secondary cells, offer another diversity level to enhance robustness in the 5G NR. In [Rao18], Rao et al. show how packet duplication jointly satisfies latency and reliability requirements in some specific scenarios. Besides PHY techniques, some MAC mechanisms are considered to reduce latency. Marsch et al. [Marsch16] have proposed a congestion control scheme providing priorities to URLLC users rather than delaying their access attempts at the service request. Pedersen et al. have proposed a punctured scheduling [Pedersen17] reducing the control latency at MAC layer by prioritizing URLLC traffic over eMBB to immediately schedule low latency traffic. While the standard scheduling is interesting when intermittent traffic is generated to guarantee the spectrum and resource efficiency at the cost of latency to establish the connection, Semi-Persistence Scheduling (SPS) [RT36881] can shorten the latency by periodically allocating resources to the users. The periodicity of the resource grant can be updated according to the channel conditions. Thus, the resources can be adequately allocated to each user in a short period of time, reducing the spectrum usage inefficiency.

3.2 Core Network Technical Enablers

The 5G Core Network (5GC) has been standardized in 3GPP specifications in order to support various new advanced services for industries and consumers and gain more operational efficiency. There are some innovative and powerful features such as the service-based architecture (SBA), using web-scale internet protocols like HTTPv2, cloud-native implementations and deployments, and network automation for service provisioning and service assurance.

In the most frequently mentioned SBA features, the 5GC provides a modular framework from which common applications can be deployed using components from different sources and suppliers. The control plane functionality and common data repositories of a 5G network are delivered by way of a set of interconnected Network Functions (NFs), each with authorization to access each other's services. By means, all the 5G core components can register themselves and subscribe to other NFs via the same Application Programming Interfaces (APIs) which is service-based interface (SBI).

In other words, the SBA is similar to micro-services, in the sense that each NF is formed by a combination of small pieces of software code. This is an important trend and can have profound implications on the next generation core network function. Some micro services can be even reused for different NFs and perform independent life-cycle management. The famous features-network slicing based on this characteristic is becoming a fundamental capability of 5G infrastructures. It enables resource allocation according to traffic characteristics and service level agreements (SLA) i.e. it provides the resources required for a specific service according to the needs of the subscribers, who send from User Plane Function (UPF), Session Management Function (SMF) and Policy Control Function (PCF).

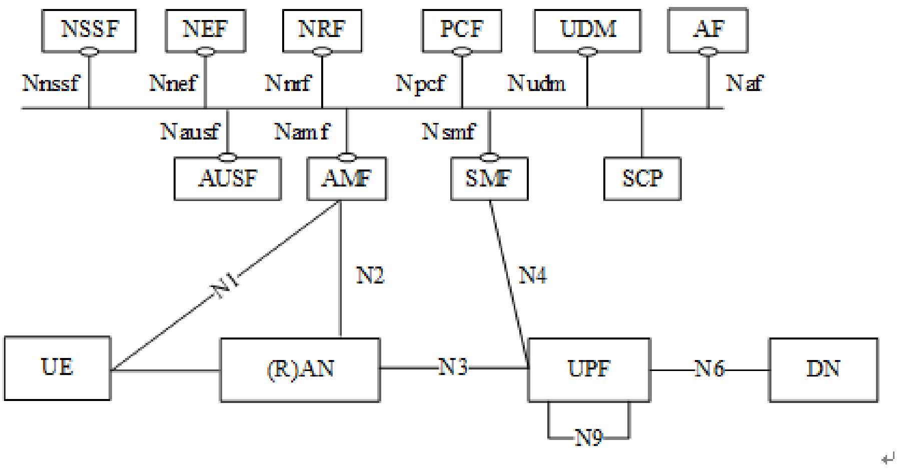

Figure 16: The Service-based Architecture in 5G core

For the control plane and user plane separation (CUPS) architecture in 5GC, multi-access computing (MEC) can distribute user plane functionality break out traffic at the edge and dynamically control the traffic break out by transmitting the edge nodes into user plane function (UPF) nodes. Especially, the Session Management Function (SMF) is implemented in pure software form running on an off-the-shelf x86 server. This implementation gains more flexibility for adding more value-added features. For example, the specialized SMF could interact with Container or Virtual Machine Hypervisor, and transparently redirect service flows to localized serving applications which are collocated on the x86 servers.

3.3 Mobile Edge Cloud Enablers

MEC is an essential solution for the 5G development as it provides computing resources close to the end users. In fact, the idea behind the MEC paradigm is that, by performing applications and processing tasks closer to the enterprise site, an enhancement of network performance is achieved from the point of view of a higher bandwidth, lower latency, increased security, and reduced network congestion.

The ETSI MEC ISG (European Telecommunications Standards Institute Multi-access Edge Computing Industry Specification Group) has published MEC specifications in the recent years. The organization has started developing the overall specification, such as technical requirements (MEC 002) and framework and reference architecture (MEC 003). Then MEC related API specifications have been developed. For example, mobile edge management platform and application (MEC 010-1, MEC 010-2) and different service API (MEC 012-MEC 015).

The multi-access edge computing system reference architecture defined in the ETSI GS MEC 003 specification is showed in Figure 17. There is a Mobile Edge Platform on the middle bottom of the picture. The Mobile Edge Platform has some features, such as data plane process, DNS handling, traffic rules controls and service registry. The applications can be run on the hypervisor in the same mobile edge host. The mobile edge platform includes mobile edge platform, mobile edge applications and virtualization infrastructure. This architecture also includes VIM, VNFM and orchestrator that is similar to NFV specifications. And other mobile edge host is a multi-site MEC module that companies have more than two regions to deploy MEC systems. These two MEC systems share the same applications.

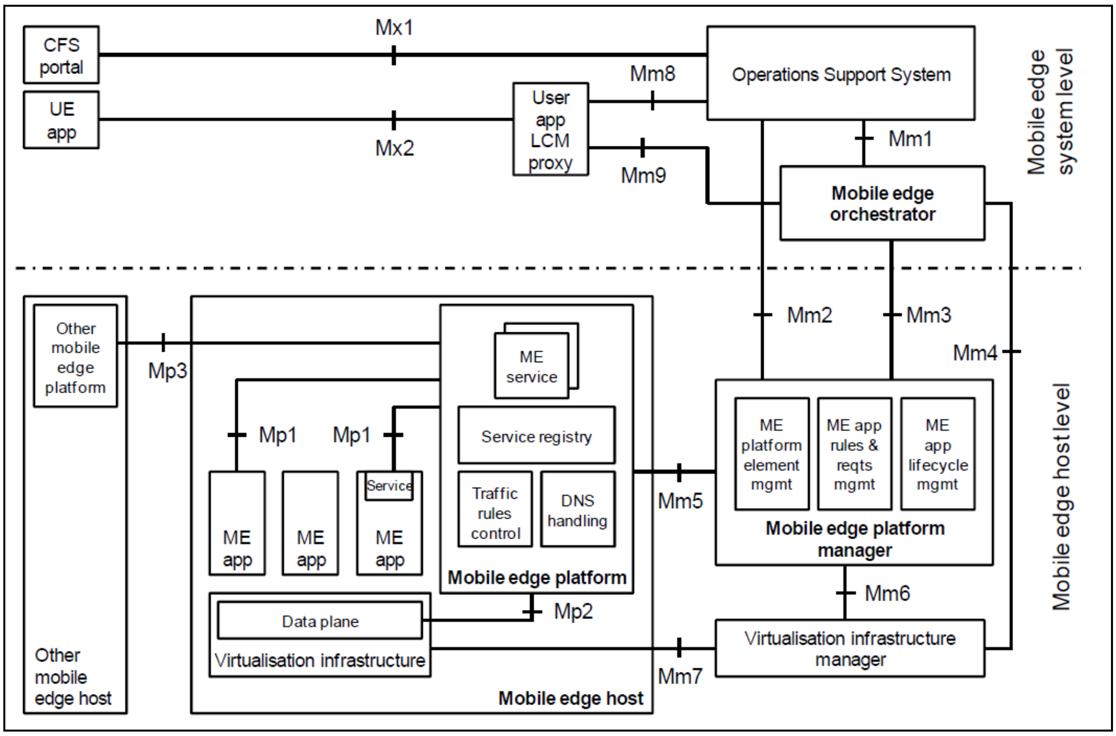

Figure 17: ETSI MEC reference architecture

[4GMEC] illustrated four deployment models of MEC in 3GPP networks. The first one is the bump-in-the-wire model. It is a convenient deployment, because it does not need the additional configurations for the core and RAN network. It has multiple deployment options. The MEC node can be placed close to the eNB or gNB, enterprise site, or aggregation site. Multisite deployment is possible too. The traffic can be setup to route to different MEC sites. A bump-in-the-wire MEC solution needs extra efforts to support similar 3GPP standard functions. The advantage of bump-in-the-wire MEC is to distinguish zoom in or zoom out easily duo to dedicate base stations for enterprises. Traditionally, the mobile network needs to be distinguished area by planning a new tracking area.

The second option is distributed EPC, where all mobile core network functions (NFs) are distributed to the network edge on a small-form-factor server or as VNFs on a MEC platform. In the pathway towards 5G, this corresponds to a full 5GC co-located with the MEC host. This concept is also described as MEC collocated with the core network functions of MEC deployment scenarios in [5GMEC].

The third option is distributed S/P-GW, where user plane (UP) NFs and few control plane (CP) NFs are distributed on a small-form-factor server or as VNFs on a MEC platform. In the pathway towards 5G, this corresponds to a hybrid 5GC solution.

A final option consists of a distributed S-GW with local break-out, where, ideally, only UP NFs are distributed on a small-form-factor server or as VNFs on a MEC platform. This is another variant of the hybrid 5GC solution. The last two options correspond to 5G as a local UPF solution described in [5GMEC].

NFV is one of the key technologies in the development of MEC. With the development of network function virtualization technology, a growing number of enterprises are studying the feasibility of replacing traditional network equipment with NFV technology. The reason is that NFV technology can provide a more flexible, low-cost, and faster pattern to establish a set of network functions than traditional network equipment. Many NFV technology developers comply with the NFV MANO (Management and Network Orchestration) defined by ETSI. NFV MANO can be used to manage the lifecycle of VNFs, deployment methods, network planning, and other requirements of NFV.

NFV solutions have many open-source solutions, such as Open Baton [BATON], OSM [OSM], and ONAP [ONAP], etc. All of them can provide the NFV technology foundation for enterprises, and each solution has different VNF installation methods and capabilities of the ETSI NFV MANO. Therefore, the VNF template cannot be shared between different NFV solutions. So MANO platform providers have to consider a feasible solution to effectively support ETSI NFV MANO and integrate multiple VNF templates of different NFV solutions.

The MANO solution specification is developed by ETSI NFV ISG, including MANO international standard architecture and API interface. The format and specifications of VNF templates are defined in the ETSI GS NFV-SOL 001 and 004 [ETSI01] [ETSI04] specifications. ETSI GS NFV-SOL 002 specification [ETSI02] is regarding monitoring VNF resources to collect the status of VNF resources for users. Manage VNFs such as lifecycle management, data model, and performance management for VNF is based on the ETSI GS NFV-SOL 003 and 005 specifications [ETSI03] [ETSI05].

3.4 Industrial Application Enablers

One of the most promising services enabled by edge computing is computation offloading, whose aim is to transfer the execution of computationally heavy applications from mobile/sensor device to nearby edge servers. A device may choose to offload an application, either because it is equipped with poor computational capacity, or because it does not have any computational resources (e.g., in the case of a sensor). This paradigm enables a (possibly energy efficient) program execution within low E2E delay constraints. Here, we refer to E2E delay as the time elapsed from the generation of a data unit, or a new task to be offloaded, until the result of its computation is sent back to the end device. Computation offloading is an application that really fits the factory environment, since the data to be offloaded could be collected by sensors within the factory, about the actual behavior of all the machineries involved in the production line. On these data, the edge server could run anomaly detection, classification, prediction algorithms to autonomously take decisions. This of course requires data exchange and computational capabilities in close proximity to the end application/user. In this deliverable, within Task 4.4, we focus on dynamic computation offloading. In this case, data/subtasks are continuously generated by the sensor/mobile devices (sometimes at unknown rate), so that they need buffering before transmission and computation. Usually, in this scenario, average and/or probabilistic latency constraints are imposed, since data usually experience a queueing delay, both from a communication and from a computation point of view. The challenge is to effectively manage resources in complex and time varying environments. Then, it is useful to present the general topics treated here, to categorize the state of the art. Within Task 4.4, we touch the following topics:

- *General dynamic computation offloading* with E2E delay guarantees, in which the objective is the mobile/sensor devices and/or the network energy consumption, with E2E delay guarantees: In this case, we consider a general application, with delay requirements, without diving into the details of the application itself;

- *Edge Machine Learning* with E2E delay and learning/inference performance guarantees, which is a particular (and natural) application of computation offloading, where end devices upload data to be processed at the edge servers running machine learning tasks (classification, prediction, estimation, etc.), and is typical of industrial environments. In this case, the specific requirements of the application, also in terms of learning/inference performance, need to be properly taken into account.

Then, we first revise the literature on general dynamic computation offloading, to then present the state of the art on the topic of Edge Machine Learning. Let us notice that the state of the art on this topic is wide, so that we will revise the most related works. The interested reader is referred to the recent surveys related to MEC and computation offloading for a more comprehensive view [Mach17], [MaoYou2017], [Jiang2019], [Pham2019].

3.4.1 Dynamic computation offloading

We now present the state of the art on general dynamic computation offloading. The dynamic formulation was initially investigated in [Huang2012], with a strategy based on Lyapunov optimization in a cloud computing framework. In [Yang2018], the authors consider a fog-enabled Device to Device (D2D) scenario and propose a strategy to associate mobile devices and offload tasks among each other. User assignment is also addressed in [SunZhou17], with the aim of minimizing the average delay under energy constraints, using a penalty function that discourages frequent handovers, while hinging on a multi-armed bandit algorithm to learn the optimal association. In [Merluzzi19], a dynamic computation offloading algorithm to optimize computation and communication resources jointly with the assignment of mobile users to APs and edge servers is proposed, hinging on tools from stochastic optimization and matching theory. In [Liu2017], the authors introduce a probabilistic constraint on the length of the computation queues, written as a bound on the probability that the queue length exceeds a certain value. Probability models based on extreme value theory (EVT) were used in [Liu2017] to handle very low probability events. Similarly, in the context of vehicular communications, the authors in [Ashraf2018] propose an algorithm aimed at minimizing the long-term average power consumption under reliability and latency constraints. Then, the work in [Liu2017] is extended in [Liu2019] to a multiple AP and edge server scenario, where an assignment strategy based on matching theory is proposed. In [Merluzzi2020], the authors propose a joint dynamic computation offloading strategy with reliability guarantees, incorporating ultra-reliable low-latency communications (URLLC) and energy harvesting devices. The authors in [Li2018] propose a deep reinforcement learning approach to minimize a weighted sum of the energy consumption and the offloading delay. [HanChen20] proposes a dynamic strategy aimed at minimizing the average power consumption of mobile devices, under a latency constraint and a constraint on the edge server average power consumption. Recent contributions consider the energy consumption of both radio access and MEC network [LiGuan19], [ChenZhou17], [Wang19], [ChangMiao18-2], [Nan17]. In particular, in [LiGuan19], a scheduling strategy is proposed to find a trade-off between task completion ratio and throughput, hinging on Lyapunov optimization. [ChenZhou17] aims at minimizing the long-term average delay under a long-term average power consumption constraint. In [Wang19], the long-term average energy consumption of a MEC network is minimized under a delay constraint, using a MEC sleep control. Also, in [ChangMiao18-2] the problem is formulated as the minimization of the energy consumption under a mean service delay constraint, optimizing the number of active base stations and the computation resource allocation at the edge server, while considering a sleep mode for both APs and edge servers. In [Nan17], Lyapunov optimization is used to reduce the energy consumption of a fog network while guaranteeing an average response time. [Mao2017] proposes a dynamic computation offloading strategy based on stochastic Lyapunov optimization to reduce a weighted sum of User Equipment (UE) and ES power consumption. In [Yu17], deep learning techniques are exploited for resource allocation in MEC, with the aim of minimizing an overall offloading cost comprising a local cost at the mobile device and a remote cost in the MEC network. In [YuPu2018], the authors exploit Lyapunov optimization, Lagrange multiplier, and sub-gradient techniques to optimize device and APs energy consumption under latency constraints, exploiting AP sleep states.

