Private 5G Networks for Connected Industries

Deliverable D4.2

Final 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:	<Date>		
Project Start Date:	01.10.2019	**Dura-tion:**	36 Months

Document Information

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

Document Number:	D4.2
Document Title:	Final specification and implementation of the building blocks
Editor:	Mickael Maman (CEA)
Authors:	Mickael Maman (CEA) Ngoc-Lam Dinh (CEA) Mohamed Sana (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) Sergio Barbarossa (SAP) Stefania Sardellitti (SAP) Paolo Di Lorenzo (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) Shally Lu (III) Frank Chih-Wei Su (III)
Dissemination Level:	Public
Contractual Date of Delivery:	31.12.2021
Work Package	WP 4
File Name:	861459-5G CONNI-D4.2-Final specification and implementation of the building blocks-V3.0.docx

Revision History

Version	Date	Comment
1.0	03.09.2021	First ToC
2.0	27.09.2021	Include D7.3 Description
3.0	22.11.2021	Final Version

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 that industrial use cases can be successfully implemented on private 5G networks for industrial requirements, including high data rates (eMBB) and low latency (URLLC).

D4.2 provides the final specification and implementation of the building blocks for private 5G networks. This deliverable is an extension of D4.1. These innovative components feed the laboratory integration reported in D5.2.

In WP4, activities are carried out on both the European and Taiwanese sides of the consortium with partners that reflect or complement the competences of both regions.

In task 4.1, Alpha Networks worked on the development of the RAN system, including the 5G CPE and gNodeB, and provided the 5G SA RAN prototype. CEA investigated how to enable deterministic URLLC. To this end, CEA worked on the design of the network orchestrator and on the combination of URLLC mechanisms using NS3 simulator. Moreover, CEA has studied a novel HARQ scheme for early decision-making. CEA is currently working on scheduling strategies for URLLC (determinist and opportunist approaches).

Within T4.2, Athonet worked on the ETSI NFV-like instantiation and orchestration of legacy 4G and then 5G mobile core network components via OSM. The framework has been successfully tested in-lab as well as during the recurrent ETSI NFV Plug tests. With Athonet VNFDs, it is possible to integrate the mobile core network with MANO implementations derived from OSM and from ONAP, provided by different vendors. III enhanced the performance and efficiency of the 5GC prototype through the specific architecture and interfaces

In T4.3, Athonet designed the solution for MEC deployments for the full-on-site and hybrid architectures and testing is underway to prepare final demonstrations. Chunghwa Telecom provided the 5G SA prototype bump-in-the-wire MEC platform and further developed multi-PDU session and multi-QoS flow functions. Chunghwa Telecom provided the ECoreCloud (ECC) NFV platform and MANO to manage Mobile Edge Enabler (MEE) VNF.

In Task 4.4, ITRI worked on three vertical use cases, namely (1) Process Diagnostics by CNC and Sensing Data Collection (2) Using Augmented/Virtual Reality for Process Diagnosis (3) Cloud-based CNC. Among these implemented use cases, (1) & (2) were implemented on a five-axis machine tool and (3) was implemented on a flexible fixture system, which is a specialized machine to test the cloud-based controller. For use case (1), 6 accelerometers were installed on the five-axis machine tool. Machining data and CNC data were collected and sent to the tool condition monitoring software deployed in MEC. For use case (2), 3D model of the machine was created. Two application were developed for machine operator and remote expert, these two applications share the same 3D machine model and are synchronized by the machine data from the Web API server developed by IMTC. The prototype of cloud CNC was developed and tested on the flexible fixture system by sending vibration suppression motion commands from cloud controller to the ground controller to evaluate the overall performance.

Last but not least, SAP developed and tested algorithms for dynamic allocation of radio and computational resources in a monitoring system where peripheral devices collect data and send them to an edge server that runs machine learning algorithms to make decisions on the observed data. The allocation is carried out to find an optimal balance between energy consumption, service delay, and accuracy of the decisions made by the edge server. Various constraints are incorporated in the method, including service delay, which includes queueing

delay in the communication and computation queues, and energy consumption. CEA collaborated with SAP on dynamic resource allocation for computation offloading. Furthermore, SAP has started to develop methods for dynamic service placement, generalizing the methods developed in WP3 to the dynamic case.

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5GC	5G Core
5G CONNI	5G for Connected Industries
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
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
DS	Delay Spread
DU	Data Unit
E2E	End-to-End
ECC	ECoreCloud
eMBB	Enhanced Mobile Broadband
EML	Edge Machine Learning
EPC	Evolved Packet Core
ENI	Experimental Networked Intelligence
ETSI	European Telecommunications Standards Institute
FoF	Factories of the Future
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
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
MIMO	Multiple Input, Multiple Output
MNO	Mobile Network Operator
NFV	Network Function Virtualization
NLOS	Non Line of Sight
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
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
SF	Shadow Fading
SMF	Session Management Function
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
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 final specification and implementation of private 5G networks building blocks. This deliverable is an extension of D4.1. These innovative components are fueling the lab integration reported in D5.1.

1.2 Structure

The document is divided in five sections according to the five defined building blocks. Section 2 investigates radio network technical enablers and provides on the one hand a 5G RAN system composed of a CPE and a gNB and on the other hand a RAN orchestrator enabling deterministic URLLC services. Section 3 focused on core network technical enablers and proposes two complementary solutions: a lightweight orchestration framework and a 5G core prototype. Section 4 investigates mobile edge cloud enablers supporting the requirements of smart factories in 5G eMBB and URLLC scenarios. Two implementations of MEC are proposed by 5G-CONNI for the European and the Taiwanese testbeds: the hybrid 5GC solution and the bump-in-the-wire solution. Section 5 details the implementation of the four selected use cases: (i) process diagnostics by CNC and sensing data collection, (ii) process diagnosis using AR/VR, (iii) cloud based controller for CNC and (iv) the multi-site use case. Section 6 rethinks the network in a holistic manner by jointly optimizing all enabling technologies and proposes several algorithms on dynamic resource allocation for wireless edge machine learning exploring energy-latency-reliability trade-offs and on energy efficient edge computing exploiting Lyapunov stochastic optimization and multi-agent reinforcement learning. Finally, Section 7 concludes this deliverable.

2 Radio Network Technical Enablers

The objectives of Task 4.1 are the development of the 5G radio access network and the investigation of deterministic URLLC services. This work is based on the industrial use cases generated by WP1.

2.1 5G RAN system composed of CPE and gNB

The main activity of Task 4.1 is the developing of 5G RAN components, specifically the CPE (Customer Premise Equipment) and gNodeB, complying with 3GPP.

The first RAN component is CPE. For our project, an industrial application CPE has been developed, with main features of 5G NR CPE listed as below:

- 5G NR Sub-6 (WAN)
- 2x GE Ethernet ports (LAN)
- Wi-Fi 5GHz (2x2, 1200Mbps)
- Wi-Fi 2.4GHz (2X2, 600Mbps)
- 1x USB 2.0 Type A
- 1x DIDO (1DI, 1DO)

The block diagram of industrial application CPE is shown in Figure 1.

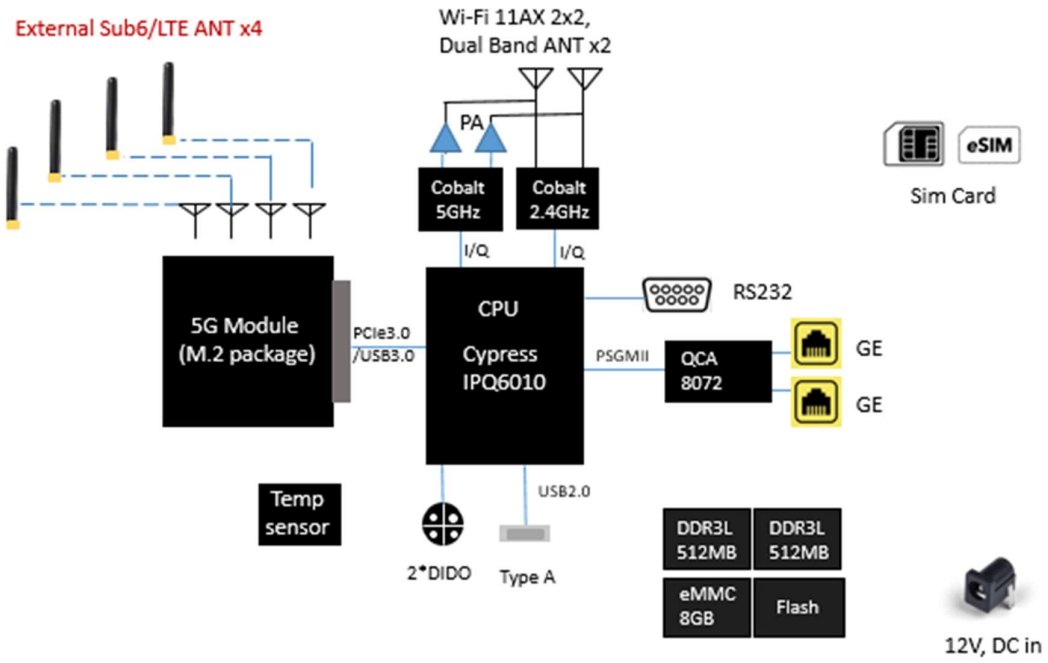

Figure 1:: Block diagram of industrial application CPE

The second RAN component is the gNodeB, a 3GPP-compliant implementation of the 5G NR base station. Alpha Networks has been developing a disaggregated gNodeB solution that consists of a RU and a CDU.

- RU: the radio unit implements a 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 the 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. The target RU specifications are listed as below:
  - 3GPP Release 15 compliant, option 7-2
  - RF Matrix: 4T4R
  - TX Power: +24dBm (250mW)/path
  - Frequency band: n79 (4800MHz-4900MHz)
  - Bandwidth: 100MHz
  - Backhaul: 10G SFP+
  - Network Synchronization: 1588v2
  - Weight: Under 3 kg
  - Dimension: 218mm x 218mm x 65mm
- CDU: 5GNR CDU consists of both 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 specifications of CDU are listed as below:
  - 3GPP Release 15 compliant
  - Support 4 RU
  - 5G NR Layer 2 and Layer 3
  - Single Cell
  - SA (Standalone) mode
  - 4x 10G Fronthaul interfaces
  - 2x25G Backhaul interfaces

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.

5G NR gNodeB Performance Brief

Several performance evaluations have been done on 5G NR gNodeB:

- RU RF performance
- RU Coverage
- Downlink Throughput
- E2E Throughput
- End to End Latency

RU RF performance including EVM, Frequency offset, ACLR, and RX sensitivity are listed as Table 1.

Table 1 : RU RF performance

Item	Spec	Chain 1	Chain 2	Chain 3	Chain 4	Result
EVM	<4.5%	3.7%	3.62%	3.93%	3.82%	Pass
Freq. error	<485 Hz	-41.87 Hz	-46.46 Hz	-46.41 Hz	-40.79 Hz	Pass
ACLR	<43.2 dBc	-48.27 dBc	-47.98dBc	-46.91dBc	-47.12dBc	Pass
Sensitivity	<-86.4 dBm	-92 dBm	-92 dBm	-92 dBm	-92 dBm	Pass

Concerning the RU Coverage, at 250 meter (LOS), the DL peak throughput is around 110 Mbps, the RU TX power is 24 dBm and the CPE RSRP is -102 dBm.

The DL peak throughput is 1Gbps and 800 Mbps in conductive and OTA respectively.

As a test result by month 23 (2021 Aug.), the E2E throughput is up to 650 Mbps. The E2E throughput test results are listed as Table 2.

Table 2 : E2E Throughput

Pattern	gNodeB MCS		Throughput UDP (Mbps)		Throughput TCP (Mbps)	
	DL	UL	DL	UL	DL	UL
P1	DMCS	DMCS	654	54.7	103	50.3
	28	25	652	55.4	127	52.5
	20	20	360	40.2	62.7	39.3
	15	15	284	29.5	194	28.5
	9	9	156	16.2	108	15.8
P2	DMCS	DMCS	324	97.7	42.1	24.4
	28	25	324	104	63.6	18.6
	20	20	192	76.3	50.3	62.6
	15	15	161	59.8	153	56
	9	9	55.4	32	22	32

As a test result by month 23 (2021 Aug.), the average E2E latency is under 12 ms. The E2E latency test results are listed as Table 3.

