Beyond 5G Multi-Tenant Private Networks Integrating Cellular, Wi-Fi, and LiFi, Powered by Artificial Intelligence and Intent Based Policy

5G-CLARITY Deliverable D2.3

Primary System Architecture Evaluation

Contractual Date of Delivery:	**June 30, 2021**
Actual Date of Delivery:	**July 31, 2021**
Editor(s):	**Anna Tzanakaki (UNIVBRIS/IASA)**
Author(s):	**Jose Ordonez-Lucena (TID), Daniel Camps-Mur (I2CAT), Alexandros Manolopoulos, Petros Georgiades, Viktoria Maria Alevizaki, Stratos Maglaris, Markos Anastasopoulos, Anna Tzanakaki (UNIVBRIS/IASA), Antonio Garcia, Kiran Chackravaram (ACC), Jonathan Prados-Garzon, Lorena Chinchilla-Romero, Pablo Ameigeiras, Pablo Muñoz (UGR), Hamada Alshaer, Anil Yesilkaya (USTRATH), Rui Bian (PLF), Tezcan Cogalan (IDCC), Ramya Vasist, Meysam Goodarzi, Jesús Gutiérrez, Vladica Sark (IHP), Mir Ghoraishi (GIGASYS)**
Work Package:	**WP2**
Target Dissemination Level:	**Public**

This document has been produced in the course of 5G-CLARITY Project. The research leading to these results received funding from the European Commission H2020 Programme under grant agreement No. H2020-871428. All information in this document is provided “as is", there is no guarantee that the information is fit for any particular purpose. The user thereof uses the information at its own risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.

Revision History

Revision	Date	Editor /Commentator	Description of Edits
0.1	11.02.2021	Mir Ghoraishi (GIGASYS)	Master document created
	15.02.2021	Anna Tzanakaki (IASA/UNIVBRIS)	ToC created
	03.03.2021	Jose Ordonez-Lucena (TID)	ToC refined and agreed contributors
0.2	05.03.2021	Anna Tzanakaki, Markos Aanstasopoulos (IASA/UNIVBRIS)	Additional per section input/description Section 3,2: First draft modelling of functional elements Section 3.3 First draft on end-to-end Modelling tools
0.3	26.03.2021	Jonathan Prados-Garzon, Lorena Chinchilla-Romero, Pablo Ameigeiras, Pablo Muñoz (UGR)	First draft of the per-component models (Section 3.2) First draft of the end-to-end models (Section 3.3) First draft use cases description (Section 4)
0.31	09.04.2021	Anna Tzanakaki, Markos Aanstasopoulos (IASA/UNIVBRIS)	First draft use cases description (Section 4) Second draft on Section 3.2
0.4	30.04.2021	Jonathan Prados-Garzon, Lorena Chinchilla-Romero, Pablo Ameigeiras & Pablo Muñoz (UGR)	Preliminary evaluation results (Section 5)
0.41	18.05.2021	Anna Tzanakaki, Markos Aanstasopoulos (IASA/UNIVBRIS)	First draft Section 2
0.5	14.05.2021	Jonathan Prados-Garzon, Lorena Chinchilla-Romero, Pablo Ameigeiras & Pablo Muñoz (UGR)	Second draft of the per-component models (Section 3.2) Second draft of the end-to-end models (Section 3.3) Second draft use cases description (Section 4)
0.51	18.05.2021	Anna Tzanakaki, Markos Aanstasopoulos (IASA/UNIVBRIS)	Third draft Section 3.2 First draft Section 3.3, Modelling of the Control Plane Functions Section 5, First draft Evaluation results
0.6	28.05.2021	Jonathan Prados-Garzon, Lorena Chinchilla-Romero, Pablo Ameigeiras, Pablo Muñoz (UGR)	Final evaluation results (Section 5)
		Anna Tzanakaki, Markos Aanstasopoulos (IASA/UNIVBRIS)	First complete version
0.7	24.06.2021	Mir Ghoraishi (GIGASYS)	Full review and edit
0.8	01.07.2021	Jose Ordonez-Lucena (TID)	D2.3 external review
0.9	14.07.2021	Daniel Camps Mur (i2CAT)	Full review
1.0	30.07.2021	Jesús Gutiérrez (IHP), Mir Ghoraishi (GIGASYS)	Final version and submission

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
5GNR	5G New Radio
5GC	5G Core
5GS	5G System
5GSM	5G Session Management
ACK	Acknowledgment
AI	Artificial Intelligence
AGV	Automated Guided Vehicle
AMF	Access and Mobility Management Function
AN	Access Network
AoA	Angle of Arrival?
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
ARP	Address Resolution Protocol
ATS	Asychronous Traffic Shaper
ATSSS	Access Traffic Steering, Switching and Splitting
AWGN	Additive White Gaussian noise
B5G	Beyong 5G
BBU	Baseband Unit
BP	???
BRF	Bayesian Recursive Filtering
CDN	Central Data Network
CLI	Command Line Interface
CP	Cyclic Prefix
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CSMA/CA	Carrier-Sense Multiple Access with Collision Avoidance
CU	Central Unit
CUPS	Control-User Plane Seperation
DCF	Distributed Coordination Function
DCI	DL Control Information
DL	DL
DL/UL-TDoA	DL/Uplink Time Difference of Arrival
DMP	Data Management Platform
DN	Data Network
DRB	Data Radio Bearer
DSCP	Differentiated Services Code Point
DU	Dicentralized Unit
E2E	End to End
eCPRI	enhanced Common Public Radio Interface
EGT	Evolutionary Game Theory
EH	Extended Header
eMBB	Enhanced Mobile Broadband
EPC	Enhanced Packet Core
EPS	Evolved Packet System
EXT-DN	External Data Network
FCFS	First-Come First-Served
FFT	Fast Fourier Transform
FG	Factor Graph
FPGA	Field Programmable Gate Array
GBR	Guaranteed Bit Rate

gNB	next-generation Node B
GOPS	Giga Operations Per Second
GPP	General Purpose Processors
GPRS	General Packet Radio Service
GTP-U	GPRS Tunnelling Protocol
GTPv2	GPRS Tunnelling Protocol version 2
HARQ	Hybrid Automatic Repeat Request
HD	High Definition
HetNet	Heterogenous wireless Network
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
HW	Hardware
ICMP	Internet Control Message Protocol
ICT	Information and Communication Technology
IE	Information Element
IFFT	Inverse Fast Fourier Transform
IoT	Internet of Things
IP	Internet Protocol
ISM	Industrial, Scientific, and Medical frequency
KPI	Key Performance Indictor
L2	Layer 2
LiFi	Light Fidelity
LLR	Logarithmic Likelihood Ratio
LoS	Line of Sight
LTE	Long Term Evolution
MAC	Media Access Control
MANO	Management and Orchestration
MA-PDU	Multi-Access PDU
MBB	Mobile Broadband
MC	Motion Control
MCS	Modulation and Coding scheme
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MME	Mobility Management Entity
mMTC	Massive Machine to Machine Communications
MN	Master Node
MQTT	Message Queuing Telemetry Transport
MTI	Measurement/simulation Time Interval
MU	Mobile User
N3IWF	Non-3GPP Interworking Function
NACK	Negative Acknowledgment
NAS	Non-access stratum
NB	NorthBound
NFV	Network Functions Virtualization
NIC	Network Interface Controller
NR	New Radio
NRF	Network Repository Function
NR-Uu	New Radio air interface
NSA	Non-Standalone
OCC	Optical Camera Communication
ODL	OpenDayLight
OF	OpenFlow
OFDM	Orthogonal Frequency Diversity Multiplexing

O-FH	Open Fronthaul
ORAN	Open Radio Access Network
OT	Operation Technology
PCF	Point Coordination Function??
PDCCH	Physical DL Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Public Data Network
PDU	Protocol Data Unit
PFCP	Packet Forward Control Packet
PHY	Physical
PI	Physical Infrastructure
PLR	Packet Loss Radio
PMF	Probability Mass Function
PNF	Phusical Network Function
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
PSA	PDU Session Anchor
PTP	Precision Time Protocol
QFI	QoS Flow Identifier
QNA	Queuing Network Analyzer
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
REST	Representational state transfer
RF	Radio Frequency
RIC	RAN Intelligent Controller
RLC	Radio Link Protocol
RMSE	Root Mean Square Error
RoE	Radio over Ethernet
RRC	Radio Resource Control
RSS	Receiver Signals Strength
RTC	Run-to-completion
RU	Remote Unit
S1AP	S1 Application Protocol??
SACK	Selective Acknowledgement
SC-FDMA	Single Carrier Frequency Division Multiple Access
SCTP	Stream Control Transmission Protocol
SCV	Squared Coefficient of Variation
SDAP	Service Data Adaptation Protocol
SDN	Software-Defined Networking
S-GW	Serving Gateway
SIM	Subscriber Information Module
SINR	Signal to Interference plus Noise Ratio
SISO	Soft-Input Soft-Output
SLA	Service Level Agreement
SM	Session Management
SMF	Session Management Function
SNPN	Standalone Non-Public Network
SNS	Softwarized Network Services
SPGWU	SGW+PGW data plane
SST	Slice/Service Type
TCP	Transmission Control Protocol
TEID	Tunnel Endpoint Identifier
TLS	Transport Layer Security

TN	Transport Network
ToF	Time of Flight
TSN	Time Sensitive Networking
TSON	Time Shared Optical Network
TWR	Two Way Ranging
UC	Use Case
UDM	Unified data management
UE	User Equipment
UE-AMBR	UE – Aggregate Maximum Bit Rate
UERANSIM	Open-source state-of-the-art 5G UE and RAN (gNodeB) implementation
UGV	Unmanned Ground Vehicle
UL	UpLink
UPF	User Plane Function
uRLLC	Ultra Reliable Low Latency Communication
V2X	Vehicle to Everything
VI	Virtual Infrarstructure
VLP	Visible Light Positioning
VM	Virtual Machine
VNF	Virtual Network Function
VR	Virtual Reality
WAT	Wireless Access Technology
Wi-Fi	Wireless Fidelity
WSP	Wireless Service Provider
X2AP	X2 Application Protocol

Executive Summary

The present deliverable provides an initial evaluation of the key features of the 5G-CLARITY system architecture reported in [2] so that its main merits and limitations can be outlined. The activities carried out in this deliverable include:

- Identification of components and features from the system architecture that will take part in the overall system evaluation.
- The modelling of selected components and features, relying on theoretical analysis adopting both analytical and numerical models.
- Definition of an evaluation plan, to specify the use case-based scenarios that will be used for the system architecture evaluation. For each scenario, this plan provides information of what the evaluation pursues and how it will be done, indicating: i) the selected components and features, together with their developed models; ii) the system level specification, by integrating individual models into end-to-end models that allows characterizing/profiling the scenario; and iii) the simulation and optimisation tools to be used for scenario evaluation.
- System architecture evaluation execution, by validating the developed end-to-end models with the selected simulation and optimisation tools. This allows assessment of 5G-CLARITY system architecture through representative use cases, indicating clear benefits with respect to the relevant state-of-the-art as well as associated trade-offs.

The outcomes from this first evaluation will be used to provide inputs to the work in WP3 and WP4, and to introduce necessary refinements in the final version of the 5G-CLARITY system architecture, to be published in the upcoming deliverable 5G-CLARITY D2.4.

1 Introduction

5G-CLARITY aims at developing a heterogeneous beyond 5G (B5G) system integrating together a variety of wireless access technologies including 5GNR, Wi-Fi and LiFi suitable for private networks. This infrastructure will be operated through Artificial Intelligence (AI)-based autonomic networking. Taking into consideration the current standardisation activities and the requirements of the services and use cases that the project aims to support, documented in 5G-CLARITY D2.1 [1], a system architecture has been proposed for 5G-CLARITY , which is reported in 5G-CLARITY [2]. The 5G-CLARITY system architecture is structured in four strata:

- **Infrastructure stratum** – including all on-premises physical network functions (PNFs), which can be wireless, transport or compute (edge and RAN compute clusters).
- **Network and application function stratum** – responsible for the 5G-CLARITY user, control and application plane functionality. This stratum includes all virtualized network and application functions that can be executed atop the 5G-CLARITY edge and RAN compute resources.
- **Management and Orchestration stratum** – responsible for the required functionality to deploy and operate the different 5G-CLARITY services (and associated resources) throughout the service lifetime, from commissioning to de-commissioning. This includes provisioning functions (for lifecycle management), monitoring functions (for data collection and processing) and other supporting functions.
- **Intelligence stratum** – hosting Machine Learning (ML) models and related policies that provide AI-driven and intent-based operation capabilities to the 5G-CLARITY solution.

In this architectural context, this deliverable provides a first preliminary evaluation of the 5G-CLARITY architecture. This evaluation aims at quantifying the benefits of the architectural features and technologies adopted in 5G-CLARITY, and also offers some benchmarking with respect to the relevant state-of-the-art solutions available thus far.

1.1 Scope and objectives of this document

This deliverable reports on the initial evaluation of the overall 5G-CLARITY architecture. The work performed in this direction includes purposely developed modelling and simulation tools, as well as initial evaluation results for a set of use cases that are relevant to the project activities and objectives. The overall methodology followed consists of:

- Definition of high-level modelling requirements. Derived by the description of the services expected to be supported and the associated Key Performance Indicators (KPIs).
- Modelling of the 5G-CLARITY architectural functional elements. These are organised according to the overall architectural structure proposed by the project, i.e., Infrastructure stratum, Network and Application Function stratum, Management and Orchestration stratum, and Intelligence stratum. The reported models rely on the development of both theoretical and simulation tools describing the performance of the corresponding elements as well as experimental profiling of specific architectural elements where this has been feasible.
- End-to-end modelling. Exploiting the functional elements' models developed and integrating these together in generic tools that can be used for the evaluation of the overall 5G-CLARITY architecture and infrastructure taking a system wide perspective. These tools take inputs from the requirements derived by the use cases defined in this deliverable and perform a Use Case-based overall architecture evaluation. This allows the assessment of the overall 5G-CLARITY solution, indicating clear benefits with respect to the relevant state-of-the-art as well as associated trade-offs.

The aim of this document is to carry out a study that will assess the enhanced, beyond 5G, features of the 5G-CLARITY architecture reported in [2], in order to identify relevant merits and limitations. In this context individual objectives targeted by this deliverable include:

- **OBJ-1:** High level requirements definition derived by the description of services expected to be supported and the associated Key Performance Indictors (KPIs) that the 5G-CLARITY architecture needs to support.
- **OBJ-2:** Identification of components and features from the overall multi-layer system architecture that need to be evaluated.
- **OBJ-3:** Definition of a detailed comprehensive and credible evaluation methodology to be followed.
- **OBJ-4:** Development of suitable models for the identified architectural functional elements relying on a mixture of theoretical and simulation tools as well as experimental profiling of specific architectural elements as appropriate.
- **OBJ-5:** Specification of evaluation scenarios that will be consider for the assessment and associated input/parameters definition.
- **OBJ-6:** System level evaluation execution integrating the developed functional element models in order to validate and benchmark the performance of the overall 5G-CLARITY architecture with respect to alternative state-of-the-art approaches.

1.2 Document structure

The rest of this document is structured as follows:

- Section 2 covers **OBJ-1**, providing an overall view of 5G-CLARITY functional scope, including supported capabilities and in-scope services, together with their KPIs.
- Section 3 covers **OBJ-2**, **OBJ-3 and OBJ-4**, reporting on the modelling of selected 5G-CLARITY architectural components relevant for this first analysis.
- Section 4 covers **OBJ-5**, focusing on end-to-end modelling tools exploiting the functional elements' models reported in section 3. These tools allow integrating the models developed in section 3 together within generic tools that can be used for the evaluation of the overall 5G-CLARITY architecture taking a system wide perspective.
- Section 5 covers **OBJ-6**, providing a Use Case-based overall architecture evaluation. This allows assessment of the overall 5G-CLARITY solution, indicating clear benefits with respect to the relevant state-of-the-art as well as associated trade-offs.
- Finally, Section 6 summarizes and concludes this document.

2 Services and KPIs

Future communication systems are expected to provide optimized support for a variety of different services, traffic loads, and end user communities. In this direction, B5G private communication networks can play a key role complementing public networks in support of multiple combinations of reliability, latency, throughput, positioning, and availability services. In this section, we provide an high level overview of the different services and requirements which are within the scope of the 5G-CLARITY ecosystem (Section 2.1). This includes a description of the functional requirements of 5G-CLARITY services (Section 2.2) and non-functional requirements (Section 2.3) used to drive the relevant modelling and evaluation studies.

2.1 The 5G-CLARITY ecosystem

B5G platforms can play an instrumental role in bringing together technology players, vendors, operators and verticals orchestrating their interaction with the aim to open up new business models and opportunities for the ICT and vertical industries and also enable cross-vertical collaborations and synergies to further enhance their value propositions. As deploying 5G solutions for vertical industries in Europe is a well-defined objective, there is a clear need to develop future-proof infrastructures to address a wide range of vertical applications. For efficiency and scalability purposes these infrastructures will have to adopt flexible architectures, offering converged services across heterogeneous technology domains deploying unified software control. In this environment, private 5G systems can play a key role complementing public 5G networks supporting the ICT and a large variety of vertical industries. In view of this need, the 5G-CLARITY project aims to position private 5G networks in the heart of the 5G vision as an integral part of the overarching 5G architecture that converges together greatly heterogeneous technologies in support of a large variety of services [55]. Typical use cases that can be supported by 5G-CLARITY include:

- Operations of Automated Guided Vehicle (AGV) supporting Industrial Internet of Things (IIoT) and Content Delivery Network (CDN) slices
- Industrial automation exploiting Wi-Fi
- Wi-Fi offloading and slicing for Ultra Reliable Low Latency Communications (URLLC) services
- Digital mobility and traffic load management in LiFi/Wi-Fi networks
- Critical services for factories
- Localization

To support these services over the 5G-CLARITY system, the functional (F) and non-functional (NF) requirements described in [2] have been used to guide the relevant evaluation studies presented in Sections 3, 4 and 5. The Functional and non-Functional requirements along with a brief description of the way these have been incorporated in the associated modelling tools is provided in Sections 2.2 and 2.3, respectively.