3.4.2 Edge Machine Learning

The aim of EML is to investigate resource allocation strategies with the final goal of running a machine learning task at the edge of the network. It is then clear that enabling edge machine learning introduces novel fundamental problems in terms of jointly optimizing communication (e.g., power, bits, source encoding, etc.), computation (e.g., CPU cycles, number of active servers/cores, etc.), and inference/training (e.g., choice and splitting of the (deep) learning architecture, model and/or data partitioning, etc.) to meet the system constraints (e.g., latency, reliability, energy) while guaranteeing a prescribed performance of the inference task. This joint optimization translates into a control action to be performed in real-time to strike an optimal trade-off between energy, latency, accuracy while coping with the time-variability of radio channels, data arrivals, computation loads, memory availability, etc. In particular, it is possible to consider the trade-off between energy consumption and delay, between accuracy and delay, energy consumption and accuracy, or their joint combination. Thus, the description of the application and the optimization of resource allocation must take into account the particular requirements of the learning task. A first general introduction to EML can be found in [Park18], where the authors present several possible trade-offs to be explored. Other recent surveys are [ChenRan19], [WangHan20], [Hellstrom2020]. The authors in [Skatchkovsky2019] consider an edge machine learning system, where an edge processor runs an algorithm based on Stochastic Gradient Descent (SGD). In particular, they investigate the trade-off between latency and accuracy by optimizing the packet payload size, given the overhead of each data packet transmission and the ratio between the computation and the communication rates. In [Mohammad2019], the authors propose an algorithm to maximize the learning accuracy under latency constraints, while the authors in [Amir2019] present a distributed machine learning algorithm at the edge, where wireless devices collaboratively minimize an empirical loss function with the help of a remote server. The authors in [Elgabli2020] propose a communication-efficient decentralized machine learning algorithm that dynamically optimizes a stochastic quantization method, with applications to regression and image classification. The authors in [WangTour18] consider generic distributed machine learning algorithms at the edge, based on SGD, investigating the trade-off between local update and global aggregation. In [Azar19], the authors present a data compression algorithm to reduce the communication burden and energy consumption of an IoT network, to enable machine learning with a desired target accuracy. EML is a research topic at its infancy, so that the research community just started investigating the possible directions. Finally, in [MDB20], a dynamic resource allocation with latency and accuracy guarantees based on Lyapunov optimization is proposed, with a model-based solution for the accuracy of a learning algorithm based on Stochastic Gradient Descent (SGD).

4 Radio Network Technical Enablers

The objective of Task 4.1 is the development of the radio network consisting of CPE and gNodeB systems that comply with 3GPP standard based on the industrial use case generated by WP1 and the investigation of RAN orchestration enabling deterministic URLLC services.

4.1 5G RAN system composed of CPE and gNB

The first system is the CPE (Customer Premises Equipment). Alpha Networks has developed a 5G NR CPE device that complies with 3GPP standard. CPE changes 5G NR to WiFi hotspot and LAN, thus field device can access 5G NR network. The main features of 5G NR CPE are listed as below:

- 5G NR Sub-6(WAN)
- 2x GE Ethernet ports(LAN)
- WiFi Dual band(WLAN)

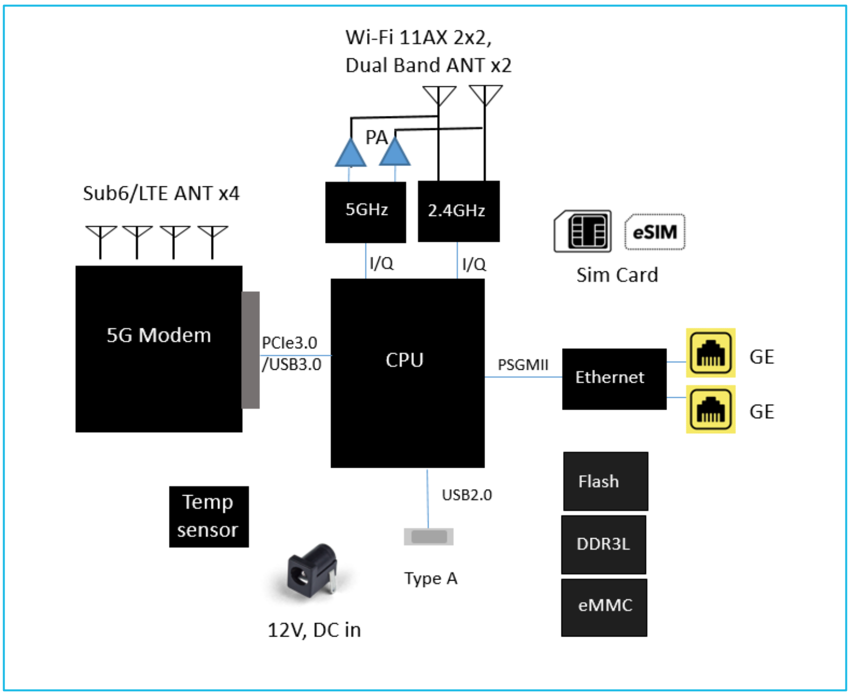

Figure 18: CPE architecture

The second system is the gNodeB. Alpha Networks has been developing a disaggregated gNodeB solution that is consisted of a RU and a CDU.

- RU: The radio unit implements lower physical layer with split option 7.2 defined in ORAN standard. It needs to support CFR (crest factor reduce) and DPD (digital pre-distortion) function to comply with 3GPP standard. In RF front-end system, the PA (power amplifier) RF component boosts RF power at the transmitter side, and the LNA (low noise amplifier) provides a better sensitivity level at the receiver side. The key considerations of RU design are size, weight, and power consumption, and target RU spec as below.
  - Matrix: x4TRx
  - TX Power: +24dBm (250mW)
  - Frequency band: n78 (3700MHz-3800MHz) / n79 (4800MHz-4900MHz)

- Dimension: 218mm x 218mm x 65mm
- PoE power supply
- Synchronization: 1588v2
- Comply with 3GPP standard TS 38.104

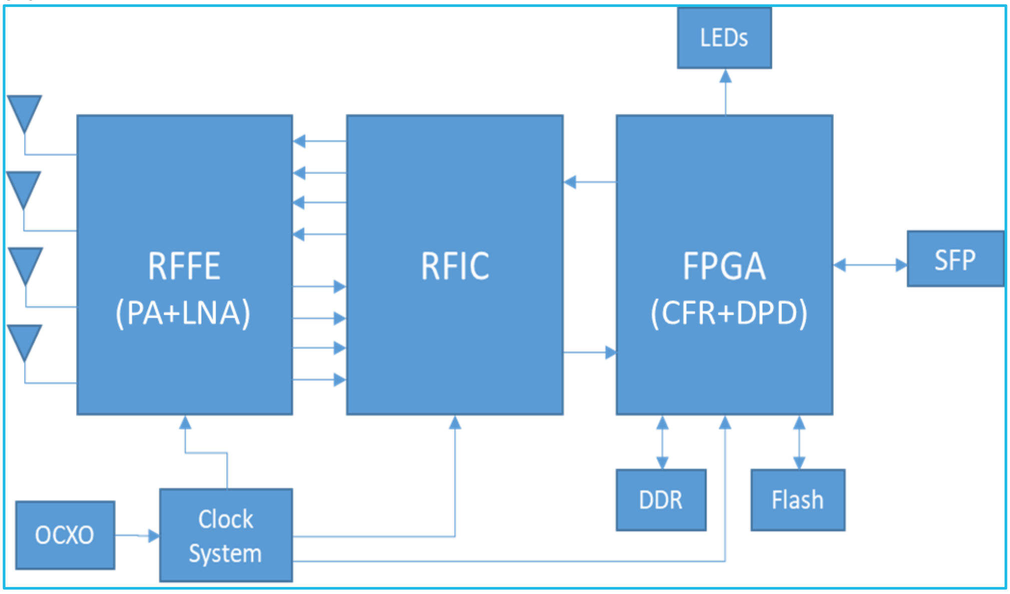

Figure 19: 5G NR RU architecture

- CDU: 5GNR CDU is consist of DU and CU. DU runs the RLC, MAC, and parts of the physical layer, and its operation is controlled by the CU, the centralized unit that runs the RRC and PDCP layers. The main features of CDU are listed as below:
  - 10G SFP x8, 25G SFP x4
  - DL 256QAM / 4 layers
  - UL 64QAM / 2 layers
  - Maximum support x4 RUs
  - X4 UE/TTI
  - Connected UE: 256+

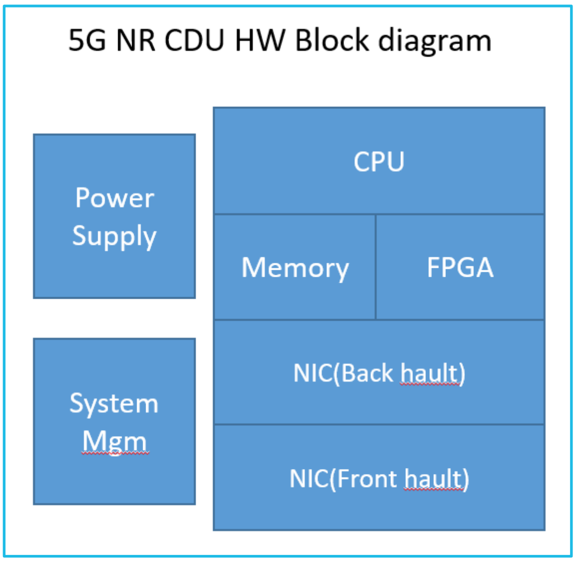

Figure 20: 5G NR CDU architecture

Since the earliest phases of the New Radio study, it was felt that splitting up the gNodeB between Central Units (CUs) and Distributed Units (DUs) would bring flexibility. A distributed unit (DU), responsible for real time L1 and L2 scheduling functions, and a centralized unit (CU) responsible for non-real time, higher L2 and L3.

3GPP started studying different functional splits between central and distributed units. They have proposed about 8 possible CU DU split options:

1. Option 1 (RRC/PCDP 1A-like split)
2. Option 2 (PDCP/RLC Split 3C-like split)
3. Option 3 (High RLC/Low RLC split, Intra RLC split)
4. Option 4 (RLC-MAC split)
5. Option 5 (Intra MAC split)
6. Option 6 (MAC-PHY split)
7. Option 7 (Intra PHY split)
8. Option 8 (PHY-RF split)

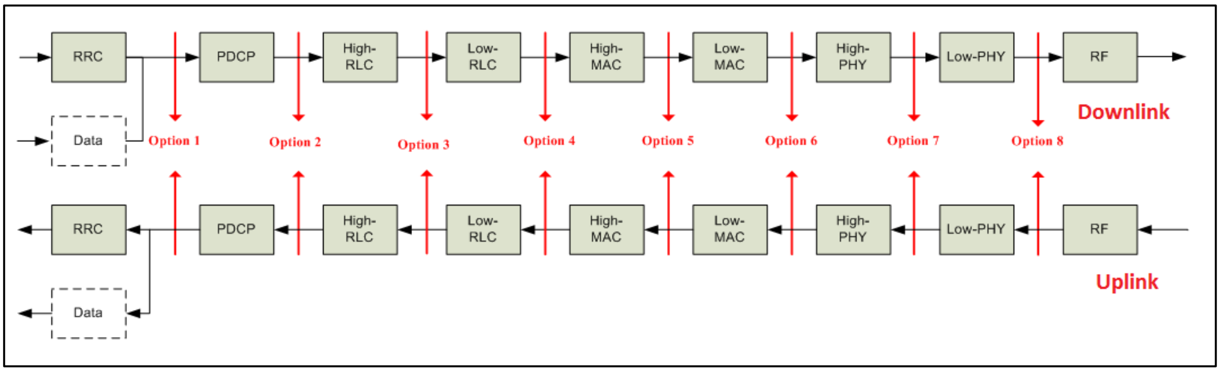

Figure 21: The proposed 8 possible CU DU split options

Some benefits of an architecture with the deployment flexibility to split and move NR functions between central and distributed units are below:

- Flexible hardware implementation allows scalable cost-effective solutions
- A split architecture (between central and distributed units) allows for coordination for performance features, load management, real-time performance optimization, and enables NFV/SDN
- Configurable functional split enables adaptation to various use cases, such as variable latency on transport

Alpha Networks has built the prototype of 5G NR gNodeB and CPE equipment, the end-to-end lab integration test is being conducted. The gNodeB and CPE will be deployed at ITRI IMTC at the end of month 26 to realize the selected use cases.

Initial End to End test result is shown as below table, the system performance is being optimized continuously. As the integration test result by month 16, the End to End throughput is up to more than 240 Mbps.

Table 4 : End-to-end test of 5G NR gNodeB and CPE equipment

gNB FW Version	OTA	gNB Config		RU config		CPE (DL Signal)			Throughput performance	
		DL MCS	UL MCS	TX Gain	RX Gain	RSRQ	RSRP	SINR	UL (Mbps)	DL (Mbps)

V1.73 P2	1M	28	20	11	3	-11	-77	31	17	241
	10M	28	15	21	15	-11	-80	29	57	235
	30M	22	9	21	15	-11	-84	29	33	164
	100M	9	9	21	20	-11	-107	13	2	25

4.2 Deterministic URLLC protocols

Based on integrating AI into unified framework as illustrating in [Jiang19] and the architecture of 5G in [TR19], we propose a framework to design a E2E orchestration to control the entire network to guarantee not only URLLC but others heterogeneous services by leveraging ML at the level of orchestrator. Furthermore, in order to overcome the extreme events caused by impairments such as non-stationary channel dynamics and exogenous uncertainties, the concept of xURLLC as depicted in [Park20] has emerged to enable proactive decision making in dynamic environment powered by ML.

The proposition of orchestrator design is to smartly coordinate UEs, BSs and network entities and to find the optimal subset of preferable mechanisms to configure the RAN, MEC and CN in order to fulfill the targeted KPIs without wasting resources. Figure 22 illustrates the orchestrator framework.

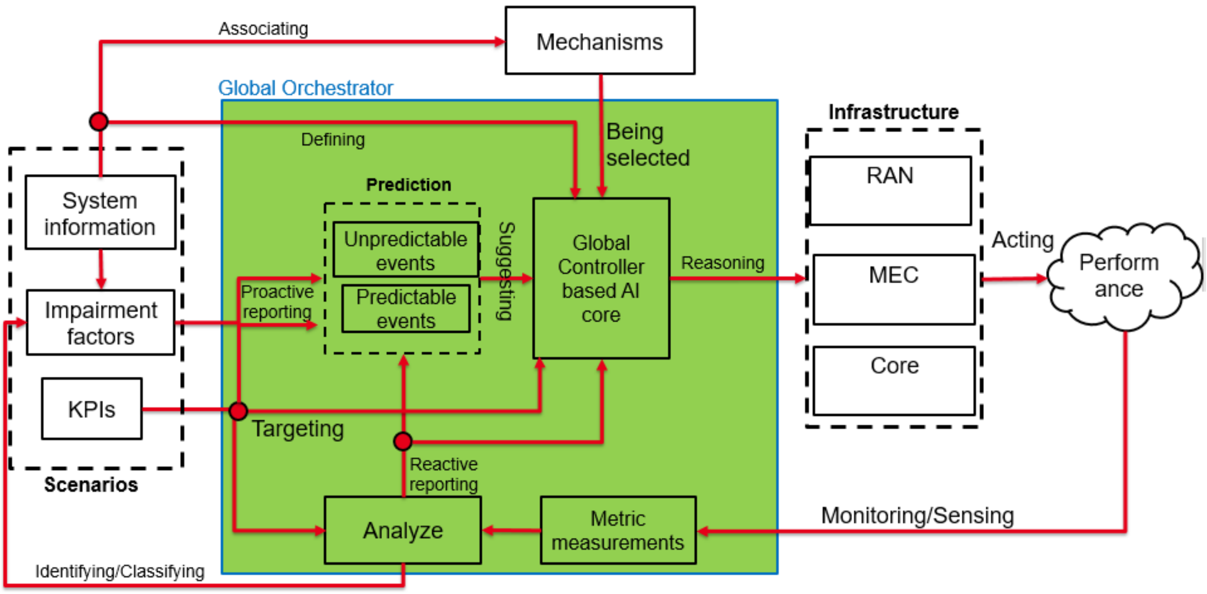

Figure 22: Framework to design a global orchestration

The inputs of the orchestrator are the following: System Information, Impairment Factors, KPIs and Mechanisms.