Table 3 : E2E Latency

Downlink UDP				Uplink UDP			
Frame size	Min Latency (ns)	Max Latency (ns)	Avg Latency (ns)	Frame size	Min Latency (ns)	Max Latency (ns)	Avg Latency (ns)
74	3613120	15314260	7390445	74	3312880	23921480	8651211
128	3454900	14999680	7433282	128	6710700	16599500	11344082
256	3976680	20879340	7609239	256	7102200	18021940	11332869
512	3315580	13169700	6269545	512	3323200	49758440	8767938
1024	3510960	14405580	5160787	1024	3840620	34908700	11217151
1280	3460160	11679900	4969842	1280	4487320	35331600	11405998
1400	3436780	11327380	4778362	1400	4233520	31996520	8939633

2.2 Deterministic URLLC protocols

The 5G and beyond network enables the exploitation of new emerging use cases, such as ultra-reliable low latency communication (URLLC). More advanced mechanisms are needed to jointly reduce latency and improve reliability while maintaining appropriate efficiency. The classical approach is to propose a systematic combination of several mechanisms to stretch

the latency below the deadline. This approach is not suitable for a dynamic environment. Proactive adaptation strategies select mechanisms for the worst case with a margin. This implies a high cost of ultra-reliable communications because worst-case impairments can be very rare. Reactive adaptation strategies propose to activate additional resources, but these strategies increase latency considerably and are based on an average single-modal latency model with jitter. The online decision maker (i.e., real-time orchestration) dynamically defines one or more decision moments to trade-off latency, reliability and efficiency. The goals are to define when to (de)activate more resources, to make an efficient trade-off between reactive and proactive approaches, and to exploit the multi-modal distribution of latency. In this study, the decision maker is applied to the well-known Hybrid Automatic Repeat Request (HARQ) procedure and is based on a strategy that allows a number (R) of parallel retransmissions than a single one. Based on the latency distribution statistic, the exploitation of early decision making (R-parallel RTXs) can be done instead of send-wait-react mode. The efficiency-latency trade-off is achieved by the timing of the decision making. The earlier the decision is made, the greater the latency gain at the cost of resource efficiency and vice versa.

In the literature, a lot of research has been conducted on the improvement of HARQ but has several limitations: (1) The tradeoff between resource efficiency, reliability and latency is not clearly defined; (2) The analytical studies do not consider many PHY/MAC technologies that are involved in the system; (3) The relationship between the traffic source rate and the system capability is missing. In our study, we further investigate the compromise between three main Key Performance Indicators (KPIs) of beyond 5G networks, namely latency, reliability and resource efficiency, which relies on a strategy that allows a certain number of retransmissions in parallel rather than a single one. By doing so, it is more likely that the latency gain can be achieved at the cost of reduced resource efficiency.
The general procedure is the following:

- Data packets are generated periodically at the application layer (App Data) with a size of L bytes and will be queued in the transmission buffer of gNB.
- The scheduler takes Tprep = a slot(s) to prepare the Transport Block (TB) in the corresponding buffer and send it over the air.
- The propagation time from the sender side to the receiver side is tp.
- The UE processes the received TB. In case of error, a NACK feedback is sent back to the gNB after Tfb = b slot(s) which represents the processing time at the UE. Otherwise, an ACK feedback is sent.
- A NACK feedback will trigger one or several retransmissions of the corrupted TB in adjacent transmission slots depending on the strategy after Tprep = a slot(s). The level of redundancy R will be made by decision maker.

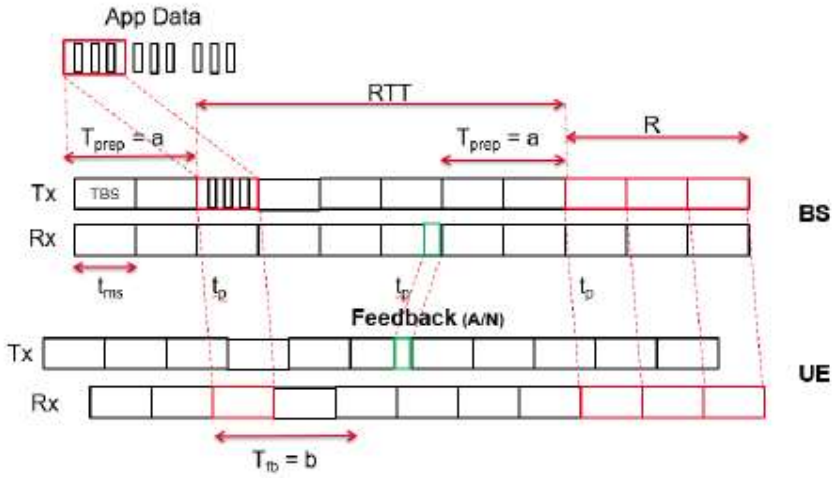

Figure 2: Parallel DL HARQ scheme

In order to evaluate the performance, we developed a system-level simulator based on NS-3 applying for 5G NR. Table 4 sums up the simulation parameters.

Table 4: Simulation parameters

Parameter	Values
L	60B
Traffic Low vs. High	λ=100 vs. 4000
(a,b)	(2, 1)
(fc, BW)	(3.5 GHz, 10 MHz)
Ptx	10 dBm
Numerology	1
(Utx, Srx)	(8×4, 4×4)
BLER	10^-4
Good vs. Bad channel	D=60 vs 120 m
(m, ηeff, CR)	(12, 1.6953, 0.42)
Kmax	6

Figure 3 shows the Complementary Distribution Function (CDF) of E2E latency for different HARQ schemes, where R = 1 and R > 1 illustrate the CDF of classical HARQ and proactive HARQ, respectively. In this case, low application traffic rate is associated with a bad channel condition to perform DL communication between the gNB and UE. Compared to the classical reactive approach (R=1), we can note that the latency gain is always obtained at a higher redundancy level and R = 3 is the optimal redundancy level to achieve reliability under a latency constraint. Settings beyond the optimum still perform better than the reactive approach, but the over-the-air queuing effect due to the overestimation of required RTXs starts to reduce performance.

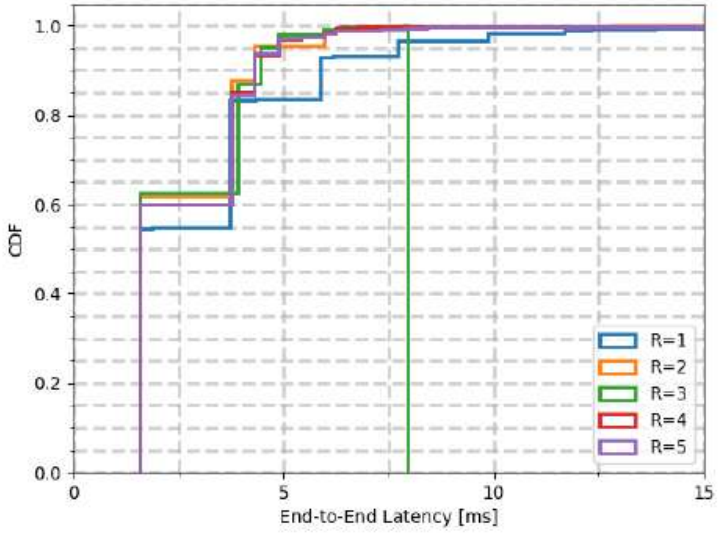

Figure 3: CDF of end-to-end latency for different redundancy levels R.

To evaluate the impact of channel conditions and arrival traffic rates, Figure 4 depicts the CDFs of E2E Latency for a redundancy level R = 3. When the channel condition is very good, a low traffic regime performs best, compared to other scenarios, as shown in particular by the probability of successful transmission during the first time communication. When packets are generated by the application at very high rates, the aggregation of the queuing effect occurs at the transmission buffer and the radio at the same time, resulting in degraded system performance. In this regard, a bad channel condition will severely spread the overall packet transmission delay.

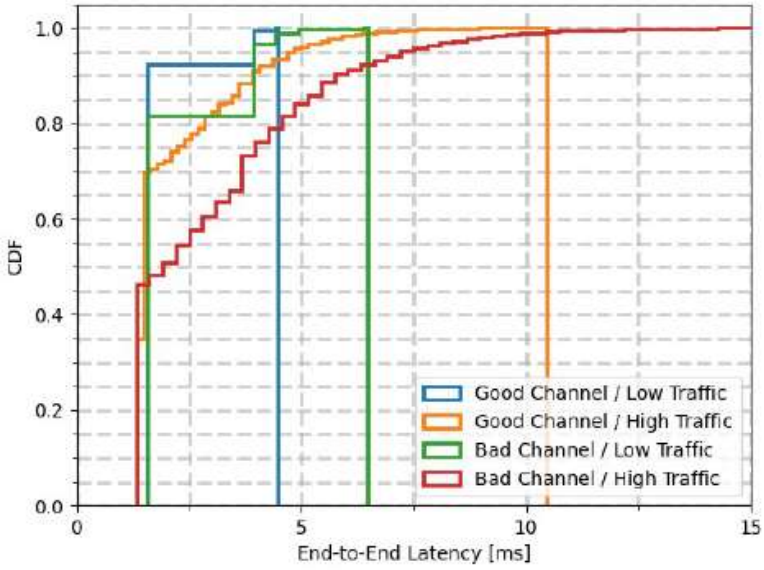

Figure 4: CDF of end-to-end latency for different channel conditions and traffic source rates under R=3

Figure 5 shows the outage probability over E2E latency. The results detail the E2E latency achieved to reach an outage of from 0.99 up to 0.99999 with different level of redundancy R. They also demonstrate how the latency gap is optimized to reach a more critical outage from

the more relaxed outage by increasing R. In general, an optimal R will facilitate the system to reach a target outage sooner than in other cases. However, in the case of a bad channel and frequent traffic rate, the aggregation queuing effect dominates the latency gain.

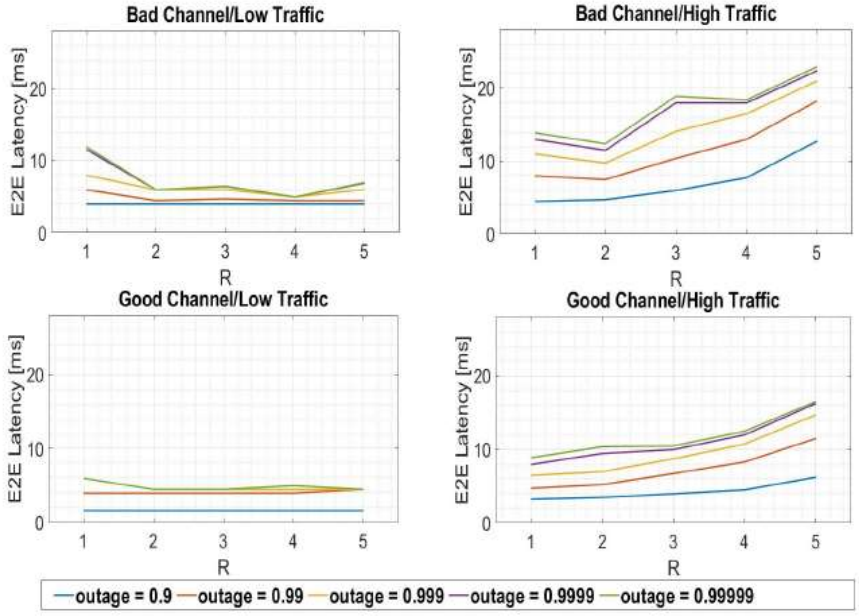

Figure 5: Outage probability on E2E Latency as a function of R, channel conditions and traffic source rates

2.3 Conclusions

In task 4.1, CEA has investigated how to enable deterministic URLLC and proposed a novel HARQ scheme for early decision-making. We evaluated the tradeoff between reliability, latency and resource efficiency by comparing the performance of classical reactive HARQ and proactive HARQ in a system level simulator, as a function of traffic source rate and channel conditions. We show that, by appropriately capturing several retransmissions before acknowledging the feedback, the end-to-end latency and jitter gains are achieved at the cost of reduced resource efficiency. The outage probability on E2E latency provided an indication of how quickly the system performance achieves a specific goal. Finally, we also showed that the side effect of inappropriate redundancy selection causes aggregation not only in the transmission buffer but also in-the-air, degrading the system performance. In future work, CEA will investigate scheduling strategies for URLLC (determinist vs. opportunist approaches).