2.2 Functional requirements

This subsection provides a summary of the main 5G-CLARITY functional requirements that have been considered (Table 2-1) to drive the 5G-CLARITY architecture performance evaluation studies reported in this deliverable report.

Table 2-1 Functional Requirements of the 5G-CLARITY System Architecture

Requirement ID	Requirement Description and Modelling Approach
CLARITY-SYST-F-R1, CLARITY-SYST-F-R2	The 5G-CLARITY system is able to support multiple services per customer as well as services with different characteristics including long-live and short-lived communication/digital services. To evaluate this functionality, a set of carefully selected scenarios have been

	considered in the in the project including critical services (e.g., factory automation services) with very high priority and availability requirements as well as services with relaxed requirements. Some of these services may have time varying requirements but may also share common QoS characteristics such as latency and packet loss rate. For example, signalling for teleoperation/remote operation services may have stringent latency and reliability requirements. Numerical evaluation studies using theoretical and experimental tools show that the 5G-CLARITY solution can flexibly support different priority services originating from the same and different customers with guaranteed QoS
CLARITY-SYST-F-R4	The 5G-CLARITY system can provide data to the customer according to the customer's requirements (e.g. relevant data, relevant time, relevant form). 5G-CLARITY offers the means to provide the required QoS (e.g., reliability, latency, and bandwidth) for a service and the ability to prioritise resource allocation when necessary to meet service level requirements. Existing QoS and policy frameworks handle latency and improve reliability by traffic engineering. To support B5G services the 5G-CLARITY solution offers advanced tools for QoS and policy control for reliable communications satisfying the latency constraints imposed by the relevant services and enable resource adaptations as appropriate. A typical technology example that is considered and evaluated in the project is Time Sensitive Networking that can differentiate and prioritize pre-emptive and express traffic. 5G-CLARITY is also expected to operate in a heterogeneous environment including a large variety of network technologies, multiple types of UE, etc. To achieve this, a harmonised QoS and policy framework that is used to control all these different technologies is analysed and evaluated.
CLARITY-SYST-F-R6	The 5G-CLARITY system shall support integration of both new and legacy functions.5G-CLARITY complements public 5G network providing data connectivity, support of legacy services (i.e voice service continuity, seamless handovers especially under high mobility, and access to a 5G core network), interoperability between different operator networks and legacy 3GPP systems (support mobility procedures between a 5G core network and an Enhanced Packet Core (EPC) with minimum impact to the user experience, e.g., QoS, QoE). This is achieved through appropriate functions and interfaces which are also evaluated using theoretical and experimental tools.
CLARITY-SYST-F-R7	The 5G-CLARITY system shall be able to provision functions using resources of different technology domains, including the wireless (LTE/5GNR/Wi-Fi/LiFi), the wired (Ethernet/TSN) and the compute domains. Allocation of resources in these domains is achieved through mechanisms and processes that ensure service continuity. Information loss during inter- and/or intra- access technology changes are minimized (especially for URLLC services). Performance evaluation studies show that 5G-CLARITY is able to minimise the user experience impact (e.g., minimization of interruption time) under mobility and efficiently allocate resources by offloading connections from one technology to another. Reconfiguration of connections is achieved with with guaranteed QoS
CLARITY-SYST-F-R8	The 5G-CLARITY system can not only offload traffic from one technology to another but also combine resources from 3GPP technologies (i.e. LTE/5G) and non-3GPP technologies (i.e. Wi-Fi/LiFi). Integration with multiple technology interfaces allows multiple access technologies (i.e., 5GNR, LiFi, Wi-Fi) to be used simultaneously for one or more services enabling flexible traffic distribution, increased throughput and reliability as well as lower latency.
CLARITY-SYST-F-R9	The 5G-CLARITY system incorporate SDN programmability in the transport network infrastructure providing. For this type of centralized controllers emphasis is given on improving control plane efficiency minimizing the signaling required prior to user data transmission and improving user plane. E.g. mMTC services associated with sensors and monitoring UEs deployed over wide geographical areas is provided with minimum amount of signaling requirements and with varying messaging size. Minimizing signaling overhead improves control plane associated resource efficiency particularly for small data

	transmissions.
CLARITY-SYST-F-R10	The 5G-CLARITY system shall be able to provide NFV support in the compute network infrastructure, allowing the deployment and operation of some network/application functions as VNFs/VAFs, when applicable.
CLARITY-SYST-F-R11	The 5G-CLARITY system supports network slicing. Network slicing allows services from different tenants, including mobile network operators and other actors to be co-hosted over the 5G-CLARITY infrastructure. In view of this, 5G-CLARITY end-to-end slicing capabilities across network domains are modelled and evaluated. Emphasis is given on the investigation of slice isolation (investigate scenarios that operators/users create, modify, and delete slices and the impact that these actions have on disrupting traffic in current slices) and provisioning of multiple service slices over the same infrastructures. Evaluation studies also focus on the assignment of multiple UEs equipped with multi-technology interfaces to applications hosted in slice, and movement of these applications from one slice to another. Creation of infrastructure slices with specific constraints in terms of capacity, users, geographic coverage, availability etc. are also investigated.
CLARITY-SYST-F-R14	The 5G-CLARITY support functionality that enables collecting and storing up-to-date data. Evaluation studies are focusing on analysing the scalability of the data management system and the processing requirements of its individual building blocks
CLARITY-SYST-F-R16	The 5G-CLARITY system shall support the capability to ensure availability of data, resources, functions and services. Evaluation studies have been focused on analysing migration mechanisms that can be used to push content to the edge cluster reducing end-to-end delay and improved user experience, enhancing availability and reliability

2.3 Non-functional requirements

Non-Functional requirements – specify quality attributes of the 5G-CLARITY system. These requirements define the properties that the functions must have, such as performance, usability, and data security needs. Key attributes that are taken into account during the design phase of the 5G-CLARITY system (either as constraints or targe design objectives) include:

- ***Data rates and traffic densities:*** The high capacity provided by 5G-CLARITY can be used to support scenarios requesting very high data rates or traffic densities (**CLARITY-SYST-NF-R8)**. The scenarios address different service areas: urban, office and factories, and special deployments (e.g., massive gatherings, broadcast, residential, and AGVs).
- ***Low latency and high reliability :*** Several scenarios require the support of very low latency and very high communications service availability implying a very high degree of reliability (**CLARITY-SYST-NF-R2)**. The overall service latency depends on the delay at the radio interface, the transmission within the 5G system, the transmission to a server which may be external to the 5G system, and the data processing. Some of these factors depend directly on the 5G system itself, whereas for others the impact can be reduced by suitable interconnections between the 5G system and services or servers outside of the 5G system, for example, to allow local hosting of the services (**CLARITY-SYST-NF-R3)**.

  Some scenarios requiring very low latency and very high communication service availability that can be supported by 5G-CLARITY are described below (**CLARITY-SYST-NF-R2)**:

  - Discrete automation in the Industry 4.0 Pilot– Discrete automation is characterised by high requirements on the communications system regarding reliability and availability (**CLARITY-SYST-NF-R5)**. Systems supporting discrete automation are usually deployed in geographically limited areas, access to them may be limited to authorised users, and they may be isolated from networks or network resources used by other cellular

customers.

- Process automation – Automation for (reactive) flows in intralogistics processes. Process automation is characterized by high requirements on the communications system regarding communication service availability. Systems supporting process automation are usually deployed in geographically limited areas, access to them is usually limited to authorised users, and served by private networks (**CLARITY-SYST-NF-R6)**.

- ***High accuracy positioning:*** High accuracy positioning is characterized by ambitious system requirements for positioning accuracy. UEs should be able to share positioning information between each other e.g., to a controller if the location information cannot be processed or used locally. Tight integration between terrestrial MEC and 5G-CLARITY will enhance positioning accuracy especially in scenarios with limited line-of-sight (LoS) connectivity.

3 Key Enablers for 5G-CLARITY Architecture Evaluation

5G-CLARITY brings forward the design of a system that addresses the wide variety of challenges identified today in private network environments, including spectrum flexibility, delivery of critical services, integration with public network infrastructures, and automated (AI-driven) and simplified (intent-based) network management with built-in slicing. The creation of simplicity out of this complex capability set requires to apply the principles of abstraction and separation of concern into the 5G-CLARITY system architecture design, as explained in 5G-CLARITY D2.2 [2]. The result is an architecture structured into different strata that can evolve independently of each other.

The initial design of 5G-CLARITY system is architected into four strata with segregated scope and different technology pace each:

- **Infrastructure stratum** – it is formed by all the on-premise hardware and software resources building up the 5G-CLARITY substrate, including user equipment and a wide variety of compute, storage and networking fabric.
- **Network and Application function stratum** – it conveys the 5G-CLARITY user, control and application plane functionality. This stratum includes all virtualized network and application functions that can be executed atop the 5G-CLARITY cloud infrastructure.
- **Management and Orchestration stratum** – it encompasses all the necessary functionality to deploy and operate the different 5G-CLARITY services (and associated resources) throughout their lifetime, from their commissioning to their de-commissioning. This includes provisioning functions (for lifecycle management), monitoring functions (for data collection and processing) and other supporting functions.
- **Intelligence stratum** – it hosts the Machine Learning (ML) models and related policies which provide AI-driven and intent-based operation capabilities to the overall 5G-CLARITY strata. This stratum allows providing usage simplicity and zero-touch experience for 5G-CLARITY system consumers, especially Operation Technology (OT) actors (e.g., industry verticals), facilitating their access to the system behaviour for Service Level Agreement (SLA) assurance purposes.

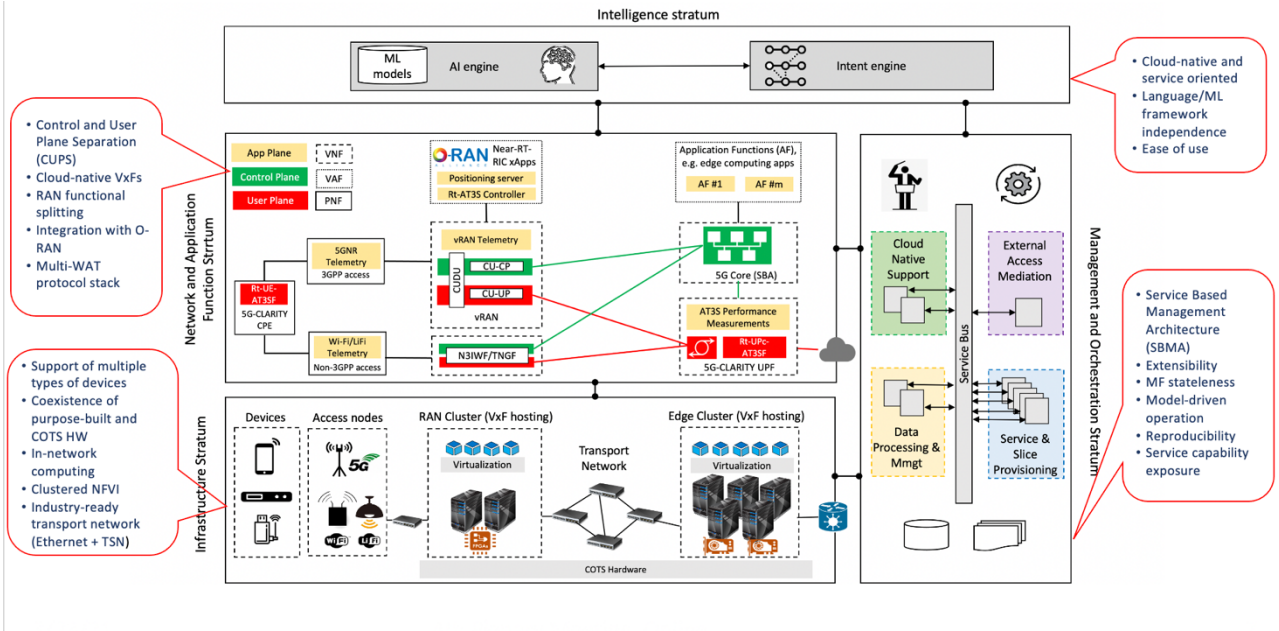


Figure 3-1 5G-CLARITY **system architecture**

Figure 3-1 illustrates the logical arrangement of the four strata into the 5G-CLARITY system architecture, including details on their individual design principles. For further details on these principles, see 5G-CLARITY D2.2, Section 4.

For this deliverable, no modifications and/or refinements are foreseen regarding the 5G-CLARITY system architecture. Once the outcomes from 5G-CLARITY D3.2 [3] and 5G-CLARITY D4.2 are available and based on the conclusions extracted from the present deliverable, a second (refined) version of the 5G-CLARITY system architecture will be provided in 5G-CLARITY D2.4.

3.1 Modelling of functional elements

In this subsection, the modelling of the functional elements of the 5G-CLARITY architecture is introduced. This is organized on a per stratum basis.

3.1.1 Infrastructure stratum

Section 3.1.1.1 introduces queuing theory-based models for evaluating two performance metrics: the response time of LiFi Access Point (AP) in the data plane and the flow setup time of SDN controller on the southbound interface. These are parts of the software-defined networking (SDN)-enabled LiFi/Wi-Fi/5G heterogenous wireless network shown in Figure 3-2.

3.1.1.1 Heterogenous wireless network modelling

An intelligent heterogenous wireless network (HetNet) control plane can support efficient service provisioning and data communications in the SDN-enabled LiFi/Wi-Fi/5G integrated network. The performance evaluation of the AP response time and the SDN controller flow setup time provide some insights regarding the network parameters and the user service requirements which should be controlled and managed by a user mobility and traffic engineering (TE) scheme. This provides a research ground to the work developed in D 3.2 to support dynamic downlink flows routing to APs and differentiated granular services across the data plane of LiFi/Wi-Fi/5G integrated network.

A LiFi AP is part of the SDN-enabled LiFi/Wi-Fi/5G HetNet architecture shown in Figure 3-2 [39]. An SDN controller has horizontal (i.e., east, west) and vertical (i.e., northbound, southbound) interfaces. A buffer on the northbound interface maintains and passes the requests generated from the SDN applications to access the data plane. Likewise, a buffer on the southbound interface passes the rule packets (i.e., short packets containing SDN rules) and incoming packets to the controller and switches, respectively, as shown in Figure 3-2. The APs in the network data plane are assumed to be OpenFlow (OF) enabled switch, which handles traffic flows according to the OpenFlow (OF) protocol [52]. Every time a traffic flow arrives at a LiFi AP or any other AP in the data plane, the OF protocol checks if their corresponding forwarding rule is already installed in the AP. If their OpenFlow rule is available, the AP immediately processes it according to their requested service. Otherwise, a rule packet, as part of the OF messages, is sent to the SDN controller to install an appropriate forwarding rule to the traffic flow in the AP. The difference between their arrival time at the AP and OF rules handling time is called the flow setup time (delay) on the southbound interface of SDN controller. The AP service response time in the data plane is defined as the elapsed of time between the time where the flows completed their OpenFlow rules handling and starting time to receive services from the LiFi AP.

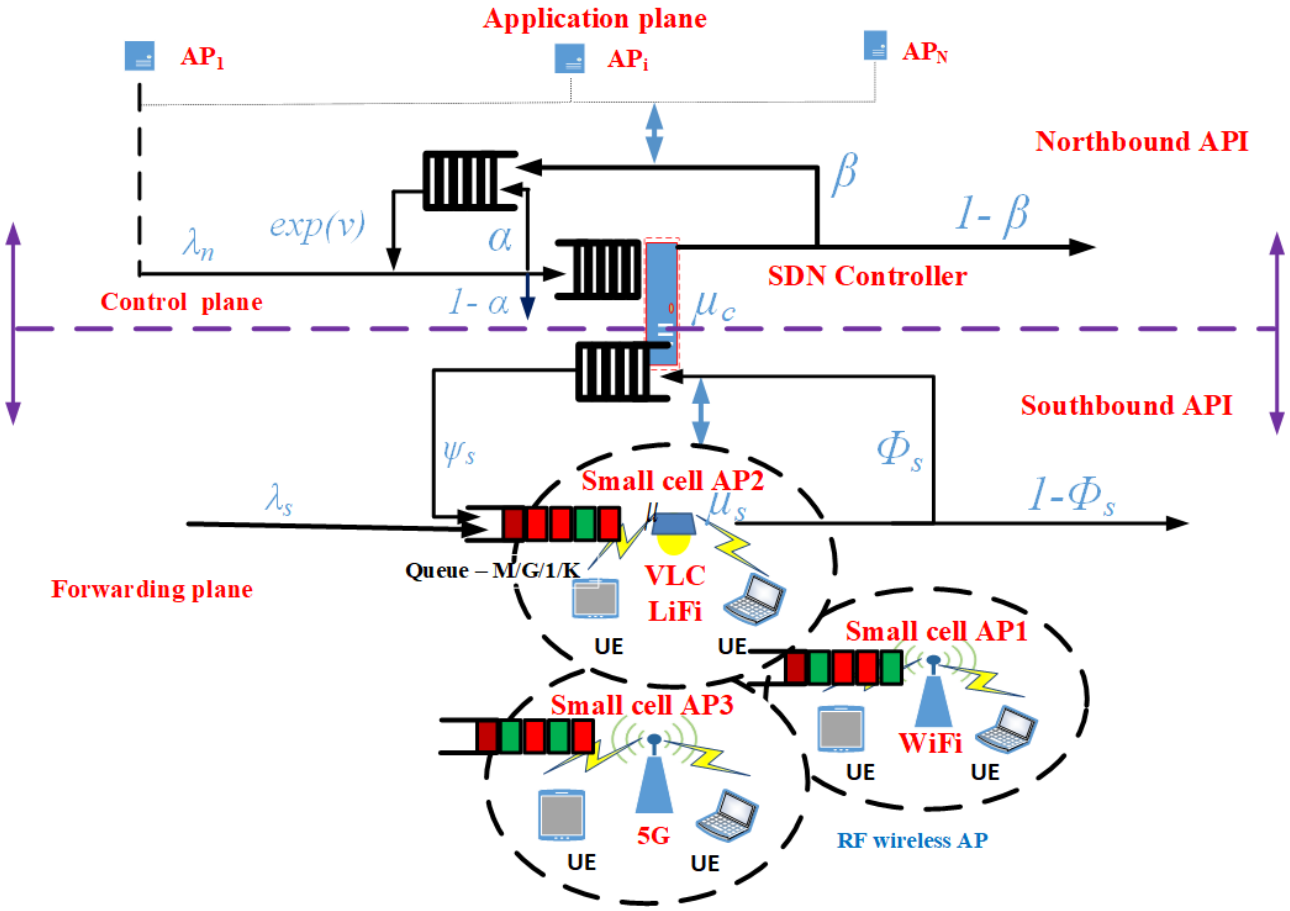


Figure 3-2 SDN-enabled HetNet architecture and applications convergence modelling

In the scenario of Figure 3-2 μ_c denotes the SDN controller service time; λ_c denotes the application requests arrival rate; $\rho_c = \frac{\lambda_c}{\mu_c}$ denotes the traffic load intensity at the controller. λ_n denotes the arrival rate of applications requests sent through the northbound interface. An application that requires a single service leaves the SDN controller with a probability β. If it requires further services like an increase of bandwidth or other resources, it re-joins the retrial queue with a probability $1 - \beta$. μ_s denotes the AP service rate; λ_s denotes the external traffic arrival rate. $\rho_s = \frac{\lambda_s}{\mu_s}$ denotes the traffic load intensity on the southbound interface of AP. If a traffic flow does not have a rule set in the AP switch, a rule packet is sent to the SDN controller, with a probability, ϕ_s, to define its forwarding rule in the AP. Otherwise, it has an already rules in the AP, which is served with a probability $1 - \phi_s$.