System Information block refers to user's and infrastructure's level which enable the appropriate mechanisms to support URLLC. The user's level information defines each user capabilities, such as the configuration of user's antenna, the supported radio access technologies, battery limitation. Based on that, the orchestrator will coordinate the appropriate strategies to adapt to the required services of each user by (de)activating available mechanisms. The infrastructure information related to the computation capabilities, storage capacities, network standards will be taken into account as a constraint of the performance that the orchestrator will reason to adapt to users requirement heterogeneity.

KPIs block presents the network states at different network layers as detailed in the Figure 23.

Layer / KPI	RAN			Edge	Core	Application
	PHY	MAC	RLC/PDCP			
Latency	Transmission time	Latency (control\Data) MAC Jitter	Latency (framing)	Computation Latency Queue length Network Jitter	Core Latency (Transport\backhaul) Core Jitter	End to end latency Survival Time Determinism
Reliability	Reliability over the air (BER, BLER)	PHY Robustness (multi-UE) Reliability (PER, PDR)	Reliability (framing)	Computational reliability	Core reliability (Transport\backhaul)	Service reliability Service availability
Usage	Spectral efficiency Throughput over the air	Resource usage Overhead MAC throughput	Spectrum reuse Network capacity	Percentile of support Computation capacity	Core capacity	Effectiveness Service bitrate Application throughput
Energy	Energy efficiency	Energy consumption		Energy consumption		Battery lifetime
Complexity	Complexity	Complexity	Complexity	Complexity	Complexity	
Coverage	Range	Mobility support	Mobility support Network coverage	Mobility support		Connection density Service area Area traffic capacity

Figure 23: Classification of KPIs according to the network layers.

As an example, the E2E latency is composed of many elements in RAN, edge and core network. In particular, the RAN latency corresponds to the transmission time (i.e. to transmit the physical block data in PHY layer) or the latency for grant establishment which depends on scheduling scheme at MAC, processing time at UE/BS, and notification messages at higher layers (RLC, PDCP). The edge latency corresponds to the computational time, queue delay, feedback and forwarding delay, and time for the reconfiguration in the MEC. The transport latency is the time that transport network needs to transfer the packet to the target direction. Similarly, the E2E reliability, usage, energy efficiency, coverage and complexity could be considered.

Concerning the Impairment Factors, several factors can degrade the performance at different network layers. With appropriate information and prediction leveraged by ML, the orchestrator can prevent predictable events degrading the system performance. In case of unpredictable events, the orchestrator can avoid performance degradation and bound the performance. From the communication point of view, reliability impairments are due to decreased power of the useful signal, uncontrollable interference, resource depletion, protocol, reliability mismatch and equipment failure. The source of perturbation can be due to time-varying channel and channel uncertainty due to mobility or changing environment, interference (e.g., collisions with other users in uncoordinated channel access, coexistence with other systems in the same frequency bands) or imperfect knowledge of the environment (e.g., inaccuracy of the measurement or outdated information). Besides, intermittent connectivity, time varying traffics or varying the number of mobile users can also cause latency extension at RAN. From computation and caching point of view, the impairments are linked to the unavailability of computation resources due to the congestion or overload of tasks. Besides, the bad coordination between several edge servers in case of multi-connectivity of user can lead to the resources wastes and latency increase. Concerning the utilization of ML in the CN, the accuracy gaps between the training phase and test phase could also generate errors.

Mechanisms block exploits the mechanisms classification in terms of time, frequency, SP, space and hardware diversity over the RAN, MEC and Core network as depicted in Figure 15. It provides a large number of solutions enabling URLLC through the network. The E2E orchestrator provides several closed-loop management processes. Firstly, the system information will inform the orchestrator about behaviors and capabilities of the network at a specific region. Afterwards, it will measure the network states and adequately applies the appropriate mechanisms at different network levels to achieve the performance of the system. In the feedback direction, the orchestrator will analyze the performance by comparing the metrics and the KPI requirements. This comparison will adjust the decision to re-orchestrate the network in a reactive way in case of difference between expected and real performance. The metric measurement and the analyzer is the key to enable close-loop management when it monitor the performance as well as the network state during the operation. The analyzer will also identify and classify impairment factors or faults in the network into predictable and unpredictable events. With appropriate information and prediction leveraged by ML, the orchestrator can prevent predictable events degrading the system performance and can avoid performance degradation from unpredictable events by bounding the performance. Based on the suggestion from the prediction, the orchestrator will proactively reconfigure the network.

First simulations of 5G NR networks were done using NS3 simulator in order to evaluate the impact of combined mechanisms and networks management for URLLC. In particular, we will focus on numerology (Frame design), micro-diversity (MIMO), HARQ (Redundancy) and adaptive MCS. We consider an E2E 5G network framework based on new radio module [Patriciello19] implemented in NS-3 network simulator. The core network is based on the design of LTE/EPC model whilst the RAN takes into account sub-6 GHz communications in 5G NR architecture. According to KPIs defined in Figure 23, the performance are evaluated in terms of E2E latency, network jitter and RAN reliability.

The considered scenario is 1 UE equipped with a (2x4) Planar Array (PLA) antenna communicating to its BS with a (4x8) PLA antenna and requesting the service from an UDP server. The space diversity is thus initially achieved by MIMO antenna at UE and BS side. The sub-6GHz 5G NR communication characteristics are the following: 20 MHz bandwidth in 3.5 GHz band, three different numerology (0, 1 or 2) and LDPC channel coding associated with various MCS (MCS5, MCS12, MCS17 and MCS25). These MCSs are respectively the four different modulation orders (4-QAM, 16-QAM, 64-QAM, 128-QAM) and the four coding rates (0.37,0.42, 0.43 and 0.8). The numerology and MCS (time, frequency and SP diversity) are applied to reduce the delay at RAN level. To improve the reliability, incremental HARQ applies redundancy in time domain. The selected propagation model is the indoor factory channel model recommended in [ITUR17]. The fast-fading channel ,which has a coherence time of 10 ms, is a source of impairment for the transmission. Another impairment is the dynamic generation of the traffic at the application layer (exponential distribution of message size and inter-arrival time with respectively a mean of 60 Bytes and 1 ms).

Table 5 : Simulation parameter of URLLC simulations using NS-3

FoF Parameters	Values
Service Type	UDP
Message Size	Exp. Distribution (Mean= 60B)
Message inter-arrival time	Exp. Distribution (Mean= 1 ms)
Device	1 UE vs 1 BS
Frequency	3.5 GHz
Bandwidth	20 MHz
Distance UE-BS	120 m
Pathloss Model	Indoor Factory
Numerology	Num0, Num1, Num2
MCS Scheme	MCS5, MCS12, MCS17, MCS25
Coding Rate	0.37, 0.42, 0.43, 0.8
HARQ	Incremental Redundancy

Figure 24 shows the distribution of the E2E latency across the different layers as a function of the numerology. The delay for the backhaul and transport network is fixed to 1ms. The E2E latency is mainly affected at PHY and MAC layers where TTI and TDMA (Time Diversity) have been successfully shortened in relation to the frame structure and the scheduler delay. Figure 24 also illustrates the jitters caused by the mismatch between message generation at the application layer and resource allocation at MAC layer.

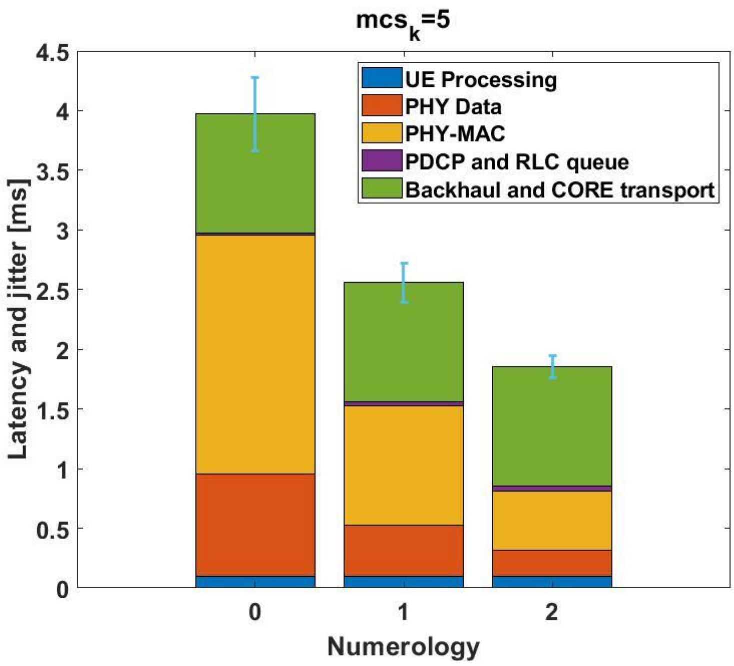

Figure 24: Distribution of E2E latency across the network layers

Figure 25 extends the study done in Figure 24 by considering the E2E latency and the reliability as a function of the numerology and the MCS. As expected, low MCSs (e.g. MCS5) have more robust communications and higher latency whereas higher MCSs provide higher throughput, lower latency but at the cost of the reliability.

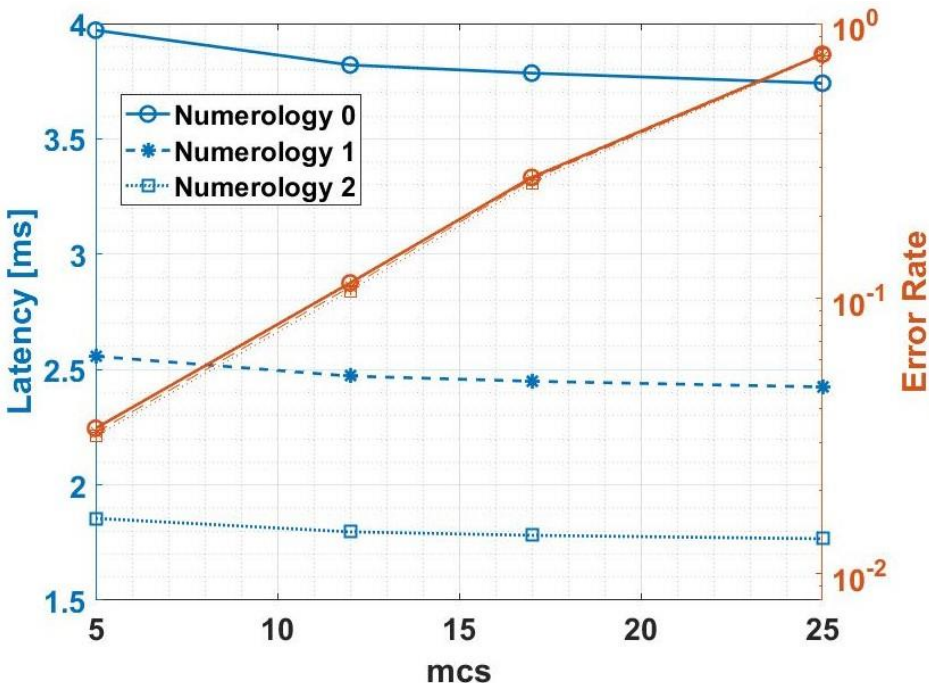

Figure 25: E2E Latency and Packet Error Rate

Figure 26 and Figure 27 respectively depict the profile and the Cumulative Distribution Function (CDF) of the E2E latency for MCS17 with the different numerology with or without HARQ. The integration of HARQ and higher numerology allows us to both improve reliability and reduce latency, however, the inevitable generation of adjacent delay cluster (time diversity of the retransmission) degrades the determinism of the communication.

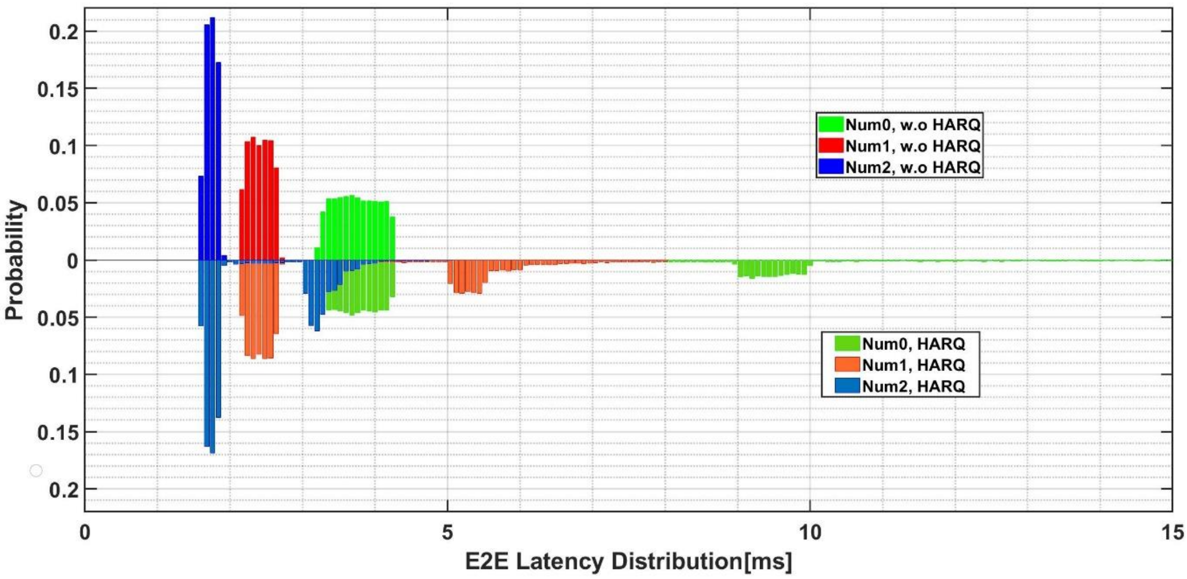

Figure 26: E2E latency distribution of MCS17 with or without HARQ

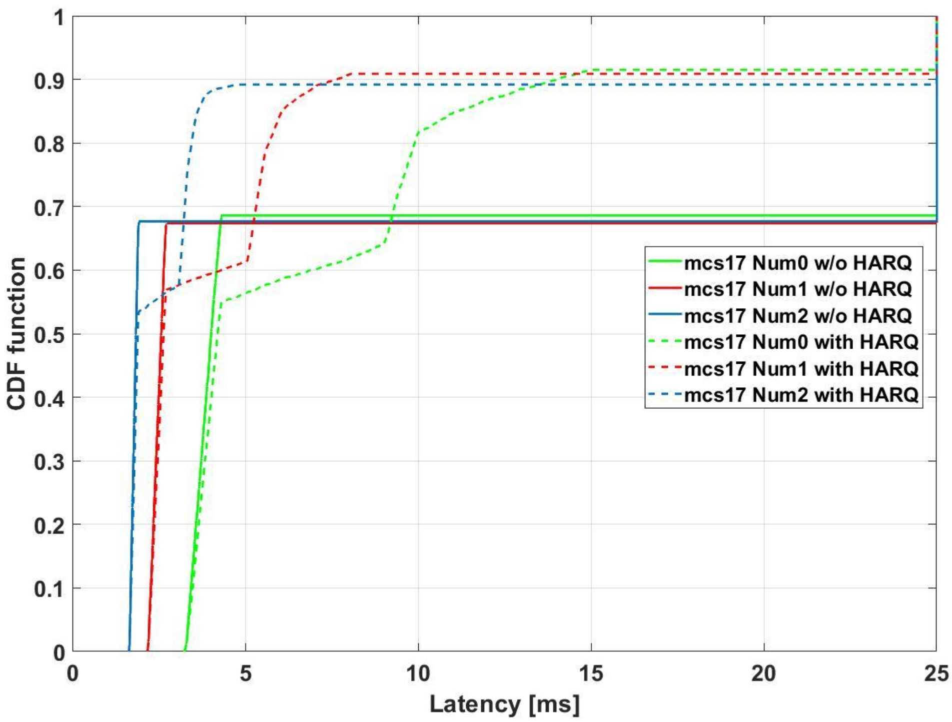

Figure 27: CDF of E2E Latency of MCS17 with or without HARQ

Figure 28 compares the performance of Adaptive Modulation and Coding (AMC) mechanism exploiting MCS diversity with robust MCS5 and high throughput MCS17 in a fast fading channel. The CDF curves show that MCS5 with HARQ outperforms AMC and reaches beyond 99% reliability for less than 5 ms latency, compared to 81% and 61% for respectively AMC and MCS17. This highlights that adapting the solution is not always the best solution to tackle signal degradation in a fast-fading channel and some margins are necessary to avoid unpredictable events and to bound the performance.