Alpha Networks has built up a RAN system composed of CDU, RU, and CPE as RAN implementation. In future work, the gNodeB and CPE will be deployed in ITRI IMTC for industrial application. The fine tune will be needed to fulfill the requirement of the selected use case going through the remaining project period.

3 Core Network Technical Enablers

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

Two complementary activities have been carried out.

- Lightweight orchestration framework
- 5G Core prototype

3.1 NFV-like lightweight orchestration framework for the core network

In the context of the laboratory activities for the EU demonstrations, a lightweight orchestration framework, capable of performing the lifecycle management of a private mobile core prototype as a VNF, is being continuously developed and integrated. As explained in Sec. 5.1 of D4.1, for such a testbed we adopted open source and ETSI standard compliant software, namely OpenStack, as far as the VIM and NFVI layers are concerned, and Open Source MANO, for the management and orchestration layer. The final objective is to deploy and configure a 5G core prototype including network functions like AMF, SMF, AUSF, UDM, and UPF meant to run as a set of VNFs. For automatic core configuration, we focused on the analysis and implementation of the semantic of the Ve-Vnfm reference point, to make OSM able to directly configure the core network.

3.1.1 Orchestration framework upgrade

The NFV-like orchestration framework has been evolved by upgrading the versions of the software components as well as adding new functionalities, like metric monitoring and alarm provisioning. The updated framework architecture is shown in Figure 2.

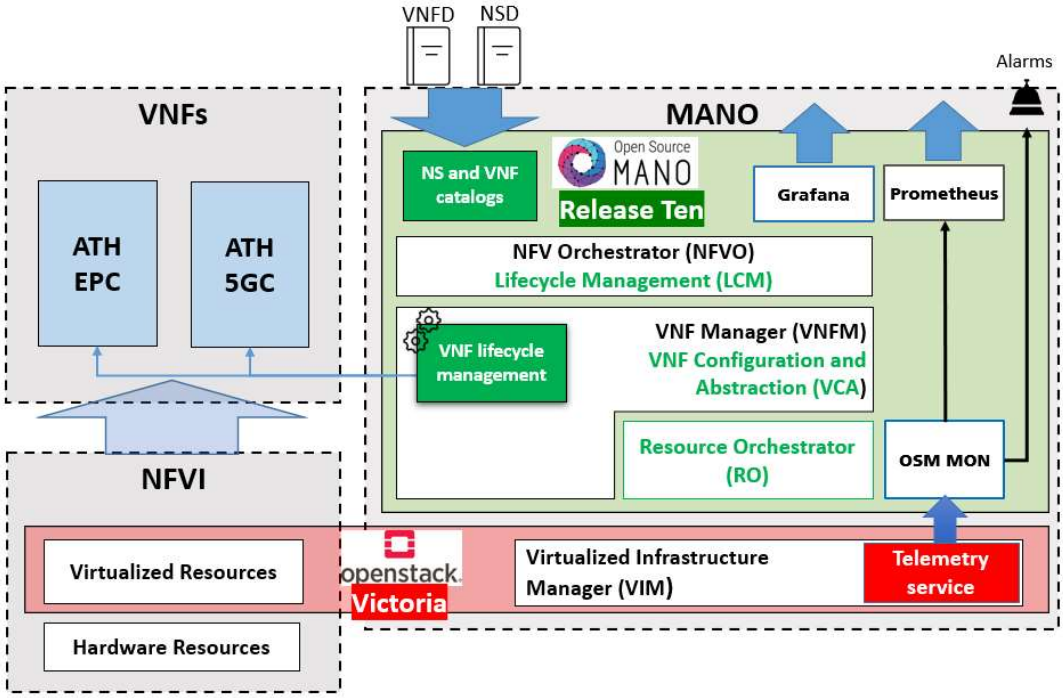

Figure 6: Updated NFV and orchestration testbed architecture.

OpenStack has been upgraded to Release Victoria. OSM has been upgraded to Release Ten, which includes new features like:

- a renewed web interface;
- new descriptors' syntax fully aligned with ETSI NFV SOL006 specifications;
- an improved support of Kubernetes, with inclusion of new functionalities such as, e.g. the scaling action for Kubernetes applications with Juju framework;
- the possibility to Scale-out/Scale-in manually through the GUI or the command line interface (CLI);
- Some enhancements in terms of alarms management.

3.1.2 VNF onboarding and configuration

The VNF onboarding procedure happens by means of VNF and network service (NS) packages, which contain the descriptors (VNFD and NSD). In turn, the descriptors gather all the characteristics (resources, connection points, scripts, policies, etc.) that a VNF shall own. From OSM Release Nine, the OSM Information Model augments ETSI NFV SOL006 YANG model, where the syntax of VNFD and NSD is completely different with respect to previous releases. Hence, we have modified all our previous descriptors in order to make them SOL006-compliant.

Within each VNF package, another element called *proxy charm* has been designed to contain all the core configuration. This configuration is provided to be sent over the Ve-Vnfm reference point between the VNFM and the VNF, adopting SOL002-compliant RESTful APIs. More details on these aspects can be found in D4.1, Section 5.1.

3.1.3 Deployment time analysis and optimization

As a follow-up of the previous investigation, the upgraded framework has been tested by onboarding and instantiating a virtual 4G core network (vEPC). The deployment time has been measured to verify the instantiation and configuration delay: the total deployment time is 587.81 seconds, as shown in Figure 7. The obtained performance is in line with what we obtained with the original setup, as detailed in D4.1, Section 5.1.

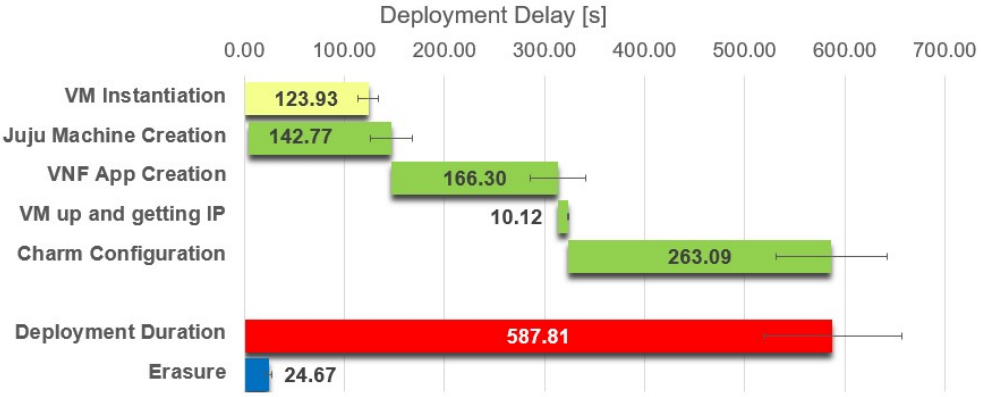

Figure 7: 4G core deployment delay without optimizations.

Moreover, in order to minimize this delay, some modifications have been applied to the Juju framework:

1. When the system starts to configure a VNF, the Juju operator framework creates a Linux container (LXC) that will handle the Juju charm. In the standard settings of OSM, Juju downloads the LXC from Internet at every VNF instantiation, to then load the charm script into this container and execute it. The download action takes time, so we

decided to modify Juju so that it downloads in advance the LXC, to have it locally and quickly available for every next instantiation request.
2. When the LXC is instantiated, its bootstrap includes the `apt-get update && apt-get upgrade` command, which takes a bunch of seconds. We disabled them.
3. Other minor modifications in Juju and proxy charms have been applied to save more seconds.

The deployment delay after these optimizations is shown in Figure 8. As one can notice, we saved more than 200 seconds with respect to the previous results, especially regarding the two operations Juju Machine Creation and VNF App Creation, whose times are significantly reduced. Concluding, we can deploy a completely configured EPC in slightly more than 6 minutes.

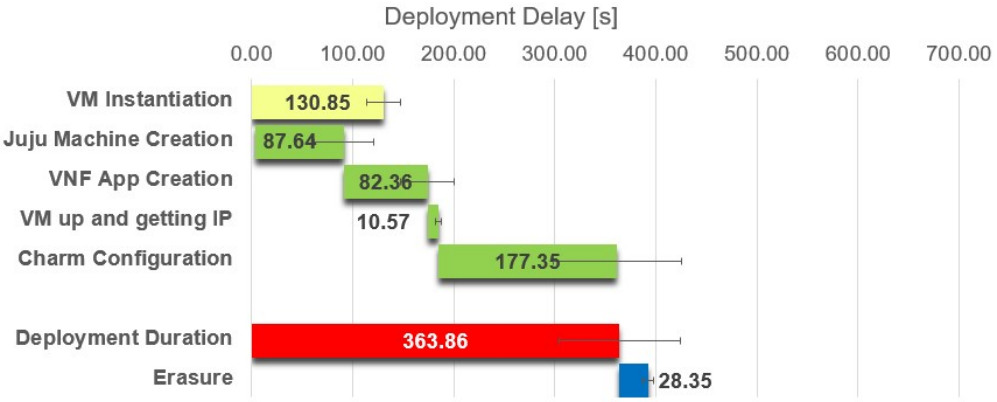

Figure 8: 4G core deployment delay with applied optimizations.

3.1.4 Metrics monitoring

Part of the upgrading work has been devoted to metric collection and their visualization. We were able to connect OSM to OpenStack to retrieve some metrics information regarding the virtualized infrastructure. An OSM module, called MON (MONitoring), is designed to interface OSM to the OpenStack Telemetry System (called Gnocchi), in order to collect VM performance metrics and expose them to other tools (e.g., for graphical visualization). In particular, we leveraged Prometheus and Grafana, both natively integrated in the OSM ecosystem. In Figure 9, we provide a screenshot of 3 VIM metrics shown on Grafana web interface for a testing 4G Core Network instantiation: CPU usage, memory usage and network traffic. We plan to update the framework to make it capable of retrieving also core-specific metrics.

Figure 9: Some vEPC performance metrics visualized in Grafana.

3.1.5 Alarms provisioning

We have also managed to implement the alarm functionality related to metric monitoring. As shown in Figure 10, the alarm policy has been defined in the VNFD, making a reference to the target metric. An alarm is triggered by the MON module when the metric value crosses a defined threshold value and generates a notification by mean of webhooks. A webhook can be sent to a pre-defined web application that monitors any anomalies of the deployed system.

```
vdu:
- alarm:
  - actions:
      alarm:
      - url: http://server:9000/hooks/alarm-webhook
      insufficient-data:
      - url: http://server:9000/hooks/no-data-webhook
      ok:
      - url: http://server:9000/hooks/ok-webhook
    alarm-id: alarm-1
    operation: GT
    value: 80
    vnf-monitoring-param-ref: vnf_cpu_metric
```

Figure 10: Alarm policy section added into the VNFD.

3.2 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. The III 5G Core provides efficient data plane and system reliability. Thus, we develop software solution and integrate hardware platform for data plane to enhance packet processing and load monitoring. By this way, the throughput can achieve to 10Gbps and to keep the data plane latency less than 1ms. For various environment enterprises use cases, III 5GC supports interworking with MEC and local breakout applications. We also support more than 100,000 UEs,

and data plane acceleration technology with DPDK or SmartNIC solution, furthermore throughput > 10Gbps and Latency < 1ms.(KPI)

Our leading product is the 5G Core for enterprise and private network scenarios. We especially focus on data plane efficiency and system reliability. Thus, we develop both software and hardware solutions for data plane to enhance packet processing and load monitoring. By this way, the throughput can achieve 10Gbps and the data plane latency is less than 1ms. For various environment enterprises use cases, III 5GC also supports interworking with MEC and local breakout applications.