3.1.1.1.1 Evaluation methodology

The method used for evaluating SDN enabled LiFi AP is queuing theory-based mathematical modelling for the response time of APs in the data plane, and the flow setup latency on the southbound interface of SDN controller.

Based on Kendall's notation [50] the downlink (DL) channel service time of LiFi AP is assumed to have a general service time distribution, as the LiFi AP DL channel varies in time and space. The arrivals of UEs and DL flows follow a Markovian Poisson distribution. A LiFi AP serves multiple UEs, and each AP has a buffer size of K packets. So, the DL channel of LiFi AP is modelled by an M/G/1/K queuing system model (G stands for general service time distribution and M indicates the Markovian Poisson distribution used in the model). This model provides performance evaluation for the buffering process of packets at the queue of southbound interface and the UE access to the channel resources of LiFi AP in the

data plane. The average delay of packets served in the LiFi AP is evaluated analytically using the following developed formula [39].

$$T_s = \frac{\rho_s+\rho_s^{K+1}(1+K\rho_s-K)}{\lambda_s(1-\rho_s^{K+1})(1-\rho_s)} \qquad (1)$$

where the service times are assumed to follow an exponential distribution with an average of $\frac{1}{\mu_s}$. Since the SDN controller manages N APs in the data plane, the queuing model, M/G/1/K/N, is proposed to investigate the impact of the number of APs in the data plane and the traffic flow rate on the southbound interface of SDN controller on the flow set up time.

The total traffic load rate at the southbound interface of controller follows a Poisson distribution with a traffic arrival rate, expressed as: $\lambda_c = \lambda_n + \phi_s\lambda_s$. A relationship is established between this parameter, buffer size and the packet service response time (delay), which allows to derive the flow set up time, as follows:

$$T_c = \frac{1}{\lambda_c} - \frac{(N-K)\mu_c P_K}{1-P_o} \qquad (2)$$

where P_o, P_K can be found, respectively, in [39], Eq. (5) and Eq. (6).

3.1.1.1.2 Numerical results

Based on Eq. (1), as shown in Figure 3-3, the APs continue to serve flows with minimal response time, irrespective of the buffer size until the traffic load intensity exceeds a specific value. For example, after the load intensity exceeds 0.6, the impact of buffer size starts to become clearer on the AP response time.

Based on Eq. (2), the flow setup time for traffic flows, that arrive at the APs without pre-assigned forwarding rules, increases in terms of the rule packet requests sent to the controller, as shown in Figure 3-4. This becomes more obvious when the number of APs exceeds a critical value. For example, when the number of APs exceeds 200, the flow set up time starts to grow rapidly. This means that the controller needs an effective mechanism that can proactively assign flow rules in the APs for traffic flows. This significantly reduces the controller setup time, which shortens the sojourn time of traffic packets arriving at the APs.

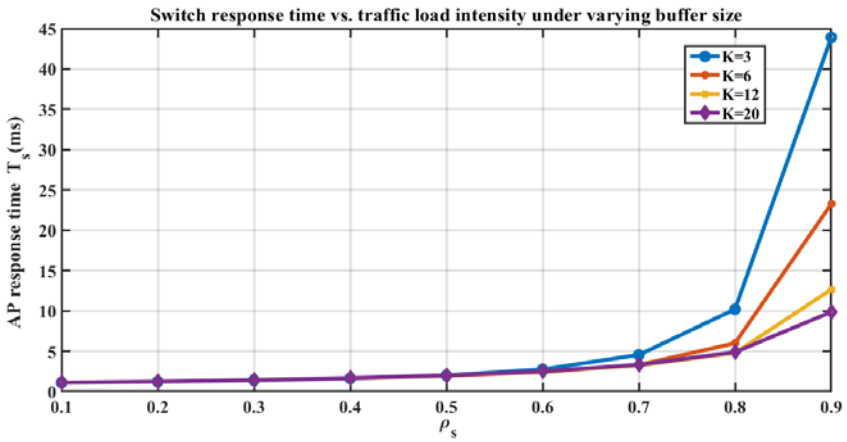


Figure 3-3 AP response time versus traffic load intensity

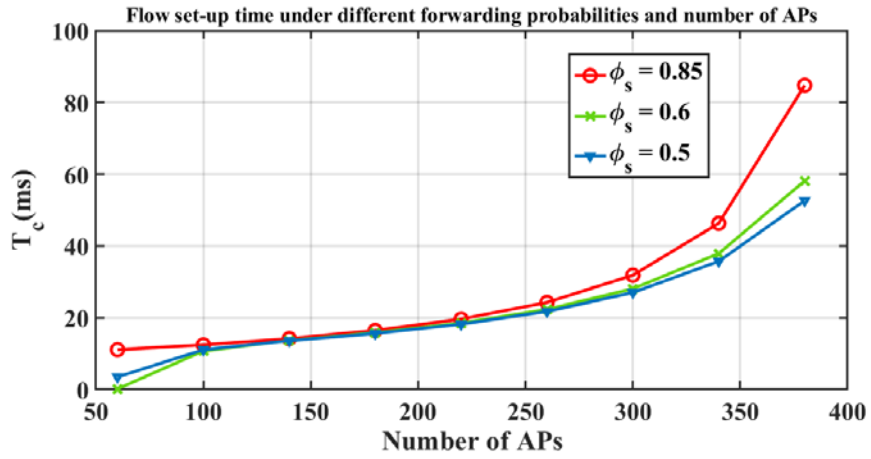


Figure 3-4 AP response time versus traffic load intensity

3.1.1.2 Asynchronous TSN bridge´s output-port packet delay model

The 5G-CLARITY project considers two layer-2 (L2) technologies to realize the transport network segments: i) standard Ethernet and ii) Time-Sensitive Networking (TSN). The former is suitable to convey best-effort traffic, whereas the latter offers deterministic QoS support suitable for critical private services. This section addresses the mean delay model for every output port of a TSN bridge, i.e., a L2 switching network device that conforms to the mandatory or optional features defined in TSN standards. Please refer to Section 5.4 in [2] for further details on the TSN nodes. The building block of the asynchronous TSN bridge is based on the Asynchronous Traffic Shaper (ATS). An ATS is allocated at each egress port of a TSN bridge handling frame transmissions at the respective physical link. The ATS consists of two queuing stages: i) the interleaved shaping which is responsible for performing a traffic regulation per flow in a cost-effective way, and ii) a set of strict priority queues.

Here, we provide a model for estimating the mean delay experienced by a given packet at the output port of an ATS. To that end, we use the non-pre-emptive multi-priority M/G/1 model from queuing theory. Then, the mean sojourn time at the output port of a TSN bridge can be estimated as [4]:

$$T_p = \frac{\sum_{i=1}^{P} \lambda_i \cdot E[s^2]}{2 \cdot (1 - \rho_1 - \cdots - \rho_{p-1}) \cdot (1 - \rho_1 - \cdots - \rho_p)} + \frac{E[l_p]}{C}$$

where:

1 and P are the highest and lowest priority, respectively.

$E[s^2]$ is the second order moment of the service time (link packet transmission time). It is mainly given by the packet length distribution, but it is also affected by deviations in the nominal transmission capacity of the link.

λ_i is the aggregated mean packet arrival rate at the priority level i.

$E[l_p]$ is the mean packet size at the priority level p.

C: *Link capacity.*

$\rho_i = \frac{E[l_i] \cdot \lambda_i}{C}$ is the link utilization.

3.1.1.3 Wi-Fi radio interface throughput for eMBB services

3.1.1.3.1 Performance metric: throughput for eMBB services in Wi-Fi air interface

This measurement provides the rate of the packets that have been successfully delivered over the Wi-Fi air interface. This metric is used to evaluate the performance of eMBB applications that require a certain Guaranteed Bit Rate (GBR).

3.1.1.3.2 Evaluation methodology

The throughput targeted for UEs requesting eMBB services through Wi-Fi technology can be approximated by an attenuated form of the Shannon´s formula, which provides the maximum theoretical throughput that can be achieved over an AWGN channel for a given SINR. This attenuated form of the Shannon's capacity formula takes into consideration the physical and MAC layer efficiency of the Wi-Fi technology.

A RAN simulator developed in MATLAB will be used to measure the SINR used as input of the throughput model mentioned above.

The considered expression to calculate the achievable rate of user j associated to the Wi-Fi access point i during the reception of the data frame is obtained based on [5][6]:

$$\beta_{i,j} = B \cdot \log_2\left(1 + \frac{SINR_{i,j}}{\eta_{phy}}\right) \cdot \eta_{mac}$$

where:

B: The system bandwidth of the Wi-Fi APs.

η_{phy}: Physical layer efficiency of the Wi-Fi system which is determined by the efficiency of the practical modulation and coding scheme. In [5] the value of this parameter is set to 1.25.

η_{mac}: System efficiency at the MAC layer of the Wi-Fi network due to the CSMA/CA and Distributed Coordination Function (DCF) mechanism.

The term of the SINR is calculated as follows:

$$SINR_{i,j} = \frac{P_{RX_{i,j}}}{I_{i,j} + N_O}$$

where:

$P_{RX_{i,j}}$ stands for the power the user j receives from AP i,

N_O is the noise that gathers the noise power and the user noise figure,

$I_{i,j}$ represents the interferences that user j receives from other APs that are transmitting within the same contention domain when it is associated to AP i.

Specifically, this term is calculated as the addition of the power that user j receives from the rest of APs that transmit in the same channel as the AP the users is attached to. This addition is weighted with a factor that indicates the level of overlapping between the channels. If m is the channel assigned to the AP user j is attached to and n is the channel assigned to the AP that is interfering AP i, then the mentioned factor indicates the level of overlapping between the channels m and n.

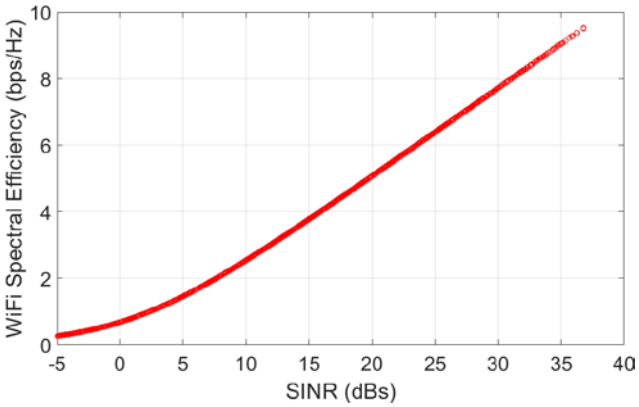


Figure 3-5 Wi-Fi Spectral efficiency versus SINR

The computation of this factor is obtained from the work in [7] and its expression is included below:

$$factor_{m,n} = \max\left(0, 1 - \frac{1}{5}|m-n|\right)$$

Notice that the values of this factor may range from 0 to 1, being 0 when there are no interferences between channels, and 1 whether the channels are entirely overlapped.

3.1.1.3.3 Numerical results

Figure 3-5 shows the spectral efficiency in Wi-Fi as a function of the SINR, which was estimated using the model included in this table. The bandwidth to be allocated to a user for guaranteeing a bit rate equals its spectral efficiency, which is given by its SINR as shown in the figure, times the bit rate to be enforced.

3.1.2 Network and application function stratum

3.1.2.1 Virtualized UPF´s mean packet delay

This section describes a queuing theory based analytical model for estimating the mean packet delay for a virtualized UPF running on a standard server. Specifically, the UPF's mean packet delay is defined as the average (arithmetic mean) sojourn time of the packet in the virtualized UPF. This metric comprises the following delay components:

a) The average back-end driver processing time and the packet transmission to the virtualization container (e.g., Virtual Machine or OS Container) through the virtual bridge, T_{BEU}^{UPF}.
b) The average protocol stack packet processing, T_{PS}^{UPF}.
c) The average queuing delay at the application layers, W_{APP}^{UPF}.
d) The average processing delay at the application layers, T_{APP}^{UPF}.
e) The back-end driver processing time and the packet transmission to the physical NIC through the virtual, T_{BED}^{UPF}.

The components a), b), and e) of the UPF's average sojourn time together with the mean and coefficient of variation of the UPF application layers' service rate will be taken from references that include the respective experimental results. The components c) and d) will be determined using queuing theory and, more precisely, the approximation of the G/G/m queue considered in [20][21]. The m queuing servers stands for the physical CPU cores allocated to the UPF's higher layers processing.

The primary assumptions considered here are extracted from the virtualized UPF's implementation operation from Intel and SK Telecom described in [22] and are listed below:

- The higher layers (e.g., GTP-U and PDU layer) processing represents the main bottleneck of the

UPF[21][22].

- The application layer serves the packets following a first-come-first-served (FCFS) discipline [21][22].
- There are as many processing threads as dedicated physical cores allocated to the UPF [21][22].
- Software-based with run-to-completion (RTC) UPF pipeline, i.e., each packet's user plane processing is executed in entirety, followed by the next packet picked for processing [21][22].
- There are CPU physical cores dedicated to the virtualisation container housekeeping [21][22]. There are m physical CPU cores destined to process packets at the UPF application layer.
- General distributions for both the arrival and service processes.

Considering all above, the UPF's mean sojourn time D_{UPF} is given by:

$$D_{UPF} = T_{BEU}^{UPF} + T_{PS}^{UPF} + W_{APP}^{UPF} + T_{APP}^{UPF} + T_{BED}^{UPF}$$

$$W_{APP}^{UPF} + T_{APP}^{UPF} = \frac{(c_a^2 + c_s^2) \cdot E_C(m, \lambda, \mu)}{2 \cdot (m \cdot \mu - \lambda)} + \frac{1}{\mu}$$

where:

c_a^2: squared coefficient of variation of the inter-arrival times to the UPF.

c_s^2: squared coefficient of variation of the UPF application layer's service times.

λ: UPF application layer mean arrival rate.

μ: UPF application layer mean service rate.

m: Physical CPU cores destined to process packets at the UPF application layer.

$E_C(m, \lambda, \mu)$: Erlang-C formula [21][22].

The accuracy of the model to estimate the average delay of a VNF was validated experimentally in [21].

3.1.2.2 Virtualized gNB-CU´s mean packet delay

This section provides a queuing theory-based approach for estimating the mean packet dealy for a virtualized gNB-CU, i.e., the average (arithmetic mean) sojourn time of the packet in the virtualized gNB-CU. This metric has similar latency components as those considered in the UPF's mean packet delay model described in Section 3.1.2.1:

a) The average back-end driver processing time and the packet transmission to the virtualisation container (e.g., Virtual Machine or OS Container) through the virtual bridge, T_{BEU}^{CU}.
b) The average protocol stack packet processing, T_{PS}^{CU}.
c) The average queuing delay at the application layers, W_{APP}^{CU}.
d) The average processing delay at the application layers, T_{APP}^{CU}.
e) The back-end driver processing time and the packet transmission to the physical NIC through the virtual, T_{BED}^{CU}.

The primary assumptions considered are listed below:

- The gNB upper layers processing represents the main bottleneck of the gNB-CU [21][22][23]. We consider the option 2 for the splitting of the gNB-CU/gNB-DU, which implies the gNB-CU is in charge of the per packet processing associated with the Radio Resource Control (RRC), Service Data Adaptation Protocol (SDAP), and Packet Data Convergence Protocol (PDCP) protocols.
- The packets are served following a FCFS discipline [21][22][23] .
- There are as many processing threads as dedicated physical cores allocated to the gNB-CU

[21][22][23].

- Software-based with run-to-completion (RTC) gNB-CU pipeline, i.e., each packet's user plane processing is executed in entirety, followed by the next packet picked for processing[21][22][23].
- There are CPU physical cores dedicated to the virtualisation container housekeeping [21][22][23]. There are m physical CPU cores destined to process packets at the gNB-CU application layer[21][22][23].
- General distributions for both the arrival and service processes.

The operation considered for the virtualized gNB-CU is the same as the virtualized UPF's implementation described in [22] and compatible with the one assumed in [23] for the Cloud RAN's BBU pool.