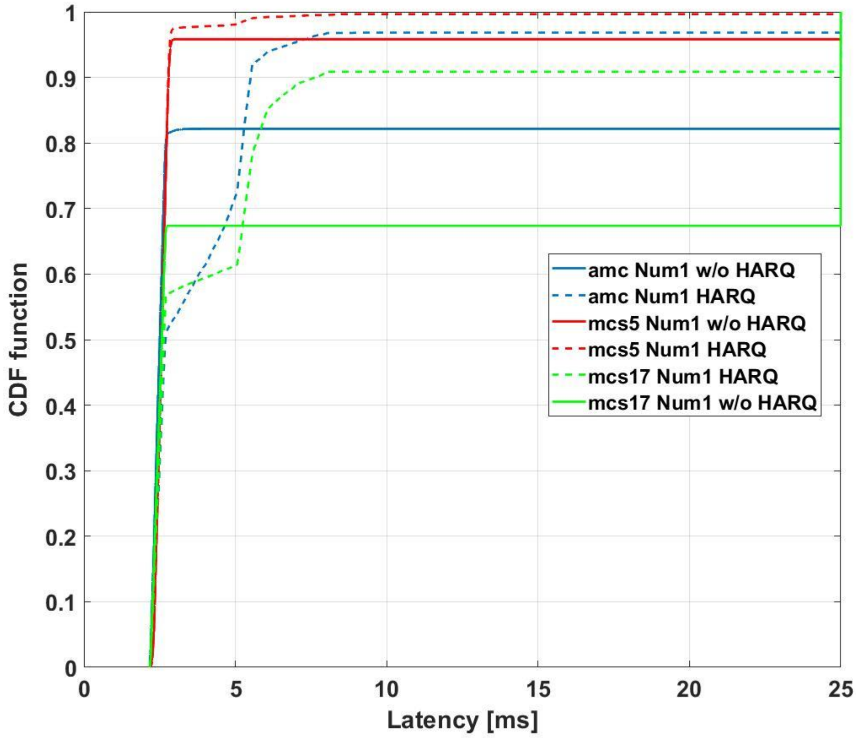

Figure 28: CDF of E2E Latency for MCS5, MCS17 and AMC

4.3 Conclusions

Thanks to NS3 network simulator, we have developed an E2E network framework able to combine several mechanisms exploiting modulation and coding diversity (LDPC, MCS, AMC), space diversity (MIMO) and time diversity (Frame Design, HARQ) at RAN level, guaranteeing URLLC communications. The performance have been evaluated in terms of reliability, E2E latency and jitter for dynamic traffics and fast-fading channel in indoor office scenario.

The obtained results allow us to know the behavior of different combinations toward URLLC performance and are a first step for further research using AI. For example, the mismatch between message generation at application layer and resource allocation at MAC layer or HARQ re-transmission can cause jitter and delay cluster and should be avoided for deterministic communications. Moreover, the adaptation of MCS provides a good trade-off between throughput, latency and reliability but may not be adapted for unpredictable events such as fast fading channel. In this case, some margins on budget link could be more relevant.

In future works, the promising findings will be extended with additional mechanisms (e.g., scheduling and multi-connectivity) and the complete evaluation of the AI based orchestrator that dynamically and efficiently manages 5D diversity mechanisms, will be evaluated.

For 5G NR RAN implementation, we have built up a RAN system composed of CDU, RU, and CPE based on the RAN design consideration of section 2.2, which is able to be tested against the 5G core to conduct the interoperability tests. In future works, we are going further to optimize the performance of gNB to fulfill the requirements of the selected use cases defined by Task 1.1.

5 Core Network Technical Enablers

Task 4.2 is focused on the development of the core network components to realize private local 5G networks to meet the envisioned industrial application requirements.

5.1 NFV-like lightweight orchestration framework for the core network

In the context of the laboratory activities for the EU demonstrations, a 5G core prototype including basic functions like AMF, SMF, AUSF, UDM, and UPF is meant to run as a set of VNFs. In this way, a lightweight orchestration framework can be investigated to implement the lifecycle management of the aforementioned virtual instances. Special focus has been reserved to the analysis and implementation of the SOL002 interface (see Figure 29), as per ETSI NFV, to enable orchestrators such as Open Source MANO (OSM) to interact with the core VNF. The implemented SOL002 interface (compliant to specification version 3.3.1, released in August 2020) provides a RESTful API and new data models which set the instantiation, configuration, and termination of VNF components. These standards are going to be studied and investigated by means of several experiments.

The adopted testbed architecture for interface and VNF instantiation and configuration tests is made of two open-source pieces of software. OpenStack acts as the VIM and NFVI, providing the host infrastructure. OSM manages and orchestrates the NFV ecosystem, by contacting the Element Manager (EM) of a 5G network function through the Ve-Vnfm interface. At the time of writing this document, the current OSM version is 8.

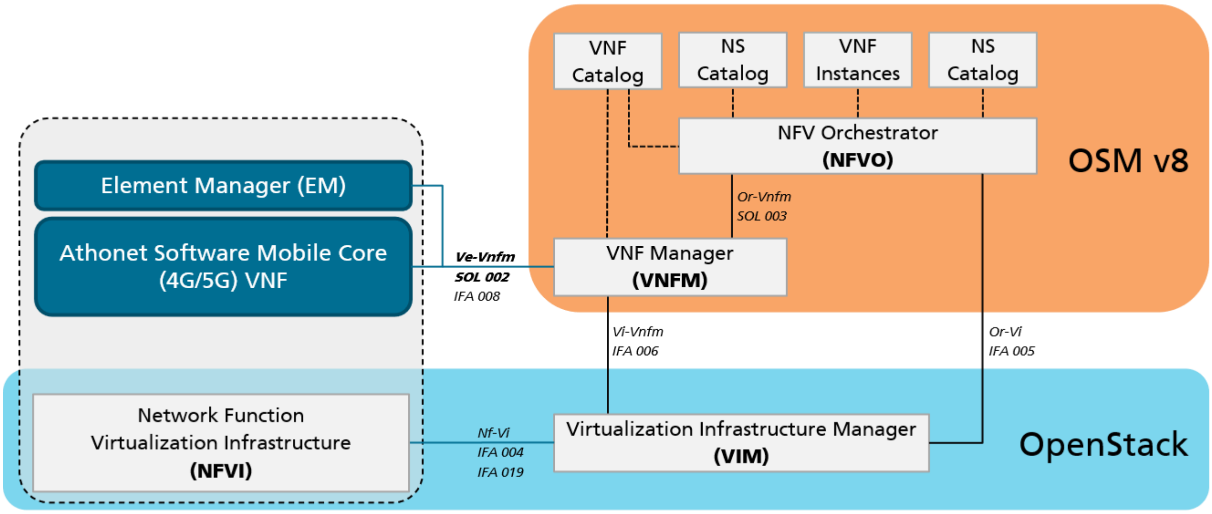

Figure 29: Overview of NFV frameworks and interfaces

In the implementation of the core as a VNF, the VNF contains its own EM and a variable number of VNF Components (VNF-C). The EM terminates the IFA 008/SOL 002 interfaces and is connected to all the other VNF-Cs of the VNF for management and configuration purposes. The VNF may contain a variable number of additional VNF-Cs, according to the desired association between a network element and its Virtual Deployable Unit (VDU). In this implementation, a VDU is a Virtual Machine (VM), and hence a VNF-C is a VM too. For instance, only one additional VNF-C, i.e., one VM, can be used to run all the main core components or one VNF-C is deployed for each of the NFs.

The 5G core lends itself to different VNF-C configurations, which are collected into the VNF Descriptor (VNFD) that is used to feed the MANO system. The VNFD contains the number and flavours of VNF-Cs that make up the VNF, along with the internal and external connection points. The former represents the virtual links among the VNF-Cs, whereas the latter are the links to other VNFs. In addition, the VNFD contains all the VNF-specific configuration data that need to be pushed to the VNF in order to execute the core functions. Such specific data range from 3GPP parameters and identifiers to vendor-specific items, e.g., logging behaviour. The VNFD is available in both the TOSCA and YANG models, as specified respectively in ETSI GS NFV-SOL 001 and ETSI GS NFV-SOL 006.

Instantiation and configuration procedures

As previously described, the VNF is onboarded into OSM via a VNFD. More precisely, the onboarding procedure happens by means of VNF and network service (NS) packages, which contain the descriptors themselves. As a second step, the initial setup of the core network (Day-1 procedure) is performed via the Juju proxy charm that is included in the VNF package and contains all the VNF configuration instructions. By owning descriptors and charms, OSM can automate the full lifecycle of the NS. A proxy charm is a collection of YAML descriptors, files and executable core designed to send the configuration to the EM compliant with the SOL002 specification over Ve-Vnfm.

For a mobile core network deployment, a large part of its configuration is not supported by the NFV standard, which is not aware of the specific core network settings. Hence, the optional field `vnfcSpecificData` has been exploited to bypass this limitation. It contains a list of generic key/value pairs, which can be used to allow specific configurations not supported by SOL002 classic parameters. The content of SOL002 (stored as a JSON file) request is illustrated in Figure 30.

```
{ "vnfcConfigurationData": [
  { "vnfcInstanceId": "<vnfc_id>",
    "CpConfig": [
    { "cpId": "<interface_name>",
      "cpdId": "<phy_interface_type>",
      "addresses": [
      { "address": {"ipAddress": "<address>/<netmask>"},
        "useDynamicAddress": "<boolean>"
      }]
    }],
    "vnfcSpecificData": "<additional_conf>"
  }]
}
```

Figure 30: Content of SOL002 request.

Environment setup

The testbed is implemented in a virtualization infrastructure host based on OpenStack. The characteristics of the host are reported in Table 6.

Table 6: Virtualization Host characteristics

HW/SW Component	Characteristic
CPU	Intel Core i7-2600, 3.40 GHz
RAM	16 GB DDR3
Storage	SSD, 480 GB
OS	Ubuntu 18.04 LTS

OpenStack	Stein

OSM is installed and deployed into a VM instantiated on OpenStack. In Table 7, a summary of virtualized OS and software tools are provided.

Table 7: Software and tools version.

Software	Version
OS	Ubuntu 18.04 LTS (bionic)
Charm-tools	2.8.2
Juju	2.8.8-bionic-amd64
OSM	8.0.4 (2020-07-01)

Outcomes

The system described above has been successfully tested with a 4G core network for compatibility against a few MANO implementations during the recurrent ETSI NFV Plugtests. Using the Athonet VNFDs, it is possible to integrate our solution with MANO implementations derived from OSM and from ONAP, provided by different vendors including Nokia, Cisco, Huawei, Whitestack, etc.

Once the system is ready, OSM acknowledges that everything is correctly instantiated and configured by a visual confirmation in its web client, as showed in Figure 31.

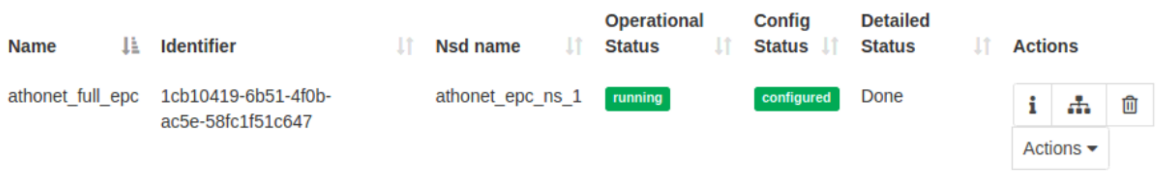

Figure 31: Confirmation of successful instantiation and configuration of the VNF EPC.

As a proof of concept, it has been decided to verify the automatic instantiation and configuration mechanism by analysing the completion time as a performance metric. The total completion required time (T_d) is the sum of the instantiation time (T_i) and the configuration time (T_c). Formally, we have

$$T_d = T_i + T_c.$$

A total of 20 instances have been deployed, in order to find an average deployment time. As shown in Figure 32, the obtained average instantiation time is 136.08 seconds and the average initial configuration time is 566.92 seconds, obtaining an average total deployment time of 703.01 seconds. The VNF average deletion time is 35.08 seconds. The configuration time largely depends on both the underlying infrastructure where OSM is running and the proxy charm structure itself. The proxy charm is instantiated inside an LXD container, based on the OS system defined inside the charm; then, a set of python packages is installed in this container.

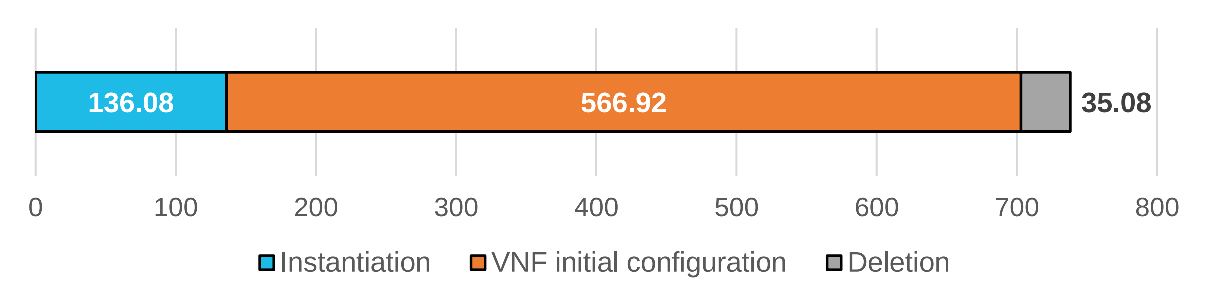

Figure 32: VNF instantiation, configuration and deletion time.

We will be testing the above architecture in 5G technology as soon as a new 5G RAN and UEs will be available, in lab and in the upcoming ETSI NFV Plugtests events.

5.2 5G Core prototype

The 5G Core Network (5GC) has been specified in 3GPP with the aim to increase the operational efficiency and support various new advanced services for industries and consumers. 5GC embraces service-based architecture (SBA) using web-scale internet protocols like HTTPv2, cloud-native implementations and deployments, and network automation for service provisioning and service assurance. SBA provides a modular framework enabling the composition of 5GC from different vendors. The control plane and common data repository are delivered by way of a set of interconnected Network Functions (NFs), each with authorization to access each other's services. All the 5GC NFs register themselves and subscribe to services from other NFs, or from Application Functions (AFs) via the Service Based Interface (SBI). The 5GC service exposure scheme is where NFs expose through the Network Exposure Function (NEF) their services to AFs and vice-versa. Today's 5GC SBA is evolving in the direction of a finer decomposition of the NFs into smaller sub-functions providing micro services to one another, where some micro services may even be reused for different NFs and have independent life-cycle management from one another.

In the Taiwanese demonstration site, the 5G core network is designed for service-based architecture (SBA) and follows 3GPP Release 15+ as a standalone (SA) solution. The III-5GC containerizes all core network functions with C/U split architecture, enabling the enterprise to distribute these functions wherever and whenever needed. All the modules can be deployed on virtual machines on top of a large number of virtualization environments, and managed as a Kubernetes platform (Figure 33).

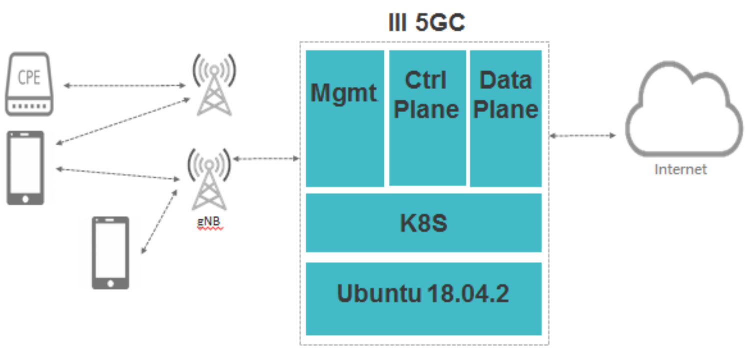

Figure 33: III-5GC Architecture

The network access and mobility management function (AMF) establishes the UE context and PDU resource allocation via slice assistance information (S-NSSAI) provided by the UE. The S-NSSAI is set up per PDU session for the policy management in the PDU session level. The session management function (SMF) controls the user plane function (UPF) and therefore directs and redirects the service flows as required for the applications. We are testing the core network basic functions like UE registration, PDU session establishment, service request and Xn & N2 handover procedure via Spirent Landslide emulator. Furthermore, for supporting the industrial applications, the development especially focuses on data plane efficiency and system reliability. Thus, we are developing both software and hardware solutions for data planes to enhance packet processing and load monitoring.