The main functional specifications are:

- 5G Core Component : AUSF / UDM / PCF / NEF
- Support Function List : Xn & N2 Handover / Multiple PDU Session / Multiple QoS Flow per Session
- OAM : Configure Management / Fault Management
- Support more than 100,000 UEs
- Data Plane acceleration technology : DPDK or SmartNIC solution
- Throughput > 10Gbps and Latency < 1ms
- Local Breakout support
- High Availability support
- Cloud-Native architecture

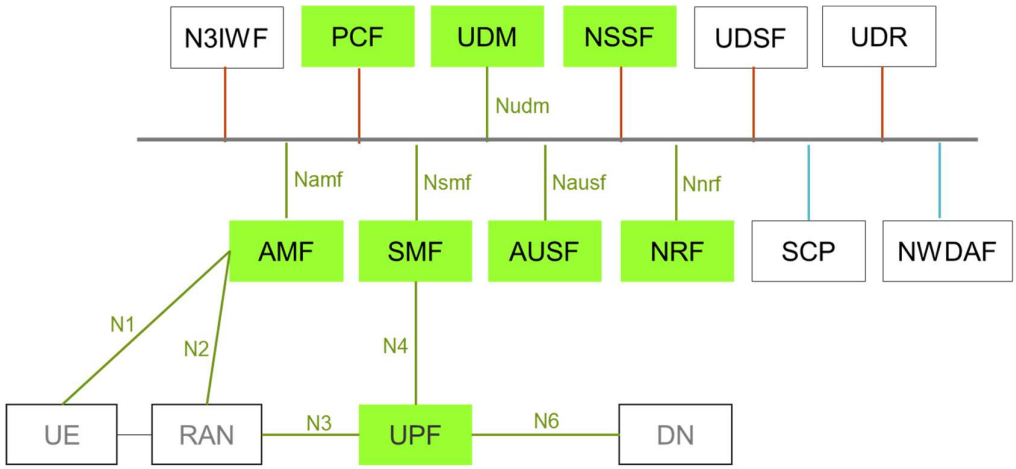

Figure 11: III 5G NG Core architecture

We keep following the 3GPP Standard to provide more 5G vertical use cases. For 2021, we focus on smart factory and industry 4.0 applications, and provide greater data throughput and latency improvements.

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.

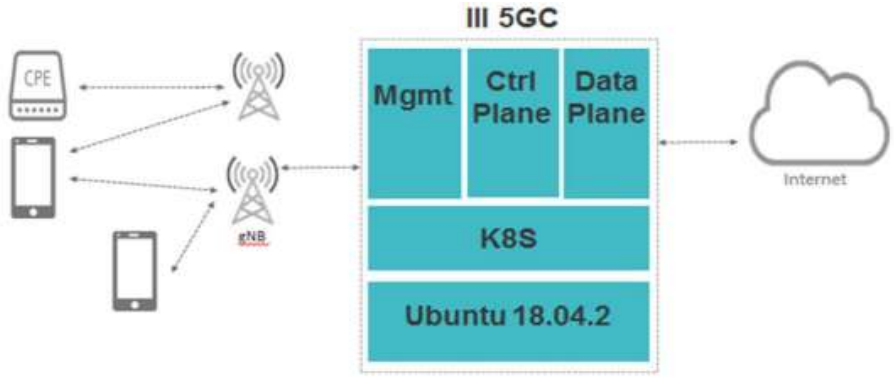

Figure 12: III 5GC Architecture

The network access and mobility management function (AMF) in III 5GC includes features of AUSF and UDM. The 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.

Table 5 : 5G Core Features

Module	Descriptions
AMF/UDM/AUSF	• Registration management • Mobility Management • Access Authentication & Authorization • Generation of 3GPP AKA Authentication Credentials • User Identification Handling • UE's Serving NF Registration Management • Authentication for 3GPP access and untrusted non-3GPP • Network Slice-Specific Authentication and Authorization
SMF	• Session Management (Establishment/Modify/Release) • UE IP address allocation & management • Selection and control of UP function
UPF	• Anchor point for Intra-/Inter-RAT mobility • Packet routing & forwarding

3.3 Conclusions

As part of T4.2, we have been working on the ETSI NFV-like instantiation and orchestration of legacy 4G first and then 5G mobile core network components via OSM. Thanks to the designed VNFDs, it is possible to deploy a mobile core network with off-the-shelf, standard-compliant MANO implementations like, e.g., OSM and ONAP. The framework is being continuously developed and integrated in-lab. The latest tests of the deployment delay showed that the proposed optimizations led to a significant delay reduction. OpenStack has been updated at release Victoria, and OSM at release Ten, with inclusion of additional functionalities. Metrics and alarms are also being investigated and tested.

In order to meet the network interconnection requirements between 5G private networks deployed in Taiwan and EU fields, respectively, the promising and expected approach is to realize the provision activities of the customer data by a unified provisioning system through a set of well-defined management APIs. The unified provisioning system shall support the flexibility to manage the 5G user data for each side of UDMs, respectively. The common UE tables with the service permissions enable the feature of the seamless access applications installed in 5G private networks on both POC fields - Taiwan and EU. Therefore, for example, after moving from EU to Taiwan, the ATHONET's UE can perform directly service registration procedures to III's UDM as a local UE and get access to the applications deployed on Taiwan site. The service scenario of the interconnection feature mentioned above will be verified on both sides to demonstrate that the same UE profile provisioned in two UDMs located in different networks can have permission to access from any side of the project network.

4 Mobile Edge Cloud Enablers

The objective on task 4.3 is to develop MEC technologies to be deployed in the project's testbeds, which support the requirements of smart factories in 5G eMBB and URLLC scenarios.. Sections 4.1 and 4.2 present the MEC implementation methods in 5G-CONNI for the European and the Taiwanese testbeds, respectively.

The hybrid 5GC solution is implemented in European testbed and the bump-in-the-wire solution is implemented in Taiwanese testbed

4.1 MEC based hybrid architecture

Among the options for the mobile network architecture discussed in Section 3 of D2.1, the splitting of CP and UP functions in the hybrid 5GC solution has been addressed. Because of its high flexibility, this solution has been chosen as the reference architecture for the European testbed of 5G-CONNI, and it entails a remote control center (which manages the CP) and an edge node (which manages the UP). This 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. Therefore, a MEC platform can easily be mapped into the 5G system architecture. This framework has also been chosen as the reference scenario for 5G-CONNI's European testbed, as described in D5.1 and the MEC host's data plane is mapped to the 5G's UPF element.

As shown in Figure 13, a multi-node NFVI cluster which mimics a multi-site scenario featuring a remote control center and an edge node has been designed.

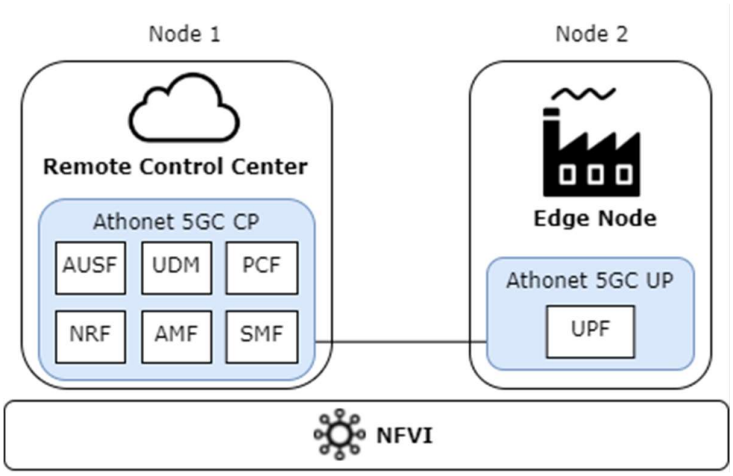

Figure 13: Multi-site scenario deployed in the NFVI cluster.

This has been achieved by adding a second physical server on our original infrastructure and upgrading Openstack to be hosted on the second node as well. This second node is a standard Desktop PC, whose characteristics have been reported in Table 6. Therefore, the first server simulates the remote cloud datacenter, hosting the 5GC CP network functions, while the second machine acts as edge node, hosting the 5G UPF, which shall provide the MEC functionalities. The edge node has been associated to OSM as a second Openstack VIM and several edge deployment tests have been performed to check the configuration.

Table 6: Physical resources of the PC Desktop node.

Component	Characteristic
CPU	Intel® Core™ i7-2600, 3.4 GHz
Memory	16 GiB, DDR3
Storage	500 GB, SSD
OS	Ubuntu 20.04.2 LTS
Openstack Services	Compute Metering-compute

The designed multi-node cluster will be the workbench for the envisioned *hybrid* deployment described in D5.1, section 2 and chosen as reference architecture for this task. The Athonet 5GC is going to be prepared in order to deploy such as distributed architecture. Preliminary experiments will include manual deployment and configuration of the UPF on the edge node, since OSM is not able to automate on both deployment and configuration of such a hybrid scenario without specific onboarding instructions in the related VNF descriptor (VNFD).

In future work, we are planning to demonstrate the NFV-like orchestration framework the envisaged hybrid deployment, where the UPF is automatically deployed in the edge node. In this respect, the VNFD is being analyzed and updated, so that it can automatically instantiate another 5GC element on the second node as the UPF. After that, the UPF configuration will be designed to be applied as the proxy charm in OSM. Then, some connectivity tests will be performed to check the end-to-end connectivity and performance between these two nodes. Finally, a MEC platform will be investigated and deployed therein as well in order to minimize the data plane latency. All the related experiments' results will be outlined in the next deliverable D4.3.

4.2 MEC based bump-in-the-wire architecture

The bump-in-the-wire MEC SA prototype has been developed that can interoperate between 5G standalone base stations and core networks and based on 3GPP SA standalone specifications. The MEC 5G SA prototype has been integrated with III's 5G core network, Alpha's 5G base station and applications on ITRI IMTC site. Augmented Reality (AR) application has been adopted on process diagnostics. The prototype of all network components integration on the Taiwanese site has been deployed at ITRI IMTC field for IEEE Globecom 2020 demo.

The bump-in-the-wire MEC SA has developed the stable version with the handover, multi-PDU sessions and multi-QoS flows functionalities and performance tuning for the 5G CONNI project. Handover includes N2 based and Xn based. The enterprises' field requires multiple base stations to cover the entire area because of the large dimensions. 5G network slices feature has been supported by multi-PDU sessions. The 5GC support S-NSSAI for the user building two PDU sessions to connect different UPFs and Data networks to achieve network slices, as shown in Figure 14. Since bump-in-the-wire MEC has to follow the 3GPP standard, it must provide multi-PDU sessions functionality to process the traffic to different UPFs.

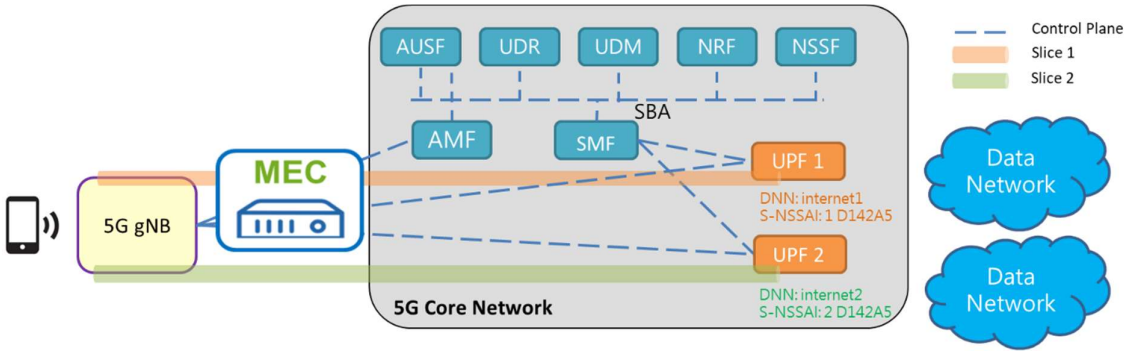

Figure 14 Network slices of two PDU sessions

In order to achieve end-to-end network slices in 5G, the QoS flow to process IP data flow from RAN to core network has been adopted. The QoS flow is compared with the bearer in 4G, as shown in Figure 15. One user can build more than one bearer to transmit different QoS traffic in 4G. On the other hand, one user can build more than one QoS flow and qosflow_id of GTP extension header to distinguish QoS flows of one PDU session in 5G.There are more than one bearers in 4G commercial network to provide different services such as VoLTE. So MEC 5G SA must provide multi-QoS flow functionalities for connecting to the commercial 5G SA network in the future.