Considering all above, the gNB-CU's mean sojourn time D_{CU} can be estimated as follows:

$$D_{CU} = T_{BEU}^{CU} + T_{PS}^{CU} + W_{APP}^{CU} + T_{APP}^{CU} + T_{BED}^{CU}$$

$$W_{APP}^{CU} + T_{APP}^{CU} = \frac{(c_a^2 + c_s^2) \cdot E_C(m, \lambda, \mu)}{2 \cdot (m \cdot \mu - \lambda)} + \frac{1}{\mu}$$

where:

λ: gNB-CU mean packet arrival rate,

μ: gNB higher-layers protocols (e.g., SDAP and PDCP) packet processing rate. This input parameter of the model depends on the carrier bandwidth, spectral efficiency, and traffic load [29][30],

c_a^2: squared coefficient of variation of the inter-arrival times to the gNB-CU,

c_s^2: squared coefficient of variation of the gNB higher-layers protocols processing times. This input parameter might depend on the signal-to-interference-plus-noise ratio (SINR) distribution of the specific scenario, the physical machine configuration (e.g., CPU governor, C-States, processor architecture and operation), and the virtualization layer,

m: Physical CPU cores and the respective threads dedicated for the per packet processing of the SDAP and PDCP protocols,

$E_C(m, \lambda, \mu)$: Erlang-C formula [20][21].

3.1.2.3 gNB-DU mean packet delay

This metric refers to the average sojourn time of a packet in the gNB-DU. We assume the splitting option #2 [4] for the gNB-CU/gNB-DU and the splitting option #7 [4] for the gNB-DU/gNB-RU. Then, the radio link control (RLC), MAC, and part of the physical layer (e.g., equalization and MIMO precoding) are in the gNB-DU. We also consider that the gNB-DU function is not virtualized. A G/G/m queueing model is used to estimate the gNB-DU mean packet delay. Thus, the gNB-DU mean packet delay D_{DU} can be computed as:

$$D_{DU} = \frac{(c_a^2 + c_s^2) \cdot E_C(m, \lambda, \mu)}{2 \cdot (m \cdot \mu - \lambda)} + \frac{1}{\mu}$$

where:

λ: gNB-DU mean packet arrival rate,

μ: gNB-DU packet processing rate associated with the RLC, MAC and part of the physical layer. This input parameter of the model depends on the carrier bandwidth and modulation and coding scheme (MCS) index [29][30][31],

c_a^2: squared coefficient of variation of the inter-arrival times to the gNB-CU,

c_s^2: squared coefficient of variation of the gNB higher-layers protocols processing times. This input parameter might depend on the computing capacity drift of the small cell and SINR distribution of the specific scenario,

m: dedicated processing units that can process packets in parallel,

$E_C(m, \lambda, \mu)$: Erlang-C formula.

3.1.2.4 gNB-RU mean packet delay

This metric corresponds to the average sojourn time of a packet in the gNB-RU. We assume the splitting option #7 for the gNB-DU/gNB-RU. Then, FFF/IFFT, resource mapping and RF functionalities resides in the gNB-RU [4][29]. We also consider that the gNB-RU function is not virtualized. A G/G/m queueing model is used to estimate the gNB-RU mean packet delay. Thus, the gNB-RU mean packet delay estimation D_{RU} is given by:

$$D_{RU} = \frac{(c_a^2 + c_s^2) \cdot E_C(m, \lambda, \mu)}{2 \cdot (m \cdot \mu - \lambda)} + \frac{1}{\mu}$$

where:

λ: gNB-RU mean packet arrival rate,

μ: gNB-RU packet processing rate associated with the base processing of the physical layer. This input parameter of the model depends on the carrier bandwidth and the virtualization layer when the function is virtualized [4][29] [30],

c_a^2: squared coefficient of variation of the inter-arrival times to the gNB-CU,

c_s^2: squared coefficient of variation of the gNB base processing time. This input parameter might depend on the computing capacity drift of the small cell and SINR distribution of the specific scenario,

m: dedicated processing units that can process packets in parallel,

$E_C(m, \lambda, \mu)$: Erlang-C formula [20][21].

3.1.2.5 NR-Uu mean packet delay

This metric refers to the average sojourn time of a packet at the radio interface. We consider G/G/m queueing model to estimate it. Then, the NR-Uu mean packet delay D_{NR-Uu} can be computed as:

$$D_{NR-Uu} = \frac{c_a^2 \cdot E_C(m, \lambda, \mu)}{2 \cdot (m \cdot \mu - \lambda)} + \frac{1}{\mu}$$

where:

λ: NR-Uu mean packet arrival rate.

μ: packet transmission rate at the radio interface. It is the inverse of the time slot τ, which is given by the specific numerology considered.

c_a^2: squared coefficient of variation of the inter-arrival times to the NR-Uu.

The squared coefficient of variation of the radio interface transmission time is roughly zero (the service process can be regarded as deterministic).

m: the number of PRBs at the radio interface divided by the average number of PRBs required to transmit a packet. More precisely, $m = round\left(\frac{W}{b}\right)$, where W is the number of PRBs and b is the

average number of PRBs required for transmitting a single packet at the radio interface. The latter depends on the per user SINR distribution of the specific scenario. The same approach for configuring the number of queuing servers is used in [31] using an M/M/c/K model to investigate how the latency threshold as well as other system parameters (e.g., bandwidth) impact on the capacity of the URLLC system. The model [31] is used as baseline to estimate the PLR at the radio interface in [18].

$E_C(m, \lambda, \mu)$: Erlang-C formula [20][21].

3.1.2.6 NR-Uu interface packet loss ratio for URLLC services

3.1.2.6.1 Performance metric: packet loss ratio (PLR) for URLLC services

This measurement provides the percentage of packet loss at the air interface, i.e., the fraction of packets that cannot be delivered over the total number of packets sent through the radio interface. Specifically, this performance metric is used to measure the packet loss ratio of URLLC applications that must fulfil stringent delay requirements.

3.1.2.6.2 Evaluation methodology

To estimate this metric, the Queueing Theory-based model proposed [18]. The main assumptions of the model are the following:

- The NR-Uu interface arrival and service processes are Poissonian.
- There is a single buffer with FIFO discipline at the radio interface to store the aggregated traffic from all the URLLC streams of the same slice.
- There are W Physical Resource Blocks (PRBs) dedicated to the URLLC slice.
- The PRB demand per packet follows an arbitrary distribution.
- The URLLC flows have two performance constraints, e.g., a maximum packet delay budget and a packet loss ratio, which are ensured all the time.

The primary notation considered for the model is included below:

- τ: Time slot duration.
- W: Number of PRB dedicated to the URLLC slice.
- λ: Average arrival rate expressed in packets per second.
- D^{QoS}: Maximum delay budget at the NR-Uu interface. This constraint is imposed by the URLLC flow with the most stringent delay requirement.
- PLR^{QoS}: Target packet loss ratio at the NR-Uu interface. This constraint is imposed by the URLLC flow with the most stringent packet loss ratio requisite.
- $b_i = \left\lceil \frac{s}{r_i} \right\rceil$: Number of PRBs required to serve a packet of size s of the UE i with data rate r_i.
- b_{min}: Minimum possible value of PRBs to serve a packet given by the PRBs demand per packet distribution.
- b_{max}: Maximum possible value of PRBs to serve a packet given by the PRBs demand per packet distribution.
- $k_{max}(b) = \left\lfloor \frac{b}{b_{min}} \right\rfloor$: Maximum number of packets that have aggregated size of b PRBs.
- $k_{min}(b) = \left\lceil \frac{b}{b_{max}} \right\rceil$: Minimum number of packets that have aggregated size of b PRBs.
- $p_{B_i}(b_i)$: It stands for the number of PRBs needed to transmit a packet addressed to the UE i.

Due to the independence among the UEs packets arrival at the gNB,

the Probability Mass Function (PMF) of the number of PRBs needed to serve an arbitrary selected packet can be expressed as follows:

$$p_b(b) = \frac{\sum_{i=1}^{N} \lambda_i \cdot p_{B_i}(b_i)}{\lambda}$$

Then, the packet loss ratio at the NR-Uu interface can be estimated as:

$$1 - \mathrm{P_c}\left(\mathrm{W}, D^{\mathrm{QoS}}, \lambda\right) = \; 1 \; - \sum_{b=b_{min}}^{L_{max}} \sum_{k=k_{min}(b)}^{k_{max}(b)} \frac{p(k,\ \lambda\tau) \cdot p_B^{(k)}(b)}{1 - p(0,\ \lambda\tau)}$$

where:

$p_B^{(k)}(b) = \sum_{i=b_{min}}^{b_{max}} p_B^{(k-1)}(b-i) \cdot p_B^{(k)}(i)$: PMF of the aggregated size of k > 1 packets.

$L_{max} = W \left\lfloor \frac{D^{QoS}}{\tau} - 1 \right\rfloor$: The maximum queue length in terms of PRBs. Enforcing this condition ensures that the most stringent delay constraint D^{QoS} is fulfilled.

$p(k,\ \lambda\tau) = \frac{(\lambda\tau)^k}{k!} \exp(-\lambda\tau)$: PMF of the Poisson distribution.

3.1.2.7 NR-Uu interface throughput for eMBB services

3.1.2.7.1 Performance metric: throughput for eMBB services

This measurement provides the rate of the packets that have been successfully delivered over the NR-Uu interface. This metric is used to measure the performance of eMBB applications that require a certain Guaranteed Bit Rate (GBR).

3.1.2.7.2 Evaluation methodology

The throughput targeted for eMBB services can be approximated by an attenuated and truncated form of the Shannon´s formula, which provides the maximum theoretical throughput that can be achieved over an AWGN channel for a given SINR. A system-level RAN simulator developed within the 5G-CLARITY project using MATLAB will be used to measure the SINR used as input of the throughput model mentioned above.

The following equations are used to approximate the throughput over a channel with a given SINR [18]:

$$\text{Throughput (SINR)}, \left(\frac{\text{bps}}{\text{Hz}}\right) = \begin{cases} 0 & \text{for } SINR \leq SINR_{MAX} \\ \alpha \cdot S(SINR) & \text{for } SINR_{MIN} \leq SINR \leq SINR_{MAX} \\ \alpha \cdot S(SINR) & \text{for } SINR \geq SINR_{MIN} \end{cases}$$

where:

$S(SINR)$: Shannon bound, $S(SINR) = log_2(1 + SINR)$ [bps/Hz]

α: Attenuation factor, representing implementation losses.

$SINR_{MIN}$: Minimum SINR of the code set [dB].

$SINR_{MAX}$: Maximum SINR of the code set [dB].

The parameters α, $SINR_{MIN}$ and $SINR_{MAX}$ can be chosen accordingly in order to represent different modem implementations and link conditions.

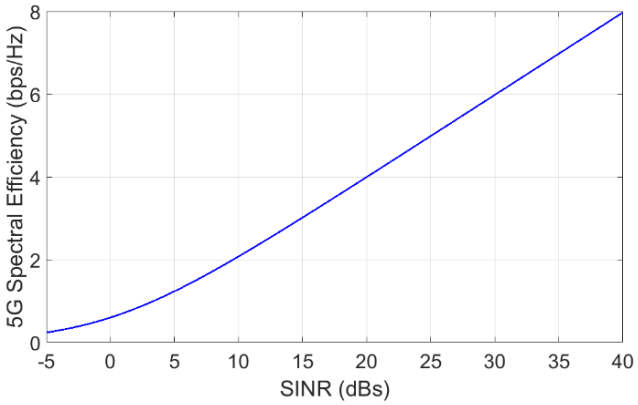


Figure 3-6 Spectral efficiency versus SINR in 5G

3.1.2.7.3 Numerical results

Figure 3-6 shows the spectral efficiency in 5G as a function of the SINR, which was estimated using the model included in this table. The bandwidth to be allocated to a user for guaranteeing a bit rate equals its spectral efficiency, which is given by its SINR as shown in the figure, times the bit rate to be enforced.

3.1.2.8 Experimental evaluation and modelling for virtualized RAN

3.1.2.8.1 Performance metric: GOPS

The main objective of this analysis is to measure the processing requirements of virtualized gNBs in terms of Giga Operations per Second (GOPS). The NG-RAN elements considered in 5G-CLARITY are hosted in the RAN cluster deployed in the private premises. The operation of these elements is supported by General Purpose Processors (GPPs) that can be accessed through appropriate interfaces (i.e., O-FH, F1). To optimally design the overall system, it is very important to identify the computational requirements of the virtualized gNB processing functions. This is important as with this input we can analyze the specificities and characteristics of the individual processing functions forming the Base Band Unit (BBU) Service Chain. Therefore, the objective of this study is to analyze the requirements of virtualized NG-RAN system running on general-purpose processors (i.e.,x86).

3.1.2.8.2 Evaluation methodology

To evaluate the performance of the NG-RAN system we rely on an open-source implementation of its protocol stack. Using this platform, the BBU processing requirement of its individual PHY elements are analyzed for various wireless access requirements and traffic load scenarios.

For our experiments we used an open source 5G RAN suite, and Intel's VTune Amplifier 2018 [Intel], a performance profiler for software performance analysis. These kernels are configurable and can be used to build applications to model wireless protocols. A summary of the BBU processing functions is presented below. This includes:

- The Single Carrier - Frequency Diversity Multiple Access is a precoded Orthogonal Frequency Diversity Multiplexing (OFDM). It is preferred compared to OFDM, for the uplink transmission, as it is less susceptible to frequency offsets and has a lower Peak-to-Average Power Ratio. The SC-FDMA Demodulation function removes the Cyclic Prefix (CP) and performs N-point Fast Fourier Transform (FFT).
- The Sub-carrier Demapper that extracts the data and the reference symbols from the subframes.
- The Frequency Domain Equalizer that estimates the Channel State Information (CSI) by the received

pilot signal through the Least Square estimation algorithm. It computes the channel coefficients, with the help of CSI, and equalizes the received data using a zero forcing MIMO detector in the frequency domain as equalizer.

- The Transform Decoder that performs M-point Inverse Fast Fourier Transfer (IFFT).
- The Constellation Demapper that receives the signal and extracts the binary stream by generating Logarithmic Likelihood Ratios (LLR).
- The Descrambler that descrambles the input sequence.
- The Rate Matcher that separates the input stream into N streams, de-interleaves each code stream and removes the redundant bits. For our experiments, one information bit is encoded into three transmitted bits, so N was constantly set to 3.
- The Turbo Decoder that takes soft information for each code, in our case LLR, and it applies iteratively the Soft-Input Soft-Output (SISO) algorithm. The Turbo Decoder consists of two SISO decoders that perform the trellis traversal algorithm, and one interleaver/de-interleaver. Higher number of iterations achieves improved error correction performance, at the expense of higher computation cost. To address this issue, for the conducted experiments 5 iterations were used.

3.1.2.8.3 Numerical results

To increase the statistical validity of the results produced by the profiler, a thorough investigation between different numbers of subframes processing was conducted, which resulted in setting the number of subframes to 1000. The set of experiments carried out was aiming at exploring the behavior of each processing function for different configurations of the gNBs PHY uplink system. Figure 3-7 presents the dependence of the instructions performed on the data rate for different modulation schemes, when processing 1000 subframes by each function.

Taking into consideration the variance of the measurements we can conclude that all functions present a linear dependence with the data rate. On the other hand, the influence of the modulation scheme, on the instructions number, differs for each function. More specifically we observe that the modulation scheme does not affect the instructions number for SC-FDMA Demodulation, Sub-carrier Demapper, Equalizer, and Transform Decoder. For the Constellation Demapper an exponential dependence of the modulation scheme is observed, while the Rate Matcher and the Turbo Decoder exhibit linear dependence.

We observe that the Turbo Decoder performs higher number of instructions, especially as the data rate increases, while the Constellation Demapper, the Rate Matcher and the Equalizer perform fewer instructions. This means that the Turbo Decoder, involving 1 to 4 orders of magnitude higher instructions compared to other functions, determines by large the total number of instructions needed to process a subframe and how this number depends on the data rate and the modulation scheme. Below are the linear expressions that fit the Turbo Decoder (1) and the Total Instructions (2) behavior.

$$\text{Instructions (million)} = 13747 * \text{Data Rate (Mbps)} \quad \text{(Eq. 3-1)}$$

$$\text{Instructions (million)} = 14575 * \text{Data Rate (Mbps)} \quad \text{(Eq. 3-2)}$$

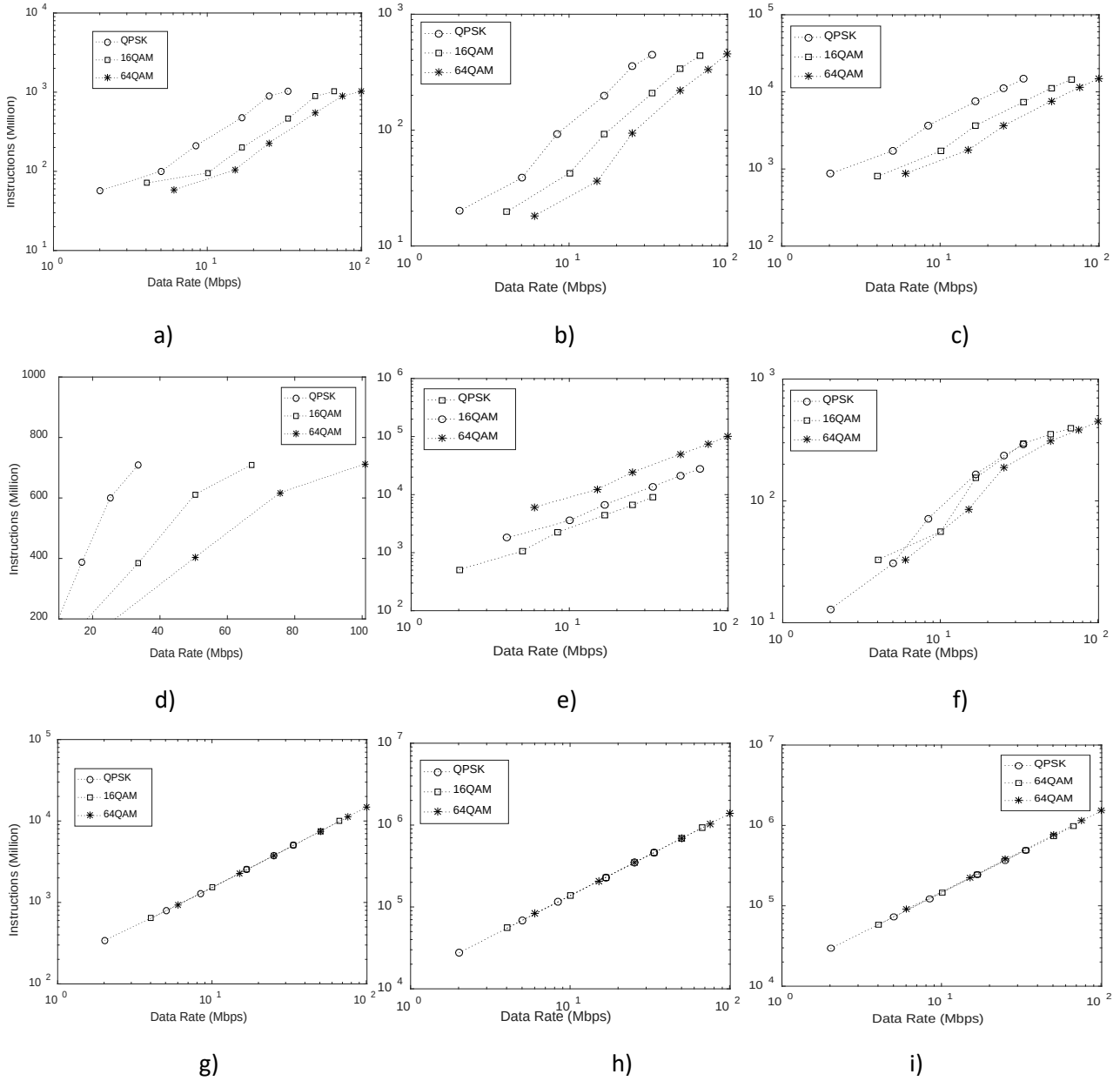


Figure 3-7 Instructions per signal processing function under various data rates for a) SC-FDMA Demodulation, b) Subcarrier Demapper, c) Equalizer, d) Transform Decoder, e) Demodulation, f) Descrambler, g) Rate Matcher, h) Turbo Decoder, and i) Total Instructions