In addition, for various environment enterprises use cases, III 5GC also supports interworking with MEC and local breakout applications. For I4.0 use cases requirement, III 5GC can provide a highly available network that works also in isolation and can prioritize data and IoT services under enterprise control.

5.3 Conclusions

In this section, 5G-CONNI's technical enablers of a mobile core network were analyzed. A lightweight orchestration framework based on the ETSI NFV standard was presented first. Then, the key technical enablers of a 5GC prototype for enhancing its performance and efficiency by the specific architecture and interfaces were described. As a future work, the orchestration framework will be improved to accommodate full-fledged deployments of 5GC networks, while the 5GC prototypes will be enhanced in throughput and latency of the in-lab testing activities envisioned for WP5.

6 Mobile Edge Cloud Enablers

The Objective of Task4.3 is developing MEC technologies that support the requirements of smart factories in 5G eMBB and URLLC scenarios, to be deployed in the project's testbeds. In Sections 6.1 and 6.2, we will present the MEC implementations proposed by 5G-CONNI for the European and the Taiwanese testbeds, respectively.

6.1 MEC based on 5G-CONNI's hybrid architecture

Among the options for the mobile network architecture discussed in Section 3 of D2.1, a special focus has been reserved to the splitting of CP and UP functions in the hybrid 5GC solution. Because of its high flexibility, this was chosen as the reference architecture for the European testbed of 5G-CONNI, and it entails 1) a remote control center, typically placed in a central cloud or datacenter, which acts as CP manager and provides the configuration, provisioning and monitoring functions; 2) an edge node, located on-premise, which includes the UPF platform and forwards user traffic to/from the RAN and local applications, thus keeping the traffic local whenever necessary, without exiting the factory premises. Such an architecture meets the requirements of 5G systems, which are conceived to allow a more flexible deployment of the data plane, aiming to natively support edge computing. As a consequence, a MEC platform can easily be mapped into the 5G system architecture, as shown in Figure 34. In this framework, chosen as the reference scenario for 5G-CONNI's European testbed, the MEC host's data plane is mapped to the 5G's UPF element.

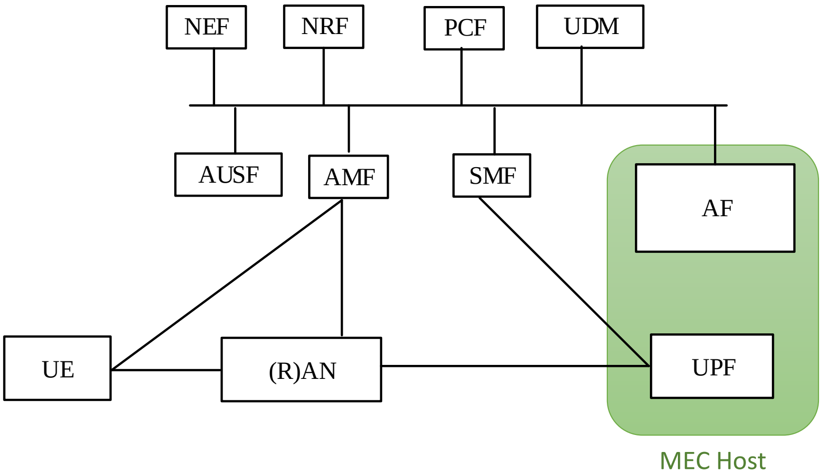

Figure 34: Integration of the MEC host into the 5G network architecture.

In particular, the MEC platform leverages the 5G network architecture and performs the traffic routing and steering function in the UPF. For example, an UL classifier of the UPF can be used to steer the UP traffic matching the filters controlled by the SMF to the local data network, where it can be consumed by the MEC application. The PCF and the SMF can set the policy to influence such traffic routing in the UPF. Also, the AF can influence the traffic routing and steering via the PCF. Therefore, MEC in 5G is able to influence the UPF through the standardized CP interface in the SMF, similarly to some of the EPC MEC deployment scenarios in 4G.

This MEC-5G architectural integration yields benefits from several points of view. In particular, the integration of 5G management, control, and orchestration processes is expected to facilitate applications/services development by providing controlled access to high-level abstractions of the 5G resources (e.g., abstractions of computing, memory/storage, and networking), thus enabling any vertical applications. Moreover, just like a true operating system, it should provide automated resources management, scheduling processes placement, facilitating inter-processes communication, and simplifying installation and management of distributed functions and services, spanning from cloud computing to MEC. This implies a shared data structure capable of supporting multi-vendor systems and applications for enabling sharing of common data amongst different protocols. Data structures include network state information, i.e., data about systems and interface state, forwarding information base (FIB) state, neighbours table and routing information base (RIB) and policies. Also, standardized data models are required using, for example, a data-modelling language such as YANG.

Implementation Focus

The MEC platform specification assumes a completely virtualized environment. This is a key requirement in order to enable seamless application lifecycle management paired with seamless platform management. Some applications, however, require hardware acceleration in order to perform certain tasks that are too difficult to achieve in a fully virtualized regime. A resulting requirement for this is the possibility to add access to the acceleration function as part of the virtualization platform. Even better, if these requirements can be fulfilled in a single "box" and can be configured upon start to allow communality of units across multiple deployments, while matching the local requirements when the unit is started.

6.2 MEC 5G SA

MNO's private CN architecture is another option of private 5G network architecture mentioned in Section 3 of D2.1. It was chosen as the reference architecture for the Taiwanese testbed of 5G-CONNI because of its low cost and rapid deployment for enterprises. Enterprises use MNO's base stations and core networks or ask MNOs to assist enterprises in establishing dedicated base stations in this architecture. There are two options on the edge site, one is UPF, and the other is bump-in-the-wire. Because of easy deployment that no additional settings are required for base stations and core networks, bump-in-the-wire was chosen as an edge breakout option for the Taiwanese testbed of 5G-CONNI.

Bump-in-the-wire" MEC 5G SA prototype is developed and continue from Chunghwa Telecom's 4G and 5G NSA MEC Cloud, as shown in Figure 35. The MEC Cloud is following the multi-access edge system reference architecture defined in ETSI GS MEC 003 specification, as mentioned in section 3.3. MEC cloud includes an NFV infrastructure and data plane functions. NFV infrastructure is also developed by Chunghwa Telecom and is named ECoreCloud (ECC). The data plane functions include SDN switch and Mobile Edge Enabler (MEE) VNF, also developed by Chunghwa Telecom. SDN switch to route and mirror the traffic. The MEE VNF provides a traffic steering function so that selected data traffic can be offloaded locally. The MEE VNF is divided into two modules: the Control Plane Analyzer module, which decodes and correlates signals, and the Data Plane Processor, which processes data plane traffic and steering.

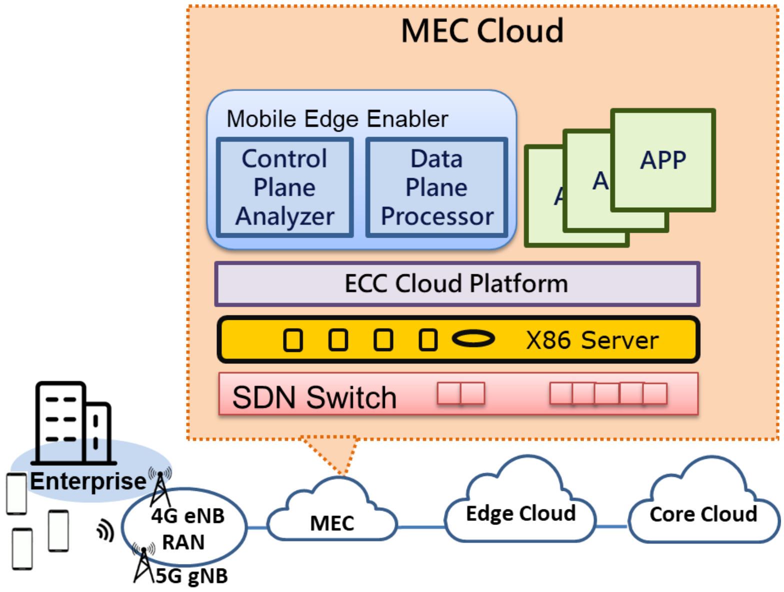

Figure 35: The architecture of bump-in-the-wire MEC cloud

The objective of the MEC SA prototype is is developing the MEC that can interoperate between 5G standalone base stations and core networks and base on 3GPP SA standalone specifications.. The architecture of MEC 5G SA prototype is shown in Figure 36. The bump-in-the-wire MEC SA prototype is deployed between 5G New Radio and 5G standalone core network. It is a convenient deployment, because it does not need the additional configurations for the core and RAN network. MEC has to handle N2 interface based on 3GPP TS36.413[TS38413] and process GTP-U extension header packet for N3 interface based on 3GPP TS29.281 [TS29281].

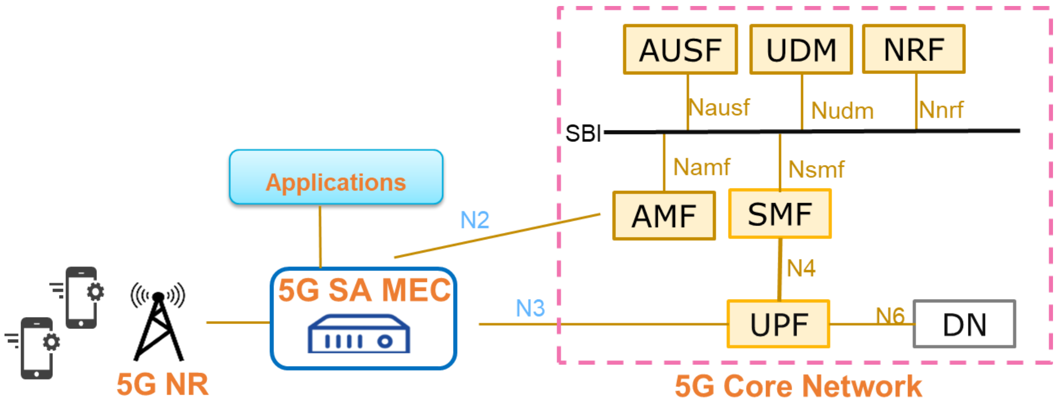

Figure 36: The architecture of MEC 5G SA version

ECC does not only provide cloud platform services but also supports ETSI NFV MANO. ECC can provide installation and disassembly of telecommunication network elements, life cycle management, performance monitoring, error management, self-repair, and various management functions, etc. And ECC can support various equipment from different vendors, open-source solutions, and customized services with high compatibility. ECC also supports various VNF templates through the ECC template function, which improves the compatibility and scalability of ECC.

The ECC MANO solution is developed based on many key technologies such as SDN, NFV, Cloud Native, and MANO. So ECC MANO can support MANO international standard framework and API interface. Therefore, ECC MANO can use this ability to build resource sharing edge clouds on standard x86 hardware devices, and the ECC template function can create

MEE VNF on the edge cloud by using VNF templates. When MEE VNF is created, ECC MANO will connect MEE VNF to edge clouds and 5G networks. All the above operations can be completed in the user interface of the ECC platform, so users can easily, quickly, and accurately complete VNF deployment. In the Taiwanese deployment of the 5G-CONNI project, the ECC manages the MEE VNF life cycle management, and the MEE VNF includes a control plane analyzer and a data plane processor components, as shown in Figure 37.

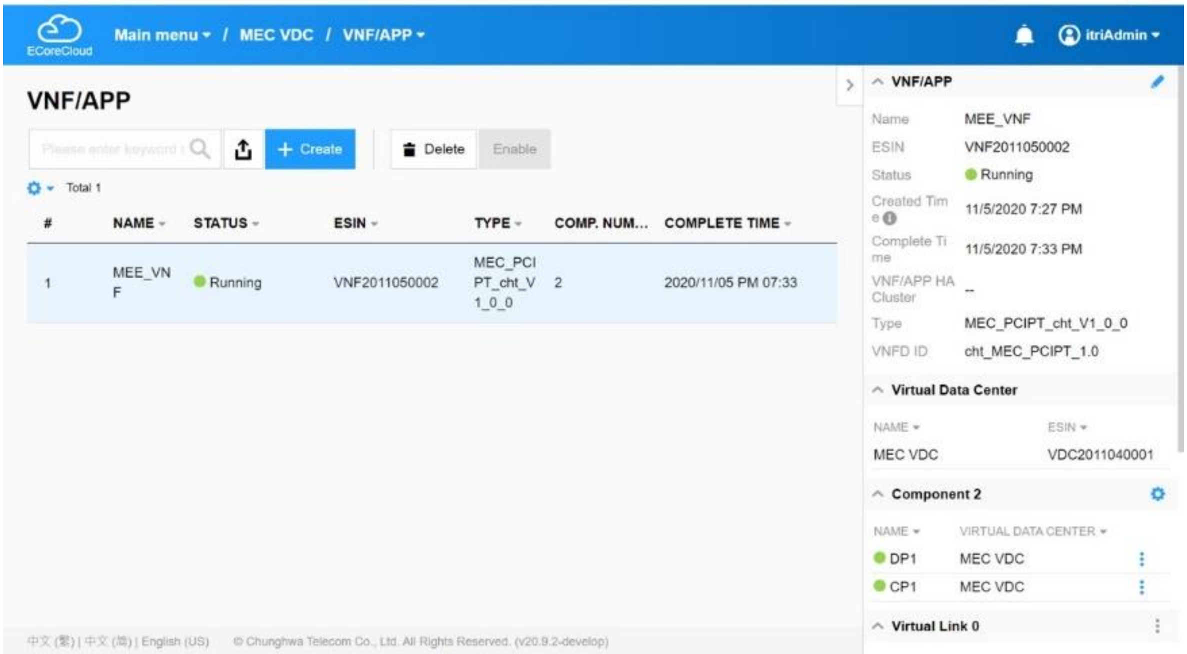

Figure 37: EcoreClould Platform Web UI and the MEE VNF deployment

MEC 5G SA prototype implementation and deployment

The MEC 5G SA prototype has been developed that can handle the signals for one PDU session and one QoS flow and GTP extension packets. The MEC 5G SA prototype is deployed at ITRI IMTC field. The MEC SA prototype has been integrated with III's 5G core network and Alpha's 5G base station, as shown in Figure 38. The MEC 5G SA prototype is performed the control plane test such as registration, deregistration, service request and AN release procedures and data plane test such as GTP extension header encapsulation and decapsulation. The ITRI IMTC's application, named Process Diagnostics using Augmented Reality (AR) has also been integrated with 5G MEC SA prototype. The traffic of Process Diagnostics using AR can be steered by MEC. Figure 39 shows that the test results that application client connects the Web API server and the UE use the AR Process Diagnostics application.