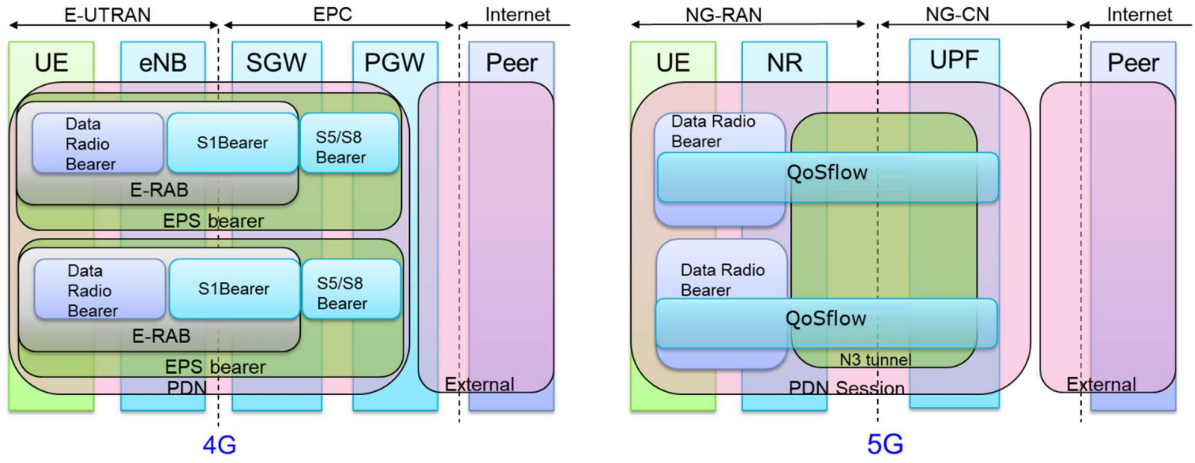

Figure 15 QoS session comparison between 4G and 5G

MEC 5G SA deployment at ITRI IMTC site

The MEC 5G SA is deployed at ITRI IMTC field. The MEC SA has been integrated with the new version of III's 5G core network and Alpha's 5G base station. The different architecture from D4.1 is that the III's 5G core is deployed at ITRI K78 as the enterprise datacenter, as shown in Figure 16. The user can access registration successfully to 5GC through the base station and MEC. MEC 5G SA has also been integrated with Augmented/Virtual Reality for Process Diagnosis industrial applications. The machine operator uses the Microsoft HoloLens to access the motion status of CNC data and sensing data via Web API. The snapshot of Microsoft HoloLens for the test result is also shown in Figure 16.

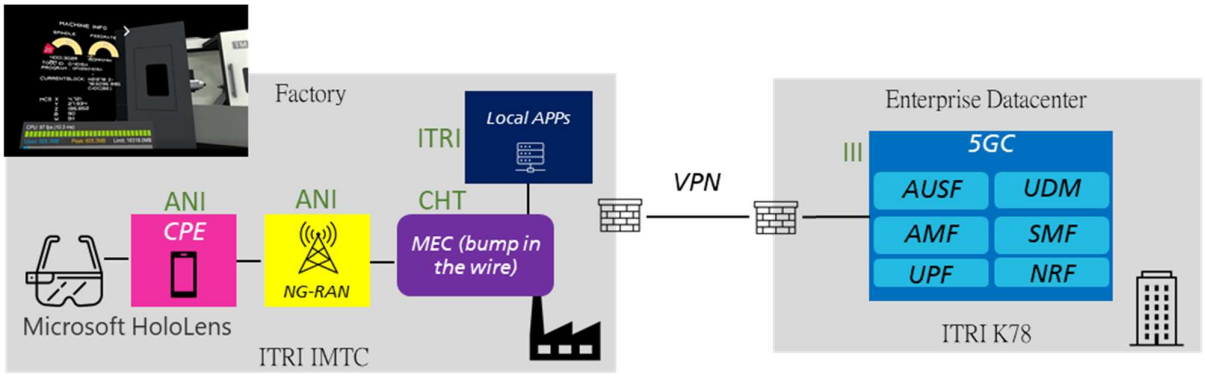

Figure 16 MEC 5G SA integrated with new version of III's 5G core network, Alpha's 5G base station and ITRI's Augmented/Virtual Reality for Process Diagnosis industrial application

ECoreCloud Cloud Network Development

With the virtualization of ITRI IMTC's functions and MEC, virtualization functions will deploy on the CHT orchestration platform (namely ECoreCloud) as VNFs. This kind of VNFs also requires a dedicated network connection to isolate different traffic for improving performance and stability. To this end, the ECoreCloud will support network connection management among VNFs based on ETSI specifications (e.g., NFV-SOL 005 and NFV-IFA 014). To realize network connection management of VNFs, the cooperative operation of physical and virtual network equipments is necessary.

The ECoreCloud plays the role of administrator for network equipment, providing routing paths for non-real-time and non-mass general traffic (such as control signals) of VNFs to virtual network equipment. When VNFs have real-time or massive communication requirements, ECoreCloud will isolate traffic by configuring independent network interfaces for VNFs. Therefore, independent network interfaces transmit isolated traffic directly to the physical network equipment to reduce interference for general traffic of virtual network equipment.

ECoreCloud has managed to realize the network connection of VNFs among virtual network equipment with ETSI specifications. Each virtual network connection is implemented by configuring routing rules on OpenvSwitch in the ECoreCloud, so ECoreCloud supports various routing paths of VNFs in the NSD or VNFD by configuring routing rules. Regarding the connection of physical network devices, ECoreCloud plans to support PCI pass through or SR-IOV to support real-time or massive network communication requirements. Regarding the network connection of physical network equipment, ECoreCloud plans to support Independent network interfaces (e.g., PCI pass through and SR-IOV) to support real-time or massive communication requirements. At present, ECoreCloud fully supports virtual network connections and will continue to plan feasible methods of physical network connections for VNFs

4.3 Conclusions

In T4.3, the solution for MEC deployments for the full-on-site and the hybrid architectures has been designed by Athonet. A PC Desktop server node has been added into the infrastructure to simulate an edge node, with Openstack instance installed and configured. Some preliminary configurations have been applied to make connectivity with the central server. Future works comprise the Athonet 5GC deployment and performance tests on the complete multi-node

infrastructure, with the instantiation of UPF on the edge node. Some deployment experiments will be scheduled in order to outline preliminary results, which will be particularly relevant for the final demonstrations.

MEC 5G SA based on bump-in-the-wire architecture has been developed handover, multi-PDU sessions and multi-QoS flows functionalities. ECoreCloud (ECC) NFV platform and MANO also provide the management of network connection of VNFs. In future work, the other industrial applications from ITRI IMTC will be integrated with MEC. One of the industrial applications will be deployed and managed on the ECC platform, and the performance of MEC SA will be tuned.

5 Industrial Application Enablers

The objective of this task is to implement the use cases selected in Task 1.1 for a proof-of-concept demonstration. A major item within this task is the initial implementation of use cases selected in Task 1.1 at ITRI shop floor, including necessary computation offloading strategies to enable complex processing of data collected from shop floor to guarantee continuous monitoring and anomaly detection during industrial processes. The progress of implementation and installation of the enablers are as follow.

5.1 Use case 1 prototype: Process Diagnostics by CNC and Sensing Data Collection

The use case 1 prototype focuses on:

- Final setup of the target machine (Litz TM2500 5-axis turn mill machine) shown in Figure 17. Six accelerometers have been installed on the tool spindle and workpiece spindle and connected to six independent single channel data acquisition devices as shown in Figure 18 and Figure 19
- Software module for data collecting installed in an industrial PC and connected with MEC. Data items supported by the data collection module are shown in Table 7
- Data analysis software installed in MEC to analyze tool condition on the target machine during machining. The user interface of the analysis software is shown in Figure 20

Figure 17: The target machine (Litz TM2500 5-axis turn mill machine)

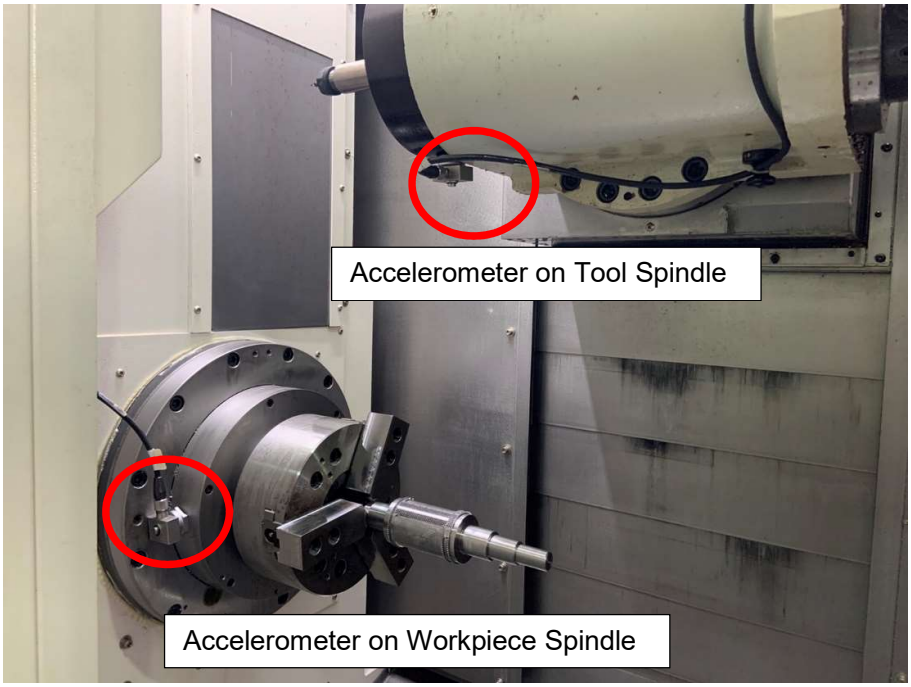

Figure 18: Accelerometers installed on machine

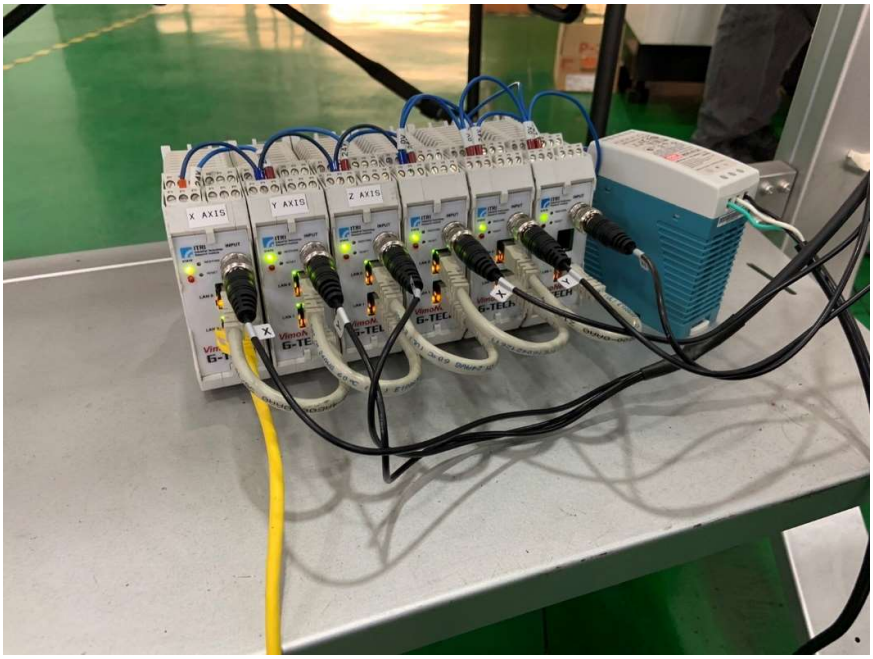

Figure 19: Six independent single channel data acquisition devices.