3.1.2.9 Experimental evaluation of the virtualized UPF

3.1.2.9.1 Performance metric: N3 interface related measurements

The main objective of this section is to experimentally evaluate the performance of UPF elements. This will be evaluated in terms of [14]:

a) Number of incoming GTP data packets on the N3 interface, from (R)AN to UPF
b) Data volume of incoming GTP data packets for different PDU session requirements
c) Data volume of outgoing GTP data packets per QoS level on the N3 interface, from UPF to (R)AN
d) Incoming/outcoming GTP Data Packet Loss

3.1.2.9.2 Evaluation methodology

To evaluate the performance UPF an experimental 5G testbed has been deployed using an open source 5G platform. It is well known that UPF acts as a termination point for several interfaces and protocol stacks including N3 (GTP-U) tunnels from the RAN, N9 for the interconnection of a chain of UPFs as well as N6 for interconnecting the system with an external data network. Based on the information that is included in the interfaces and the information that it receives from the SMF, the UPF can take several actions including:

- Mapping of traffic to the appropriate tunnels based on the QoS Flow Identifier (QFI) information [ETSI TS 123 501] [15]. This requires UPFs to be able to perform Deep Packet Inspection (DPI) and identify the necessary values in the GTP-U header, associate QFIs with the appropriate Differentiated Services Code Point (DSCP) codes in the external IP network and perform the relevant protocol adaptations (encapsulations/decapsulations) at line rate.
- Steering of packets to the appropriate output port and take the necessary packet forwarding actions.
- Packet counting for charging and policy control purposes.
- DPI for security and anomaly detection purposes.
- Buffering and queuing management for traffic service differentiation and assurance of end-to-end delays.

A high-level view of the multiprotocol functionalities that a UPF can support is shown in Figure 3-8.

Where it is observed that the UPF should be capable of an extensive set of protocols including GTP-U, DFCP, IP, etc., assisting in the operation of SDAP and PDCP through mapping of DSCP classified IP traffic coming from the external data network to the appropriate QFI classes. It should be also capable of handling legacy and new protocols such as eCPRI/ORAN, Radio over Ethernet (RoE). Programmable HW (such as FPGAs and SmartNICs) can effectively classify and steer traffic within the server based on control plane (N1/2, N4), user plane (N3, N6) or UPF -to-UPF (N9) interfaces. For example, the NIC can steer control plane protocols such as PFCP into the SMF or control plane part of UPF and can steer UE session either based on PDU session, flow, QoS class etc. on N3 and N6. Furthermore, through programming it may be used to support extended header (EH) for 5G user plane traffic.

To evaluate the performance of the UPF, a model based on queuing theory is developed. This model is compared with experimental evaluation results. The flowchart of this model is shown in Figure 3-9, where in the first block traffic can be prioritized and steered via configurable policies into one of the available queues.

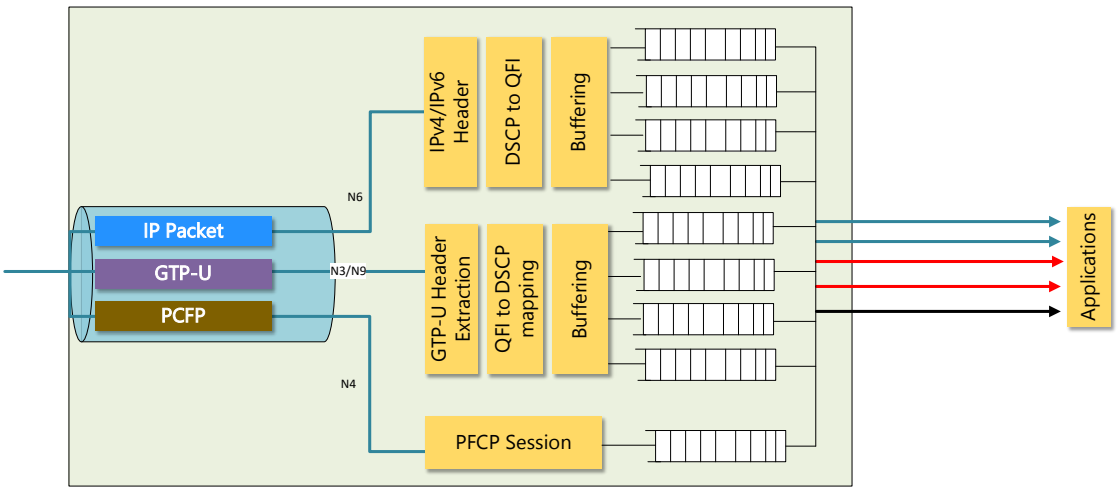


Figure 3-8 UPF multiprotocol interfaces

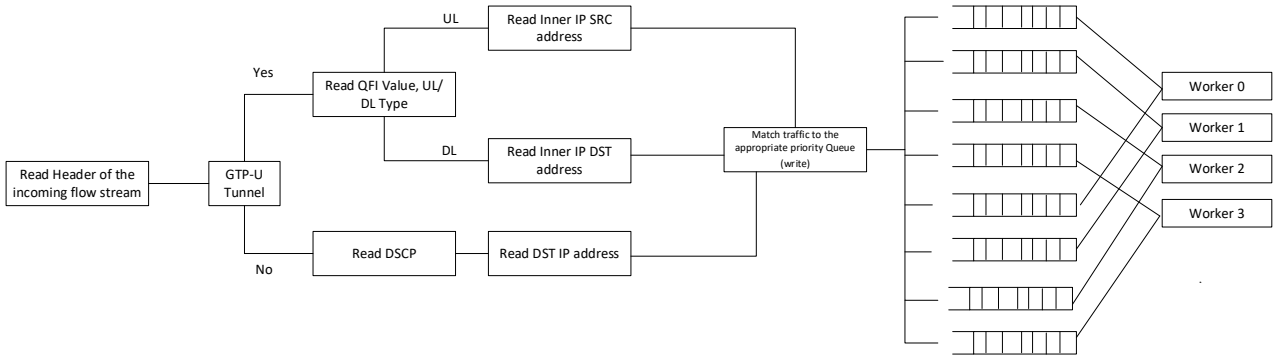


Figure 3-9 Detailed queuing model of UPF

3.1.2.9.3 Numerical results

We initially evaluate the computational requirements of the UPF a function of throughput assuming that the system is hosted in virtualized machines with different allocated computational resources. The results are based on the free5GCore platform but as this platform is based on 3GPP Rel 16, it can be extended to any other system compatible with this standard. The different VM configurations used to host the 5GC platform are summarized in Table 3-1.

Figure 3-10 shows the impact of PDU session throughput on CPU resource utilization. It is clear that as we allocate more resources to the Core Network, the CPU utilization is reduced. For instance, for the same data rate (around 90Mbps), the small VM consumes 16% of its CPU, while the medium consumes 7.5%, while the VMs with more available resources (i.e., large and x large) consume less than 5%.

Table 3-1 VM Configurations Used to Host the Virtualized 5GC Platform

VM	No of CPU Cores	RAM [GB]	Storage [GB]	Network [Gbps]
Small	1	2	20	1
Medium	2	4	40	1
Large	4	8	80	1
X Large	8	16	160	1

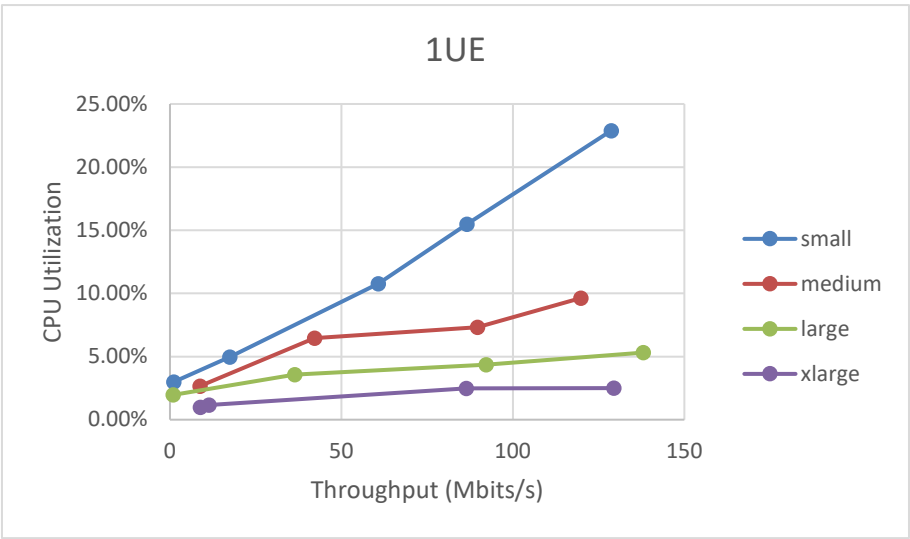


Figure 3-10 CPU consumption for various data rates under different VM configuration options

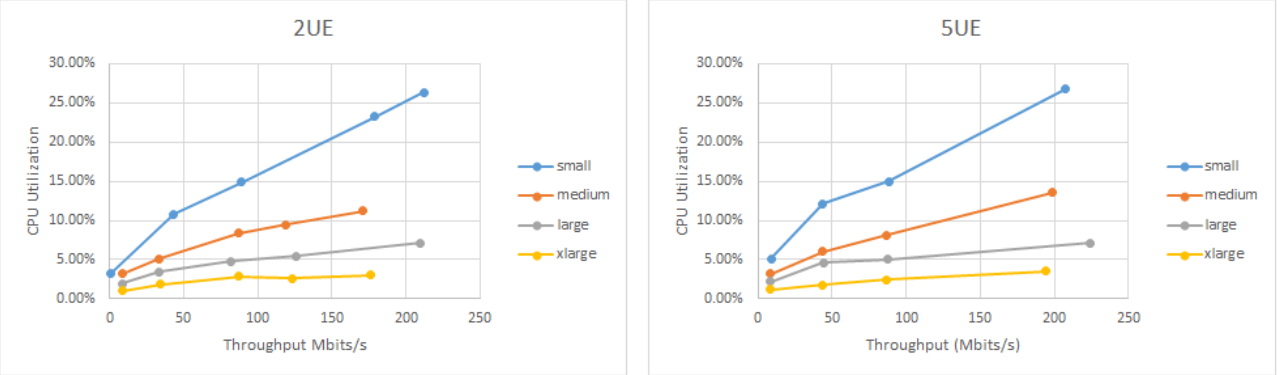


Figure 3-11 Multiple connected UEs

Since our ultimate goal was to evaluate the performance under high loading conditions, we started connecting more UEs to the Core Network while running the same test. The results for multiple connected UEs are illustrated in Figure 3-11.

The overall trend remains the same as with the one connected UE. One difference is that we observe an increase in the aggregated data rate. From Figure 3-12, it is concluded that increasing the number of connected UEs to 5, does not have an impact to the CPU consumption. Here, we can observe that for all four hosts, the same aggregated data rates result in the same CPU consumption regardless the number of connected UEs. This is expected as the UPF performance depends on the aggregated number of packets transferred.

As additional metric that has been also evaluated is related to the impact of allocated CPU resources on packet latency. Results in Figure 3-13 show that an increase of traffic terminated at the UPF leads to an increased number of hardware interrupts and, therefore, increased CPU utilization.

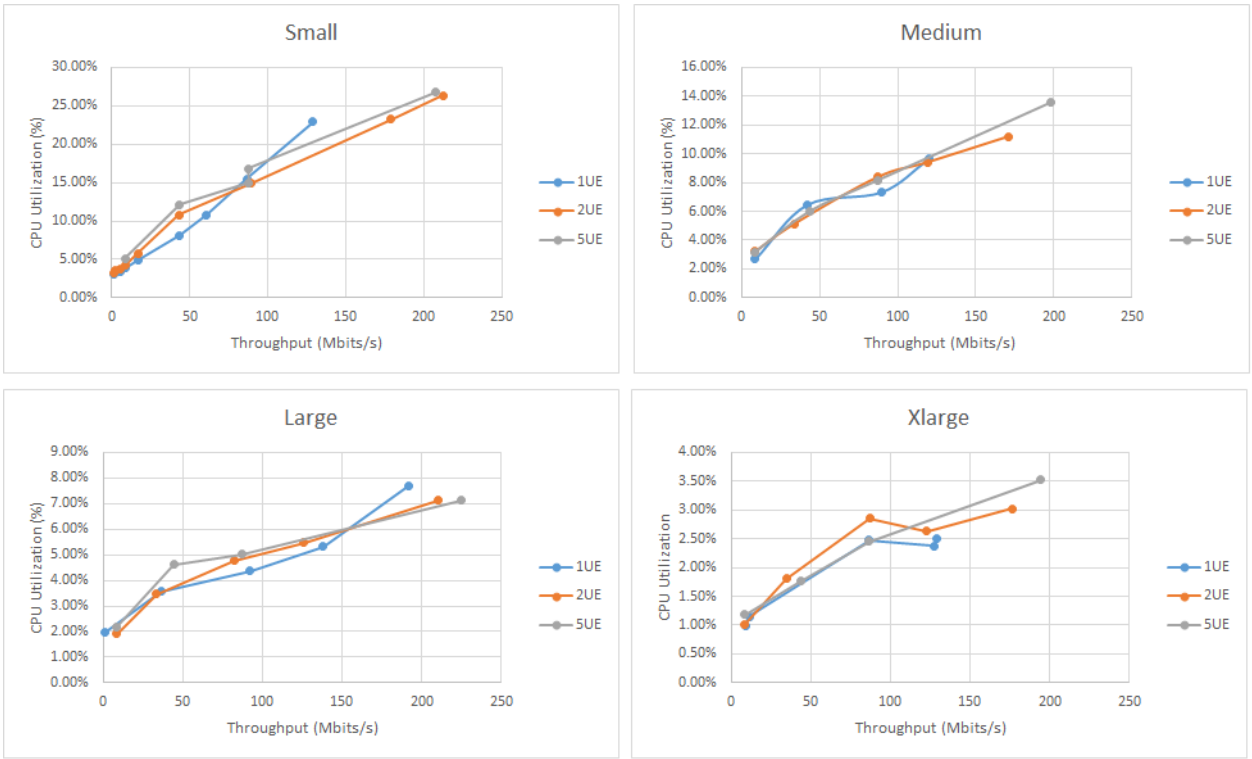


Figure 3-12 CPU utilization vs throughput for different number of UEs

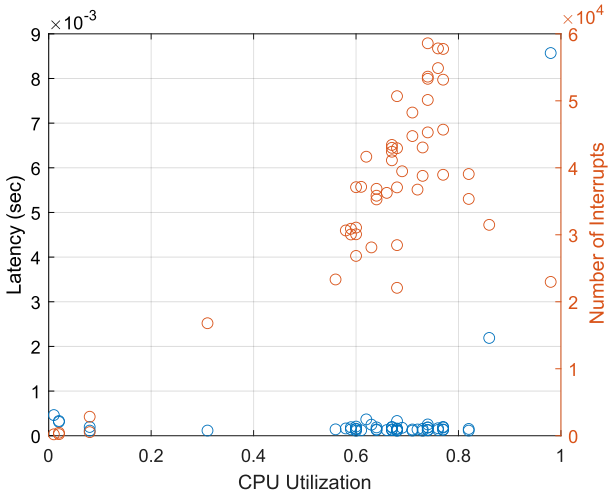


Figure 3-13 Correlation between IRQs and CPU utilization (right axis), CPU utilization and packet latency (left axis) as it has been measured over the experimental platform

3.1.3 Management and Orchestration stratum: modelling and performance evaluation for SDN controller

This subsection focuses on the experimental evaluation of the SDN controller entity and the component responsible for the *data lake* and its associated *data management* and data processing services. Both entities have key role in 5G-CLARITY as the former is used to control the converged access network infrastructures while the latter is responsible for the collection of the monitoring data and feed the components that are responsible for the Lifecyle management of the 5G-CLARITY architecture. For these entities critical parameters that are evaluated affecting the scalability of the system include control plane latencies and operations per second under various input traffic.

3.1.3.1.1 Performance metric: data to Control plane latency

The main concept of SDN is the separation of the control from the data forwarding by providing a centralized control over the network through the controller. A key problem with the nature of the SDN architecture is the centralized approach for the placement of the controller. Network topologies with a large number of devices and long distances between them (creating corresponding delays in communication between the switch and the central controller) require the creation of a more distributed architecture that allow the continuing operation of time sensitive applications. The ultimate goal is to create SDNs that are easily expandable and cover large geographical areas.