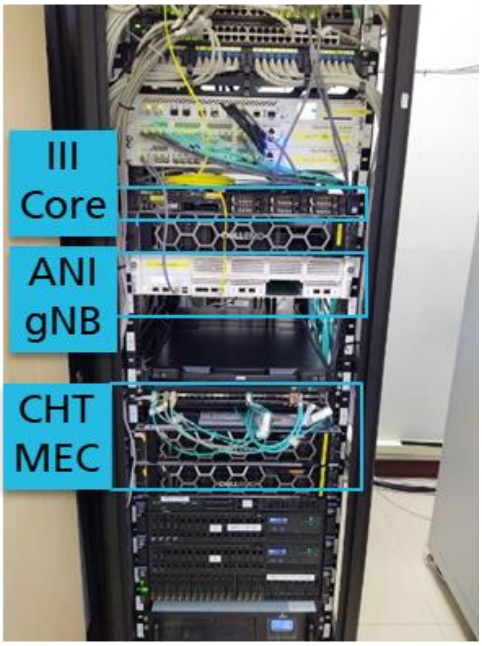

Figure 38: MEC 5G SA prototype integrated with III's 5G core network and Alpha's 5G base station

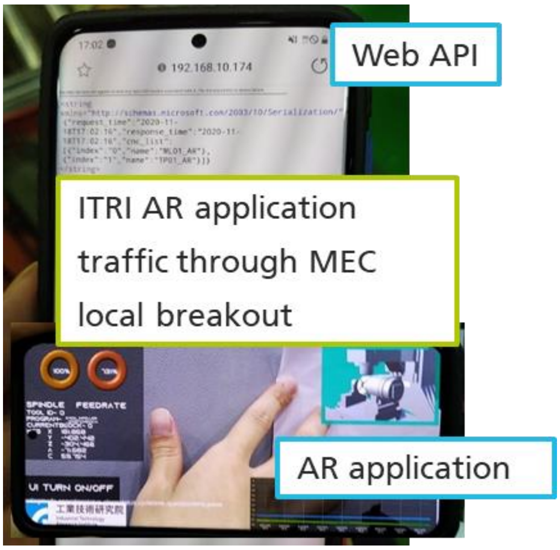

Figure 39: Test result of MEC 5G SA prototype integrated with AR Process Diagnostics application

6.3 Conclusions

Two types of MEC developed in the 5G-CONNI project were presented. One is deployed in conjunction with a hybrid 5GC architecture, whereas the other is deployed along with an MNO's Private Core Network architecture. The former MEC architecture is based on a distributed UPF, whereas the latter is based on the bump-in-the-wire approach. MEC 5G SA is developed as a prototype now. As future work, on the EU side the MEC configuration based on the hybrid architecture will be tested in-lab before being deployed, as part of WP5 activities. On the other hand, the MEC 5G SA will develop multiple PDU sessions, multiple QoS flows and support of the Xn handover feature. It will be integrated with III's 5G core network and ATH's 5GC when new features will be developed. The industrial applications will be deployed and managed on the ECC platform.

7 Industrial Application Enablers

The objective of this task is to rethink the network in a holistic manner by jointly optimizing all enabling technologies, namely radio (Task 4.1), core (Task 4.2), MEC platform (Task 4.3) and the use cases selected in Task 1.1 for a proof-of-concept demonstration.

7.1 Three use cases prototype applications

The initial plan of implementation of selected use cases has been proposed and the progress is described as follows:
For use case to be implemented at BOSCH site
For the use case “Robot platforms with edge intelligence and control”, first conceptual designs have been created and initial investigations into more robust and reliable operation principles of safety-critical robot actions (e.g. arm movements), in case the 5G system performance varies over time or when network failures occur, have been conducted and documented. The further derivation of implementation details and the actual implementation are planned to start during the second year of the project.
For use cases to be implemented at ITRI site

The overall 5G network architecture for the implementation of the use cases at ITRI site is shown in Figure 40. Software modules for data collection, analysis have been installed in separate industrial computers and connected with MEC. Remote rendering server with virtualized GPU will also be established and connected with MEC to conduct rendering of AR/VR images and transfer to end user device such as cell phone, tablet and head-mount display. Description of the use cases are as follows:

- Process Diagnostics by CNC and Sensing Data Collection: The data collection and process diagnosis has been investigated at one target cell and a three-axis accelerometer was attached to collect CNC data as well as sensing data for process data analysis. As shown in Figure 41, there are two modules in the process diagnostics system: (1) in-line monitoring system used to collect process data and detect tool condition or abnormal vibration according to predefined model and threshold values. (2) the data analysis system which uses the labeled process data to train machine learning models and update the model parameters or threshold values for tool condition and abnormal vibration in the in-line monitoring system. The tool condition monitoring and abnormal vibration detection software that installed at shop floor machine is shown in Figure 42.
- Using Augmented/Virtual Reality for Process Diagnosis: The 3D model has been constructed for the target machine. The association of 3D model and CNC/sensing data via APIs has been under development. In particular, sensing data will be associated with corresponding 3D components to show process status information. For example, vibration levels can be shown on the 3D model of the spindle as color-coded contour. The prototype of this use case has been demonstrated during IEEE Globecom 2020. Screen capture of the AR app is shown in Figure 42, CNC parameters and sensing data are collected from shop floor machine and send to end user device via WebApi. The 3D kinematic mode of the machine tool has been constructed and driven by machine coordinate values collected from shop floor machine so that the motion of 3D model can be synchronized with real machine. Figure 43 shows how the AR app is used by machine operator.

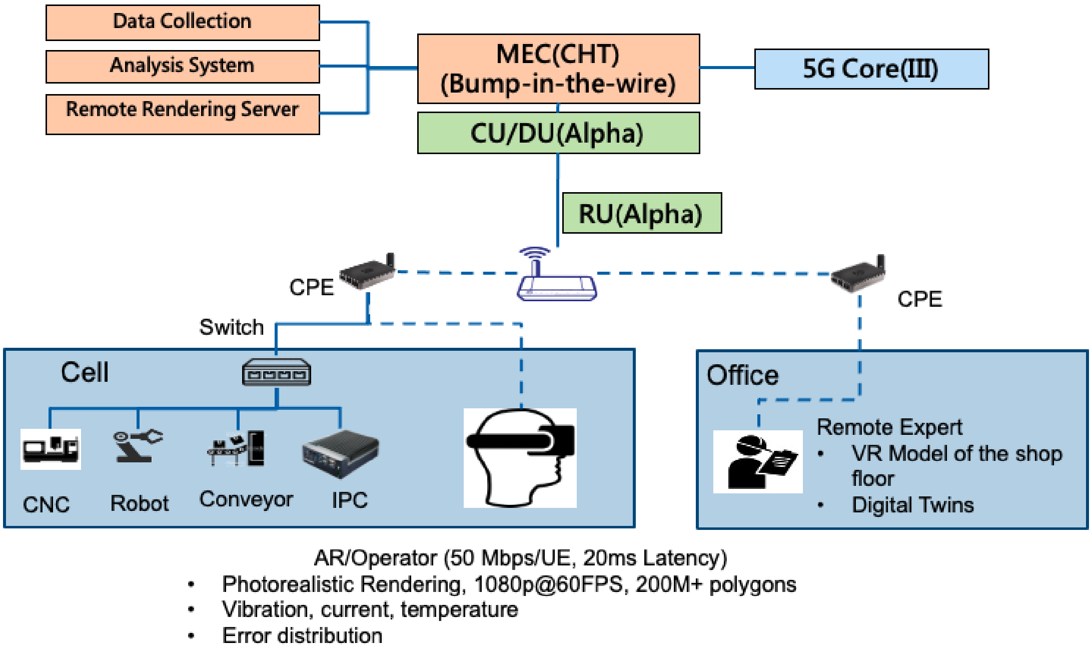

Figure 40: Prototype for the Using Augmented/Virtual Reality for Process Diagnosis use case

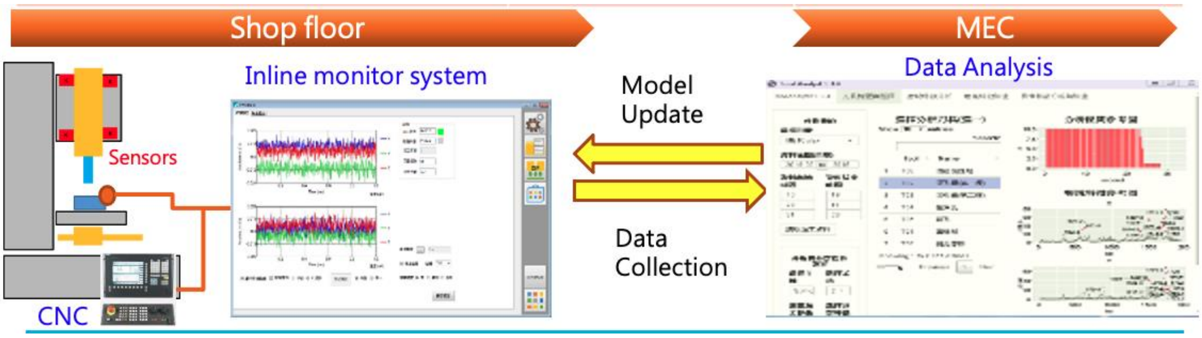

Figure 41: The implementation architecture of the Process Diagnostics by CNC and Sensing Data Collection

Figure 42: The tool condition monitoring system used in "Process Diagnostics by CNC and Sensing Data Collection" use case

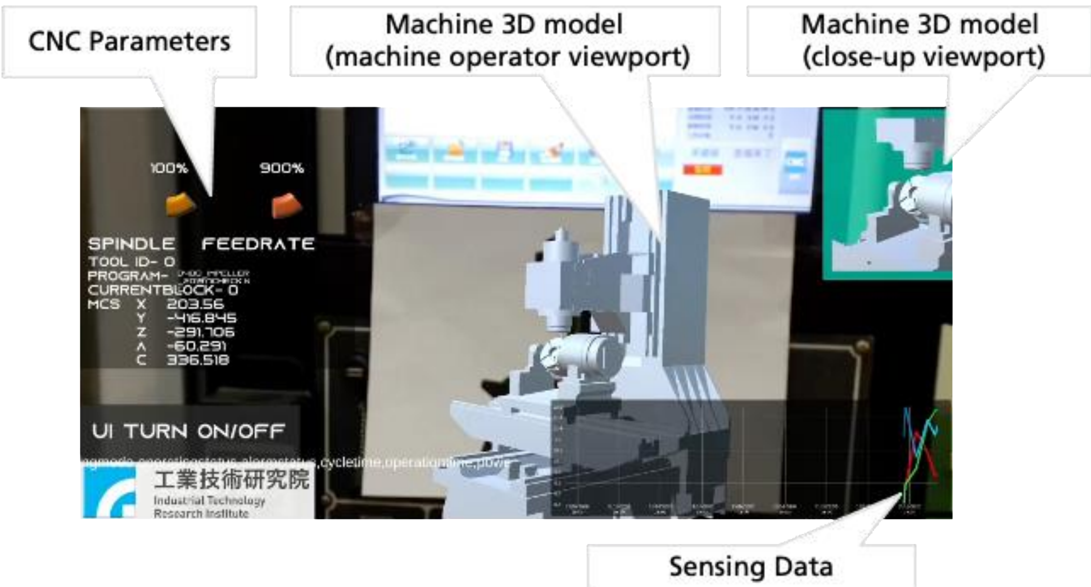

Figure 43: The augmented reality app developed for the "Using Augmented/Virtual Reality for Process Diagnosis" use case

Figure 44: Demonstration of the "Using Augmented/Virtual Reality for Process Diagnosis" use case

- For the use case "Cloud-Based Controller for CNC", the first version of the cloud-based CNC software has been tested using the flexible fixture system under distributed network architecture as shown in Figure 45. In the cloud-based controller for CNC, the motion command generation and motion command execution modules are implemented as separate software modules and interact with each other by the MQTT broker, which could be replaced with industrial protocol such as EtherCAT or Profinet.

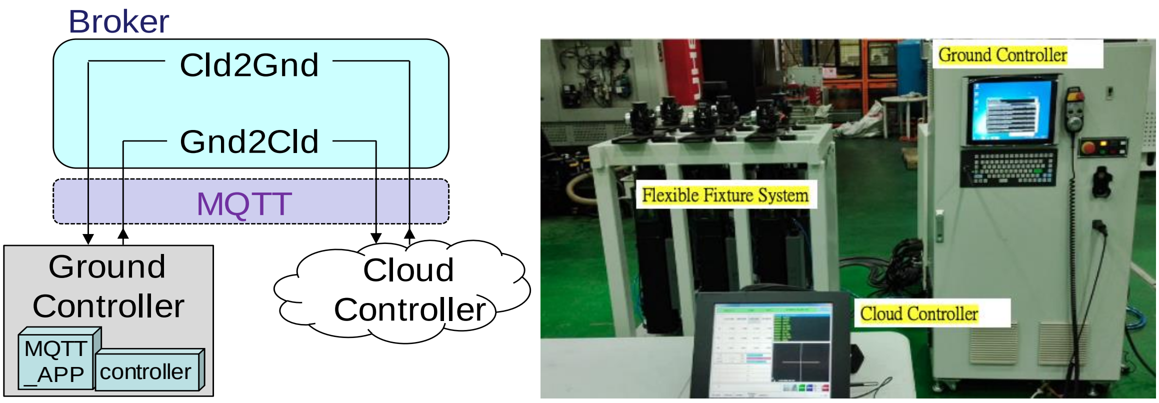

Figure 45: Architecture of the cloud CNC.

.

7.2 Joint optimization of enabling technologies

In task 4.4, devising computation offloading strategies to enable complex processing of data collected by mobile inspector is necessary in order to guarantee continuous monitoring and anomaly detection during industrial processes. The progress is described as follows:

7.2.1 Dynamic resource allocation for computation offloading

One of the applications enabled by edge computing is the offloading of computational demanding programs from resource-poor sensors/machinery/mobile devices to nearby edge servers. A typical example in the smart factory is the processing of data collected by the industrial sensors for anomaly/fault detection, predictive maintenance, etc. Dynamic computation offloading refers to the case in which the sensors continuously collect data to be processed, which perfectly fits the industry use cases. Differently from the central cloud, MEC offers access to limited computational capacity at the edge of the network. The resulting resource management is a challenging problem, also due to the fact that computation resource optimization is naturally coupled to radio access resources. Indeed, when an application is offloaded, both communication and computation delays are experienced, which suggests joint management of these resources in a holistic view of the system. Another challenge of this research topic is the fact that wireless channels and data arrivals are typically time-varying, with complex statistics which could be unknown in advance. In this section, we present algorithms for dynamic computation offloading, based on stochastic Lyapunov optimization, with the aim of reducing the energy consumption of the sensors (and edge servers), with guarantees on the average E2E delay, and reliability constraints on the probability that the E2E delay exceeds a predefined threshold. In this case, we refer to the E2E delay, as the time elapsed from the generation of a new data unit/subtask locally at the devices, until the result of this subtask is provided by the edge server. Stochastic Lyapunov optimization [Neely10] allows to solve long-term optimization problems without assuming any knowledge on the statistics of data arrivals and radio channels. Without going into the mathematical details, the dynamic problem is approached by defining local communication queues at the sensors, containing data to be transmitted, and remote queues at the edge server, containing the amount of computations to be performed. We developed a low complexity algorithm able to:

1. Minimize the average energy consumption of devices
2. Guarantee a constraint on the average E2E delay
3. Guarantee a constraint on the probability that the E2E delay exceeds a predefined threshold

The interested reader is referred to [Merluzzi2020] for more technical details. The algorithm requires the solution of a simple optimization problem on a per-slot basis, and it is capable of guaranteeing the constraints without any prior knowledge of the statistics of radio channels and data arrivals. The variables involved in the optimization are the transmit powers of devices and the CPU scheduling at the edge server. We now present some numerical results to assess the performance of the algorithm. As a first result, we show the energy delay trade-off. In particular, theoretically speaking, the energy consumption decreases while the average E2E delay increases until reaching the constraint imposed by the application. Then, in Figure 46, we show the trade-off, obtained with our algorithm termed as *DyCO*. The simulation parameters are shown in Table 8, with the indoor channel model as described in section 2.3.2 for path loss and shadowing, and a statistical model for fast fading, namely a Rayleigh fading with unit variance. The model for the traffic is Poisson with random parameter as shown in Table 8.

Table 8 : Simulation parameters of dynamic resource allocation evaluation

Simulation Parameters	Values
Area of the proposed scenario	150 x 150 m^2
Number of UEs	5, uniformly randomly distributed in the area
Base station	Place at the center, operating at 3.5 GHz
Channel model	3GPP InF-DL for path loss and shadowing
Fast fading	Rayleigh fading with unit variance 10 ms coherence time
Bandwidth	20 MHz

Traffic	Poisson arrival rate with parameter $\lambda = 10^x$, with x randomly chosen in [1,3]
Maximum UE transmit power	20 dBm
Noise power spectral density	-174 dBm/Hz, 5dB noise figure at the receiver

As we can notice, the energy consumption decreases until reaching the desired constraint on the delay. The method is compared to a benchmark method that requires the stability of the communication and computation queues, while not enforcing an explicit constraint on the average queue length, which is directly related to the average delay of the service by Little's law [Little1961]. We can notice how this benchmark solution keeps on reducing the energy consumption, but not meeting the constraint on the average E2E delay, differently from our method, that finds the minimum energy that guarantees the constraints. Also, we can notice how our method is able to obtain a better trade-off. Namely, given a fixed energy consumption, our solution is capable of keeping the delay stable around a lower value than the benchmark strategy. The numerical results are shown for different values of the required E2E delay. As a second result, we consider an additional constraint on the probability that the E2E delay exceeds a predefined threshold. For this simulation, we use the same parameters as the previous one, with an average delay bound of 30 ms, while the maximum probabilistic delay is shown in the legend of the figure for each device. Then, in Figure 47 we introduce the reliability constraint, and thus we plot the reliability function of the E2E delay, defined as $1 - CDF(E2E\ Delay)$, where CDF denotes the cumulative distribution function. Then, all the curves in Figure 47 represent the probability that the E2E delay exceeds the value on the abscissa. As we can notice, all devices meet their constraints, as shown in the plot with the circles around their respective values. Indeed, in this case, all devices have different requirements on the maximum E2E delay, but they all require not to exceed it 99 % of the time ($\epsilon = 10^{-2}$ reliability constraint).