Table 7 : Data items supported by the data collection module

Item	Name
1	Feed Rate (speed of tool w.r.t workpiece)
2	Current NC Program Name
3	Current NC Code Block
4	Machine Coordinate Value
5	Absolute Coordinate Value
6	Actual Spindle Speed(RPM)
7	Active Tool ID
8	Spindle Load(Torque)
9	Autocorrelation of acceleration for each sensing channel
10	Average of acceleration for each sensing channel
11	Crest Factor of acceleration for each sensing channel
12	Frequency response of acceleration for each sensing channel
13	Root Mean Square of acceleration for each sensing channel
14	Skewness of acceleration for each sensing channel
15	Standard Deviation of acceleration for each sensing channel

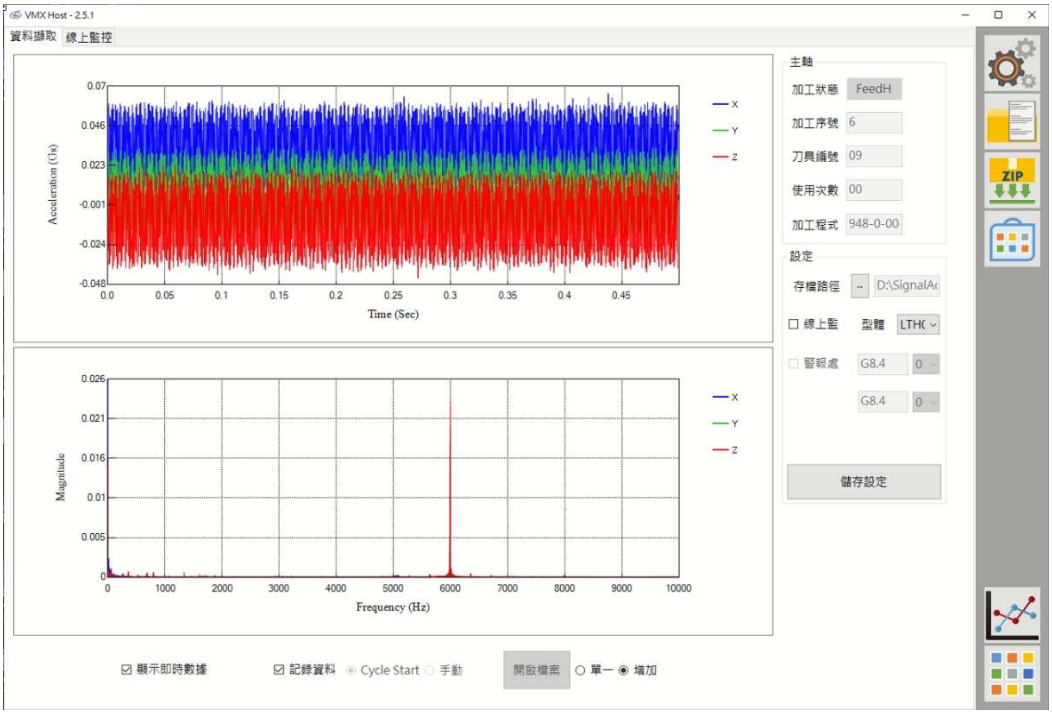

Figure 20: Data analysis software for tool condition monitoring

5.2 Use case 2 prototype: Process Diagnosis Using Augmented/Virtual Reality

The use case 2 prototype focuses on:

- Implementation architecture for the Using Augmented/Virtual Reality for Process Diagnosis use case, shown in Figure 21. A 3D model of the machine tool has been constructed for applications for machine operator and remote expert. Both applications can access machine CNC data and sensing data via web API so that 3D models are synchronized.

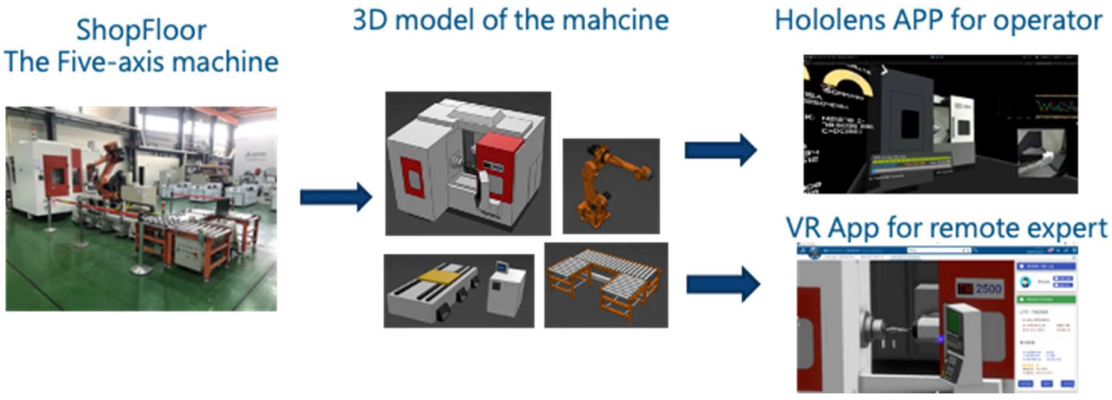

Figure 21: Implementation for the Using Augmented/Virtual Reality for Process Diagnosis use case

- The snapshot of the hololens app for operator, shown in Figure 22. With this app, machine operator can observe the motion status and key operational parameters as well as sensing data without opening the machine door or interrupting machining task
- The snapshot of the VR app for remote expert, shown in Figure 23. When there are difficulties in resolving shop floor problems, the machine operator can call out to the remote expert. The VR app for remote expert can connect to the physical machine via the data collection module implemented in use case 1 and synchronize the motion and operational parameters and sensing data and resolve the process issue by using sophisticated simulation and diagnosis software with collected shop floor data.

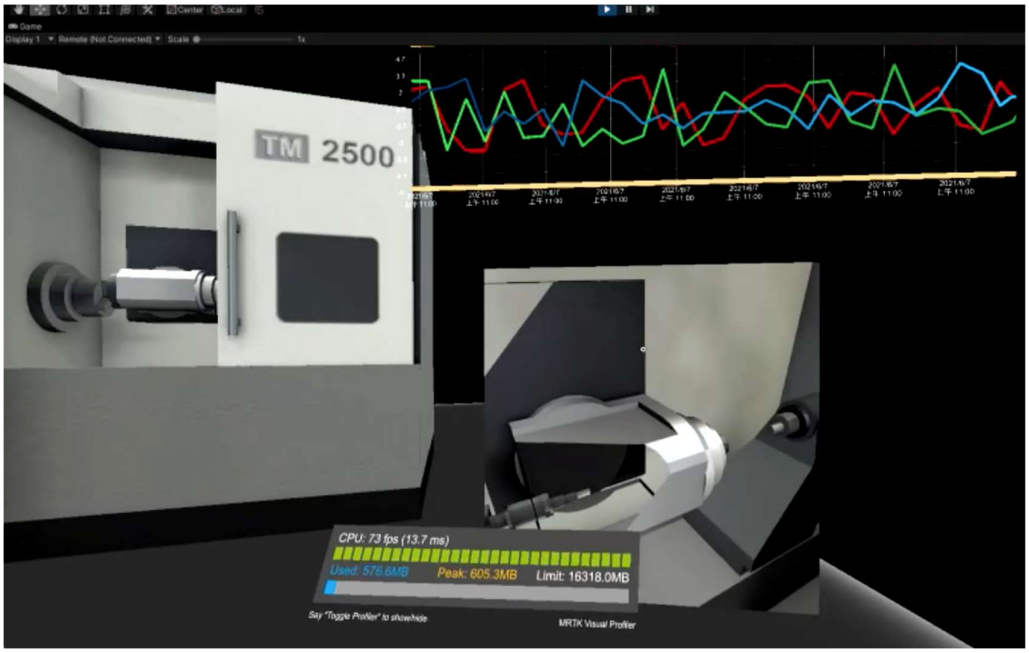

Figure 22: The snapshot of the hololens app for operator

Figure 23: The snapshot ofr the VR app for remote expert

5.3 Use case 3 prototype: Cloud-Based Controller for CNC

The cloud-based CNC software and the test machine has been constructed and tested under distributed network architecture (shown in Figure 24) where the motion command generation, motion command execution modules are separated. The test machine is shown in Figure 25, which is a flexible fixture system used in aerospace part machining. In the test scenario, a steel plate workpiece is installed on the fixture system and excited by a shaker to simulate the vibration during machining process. The cloud controller sends a series of motion command to particular moving axis to reduce the vibration.

The objectives of the testing is as follows:

- The suppression of vibrations under 500Hz (structural, shell workpiece)
- The sending of the vibration suppression command package every 10ms, each contains 40 motion commands
- Latency and reliability are critical as packet loss causes jitter and phase error on motion command