In the following, an approach for measuring the control plane latency is being discussed. The main idea is the creation of a northbound application over the OpenDayLight (ODL) platform - as a network administrator - that communicates externally with the controller, using the REST APIs. The application implements at first step a mechanism for collecting data about the network topology and at second step a mechanism for sending echo messages to all switches simultaneously, in order to measure the time response of ODL controller.

3.1.3.1.2 Evaluation methodology

To measure control plane latency an emulation environment based on mininet has been set up. A custom application has been developed that allows to measure data to control plane latencies for various network topologies and system deployment options. The application takes advantage of the ODL's REST architecture

by using the REST API at http://<ODL-IP>:8181/restconf/. In general, modifications are only made in the config state, which automatically updates the operational state of the controller. From the operational state, the network manager receives the desired information.

Python was chosen as the programming language because of its ability of executing bash commands. Python was combined with bash shell scripting to take advantage of the ODL REST API with the usage of the curl command.

In the beginning, through REST API we extract the topology of the network. Specifically, using the curl (bash command), with the appropriate headings and the GET method in the following URL:

curl -u <USERNAME>:<PASSWORD> -X GET -s http://<ODL-IP>:8181/restconf/operational/network-topology:network-topology/

The information about the network is obtained in json (or xml) format. In particular, the obtained json file includes the id of every link in the network, as well as the source and destination node and port of each link. From the above, the number of switches is exported.

For measuring the control plane delay, the application sends simultaneously echo messages to all the switches of the network through the NBI REST interface and records the time elapsed for receiving a reply. Thus, curl commands with POST method are employed in the following URL:

curl -u <USERNAME>:<PASSWORD> -H 'Content-Type: application/yang.operation+json' -X POST -s -d @data.json -w %{time_total} -o /dev/null http://195.134.79.53:8181/restconf/operations/sal-echo:send-echo

The REST API above requires an input in json format (data.json), that specifies the destination network node of the echo message. Thus, the application creates as many data.json files as the number of the nodes, that will serve as the input of each POST request. The curl commands are executed in parallel and their number is equal to the number of the nodes in the network, in order to see how the number of nodes affects the time responsiveness of ODL controller. For higher accuracy the above procedure is repeated 100 times, and the delay is considered as the average delay of each repetition.

3.1.3.1.3 Numerical results

The results for linear topologies with different number of network nodes are shown in Figure 3-14.

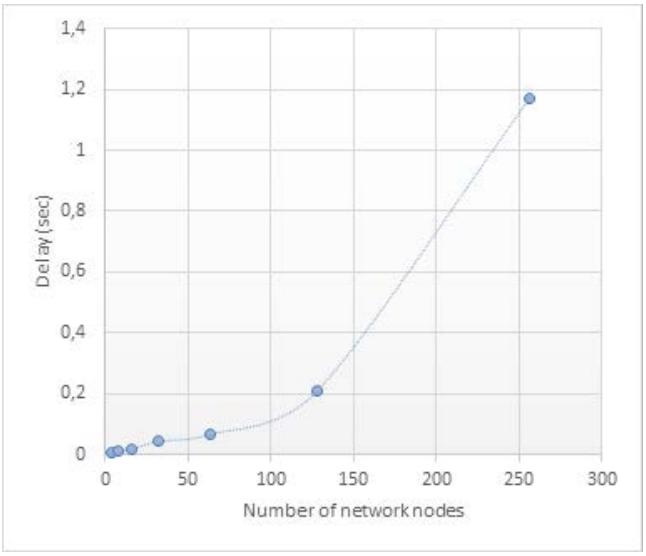


Figure 3-14 Dependence of processing time of SDN controller on the number of the network nodes

3.1.3.2 Modelling and performance evaluation for data management platform

3.1.3.2.1 Performance metric: Giga operations per second for the data management platform

The main objective of this section is to evaluate the performance of the data lake (Data Management Platform -DMP) used to collect measurements[2]. To achieve this, we rely on VTune profiler to measure the computational requirements of its building blocks as a function of the incoming data rate.

3.1.3.2.2 Evaluation methodology

To analyse the performance of the DMP platform, we rely on VTune profiler to measure the computational requirements of its building blocks as a function of the incoming data rate. The DMP is response to perform data collection, storage and processing. For the DMP, we analyzed the processing requirements of the message brokering servers (MQTT), ii) the Control Server that receives the data either from the MQTT broker or HTTP requests and forwards them for storage and iii) the time series database used for storage. The processing requirements of the DMP were derived by calculating the instructions per second required by its components as a function of the total IoT load.

3.1.3.2.3 Numerical results

The performance of the system in terms of operations per second is shown in Figure 3-15.

3.2 End-to-end modelling tools

In the previous section, a brief description of the main building blocks used in 5G-CLARITY has been provided. This section will describe the main tools that will be used to evaluate the performance of the overall system. This will include tools based on queuing theory, experimental platforms and emulation systems.

3.2.1 Modelling of the 5G DL URLLC slice's E2E mean processing time

We model the 5G system (5GS) and the transport network (TN) to interconnect its different components as an open queuing network. To solve this network, i.e., to estimate the E2E mean delay, we use the queuing network analyser (QNA) method proposed in [20]. This method can be regarded as an extension of the methodology to solve Jackson's open networks, which consists of M/M/c queuing nodes, to general open networks composed of G/G/c queuing nodes. The model presented here provides the E2E mean response time of the downlink of 5G-CLARITY slices. To that end, it relies on the specific delay models of the different functional elements described in Sections 3.1.2.1, 3.1.2.2, 3.1.2.3, 3.1.2.4, and 3.1.2.5. More precisely, QNA serves to estimate the first and second order moments of the aggregated internal arrival process to every functional element and estimate the E2E mean latency from the per-component delays, as detailed next.

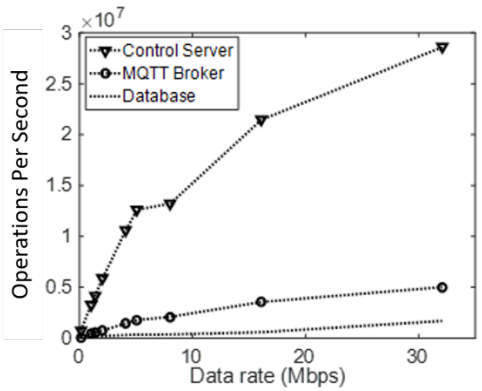


Figure 3-15 Instructions per second under various incoming data rates for the DMP components

The primary notation is defined in Table 3-2. The procedure to solve a network following the QNA method comprises the following steps:

1) Computation of the first and second order moments of the internal arrival processes:

   First, similar to the methodology to solve Jackson's open networks, we compute the aggregated arrival rate for each queuing facility by solving the following set of linear equations:

$$\lambda_k = \lambda_{0k} + \sum_{i=0}^{K} \lambda_i \nu_i p_{ik}$$

   Next, we compute the squared coefficient of variation (SCV) of the inter-arrival times for each queuing facility by solving the following set of linear equations:

$$c_{ak}^2 = a_k + \sum_{i=0}^{K} c_{ai}^2 \cdot b_{ik}$$

$$a_k = 1 + \omega_k\{(q_{ok} \cdot c_{0k}^2 - 1) + \sum_{i=1}^{K} q_{ik}[(1 - p_{ik}) + \nu_i \cdot p_{ik} \cdot \rho_i^2 \cdot x_i]\}$$

$$b_{ik} = \omega_k \cdot q_{ik} \cdot p_{ik} \cdot \nu_i\left(1 - \rho_i^2\right)$$

$$x_i = 1 + m_i^{-0.5} \cdot \left(max\left(C_{si}^2, 0.2\right) - 1\right)$$

$$\omega_k = \left(1 + 4 \cdot (1 - \rho_k)^2 \cdot (\gamma_k - 1)\right)^{-1}$$

$$\gamma_k = \left(\sum_{i=0}^{K} q_{ik}^2\right)^{-1}$$

2) Mean delay per queue computation:

   The set of linear equations in step 1 enable us to estimate the mean arrival rate and SCV of the inter-arrival packet times for each queue in the network, i.e., the first and second order moments of the aggregated packet arrival process to each queuing node. These moments are used as inputs to compute the mean sojourn time using the models described in Section II. Specifically, the mean sojourn times of the UPF, gNB-CU, gNB-DU, gNB-RU, and radio interface instances are estimated using the models described in Sections 3.1.2.1, 3.1.2.2, 3.1.2.3, 3.1.2.4, and 3.1.2.5, respectively.

3) E2E mean delay computation:

   Finally, we estimate the E2E mean delay of the DL 5G system by adding the individual mean response time contributions of the different queues in the system:

$$\mathrm{T}^{\mathrm{e2e}} = \Phi + \sum_{i=1}^{I_{UPF}} T_i^{UPF} \cdot V_i^{UPF} + T^{N3} + \sum_{i=1}^{I_{CU}} T_i^{CU} \cdot V_i^{CU} + \mathrm{T}^{\mathrm{F1}} + \sum_{i=1}^{I_{DU}} T_i^{DU} \cdot V_i^{DU} + \sum_{i=1}^{I_{RU}} T_i^{RU} \cdot V_i^{RU} + \sum_{i=1}^{I_{NR-Uu}} T_i^{NR-Uu} \cdot V_i^{NR-Uu}$$

   where:

   - Φ denotes the constant delays in the system that can be modelled as queuing facilities with an infinite number of servers. In other words, it refers to those delay components that does not depend on the traffic load (e.g., propagation delays at the links) and those resources that does depend slightly on it because they are not the primary bottlenecks in the system (e.g., switching

fabric processing time of the physical L2 bridges, virtual switches packet processing, L2-L4 protocol stack processing at the physical machines, physical network functions constant processing delays…).

- $I_{UPF}, I_{CU}, I_{DU}, I_{RU}$ and I_{NR-Uu} are respectively the number of UPF, gNB-CU, gNB-DU, gNB-RU, and NR-Uu instances of the URLLC slice. For instance, the virtualized UPF and CU functions might have several replicas (instances) running on independent virtualization containers due to the limitation on the processing capacity of a single physical machine or server. On the other side, considering an NR deployment with small cells that integrate the DU+RU functionalities, there will be a DU instance, an RU instance and a NR-Uu instance per small cell.
- $T_i^{UPF}, T_j^{CU}, T_k^{DU}, T_l^{RU}$ and T_m^{NR-Uu} denote the UPF, gNB-CU, gNB-DU, gNB-RU and radio interface mean sojourn time of the instances i, j, k, l, and m, respectively.
- $V_i^{UPF}, V_j^{CU}, V_k^{DU}, V_l^{RU}$, and V_m^{NR-Uu} stand for the visit ratio of the UPF, gNB-CU, gNB-DU, gNB-RU, and radio interface instance i, j, k, l, and m respectively. The visit ratio is defined as the average number of visits to a given node by a packet during its lifetime in the system.
- T^{N3} and T^{F1} are the mean packet transmission delays at the N3 and F1 interfaces introduced by the backhaul and midhaul networks, respectively. These delays depend on the TN setup when TSN is used as L2 technology. For instance, when the TN is realized as an asynchronous TSN network, we might use the mixed integer convex non-linear program formulated in [28]. Although the optimization goal considered in [28] is the minimization of the flow rejection probability, we can easily adapt it for other goals such as the minimization of the percentage of E2E delay budget consumed by the corresponding TN segment. Finding the optimal solution for a given optimization goal using the optimization program in [28] might serve to model the benefits brought by a ML algorithm running at the 5G-CLARITY system's AI engine for the transport network configuration optimization. Once the different TN segments are configured, e.g., the paths interconnecting the TN segments endpoints (e.g., UPF and CU in the backhaul network, and CU and small cells in the midhaul networks), the priority levels for the different 5G-CLARITY slices, and the reserved aggregated bandwidth at every link for each 5G-CLARITY slice, the mean delay time at the respective TN segment is estimated as average of the mean packet delay offered by all the available paths interconnecting the endpoints. The mean packet delay for each path is estimated as the sum of the delay contributions of all the ATSs/links included in the path using the model described in Section 3.1.1.2. Please note that the delay contributions related to the signal propagation through the wires and TSN bridges processing time are included in Φ variable described previously.

It shall be noted that the QNA methodology is an approximation method that generalizes the ideas of independence and product-form solutions to general systems. Then, the experimental validation of the model is of utmost importance as otherwise its validity is questionable. In [21], the QNA method is experimentally validated for predicting the E2E response time of softwarised network services. Furthermore, that work also demonstrates the usefulness of the QNA method to perform the dynamic resource provisioning of SNSs while ensuring a maximum E2E response time [21] .

Table 3-2 Primary Notation Used in the E2E Model for Assessing the Mean Response Time of 5G-CLARITY Slices

Notation	Description
K	Number of queues in the model
P	The steady-state transition probability matrix
k, i	Network nodes indexes
p_{ki}	The probability of a packet leaving a node k to node i

p_{0k}	The probability that a packet leaves the network at queue k
λ_{0k}	Mean arrival rate of the external arrival process at queue k
c_{0k}^2	SCV of the inter-arrival packet times for the external arrival process at queue k
μ_k	Mean service rate of each server at queue k
c_{sk}^2	SCV of the service time at queue k
λ_k	Aggregated arrival rate at queue k
c_{ak}^2	SCV of the inter-arrival packet times at queue k
m_k	Number of servers at queue k
a_k, b_{ik}	Coefficients of the set of linear equations to estimate the SCV of the inter-arrival packet times at each queue k
ω_k,x_i, γ_k	Auxiliary variables to compute a_k and b_{ik}
q_{0k}	The proportion of arrivals to node k from its external arrival process
q_{ik}	The proportion of arrivals to node k from node i

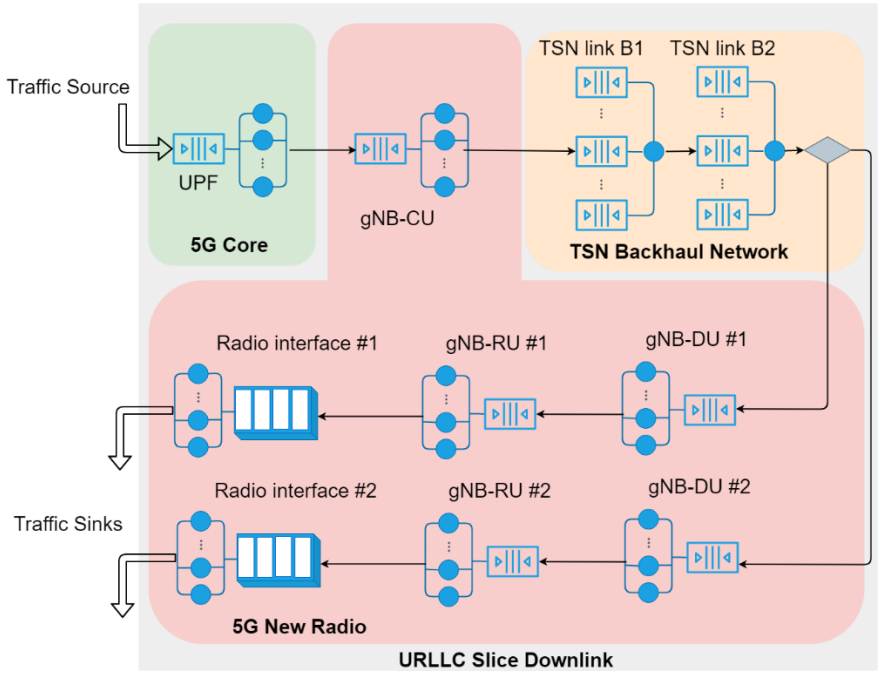


Figure 3-16 Queuing model of the DL of a 5G-CLARITY URLLC slice

Figure 3-16 shows the queueing model of a URLLC slice DL. Without loss of generality, the figure includes only the main bottlenecks and the rest of delay components are considered deterministic/constant (e.g., processing delay of each TSN bridge, propagation delay at every link, processing delay of the L1-L4 protocol stack in the VNFs...). There is only one instance for each VNF (e.g., UPF and gNB-CU) and two small cells (gNB-DU + gNB-RU + radio interface). The queuing servers at the VNFs, gNB-DU, and gNB-RU stand for processing units (e.g., physical CPU cores) and the respective processes or threads running the tasks associated with a packets processing in parallel. For instance, the service time of every queuing server at the UPF queuing node stands for the processing time required by a processing unit/thread to run the tasks associated with a single packet processing, which is ultimately given by the total number of instructions to be executed and the processor computing power. The radio interface is modelled as a multi-server with a finite queue, where each queuing server represents a PRB whose service time is a time slot duration. Observe that there might be packet losses at the radio interface. For the asynchronous TSN transport network, there is a bottleneck

(queuing node) at each TSN bridge output port that handles the frames of a given link. Each TSN bridge port is modelled as a non-pre-emptive multi-priority queuing node, where the server stands for the link packet transmission process whose service time is given by the nominal transmission capacity of the link. The only external packet arrival process to the URLLC slice DL is at the UPF and the packets leave the queuing network right after they are transmitted through the radio interface.

3.2.2 Modelling advantages of multi-WAT

In order to evaluate the Wi-Fi capacity to accommodate eMBB traffic from 5G, which is one of the primary advantages of using multi-WAT as proposed in 5G-CLARITY, RAN simulations are carried out. The relevant RAN simulator includes the analytical performance models described in previous sections, modelling different network functional elements. The specific performance metrics modelled include the throughput achieved by eMBB users in Wi-Fi and 5G, and the packet loss ratio suffered by URLLC slices in NR-Uu interface (Sections 3.1.1.3, 3.1.2.6, 3.1.2.7). We consider a multi-WAT scenario in which 5G NR and Wi-Fi technologies coexist, and two different services (e.g., eMBB and URLLC) are provided through different slices. Using the abovementioned simulator, we evaluate the offloading Wi-Fi capacity of eMBB users in terms of bandwidth released from 5G technology.