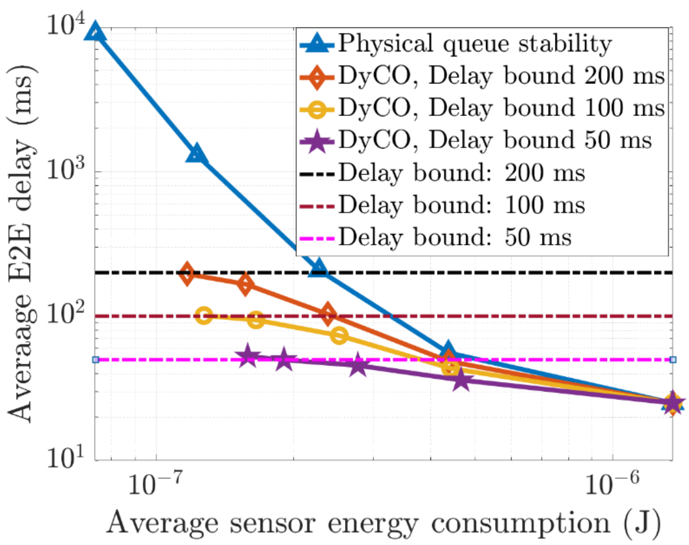

Figure 46: Average E2E delay vs. average energy consumption

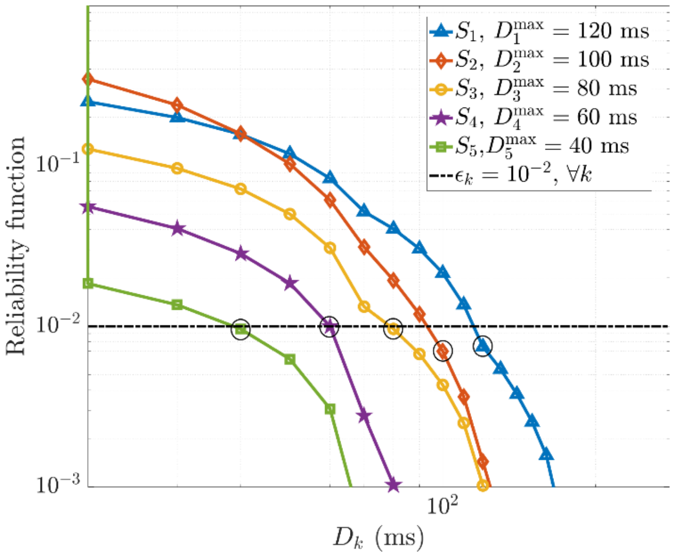

Figure 47: Reliability function of the E2E delay

7.2.2 Dynamic resource allocation for edge machine learning

The work dedicated to dynamic resource allocation for computation offloading does not go into the details of the specific application, thus concentrating on the energy-delay trade-off. Instead, diving into the application, it is possible to introduce new performance measure, such as the accuracy of a certain learning task running at the edge server on data collected and transmitted by sensor devices. This opens a new research direction known as Edge Machine Learning (EML) [Park19] [MDB20] whose aim is to enable machine learning algorithms at the edge, by exploring the best trade-off between energy, latency and learning accuracy/reliability. Using the tools of stochastic Lyapunov optimization, we devised an online algorithm able to guarantee an average E2E delay and an average accuracy of an estimation task based on Least Mean Squares (LMS). The method does not require any prior knowledge of the statistics of data arrivals, radio channels, and data distributions. In particular, in this case, the accuracy is affected by the number of quantization bits used to encode the data. More bits lead to better accuracy but higher energy consumption, due to the longer payloads to be transmitted. Thus, the optimization variables are the devices' transmit power, the CPU scheduling at the edge server, and the number of quantization bits used to represent the data. In this particular case, also the energy consumption of the edge server is optimized. Thus, the aim is to devise a low energy strategy able to guarantee a predefined E2E delay and accuracy constraints on the learning task. The interested reader is referred to [MDB20] for more technical details.

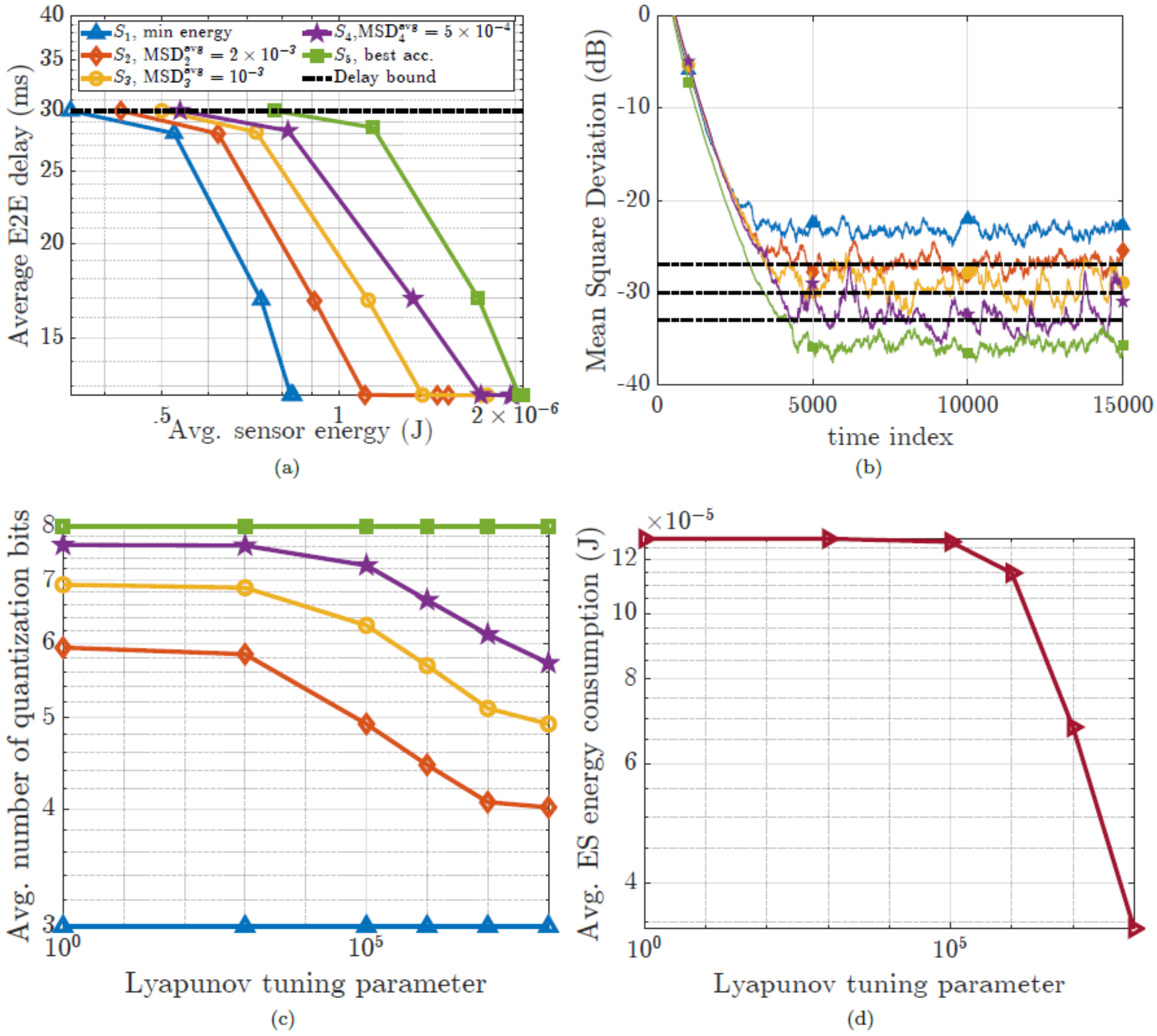

Figure 48: Energy-delay-accuracy trade-off

In Figure 48, we illustrate the performance of our solution in terms of trade-off between energy, delay, and learning accuracy. In particular, we want to highlight how given a certain E2E delay constraint (set to 30 ms), the accuracy affects the performance in term of energy consumption. To this aim, we consider 5 sensors at the same distance from the AP, with the same arrival rate, the same average E2E delay constraint, but different constraints on the MSD. In particular, we assume that two sensors represent two benchmarks: i) Minimum energy, obtained by the device always transmitting with 3 quantization bits (i.e., the minimum number of bits), for all t; ii) Best accuracy, obtained by the device with 8 quantization bits (i.e., the maximum number of bits), for all t. The other three devices have different intermediate requirements for the MSD. In Figure 48 (a), we show the average E2E delay as a function of the sensor energy consumption, obtained by tuning a Lyapunov trade-off parameter. In particular, the parameter increases from right to left, as shown in the figure. From Figure 48 (a), we can notice how the energy consumption decreases while the E2E delay increases. However, this trade-off is different among the different devices due to the different accuracy constraints. In particular, let us first comment on the results for the two benchmarks. The green curve (squared marker) shows the best accuracy case, which indeed achieves the highest minimum energy among all devices, given the E2E delay (the legend of all figures is shown in Figure 48 (a)). At the same time, from Figure 48 (b), which shows the MSD vs. the time index, we can notice how this device achieves the minimum MSD. These first two results are due to the maximum number of quantization bits shown in Figure 48 (c), which reaches the highest value for this device. Note that the average number of quantization bits is shown as a function of the Lyapunov tuning parameter, using the same values of as for Figure 48 (a). On the other hand, the blue curve (triangle marker) represents the minimum energy case, and it achieves the worst accuracy due to the minimum number of quantization bits adopted, as can be noticed from Figure 48 (a),(b),(c). The other curves represent intermediate energy cases obtained fixing a target MSD constraint, and can be interpreted via similar analysis over Figure 48 (a),(b),(c). Finally, the energy consumption of the ES decreases as the Lyapunov parameter increases, as expected, until reaching a floor as the device energy consumption (Figure 48 (d)). In summary, the take-home message of Figure 48 is twofold:

- Our method is able to obtain a low system energy solution with accuracy and E2E delay guarantees;
- By relaxing the accuracy constraint, lower energy can be achieved due to the lower number of quantization bits, which translates into a lower average data rate over the wireless interface.

The required accuracy strongly depends on the particular scenario/use case. It is not always desirable to obtain the best possible accuracy, if this leads to a higher energy cost. With our method, setting a target accuracy, it is possible to achieve lower energy consumption without degrading the performance of the application below the desired threshold.

7.3 Conclusions

In this section, we analyzed possible solutions for the joint optimization of the enabling technologies, i.e. radio and MEC. In particular, we first presented three use cases in ITRI site to demonstrate process diagnostics by massive data collection and AR/VR devices. And cloud-based CNC has been tested under distributed architecture. Network architecture for these use cases has been planned and implemented based on enabling technologies from project partners. Furthermore, we presented the problem of computation offloading, used to transfer the execution of computationally heavy applications from sensor/mobile devices to nearby edge servers. We have shown that computation offloading naturally couples the optimization of radio (transmit power, bandwidth, etc.) and computation (e.g. CPU scheduling), due to the fact that the overall delay of the services comprises a transmission delay and a computation delay at the edge server. We presented the problem of dynamic computation offloading, in which the application continuously generates data to be processed. Using the tools of stochastic Lyapunov optimization, we developed a low complexity algorithm able to guarantee E2E delay (both average and probabilistic) with low energy consumption, by not assuming any knowledge of the statistics of data arrivals and radio channels. As a natural application of computation offloading, we extended this analysis to the particular case of edge machine learning, in which the application to be offloaded is a learning task. Thus, diving into the characteristic of the application, we have shown how it is possible to explore a trade-off between energy, delay and learning/inference accuracy. Numerical results on both general computation offloading and edge machine learning assess the performance of our proposed strategy in exploring the new trade-offs of future communication networks. Future works include (1) the exploration of other machine learning tasks (e.g. classification), as well as more efficient strategies for data quantization (e.g. vector quantization); (2) Interaction between end user and digital twins; (3) Correlation between network performance and analytical performance; (4)Finally, scenarios with multiple APs and/or multiple ES are an interesting research direction.

8 Conclusions

WP4 (Technical Enablers for Industrial Applications) covers Mobile Edge Computing (MEC) cloud development, industrial application technical development, radio network technical development, and core network technical development for industrial field. The main goal of this work package is to ensure industrial use cases can be implemented on private 5G networks successfully for industrial requirements, including high data rates (eMBB) and low latency (URLLC).

D4.1 provides initial specification and implementation of private 5G networks building blocks. This deliverable will be extended in D4.2. These innovative components will fuel the lab integration reported in D5.1.

During WP4, activities are carried out on both the European and the Taiwanese side of the consortium with partners either mirroring or complementing competences across the two regions.

Section 2 aims to harmonize the research and development of innovative new technologies required for the target use cases. First, WP4 is linked to WP2 with a short recall of 5G-CONNI private network architecture and the different combination of public and privates networks. Then some design considerations are made according to WP1 selected use cases, the functional and non-functional requirements to be achieved by current developments. Finally, some modelling methodologies are discussed in order to align the simulation frameworks parameters (e.g., deployment scenarios, traffic models and channel models for industrial environments). In future work, the channel models extracted in WP3 from the measurement campaigns conducted at Bosch facilities could be integrated in these simulation frameworks.

Section 3 provides an update of the state of the art for the four building blocks. This section benchmarks the project's investigation against the state of the art.

Section 4 presents the investigation of RAN orchestration enabling deterministic URLLC services and the development of the radio network consisting of CPE and gNodeB systems. Thanks to NS-3 network simulator, we have developed an E2E network framework able to combine several mechanisms exploiting modulation and coding diversity (LDPC, MCS, AMC), space diversity (MIMO) and time diversity (Frame Design, HARQ) at RAN level, guaranteeing URLLC communications. The performance have been evaluated in terms of reliability, E2E latency and jitter for dynamic traffics and fast-fading channel in indoor office scenario. In future works, we will extend this work with the complete evaluation of the orchestrator that dynamically and efficiently manages 5D diversity mechanisms. For 5G NR RAN implementation, we have built up a RAN system composed of CDU, RU, and CPE based on the RAN design consideration and we will realize the selected use case defined by 5G CONNI in future works.

Section 5 presents the development of the core network components to realize private local 5G networks. A lightweight orchestration framework based on the ETSI NFV standard was presented first. Then, the key technical enablers of a 5GC prototype for enhancing its performance and efficiency by the specific architecture and interfaces were described. As a future work, the orchestration framework will be improved to accommodate full-fledged deployments of 5GC networks, while the 5GC prototypes will be enhanced in throughput and latency of the in-lab testing activities envisioned for WP5.

Section 6 presents the two types of MEC developed in the 5G-CONNI project. One is deployed in conjunction with a hybrid 5GC architecture, whereas the other is deployed along with an MNO's Private Core Network architecture. The former MEC architecture is based on a distributed UPF, whereas the latter is based on the bump-in-the-wire approach. MEC 5G SA is developed as a prototype now. As future work, on the EU side the MEC configuration based on the hybrid architecture will be tested in-lab before being deployed, as part of WP5 activities. On the other hand, MEC 5G SA will develop multiple PDU sessions, multiple QoS flows and support of the Xn handover feature. It will be still integrated with III's 5G core network and ATH's 5GC when new features will be developed. The industrial applications will be deployed and managed on the ECoreCloud platform.