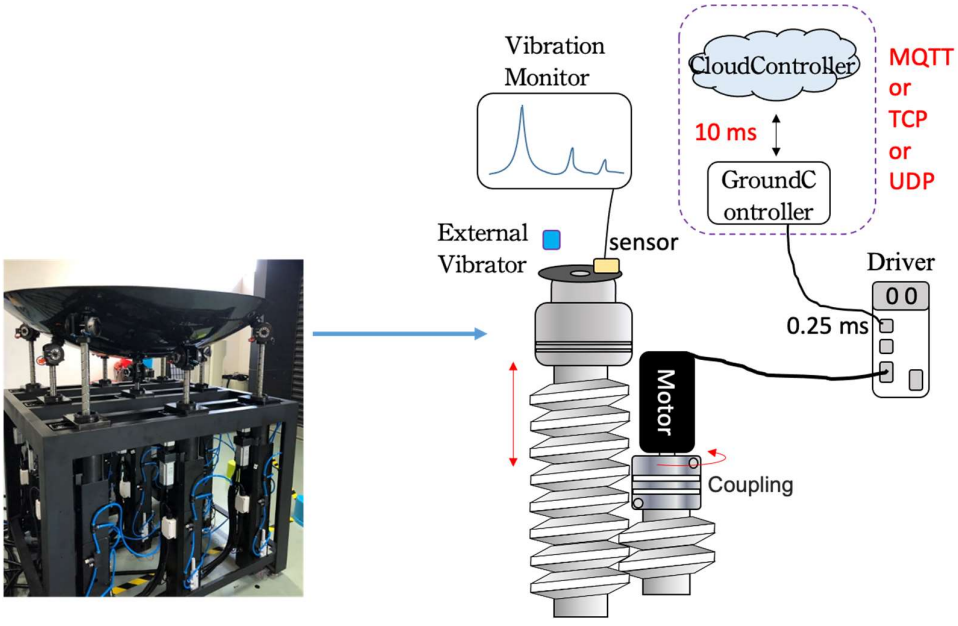

Figure 24: Architecture of the cloud CNC

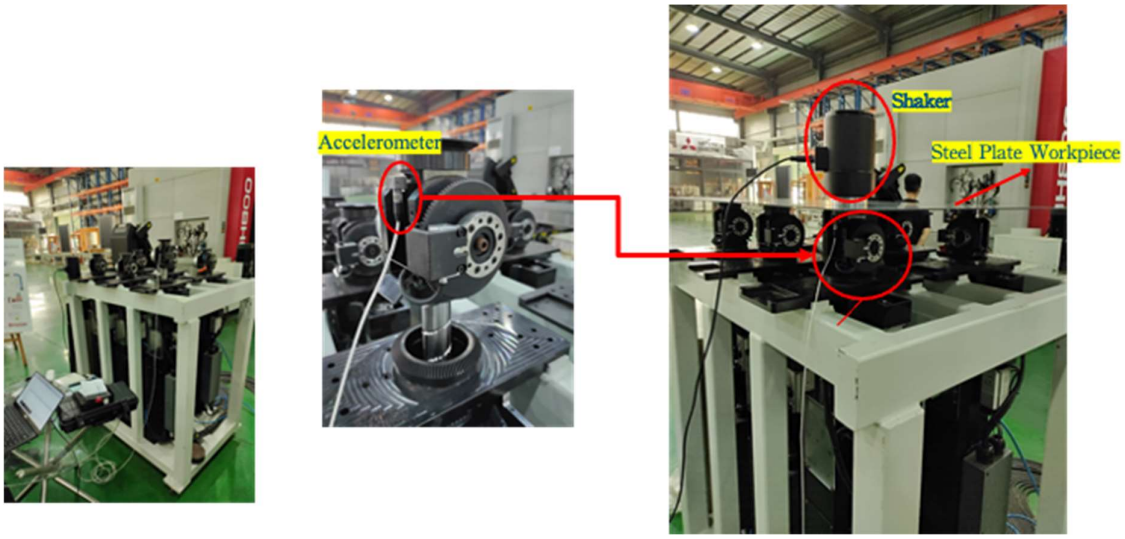

Figure 25: Test machine for the cloud CNC

5.4 Multi-site use cases

Deploying manufacturing sites overseas has been a common practice for companies that aim to reduce cost on production and logistics. Collaboration of engineers from various location and transfer expertise to production sites calls for a lot of traveling cost. With the pandemic of COVID 19, things get worse as almost all international traveling has been shut down. 5G technology brings new opportunity to resolve the above difficulty as we can link engineers with digital twins and interact with each other via cutting edge AR/VR technology to provide immersive environment so that they can design, plan, and troubleshooting in the same virtual factory. By doing so, experts from enterprise headquarters can (i) support manufacturing sites all around the world, (ii) reduce traveling cost for collaboration and (iii) quickly deploy new manufacturing sites while keeping the core technology within enterprise and provide necessary support by using digital twins in the cloud computing platform.

There are two proposed implementation architecture as shown in Table 8

Table 8 : Proposed architectures for multi-site use case

Architecture	Implementation Requirements
Option 1:Remote Rendering with nVidia CloudXR SDK (with Dassault Soft license) (*CloudXR 3.0 not support Hololens 2) (Shown in Figure 26)	• Workstation PC (nVidia Pascal GPU) with Dassault 3DExperience, linked with MEC. Use the Dassault product as-it-is. • Development effort base on Dassault SDK is required to link 3Dexcite with Web API from ITRI • CloudXR developer license is required
Option 2:Remote Rendering with nVidia CloudXR SDK (without Dassault Soft license) (*CloudXR 3.0 not support Hololens 2)	• Workstation PC or rack server (nVidia Pascal GPU), linked with MEC. • CloudXR developer license is required

(Shown in Figure 27)	• Development effort based on CloudXR SDK • Development effort based on Unity or Unreal to develop SteamVR compatible Apps to show 3D scene of ITRI site

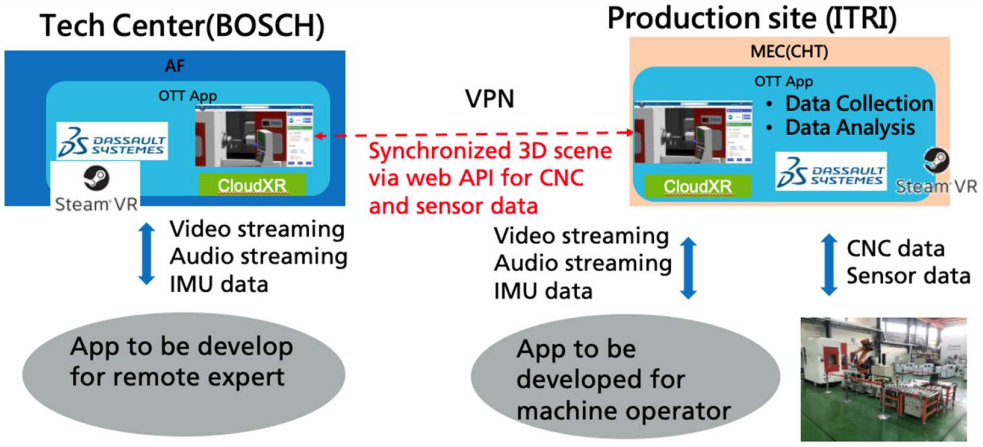

Figure 26: The implementation architecture for option 1

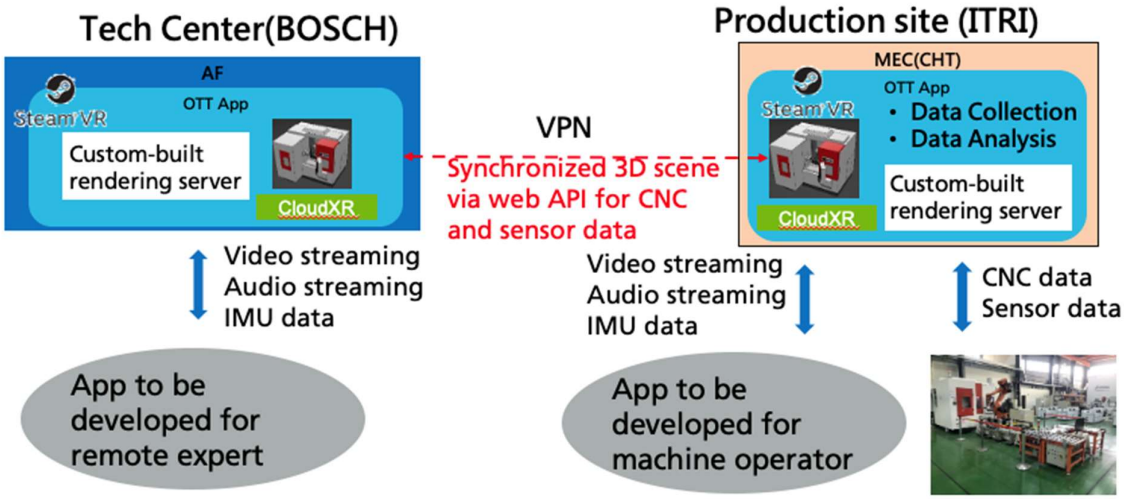

Figure 27: The implementation architecture for option 2

5.5 Conclusions

In Task 4.4, three vertical use cases have been implemented in ITRI site, namely (1) Process Diagnostics by CNC and Sensing Data Collection (2) Using Augmented/Virtual Reality for Process Diagnosis (3) Cloud-based CNC. Among these implemented use cases, (1) & (2) were implemented on a five-axis machine tool and (3) was implemented on a flexible fixture system, which is a specialized machine to test the cloud-based controller.

For use case (1), 6 accelerometers were installed on the five-axis machine tool. In addition to these attached sensors, there is also a thermal sensor built in the machining spindle, which will also be collected. Machining data and CNC data were collected and sent to the tool condition monitoring software deployed in MEC. For use case (2), 3D model of the machine have

been created. Two applications were developed for machine operator and remote expert, these two applications share the same 3D machine model and synchronized by the machine data from the Web API server developed by IMTC.

The prototype of cloud CNC has been developed and tested on the flexible fixturing system by sending vibration suppression motion commands from cloud controller to the ground controller to evaluate overall performance.

All these use cases are currently being tested under the existing WiFi infrastructure in the ITRI site and should be available for testing under 5G private network by the end of November. Performance data and comparison between WiFi and 5G will be shown in the final report.

6 Joint optimization of enabling technologies

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). 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:

6.1 Energy Efficient Edge Computing

Wireless communication networks are experiencing an unprecedented revolution, evolving from pure communication systems towards a tight integration of communication, computation, caching, and control. Edge Computing was conceived to enable energy efficient, low-latency, highly reliable services by bringing cloud resources close to users. In this context, dynamic computation offloading allows resource-poor devices to reliably transfer application execution to a nearby Edge Server (ES), with the purpose of reducing energy consumption and/or latency. From a network management point of view, this task is complex and requires the joint optimization of radio and computation resources. Thus, CEA collaborated with SAP on energy efficient computation offloading enabled by edge computing. A paper called “Lyapunov Meets Distributed Reinforcement Learning for Energy Efficient Edge Computing” have been published at 2021 IEEE International conference on communications.

In this study, we combine the convenience of a model-based solution that exploits Lyapunov stochastic optimization, with the power of data-driven solutions based on multi-agent reinforcement learning (MARL), aiming at energy efficient computation offloading from an overall network perspective. We consider a scenario where multiple users compete for radio resources, generating interference onto each other’s communications, and for computation resources in a single core of the ES. Thus, multiple users simultaneously compete for limited radio and edge computing resources to get offloaded tasks processed under a delay constraint. The radio resource allocation takes into account inter-cell and intra-cell interference, and the duty cycles of the radio and computing equipment have to be jointly optimized to minimize the overall energy consumption To address this issue, we formulate a dynamic offloading problem as a long-term system energy minimization with average end-to-end delay constraints, in a network comprising multiple Access Points (APs) and one ES, all capable of exploiting low-power sleep operation modes. Although we do not assume any knowledge on radio channels and data arrival statistics, we propose an online solution that, in each time slot, optimizes the user-AP association in a distributed way, and the ES’s CPU scheduling via a fast iterative algorithm

whose solution scales linearly with the number of user equipment (UEs). Whereas the first one can be efficiently solved using a fast iterative algorithm, the second one is solved using distributed multi-agent reinforcement learning due to its non-convexity and NP-hardness. The originality of our strategy with respect to the state of the art, lies in the capability of simultaneously: i) minimizing the duty cycles of all the network elements under delay constraints; ii) effectively managing radio interference; iii) being low-complexity; iv) combining Lyapunov optimization with DRL; v) being distributed and compatible with UE's mobility.

We consider a network scenario as depicted in Figure 28, where K UEs offload computational tasks to an ES, via one out of N possible APs. For the radio channel, we consider a spatial division multiple access system, in which all UEs are served by the APs over the same time-frequency resources but with different beams. In our setting, computation offloading involves two steps: i) an uplink transmission phase of input data from the UEs; ii) a computation phase at the ES. New data units are continuously generated from an application at the UE's side and consequently offloaded and processed at the ES.

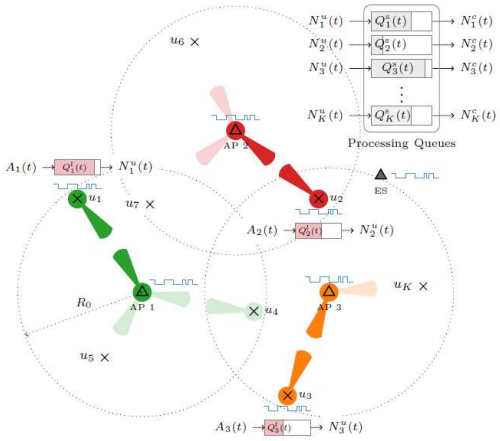

Figure 28: Network model with 3 APs deployed with K UEs.

We compare our solution, labeled L2OFF (Learning to Offload) in Figure 29 and Figure 30, to two benchmarks:

- Exhaustive search: at each slot, we perform an exhaustive search over all possible solutions.
- Max-SNR: each UE is associated with a Bernoulli random variable with probability p of being in active state (corresponding to the average duty cycle of UEs). Then, at each t, an active UE gets associated with the AP providing the maximum signal-to-noise ratio (SNR).