To do so, given a distribution of users in the scenario under consideration, the simulator estimates the Signal to Interference Noise Ratio (SINR) experienced by each user considering standard propagation models. Then, we estimate the bandwidth required by each eMBB user. The bandwidth required by each user depends on the estimated SINR and its particular traffic characteristics. Then, the slice aggregated bandwidth is computed as the sum of the bandwidth required by all the eMBB users that belong to the same slice.

More precisely, given the assumed scenario layout of a private industrial network (see Figure 5-3) we consider different number of UEs, in order to measure the bandwidth released from 5G for different workloads. The eMBB users' locations are randomly generated following a uniform distribution to sample the SINR they experience. In addition, it is worth saying that we generate a sufficient number of samples to ensure statistical stability in the results.

To estimate the bandwidth freed up from 5G technology, we proceed as follows:

As illustrated in the diagram depicted in Figure 3-17, first we determine the users that can be served through 5G technology in a realistic way, so that the SINR perceived by these users is above a certain minimum threshold that ensures an acceptable quality of service. Also, we take into account the maximum bandwidth available at every gNB. In other words, the bandwidth demand of the users attached at a given gNB does not exceed its available bandwidth. The users located far from the gNBs will be considered as outlier samples. Then, we iteratively check the gNB in which a given UE is attached to in order to compute the bandwidth consumed by each of the considered UEs. Next, we look for the best candidate Wi-Fi AP among the ones deployed on the scenario (i.e., it is the one that offers the highest SINR given the position of the UE). If the SINR is greater than a defined minimum threshold and the Wi-Fi AP has enough capacity to accommodate the UE under consideration, this UE will be offloaded to Wi-Fi technology.

The propagation model used to compute the path losses for the SINR computation for both technologies 5G and Wi-Fi is Indoor Hotspot [6]. The specification of the most relevant simulation parameters is included in Table 5-1.

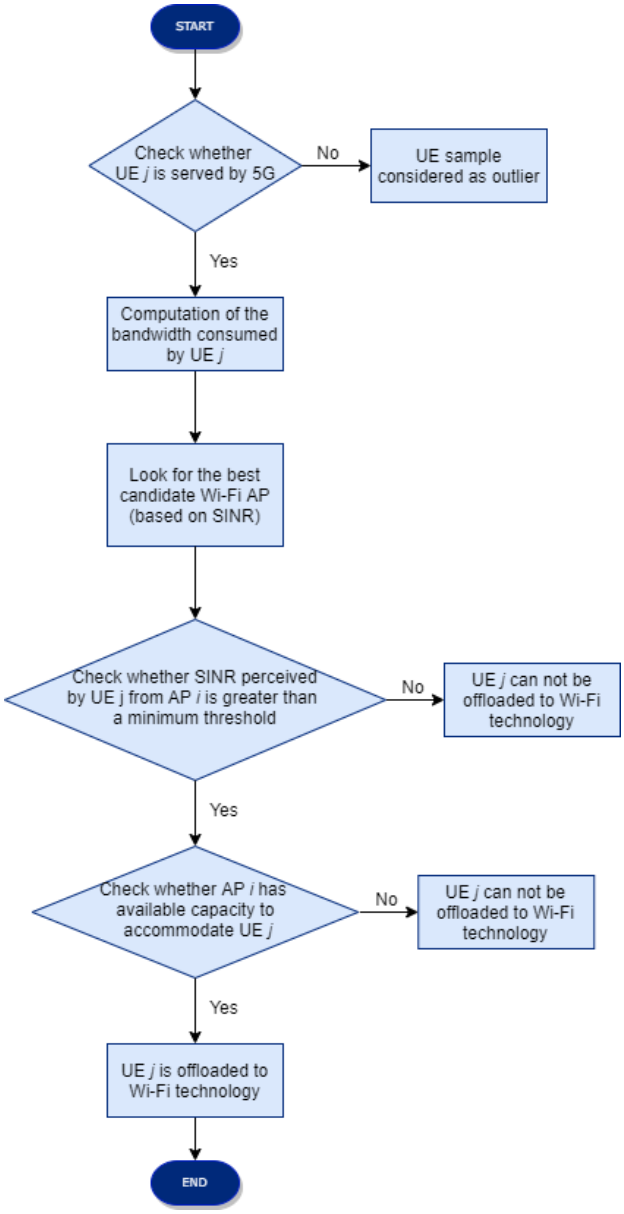


Figure 3-17 Wi-Fi offloading procedure description

3.2.3 Modelling of multi-WAT RAN for network resilience

An additional benefit to be taken into account during the operation of multi-WAT network is resilience. To ensure resilience, the Markov chain model shown in Figure 3-18 can be extended to cover the case of failure of gNB, LiFi or Wi-Fi APs. The key idea behind the proposed protection scheme is that in case of failure of an AP, services are redirected to the remaining operational APs. The overall process is modelled through the Markov Chain shown in Figure 3-18. As previously described, under normal operational conditions users are served by all APs. However, in case of failure of Wi-Fi, LiFi or gNB APs demands are served by the other APs. For example, assuming that the system is in state (i, j, k), in case of gNB failure the i service flows of the gNB AP will be redirected to the Wi-Fi AP and the new state of the system will be $(0, j, k + i)$. Similarly, in case of failure of LiFi the new state of the system will be $(i + j, 0, k)$. The failed AP can be either repaired after a predefined interval or remain out of operation. In case of failure of another AP, all demands will be served by a single access technology. Finally, an immediate repair is scheduled in case of failure of all APs.

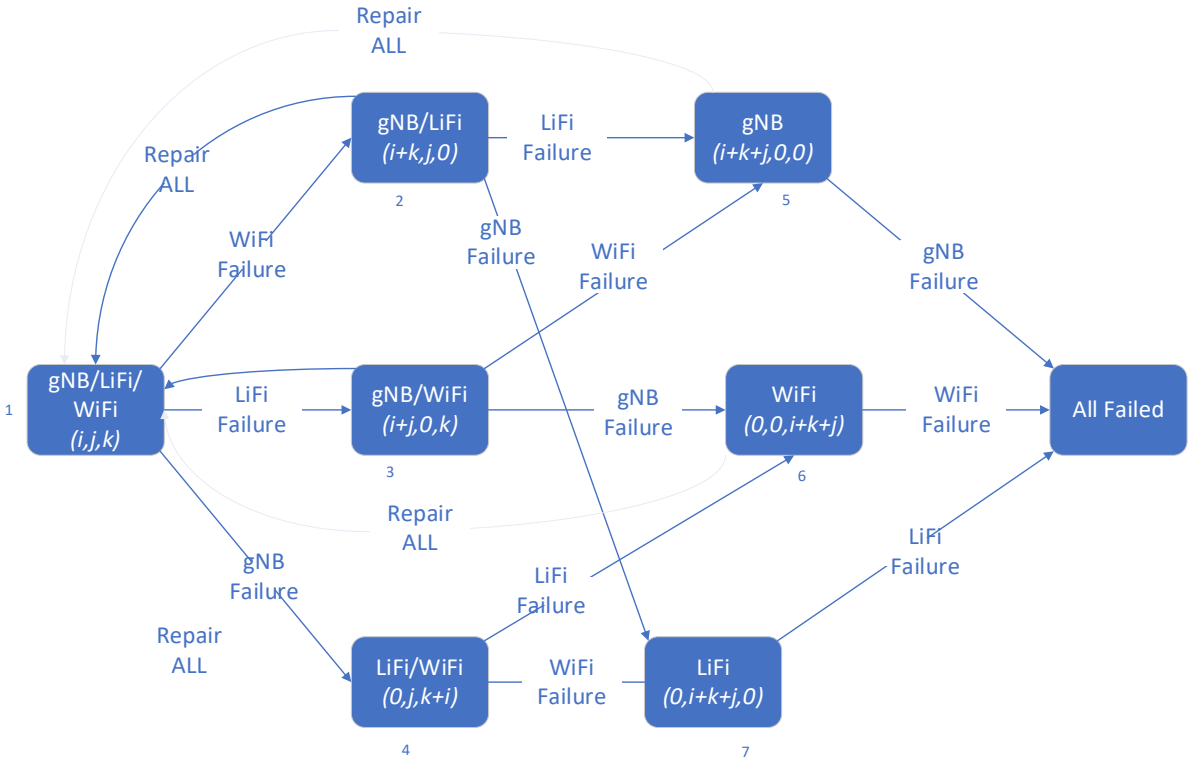


Figure 3-18 Repair/failure transition states of the on-board multi-technology access network comprising gNB/Wi-Fi/LiFi

3.2.4 Control Plane modelling

The present section focuses on the modelling and evaluation of the basic control and data plane procedures that are instantiated to provide the end-to-end services. To model the associated processes the analysis combines measurements from an experimental platform that has been deployed to support the envisioned use case as well as theoretical modelling tools. In this preliminary study, the radio access network is based on UERANSIM which will be replaced in the upcoming studies by ORAN. For the 5GC, both a 5G non-standalone (NSA) as well as a standalone (SA) version has been hosted in a cloud environment based on OpenStack.

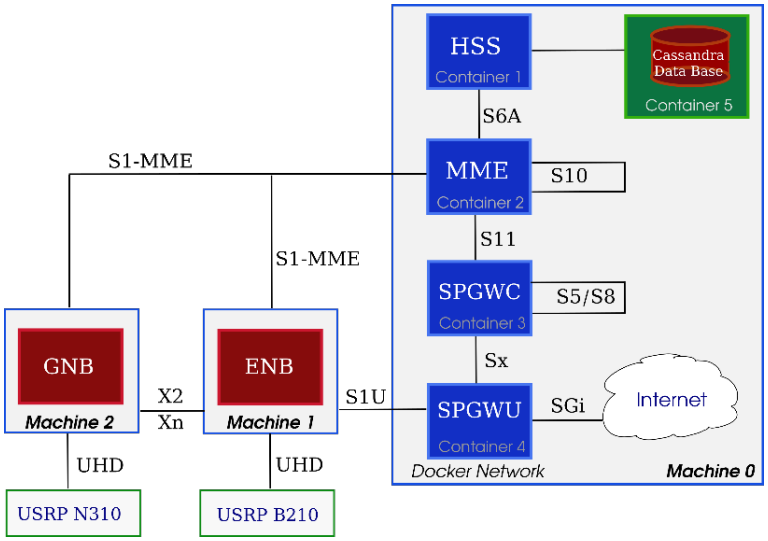


Figure 3-19 5G NSA architecture

3.2.4.1 5G non-standalone

We initially build a model to analyse the performance of a 5G NSA system in terms of service deployment times using measurements collected from an actual experimentation system. To achieve this, a 5G platform has been deployed providing slices supporting the use cases described in Section 4. The NSA version of the OpenAir Interface (OAI) platform has been deployed at IASA's/National and Kapodistrian University of Athens' (NKUA) private cloud facilities as shown in Figure 3-19. We initially evaluate the time needed for the association of a UE with the core.

During the experimental process it was observed that the UE initially accepts the configuration provided by the eNB which means that the RRC and X2AP are validated. There is also a successful random-access process that interprets that PRACH has been decoded correctly in gNB and the UE receives and decodes correctly msg2 (NR PDCCH Format 1_0 and NR PDSCH). Msg3 is transmitted to the gNB according to the configuration sent to msg2 and received correctly in gNB. It also successfully switches user-level traffic from the 4G cell to 5G (E-RAB modification message) where it is confirmed by S1AP.

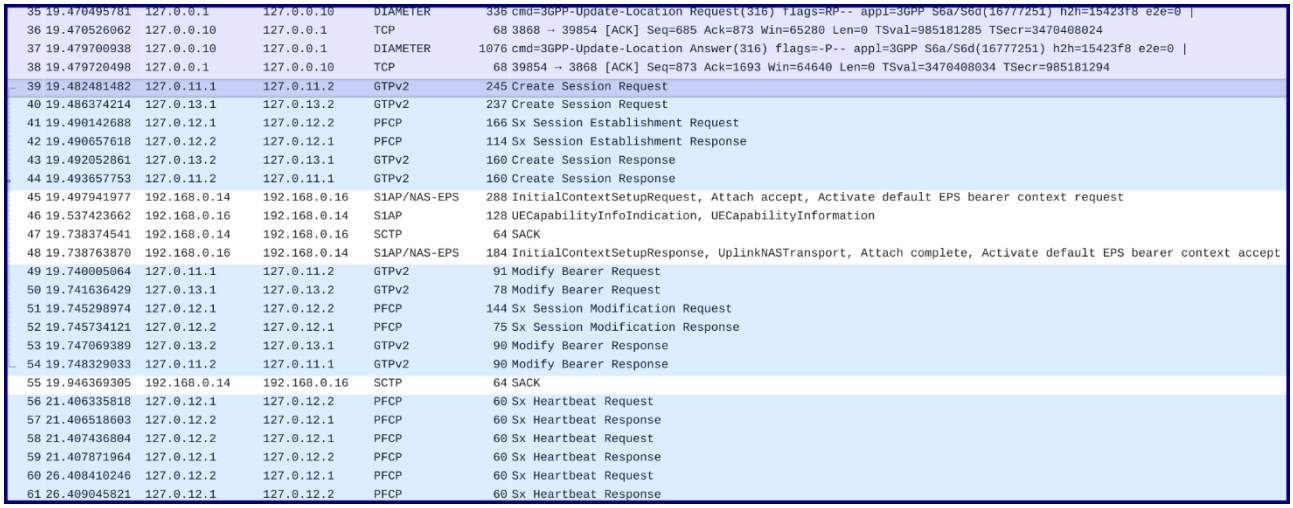
In terms of DL traffic, the PDCCH DCI format 1_1 and the corresponding PDSCH are decoded from the telephone and ACK/NACK signals (PUCCH format 0) are received on the gNB.

Finally, the UL / DL traffic is done with validated HARQ (ping, iperf) procedures. It is worth noting that the maximum data traffic valid for the DL is 3Mbps while for the uplink 1Mbps as it is a 5G NSA test platform although some packet losses may still occur even in ideal channel conditions.

Figure 3-20 shows a snapshot of packet traces as exported by Wireshark capturing the backbone traffic. The process is initiated by interconnecting the MME and HSS entities supported by the TCP protocol and Diameter. Diameter is an authentication, authorization and counting protocol for computer networks. After the handshake is achieved, it is time to activate the Sx interface where it connects the SPGWC and SPGWU via the Packet Forwarding Control Protocol (PFCP) introduced in CUPS to interfacing the control plane with the user plane, as seen in the packets numbering 8 and 9. Then (packets from 10 to 17) through the SCTP protocol, the core network communicates with the external networks, which in this case is the access network.

No.	Time	Source	Destination	Protocol	Lengt	Info
1	0.000000000	127.0.0.1	127.0.0.10	TCP	76	39854 → 3868 [SYN] Seq=0 Win=65495 Len=0 MSS=65495 SACK_PERM=1 TSval=3470388554 TSecr=0 WS=128
2	0.000026935	127.0.0.10	127.0.0.1	TCP	76	3868 → 39854 [SYN, ACK] Seq=0 Ack=1 Win=65483 Len=0 MSS=65495 SACK_PERM=1 TSval=985161814 TSecr=3470388554 WS=128
3	0.000043250	127.0.0.1	127.0.0.10	TCP	68	39854 → 3868 [ACK] Seq=1 Ack=1 Win=65536 Len=0 TSval=3470388554 TSecr=985161814
4	0.013506874	127.0.0.1	127.0.0.10	DIAMETER	392	cmd=Capabilities-Exchange Request(257) flags=R--- appl=Diameter Common Messages(0) h2h=15423f6 e2e=5b32c109 |
5	0.013530964	127.0.0.10	127.0.0.1	TCP	68	3868 → 39854 [ACK] Seq=1 Ack=325 Win=65280 Len=0 TSval=985161828 TSecr=3470388567
6	0.016749687	127.0.0.10	127.0.0.1	DIAMETER	456	cmd=Capabilities-Exchange Answer(257) flags=---- appl=Diameter Common Messages(0) h2h=15423f6 e2e=5b32c109 |
7	0.016769457	127.0.0.1	127.0.0.10	TCP	68	39854 → 3868 [ACK] Seq=325 Ack=389 Win=65152 Len=0 TSval=3470388571 TSecr=985161831
8	1.402875230	127.0.12.2	127.0.12.1	PFCP	75	Sx Association Setup Request
9	1.403938803	127.0.12.1	127.0.12.2	PFCP	79	Sx Association Setup Response
10	5.777340678	192.168.0.16	192.168.0.14	SCTP	84	INIT
11	5.777510631	192.168.0.14	192.168.0.16	SCTP	308	INIT_ACK
12	5.777862413	192.168.0.16	192.168.0.14	SCTP	280	COOKIE_ECHO
13	5.777941493	192.168.0.14	192.168.0.16	SCTP	52	COOKIE_ACK
14	5.778596565	192.168.0.16	192.168.0.14	S1AP	124	S1SetupRequest
15	5.778644124	192.168.0.14	192.168.0.16	SCTP	64	SACK
16	5.781067722	192.168.0.14	192.168.0.16	S1AP	92	S1SetupResponse
17	5.781656774	192.168.0.16	192.168.0.14	SCTP	64	SACK
18	6.404486740	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Request
19	6.405140298	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Response
20	11.404932933	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Request
21	11.405148590	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Response
22	11.405552741	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Request
23	11.405928518	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Response
24	16.406451627	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Request
25	16.406971387	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Response
26	19.417380703	192.168.0.16	192.168.0.14	S1AP/NAS-EPS	160	InitialUEMessage, Attach request, PDN connectivity request
27	19.421007662	127.0.0.1	127.0.0.10	DIAMETER	348	cmd=3GPP-Authentication-Information Request(318) flags=RP-- appl=3GPP S6a/S6d(16777251) h2h=15423f7 e2e=0 |
28	19.421046309	127.0.0.10	127.0.0.1	TCP	68	3868 → 39854 [ACK] Seq=389 Ack=605 Win=65280 Len=0 TSval=985181235 TSecr=3470407975
29	19.424295492	127.0.0.10	127.0.0.1	DIAMETER	364	cmd=3GPP-Authentication-Information Answer(318) flags=-P-- appl=3GPP S6a/S6d(16777251) h2h=15423f7 e2e=0 |
30	19.424325789	127.0.0.1	127.0.0.10	TCP	68	39854 → 3868 [ACK] Seq=605 Ack=685 Win=65280 Len=0 TSval=3470407978 TSecr=985181238
31	19.426994654	192.168.0.14	192.168.0.16	S1AP/NAS-EPS	144	DownlinkNASTransport, Authentication request
32	19.444419115	192.168.0.16	192.168.0.14	S1AP/NAS-EPS	140	UplinkNASTransport, Authentication response
33	19.448165469	192.168.0.14	192.168.0.16	S1AP/NAS-EPS	120	DownlinkNASTransport, Security mode command
34	19.467433656	192.168.0.16	192.168.0.14	S1AP/NAS-EPS	148	UplinkNASTransport, Security mode complete