Section 7 rethinks the network in a holistic manner by jointly optimizing all enabling technologies for a proof-of-concept demonstration. In particular, we first presented three use cases in ITRI site to demonstrate process diagnostics by massive data collection and AR/VR devices. And cloud-based CNC has been tested under distributed architecture. Network architecture for these use cases has been planned and implemented based on enabling technologies from project partners. Moreover, we analyzed possible solutions for the joint optimization of the enabling technologies, i.e. radio and MEC. We presented the problem of computation offloading, used to transfer the execution of computationally heavy applications from sensor/mobile devices to nearby edge servers and the problem of dynamic computation offloading, in which the application continuously generates data to be processed. Future works include the exploration of other machine learning tasks (e.g. classification), as well as more efficient strategies for data quantization (e.g. vector quantization).

References

[4GMEC] MEC Deployments in 4G and Evolution Towards 5G

[5GMEC] MEC in 5G networks

[Aijaz2012] A. Aijaz, O. Holland, P. Pangalos and H. Aghvami, "Energy savings for cellular access network through Wi-Fi offloading," 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, 2012, pp. 4369-4373, doi: 10.1109/ICC.2012.6363815.

[Amir2019] M. Mohammadi Amiri and D. Gündüz, "Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air," in IEEE Transactions on Signal Processing, vol. 68, pp. 2155-2169, 2020, doi: 10.1109/TSP.2020.2981904.

[Ashraf15] S. A. Ashraf, F. Lindqvist, R. Baldemair, and B. Lindoff, “Control channel design trade-offs for ultra-reliable and low-latency communication system,” in 2015 IEEE Globecom Workshops (GCWkshps), IEEE, dec 2015.

[Ashraf2018] M. I. Ashraf, C. Liu, M. Bennis, W. Saad and C. S. Hong, "Dynamic Resource Allocation for Optimized Latency and Reliability in Vehicular Networks," in IEEE Access, vol. 6, pp. 63843-63858, 2018, doi: 10.1109/ACCESS.2018.2876548

[Azar19] Joseph Azar and Abdallah Makhoul and Mahmoud Barhamgi and Raphaël Couturier, “An energy efficient IoT data compression approach for edge machine learning”, Future Generation Computer Systems, 2019

[BATON] Open Baton. https://openbaton.github.io/

[ChangMiao18-2] P. Chang and G. Miao, "Resource Provision for Energy-Efficient Mobile Edge Computing Systems," 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-6, doi: 10.1109/GLOCOM.2018.8648008.

[Chen18] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hard-jawana, W. Bao, Y. Li, and B. Vucetic, “Ultra-reliable low latency cellular networks: Use cases, challenges and approaches,” IEEE Communications Magazine, vol. 56, pp. 119–125, dec 2018.

[ChenRan19] J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," in Proceedings of the IEEE, vol. 107, no. 8, pp. 1655-1674, Aug. 2019, doi: 10.1109/JPROC.2019.2921977.

[ChenZhou17] L. Chen, S. Zhou and J. Xu, "Energy efficient mobile edge computing in dense cellular networks," 2017 IEEE International Conference on Communications (ICC), Paris, 2017, pp. 1-6, doi: 10.1109/ICC.2017.7997128.

[Elgabli2020] A. Elgabli, J. Park, A. S. Bedi, C. B. Issaid, M. Bennis and V. Aggarwal, "Q-GADMM: Quantized Group ADMM for Communication Efficient Decentralized Machine Learning," in IEEE Transactions on Communications, doi: 10.1109/TCOMM.2020.3026398.

[ETSI01] ETSI GS NFV-SOL 001 V2.6.1 (2019-05) Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; NFV descriptors based on TOSCA specification

[ETSI04] ETSI GS NFV-SOL 004 V2.7.1 (2019-12) Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; VNF Package and PNFD Archive specification

[ETSI02] ETSI GS NFV-SOL 002 V2.6.1 (2019-04) Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; RESTful protocols specification for the Ve-Vnfm Reference Point

[ETSI03] ETSI GS NFV-SOL 003 V2.7.1 (2019-12) Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; RESTful protocols specification for the Or-Vnfm Reference Point

[ETSI05] ETSI GS NFV-SOL 005 V2.6.1 (2019-04) Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; RESTful protocols specification for the Os-Ma-nfvo Reference Point

[HanChen20] D. Han, W. Chen and Y. Fang, "Joint Channel and Queue Aware Scheduling for Latency Sensitive Mobile Edge Computing With Power Constraints," in IEEE Transactions on Wireless Communications, vol. 19, no. 6, pp. 3938-3951, June 2020, doi: 10.1109/TWC.2020.2979136.

[Hellstrom2020] Henrik Hellström, {José Mairton B.} da Silva Jr and Viktoria Fodor and Carlo Fischione, "Wireless for Machine Learning", Available online at: https://arxiv.org/pdf/2008.13492.pdf, 2020

[Huang2012] D. Huang, P. Wang and D. Niyato, "A Dynamic Offloading Algorithm for Mobile Computing," in IEEE Transactions on Wireless Communications, vol. 11, no. 6, pp. 1991-1995, June 2012, doi: 10.1109/TWC.2012.041912.110912

[ITUR17] I.-R. Recommendation, "Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 mhz to 100 ghz," tech. rep., ITU-R, June, 2017

[Jiang19] W. Jiang, S. D. Anton, and H. D. Schotten, "Intelligence slicing: A unified framework to integrate artificial intelligence into 5g networks," in2019 12th IFIP Wireless and Mobile Networking Conference (WMNC), IEEE, sep 2019.

[Jiang2019] C. Jiang, X. Cheng, H. Gao, X. Zhou and J. Wan, "Toward Computation Offloading in Edge Computing: A Survey," in IEEE Access, vol. 7, pp. 131543-131558, 2019, doi: 10.1109/ACCESS.2019.2938660.

[Li17] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5g ultra-reliable and low-latency systems design," in 2017 European Conference on Networks and Communications (EuCNC), IEEE, jun 2017.

[Li2018] J. Li, H. Gao, T. Lv and Y. Lu, "Deep reinforcement learning based computation offloading and resource allocation for MEC," 2018 IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, 2018, pp. 1-6, doi: 10.1109/WCNC.2018.8377343.

[LiGuan19] L. Li, Q. Guan, L. Jin and M. Guo, "Resource Allocation and Task Offloading for Heterogeneous Real-Time Tasks With Uncertain Duration Time in a Fog Queueing System," in IEEE Access, vol. 7, pp. 9912-9925, 2019, doi: 10.1109/ACCESS.2019.2891130.

[Little1961] J. D. C. Little, "A proof for the queuing formula: L=λW," Oper. Res. ,vol. 9, no. 3, p. 383–387, Jun. 1961

[Liu2017] C. Liu, M. Bennis and H. V. Poor, "Latency and Reliability-Aware Task Offloading and Resource Allocation for Mobile Edge Computing," 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 2017, pp. 1-7, doi: 10.1109/GLOCOMW.2017.8269175.

[Liu2019] C. Liu, M. Bennis, M. Debbah and H. V. Poor, "Dynamic Task Offloading and Resource Allocation for Ultra-Reliable Low-Latency Edge Computing," in IEEE Transactions on Communications, vol. 67, no. 6, pp. 4132-4150, June 2019, doi: 10.1109/TCOMM.2019.2898573.

[Liva16] G. Liva, L. Gaudio, T. Ninacs, and T. Jerkovits, "Code design for short blocks: A survey,"11th International ITG Conference on Systems, Communications and Coding, 2016-10-04.

[Mach17] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," in IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628-1656, thirdquarter 2017, doi: 10.1109/COMST.2017.2682318.

[Marsch16] P. Marsch, I. D. Silva, ̈Omer Bulakci, M. Tesanovic, S. E. E.Ayoubi, and M. S ̈aily, "Emerging network architecture and functional design considerations for 5g radio access," Transactions on Emerging Telecommunications Technologies, 2016.

[Marvi2019] M. Marvi, A. Aijaz and M. Khurram, "On the Use of ON/OFF Traffic Models for Spatio-Temporal Analysis of Wireless Networks," in IEEE Communications Letters, vol. 23, no. 7, pp. 1219-1222, July 2019, doi: 10.1109/LCOMM.2019.2917681

[Mao2017] Y. Mao, J. Zhang, S. H. Song and K. B. Letaief, "Stochastic Joint Radio and Computational Resource Management for Multi-User Mobile-Edge Computing Systems," in IEEE Transactions on Wireless Communications, vol. 16, no. 9, pp. 5994-6009, Sept. 2017, doi: 10.1109/TWC.2017.2717986.

[MaoYou2017] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," in IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, Fourthquarter 2017, doi: 10.1109/COMST.2017.2745201.

[MDB20] M. Merluzzi, P. D. Lorenzo and S. Barbarossa, "Dynamic Resource Allocation for Wireless Edge Machine Learning with Latency And Accuracy Guarantees," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 9036-9040, doi: 10.1109/ICASSP40776.2020.9052927.

[Merluzzi19] M. Merluzzi, P. D. Lorenzo and S. Barbarossa, "Dynamic Joint Resource Allocation and User Assignment in Multi-access Edge Computing," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 4759-4763, doi: 10.1109/ICASSP.2019.8683499.

[Merluzzi2020] M. Merluzzi, P. D. Lorenzo, S. Barbarossa and V. Frascolla, "Dynamic Computation Offloading in Multi-Access Edge Computing via Ultra-Reliable and Low-Latency Communications," in IEEE Transactions on Signal and Information Processing over Networks, vol. 6, pp. 342-356, 2020, doi: 10.1109/TSIPN.2020.2981266.

[Mohammad2019] U. Mohammad and S. Sorour, "Adaptive Task Allocation for Mobile Edge Learning," 2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW), Marrakech, Morocco, 2019, pp. 1-6, doi: 10.1109/WCNCW.2019.8902527.

[Moreira10] P. Moreira, P. Alvarez, J. Serrano, I. Darwezeh and T. Wlostowksi, "Digital dual mixer time difference for sub-nanosecond time synchronization in Ethernet," in 2010 IEEE International Frequency Control Symposium, Newport Beach, CA, USA, 2010

[Nan17] Y. Nan et al., "Adaptive Energy-Aware Computation Offloading for Cloud of Things Systems," in IEEE Access, vol. 5, pp. 23947-23957, 2017, doi: 10.1109/ACCESS.2017.2766165.

[Neely10] M. J. Neely, Stochastic Network Optimization with Application to Communication and Queueing Systems, Morgan & Claypool Publishers, 2010

[Nielsen18] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," IEEE Transactions on Communications, vol. 66, pp. 1322–1334, mar 2018.

[NS3APP]https://www.nsnam.org/doxygen/classns3_1_1_on_off_application.html#details

[ONAP] ONAP. https://www.onap.org/tag/nfv

[ORANCUS] O-RAN Alliance Technical Specification, WG4：O-RAN.WG4.CUS.0-v03.00. 3, 2020

[OSM] OSM. https://osm.etsi.org/

[Park19] J. Park, S. Samarakoon, M. Bennis and M. Debbah, "Wireless Network Intelligence at the Edge," in Proceedings of the IEEE, vol. 107, no. 11, pp. 2204-2239, Nov. 2019, doi: 10.1109/JPROC.2019.2941458.

[Park20] J. Park, S. Samarakoon, H. Shiri, M. Abdel-Aziz, T. Nishio,A. Elgabli, and M. Bennis, "Extreme urllc: Vision, challenges,and key enablers," 1 2020.

[Patriciello19] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An e2esimulator for 5g NR networks,"Simulation Modelling Practiceand Theory, vol. 96, p. 101933, nov 2019

[Pedersen16] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5g frame structure design for frequency-division duplex cases," IEEE Communications Magazine, vol. 54,pp. 53–59, mar 2016.

[Pedersen17] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, “Punctured scheduling for critical low latency data on a shared channel with mobile broadband,” in 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), IEEE, sep 2017.

[Pham2019] Q. Pham et al., "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," in IEEE Access, vol. 8, pp. 116974-117017, 2020, doi: 10.1109/ACCESS.2020.3001277.

[Pocovi15] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen, and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks,” in 2015 IEEE Globecom Workshops (GC Wkshps), IEEE, dec 2015.

[Qadir15] J. Qadir, A. Ali, K.-L. A. Yau, A. Sathiaseelan, and J. Crowcroft, “Exploiting the power of multiplicity: A holistic survey of network-layer multipath, ”IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2176–2213, 2015.

[Rao18] J. Rao and S. Vrzic, “Packet duplication for URLLC in 5g: Architectural enhancements and performance analysis,” IEEE Network, vol. 32, pp. 32–40, mar 2018.

[RT36881] R. T. 36.881, “Study on latency reduction techniques for lte(release 14), v14.0.0,” tech. rep., 3GPP, Jun. 2016

[She16] C. She, C. Yang, and T. Q. S. Quek, “Uplink transmission design with massive machine type devices in tactile internet,” in 2016 IEEE Globecom Workshops (GC Wkshps), IEEE, dec 2016.

[Skatchkovsky19] N. Skatchkovsky and O. Simeone, "Optimizing Pipelined Computation and Communication for Latency-Constrained Edge Learning," in IEEE Communications Letters, vol. 23, no. 9, pp. 1542-1546, Sept. 2019, doi: 10.1109/LCOMM.2019.2922658

[SunZhou17] Y. Sun, S. Zhou and J. Xu, "EMM: Energy-Aware Mobility Management for Mobile Edge Computing in Ultra Dense Networks," in IEEE Journal on Selected Areas in Communications, vol. 35, no. 11, pp. 2637-2646, Nov. 2017, doi: 10.1109/JSAC.2017.2760160.

[TR19] G. P. A. W. Group, “View on 5g architecture, version 3.0,” tech.rep., 5GPP, 2019.

[TR38901] ETSI, 3GPP TR 38.901 Study on channel model for frequencies from 0.5 to 100 GHz, version 16.1.0 Release 16, 2020-11

[TS29281] 3rd Generation Partnership Project (3GPP), TS 29.281 (V15.7.0): General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U), 01, 2020

[TS38104] 3rd Generation Partnership Project (3GPP) TS 38.104 (V16.0.0)： Base Station (BS) radio transmission and reception. 6, 2019

[TS38300] 3rd Generation Partnership Project (3GPP), TS 38.300 (V16.0.0)：NR and NG-RAN Overall Description, 12, 2019

[TS38401] 3rd Generation Partnership Project (3GPP), TS 38.401 (V16.0.0)：NG-RAN; Architecture description. 12, 2019

[TS38413] 3rd Generation Partnership Project (3GPP), TS 38.413 (V15.10.0): NG Application Protocol (NGAP), 01, 2021.

[Wang19] S. Wang, X. Zhang, Z. Yan and W. Wenbo, "Cooperative Edge Computing With Sleep Control Under Nonuniform Traffic in Mobile Edge Networks," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4295-4306, June 2019, doi: 10.1109/JIOT.2018.2875939.

[WangHan20] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan and X. Chen, "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 869-904, Secondquarter 2020, doi: 10.1109/COMST.2020.2970550.

[WangTour18] S. Wang et al., "When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning," IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, Honolulu, HI, 2018, pp. 63-71, doi: 10.1109/INFOCOM.2018.8486403.

[Wittig20] S. Wittig, W. Keusgen, M. Peter and T. Eichler, "On the Dynamic Range of Digital Correlative Time Domain Radio Channel Measurements," arXiv preprint arXiv:2008.07805, 2020.

[Yang2018] Y. Yang, S. Zhao, W. Zhang, Y. Chen, X. Luo and J. Wang, "DEBTS: Delay Energy Balanced Task Scheduling in Homogeneous Fog Networks," in IEEE Internet of Things Journal, vol. 5, no. 3, pp. 2094-2106, June 2018, doi: 10.1109/JIOT.2018.2823000.

[Yu17] S. Yu, X. Wang and R. Langar, "Computation offloading for mobile edge computing: A deep learning approach," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, 2017, pp. 1-6, doi: 10.1109/PIMRC.2017.8292514.

[YuPu2018] Yu B., Pu L., Xie Q., Xu J., Zhang J. (2018) U-MEC: Energy-Efficient Mobile Edge Computing for IoT Applications in Ultra Dense Networks. In: Chellappan S., Cheng W., Li W. (eds) Wireless Algorithms, Systems, and Applications. WASA 2018. Lecture Notes in Computer Science, vol 10874. Springer, Cham. https://doi.org/10.1007/978-3-319-94268-1_51