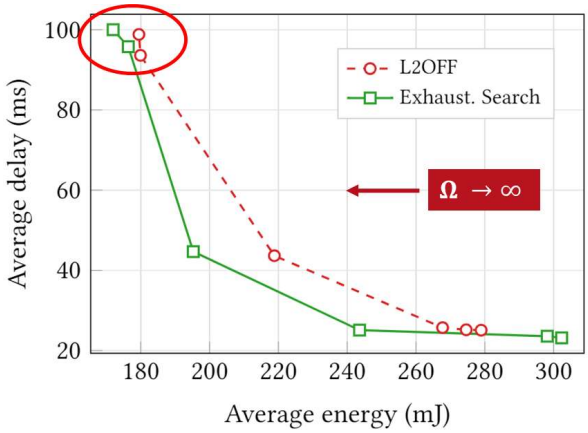

Figure 29: Energy-delay trade-off for K = 6 UEs and for a fixed delay constraint of 100 ms

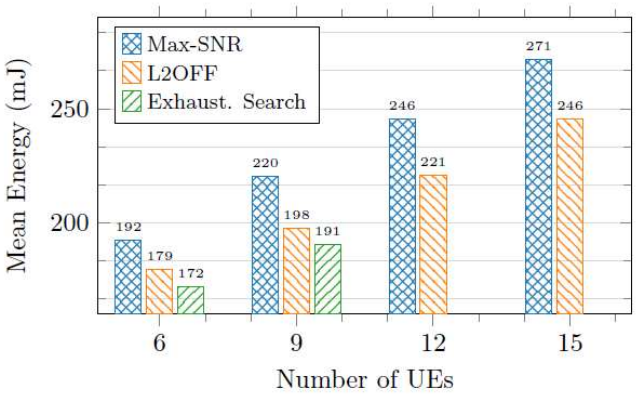

Figure 30: Average energy for fixed average delay of 100 ms

The resulting framework achieves up to 96.5% performance of the optimal strategy at 100 ms delay as highlighted by the red circle in Figure 29, while drastically reducing complexity. The proposed solution also allows to increase the network's energy efficiency compared to a benchmark heuristic approach.

6.2 Dynamic resource allocation for edge learning

In this section, we describe a dynamic algorithm developed from SAP, whose goal is to schedule and allocate the radio and computation resources in the edge cloud, in order to strike an optimal balance between energy consumption (both for communication and computation), end-to-end (E2E) service delay, and learning/inference accuracy, enabling training and inference tasks at the edge of the wireless network. The scenario of interest is composed of a set of devices collecting measurements and sending the data to an edge server that needs to take decisions about the observed data. Differently from previous works on computation offloading and edge machine learning, we assume a goal-oriented communication perspective, where the scope of the communication is not necessarily to convey all bits reliably within a given time constraint, but to send just enough data to the edge server to enable it to take decisions with the desired accuracy, striking the best trade-off between energy consumption, E2E service delay and accuracy. To achieve this goal, we dynamically act on the source encoder to adjust

the transmission rate, while still fulfilling the goal of the learning task. The idea is to tolerate a small amount of distortion on the received data, to achieve a better energy-delay trade-off, but still being able to satisfy the accuracy requirements of the learning task. In particular, we focused on two different resource allocation strategies:

- minimum energy consumption under delay and learning accuracy constraints; for this class of algorithms, we consider two different sub-classes: Model-based EML and Data-driven EML; to test this strategy, we considered an application involving estimation/prediction based on Least Mean Squares (LMS);
- best learning/inference accuracy under latency and energy constraints; in this case, there is no prior model available and the performance cannot be measured online. This is typical of some learning tasks such as classification. To test this strategy, we considered a classification problem running over two different real datasets, involving a Support Vector Machine (SVM) and a Neural Network (NN).

We considered a continuous flow of data that are generated locally by the sensor devices, and uploaded to the ES, which processes the received data running an online learning algorithm. The overall delay experienced by a data unit from its generation up to the decision (estimation or classification) taken at the ES is given by the sum of: i) the uplink queueing delay, ii) the transmission delay, iii) the queueing delay at the ES, and iv) the computation time at the ES. Our goal was to devise an optimal scheduler that allocates radio and computation resources in order to achieve the best trade-off between energy consumption, accuracy and delay. We developed algorithms based on stochastic optimization, which is a convenient model to handle situations where in every time slot there are parameters that are unknown (or only imperfectly known), such as the channel state, data arrival times, etc.

To test the algorithms, SAP considered as an application image recognition using a Support Vector Machine (SVM) algorithm that runs on the edge server, based on the images transmitted by the mobile devices. At the ES, an SVM with polynomial kernel classifies the data. We assume that each sensor has a different requirement on the energy spent for transmission, to simulate the situation in which the devices have different battery energy levels and then adapt their requirement in terms of energy consumption in order to prolong the battery lifetime.
In Figure 31, we illustrate the performance of the developed algorithms, in terms of trade-off between energy, delay, and learning accuracy. In particular, the four curves refer to four devices accessing the edge server to run an image classification task: S1, S2, and S3 refer to devices with different energy constraints, whereas S4 refers to a device with no energy constraint. The dashed line reports the average delay constraint. We can check that all devices meet the average delay constraint. Furthermore, we can see how the devices with fewer restrictions on energy constraints are able to achieve a given classification rate with a smaller average delay.

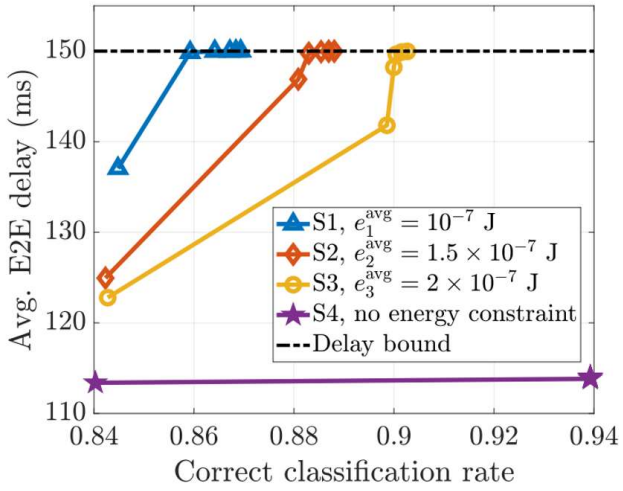

Figure 31: Average delay vs. correct classification rate, using an SVM algorithm running at the edge server, in a scenario composed of four mobile devices having different energy constraints.

As a further application, we also considered a classification task performed using a neural network running on real data associated to a hydraulic system monitoring (HSM) dataset. The results are reported in Figure 32. As expected, all devices meet the required energy and delay constraints. The device without energy constraint (green curve) shows the best accuracy, with the lowest average E2E delay.

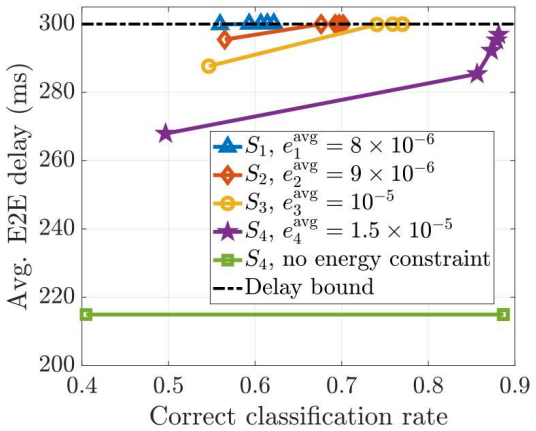

Figure 32: Average delay vs. correct classification rate, using a neural network running on a hydraulic monitoring dataset.

6.3 Conclusions

In task 4.4, CEA has collaborated with SAP on the problem of energy-efficient computation offloading enabled by edge computing in dense mmWave networks. In the considered scenario, multiple users simultaneously compete for limited radio and edge computing resources to get offloaded tasks processed under a delay constraint. The radio resource allocation takes into account inter-cell and intra-cell interference, and the duty cycles of the radio and computing equipment have to be jointly optimized to minimize the overall energy consumption. To address this issue, we formulate the underlying problem as a dynamic long-term optimization. Then, based on Lyapunov stochastic optimization tools, we decouple the formulated problem into a CPU scheduling problem and a radio resource allocation problem. Whereas the first one can be efficiently solved using a fast iterative algorithm, the second one is solved using distributed multi-agent reinforcement learning due to its non-convexity and NP-hardness. The resulting framework achieves up to 96.5% performance of the optimal strategy based on exhaustive search, while drastically reducing complexity. The proposed solution also allows to increase the network's energy efficiency compared to a benchmark heuristic approach.

In Task 4.4, SAP developed and tested algorithms for the dynamic allocation of radio and computational resources in a monitoring system where peripheral devices collect data and send them to an edge server that runs machine learning algorithms to take decisions about the observed data. The allocation is carried out in order to find an optimal balance between energy consumption, service delay, and the accuracy of the decisions taken by the edge server. Different constraints are incorporated in the method, including service delay, which

incorporates queueing delay in the communication and computation queues, and energy consumption. Furthermore, SAP has started developing methods for dynamic service placement, generalizing the methods developed in WP3 to the dynamic case.

7 Conclusions

D4.2 provides the final specification and implementation of private 5G networks building blocks. This deliverable is an extension of D4.1. These innovative components are fueling the lab integration reported in D5.2.

In task 4.1, the 5G CONNI project built a RAN system composed of CDU, RU, and CPE as RAN implementation and proposed a novel HARQ scheme for fast decision-making. The tradeoff between reliability, latency and resource efficiency was evaluated by comparing the performance of classical reactive HARQ and proactive HARQ in a system level simulator, as a function of the traffic source rate and channel conditions. In future work, scheduling strategies for URLLC are being investigated and gNodeB and CPE will be deployed in ITRI IMTC for industrial application. Development will be required to meet the requirements of the selected use case throughout the remaining project period.

In task 4.2, 5G CONNI project worked on the ETSI NFV-like instantiation and orchestration of legacy 4G and 5G mobile core network components via OSM. With the designed VNFDs, it is possible to deploy a mobile core network with off-the-shelf, standard-compliant MANO implementations such as OSM and ONAP. The framework is continuously developed and integrated in the laboratory. In order to meet the network interconnection requirements between 5G private networks deployed in Taiwan and EU, the promising and expected approach is to realize the provision activities of the customer data by a unified provisioning system through a set of well-defined management APIs. The common UE tables with the service permissions enable the feature of the seamless access applications installed in 5G private networks on both POC fields - Taiwan and EU. The service scenario of the interconnection feature detailed in this deliverable will be verified on both sides to demonstrate that the same UE profile is provisioned in two UDMs located in different networks.

In task 4.3, two implementations of MEC are proposed by 5G-CONNI for the European and Taiwanese testbeds: the hybrid 5GC solution and the bump-in-the-wire solution. For the full-on-site and the hybrid architectures, a PC Desktop server node was added to the infrastructure to simulate an edge node, with an Openstack instance installed and configured. Some preliminary configurations were applied to provide connectivity with the central server. MEC 5G SA based on a bump-in-the-wire architecture, developed handover, multi-PDU sessions and multi-QoS flows functionalities. The ECoreCloud (ECC) NFV platform and MANO also provide network connection management for VNFs. Future work includes Athonet 5GC deployment and performance tests on the complete multi-node infrastructure, with the instantiation of UPF on the edge node. Some deployment experiments will be scheduled in order to outline preliminary results, which will be particularly relevant for the final demonstrations. On the Taiwanese side, the industrial applications from ITRI IMTC will be deployed and managed on the ECC platform and MEC SA will tune the performance for the 5G CONNI project.

In Task 4.4, 5G CONNI project worked on three vertical use cases, namely (1) Process Diagnostics by CNC and Sensing Data Collection (2) Using Augmented/Virtual Reality for Process Diagnosis (3) Cloud-based CNC. Among these implemented use cases, (1) & (2) were implemented on a five-axis machine tool and (3) was implemented on a flexible fixture system, which is a specialized machine to test the cloud-based controller. 5G CONNI project also investigated the energy-efficient problem of offloading computation through edge computing in dense

mmWave networks. In the considered scenario, multiple users simultaneously compete for limited radio and edge computing resources to get offloaded tasks processed under a delay constraint. Finally, 5G CONNI project developed and tested algorithms for dynamic allocation of radio and computational resources in a monitoring system where peripheral devices collect data and send them to an edge server that runs machine learning algorithms to make decisions on the observed data. The allocation is carried out to find an optimal balance between energy consumption, service delay, and accuracy of the decisions made by the edge server. In future work, 5G CONNI project has started to develop methods for dynamic service placement, generalizing the methods developed in WP3 to the dynamic case.