Figure 3-20 5G NSA packet traces



35	19.470495781	127.0.0.1	127.0.0.10	DIAMETER	336	cmd=3GPP-Update-Location Request(316) flags=RP-- appl=3GPP S6a/S6d(16777251) h2h=15423f8 e2e=0 |
36	19.470526062	127.0.0.10	127.0.0.1	TCP	68	3868 → 39854 [ACK] Seq=685 Ack=873 Win=65280 Len=0 TSval=985181285 TSecr=3470408024
37	19.479700938	127.0.0.10	127.0.0.1	DIAMETER	1076	cmd=3GPP-Update-Location Answer(316) flags=-P-- appl=3GPP S6a/S6d(16777251) h2h=15423f8 e2e=0 |
38	19.479720498	127.0.0.1	127.0.0.10	TCP	68	39854 → 3868 [ACK] Seq=873 Ack=1693 Win=64640 Len=0 TSval=3470408034 TSecr=985181294
39	19.482481482	127.0.11.1	127.0.11.2	GTPv2	245	Create Session Request
40	19.486374214	127.0.13.1	127.0.13.2	GTPv2	237	Create Session Request
41	19.490142688	127.0.12.1	127.0.12.2	PFCP	166	Sx Session Establishment Request
42	19.490657618	127.0.12.2	127.0.12.1	PFCP	114	Sx Session Establishment Response
43	19.492052861	127.0.13.2	127.0.13.1	GTPv2	160	Create Session Response
44	19.493657753	127.0.11.2	127.0.11.1	GTPv2	160	Create Session Response
45	19.497941977	192.168.0.14	192.168.0.16	S1AP/NAS-EPS	288	InitialContextSetupRequest, Attach accept, Activate default EPS bearer context request
46	19.537423662	192.168.0.16	192.168.0.14	S1AP	128	UECapabilityInfoIndication, UECapabilityInformation
47	19.738374541	192.168.0.14	192.168.0.16	SCTP	64	SACK
48	19.738763870	192.168.0.16	192.168.0.14	S1AP/NAS-EPS	184	InitialContextSetupResponse, UplinkNASTransport, Attach complete, Activate default EPS bearer context accept
49	19.740005064	127.0.11.1	127.0.11.2	GTPv2	91	Modify Bearer Request
50	19.741636429	127.0.13.1	127.0.13.2	GTPv2	78	Modify Bearer Request
51	19.745298974	127.0.12.1	127.0.12.2	PFCP	144	Sx Session Modification Request
52	19.745734121	127.0.12.2	127.0.12.1	PFCP	75	Sx Session Modification Response
53	19.747069389	127.0.13.2	127.0.13.1	GTPv2	90	Modify Bearer Response
54	19.748329033	127.0.11.2	127.0.11.1	GTPv2	90	Modify Bearer Response
55	19.946369305	192.168.0.14	192.168.0.16	SCTP	64	SACK
56	21.406335818	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Request
57	21.406518603	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Response
58	21.407436804	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Request
59	21.407871964	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Response
60	26.408410246	127.0.12.2	127.0.12.1	PFCP	60	Sx Heartbeat Request
61	26.409045821	127.0.12.1	127.0.12.2	PFCP	60	Sx Heartbeat Response

Figure 3-21 5G NSA core network analyse packets-attaching UE in the network

Typically, the distinguish the handshake of the SCTP protocol and how it differs from the TCP handshake can be distinguished. Regarding packets from 18 to 25 in the Wireshark analysis we notice that the PFCP protocol is in the heartbeat protocol process.

A heartbeat protocol is generally used to negotiate and monitor the availability of a resource, such as a floating IP address. Usually when a heartbeat starts on a machine, it will perform a selection process with other machines on the heartbeat network to determine which machine, if any, owns the resource. Thus, there will be switching so that the SPGWU can freely use the system resources to move the packet, giving substance to what we have referred to as the separation of control and data planes. In packet 26 we have the first NAS signal where a device requests input to the core network and PDN connectivity. The PDN connection procedure is used by the UE to request the setting of a default EPS carrier on a PDN. The EU requests connectivity to a PDN by sending a PDN CONNECTIVITY REQUEST message to the network. The core network authenticates the UE by cross-referencing SIM card information with information entered into the core network database with the help of the HSS entity as it works with the MME entity. Swapping certification messages and security between the two networks (core and access) is perceived in packets 31 to 38.

In LTE, GTPS (GPRS Tunnelling Protocol) tunnels are used between two nodes that communicate via a GTP-based interface to separate traffic into different communication streams. A GTP tunnel is identified at each node by a TEID (Tunnel Endpoint Identifier), an IP address, and a UDP port number.

The receiving side of a GTP tunnel locally assigns the TEID value to be used by the transmission side. GTPv2 includes an updated control plane that allows the transmission of control messages between MME, S-GW, PDN GW, etc. Thus, packets 39 to 44 activate the GTPv2 protocol for the S11, S5 / S8 and SX interfaces as shown in the adjacent diagram.

Finally, the attachment of the UE is completed with package 49, activating the access to the EPS vector, where through the bearers and through the GTPv2 protocol streams are created serve both signalling and packet handling. As we can see, the successful connection of a device is created after 19,738 seconds and with a total number of 48 packets.

3.2.4.2 5G standalone

The performance of the proposed use cases is also evaluated using a standalone version of the core functions. The implementation of the standalone system is based on the architecture shown in Figure 3-22 where the NRF, AMF and SMF functions are used for the control plane and the UPF function for the user plane.

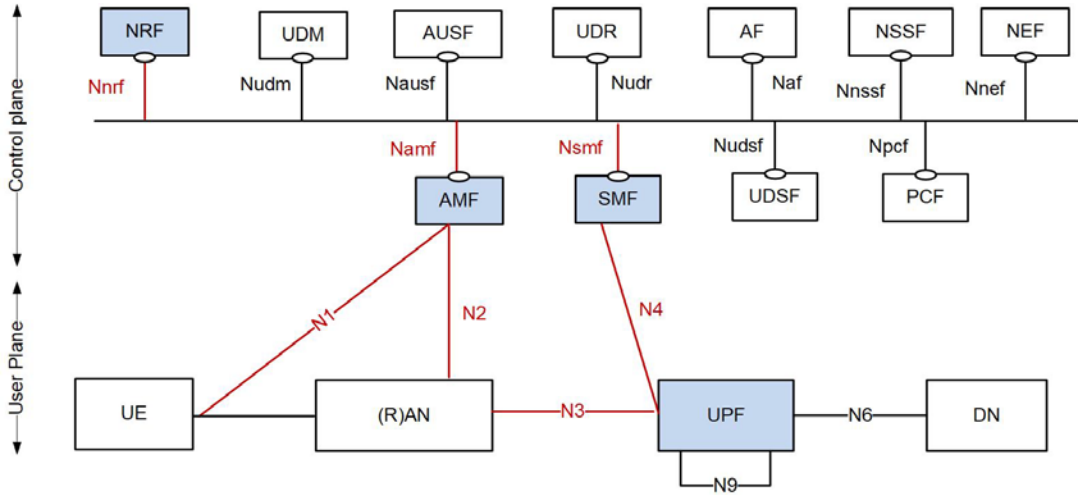


Figure 3-22 5G SA architecture

The system has been deployed in a containerized environment and monitored through "tShark", i.e., a CLI version of Wireshark. This process helps us to monitor packet exchange within the core network of the 5G SA system. An overview of the network and the addresses received assigned to the various 5G elements is shown in Table 3-3.

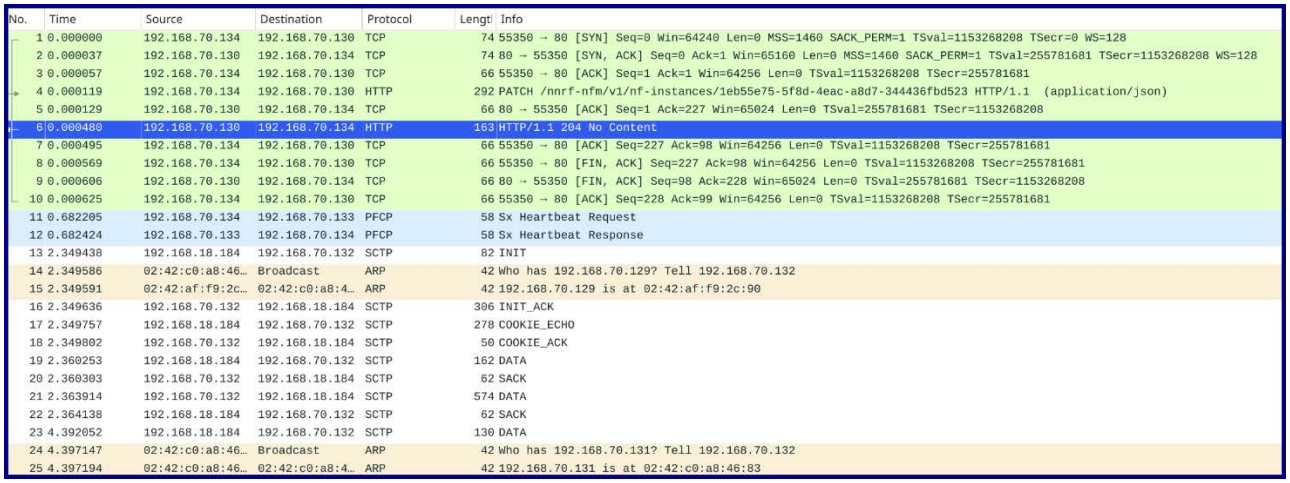
The scenario under investigation sets up an infrastructure slice interconnecting the UE with an external MEC node (referred to as External Data Network – EXT-DN). To set this connection, a set of messages are initially exchanged between the smf, nrf and upf as shown in Figure 3-23. Specifically, in this message it is shown that the SMF sends a POST message to the NRF to record entry/ deletion events as shown in packet 13. Then, the SMF and the SPGWU is registered to the NRF (PUT requests in packet 23 and packet 35, respectively). In packet 40 a POST request is received and the NRF informs the SMF about the registration of the SPGWU. The request and the response for the SPGWU PFCP is shown in packages 42 and 46.

As shown in Figure 3-24, a TCP connection is initially established between the SPGWU and the NRF. Then, using the Heartbeat protocol, connectivity between SPGWU and SMF is verified. This is implemented through an ARP protocol that is corresponded to the MAC addresses of the interfaces with the IP addresses (class C) of the network (packet 13). The SCTP protocol is initialized, then a set of messages is exchanged between packets 16 and 23 implementing a handshake process using a four-way message exchange to enhance security. SCTP is responsible for connecting to the kernel network and specifically to the AMF function.

Table 3-3 System Configuration

CONTAINER	IP-ADDRESS	DESCRIPTION
Mysql	192.168.70.131	Data Base
AMF	192.168.70.132	Amf entity
SMF	192.168.70.133	Smf entity
NRF	192.168.70.130	Nrf entity
SPGWU	192.168.70.134	Has a UPF role entity
EXT-DN	192.168.70.135	External network for testing
HOST	192.168.70.129	Containers host machine

Figure 3-23 Message communication between smf, nrf and upf



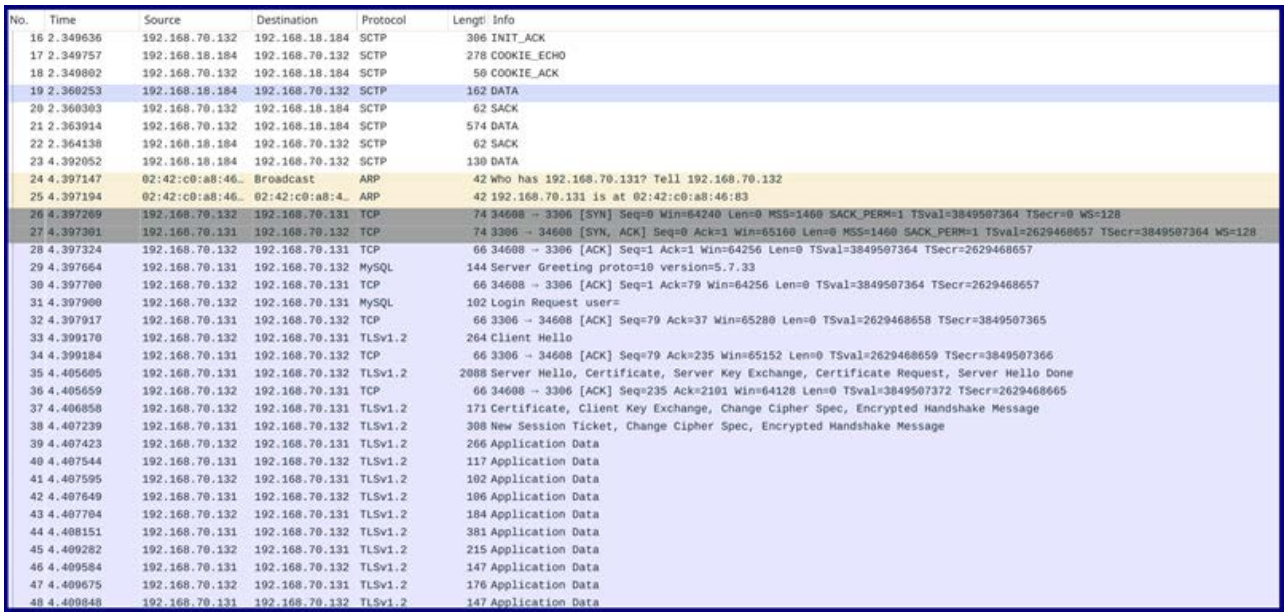
No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	192.168.70.134	192.168.70.130	TCP	74	55350 → 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 TSval=1153268208 TSecr=0 WS=128
2	0.000037	192.168.70.130	192.168.70.134	TCP	74	80 → 55350 [SYN, ACK] Seq=0 Ack=1 Win=65160 Len=0 MSS=1460 SACK_PERM=1 TSval=255781681 TSecr=1153268208 WS=128
3	0.000057	192.168.70.134	192.168.70.130	TCP	66	55350 → 80 [ACK] Seq=1 Ack=1 Win=64256 Len=0 TSval=1153268208 TSecr=255781681
4	0.000119	192.168.70.134	192.168.70.130	HTTP	292	PATCH /nnrf-nfm/v1/nf-instances/1eb55e75-5f8d-4eac-a8d7-344436fbd523 HTTP/1.1 (application/json)
5	0.000129	192.168.70.130	192.168.70.134	TCP	66	80 → 55350 [ACK] Seq=1 Ack=227 Win=65024 Len=0 TSval=255781681 TSecr=1153268208
6	0.000480	192.168.70.130	192.168.70.134	HTTP	163	HTTP/1.1 204 No Content
7	0.000495	192.168.70.134	192.168.70.130	TCP	66	55350 → 80 [ACK] Seq=227 Ack=98 Win=64256 Len=0 TSval=1153268208 TSecr=255781681
8	0.000569	192.168.70.134	192.168.70.130	TCP	66	55350 → 80 [FIN, ACK] Seq=227 Ack=98 Win=64256 Len=0 TSval=1153268208 TSecr=255781681
9	0.000606	192.168.70.130	192.168.70.134	TCP	66	80 → 55350 [FIN, ACK] Seq=98 Ack=228 Win=65024 Len=0 TSval=255781681 TSecr=1153268208
10	0.000625	192.168.70.134	192.168.70.130	TCP	66	55350 → 80 [ACK] Seq=228 Ack=99 Win=64256 Len=0 TSval=1153268208 TSecr=255781681
11	0.682205	192.168.70.134	192.168.70.133	PFCP	58	Sx Heartbeat Request
12	0.682424	192.168.70.133	192.168.70.134	PFCP	58	Sx Heartbeat Response
13	2.349438	192.168.18.184	192.168.70.132	SCTP	82	INIT
14	2.349586	02:42:c0:a8:46…	Broadcast	ARP	42	Who has 192.168.70.129? Tell 192.168.70.132
15	2.349591	02:42:af:f9:2c…	02:42:c0:a8:4…	ARP	42	192.168.70.129 is at 02:42:af:f9:2c:90
16	2.349636	192.168.70.132	192.168.18.184	SCTP	306	INIT_ACK
17	2.349757	192.168.18.184	192.168.70.132	SCTP	278	COOKIE_ECHO
18	2.349802	192.168.70.132	192.168.18.184	SCTP	50	COOKIE_ACK
19	2.360253	192.168.18.184	192.168.70.132	SCTP	162	DATA
20	2.360303	192.168.70.132	192.168.18.184	SCTP	62	SACK
21	2.363914	192.168.70.132	192.168.18.184	SCTP	574	DATA
22	2.364138	192.168.18.184	192.168.70.132	SCTP	62	SACK
23	4.392052	192.168.18.184	192.168.70.132	SCTP	130	DATA
24	4.397147	02:42:c0:a8:46…	Broadcast	ARP	42	Who has 192.168.70.131? Tell 192.168.70.132
25	4.397194	02:42:c0:a8:46…	02:42:c0:a8:4…	ARP	42	192.168.70.131 is at 02:42:c0:a8:46:83

Figure 3-24 5G SA core network packets-initial connection

After the core network receives user data, it contacts the database to verify its subscription. Therefore, through the TCP protocols the connection is established and with the MySQL protocol (package 29) the two parties communicate, i.e., the server that is the container that implements the database and the client that requests access to the data and in this case is the AMF entity.

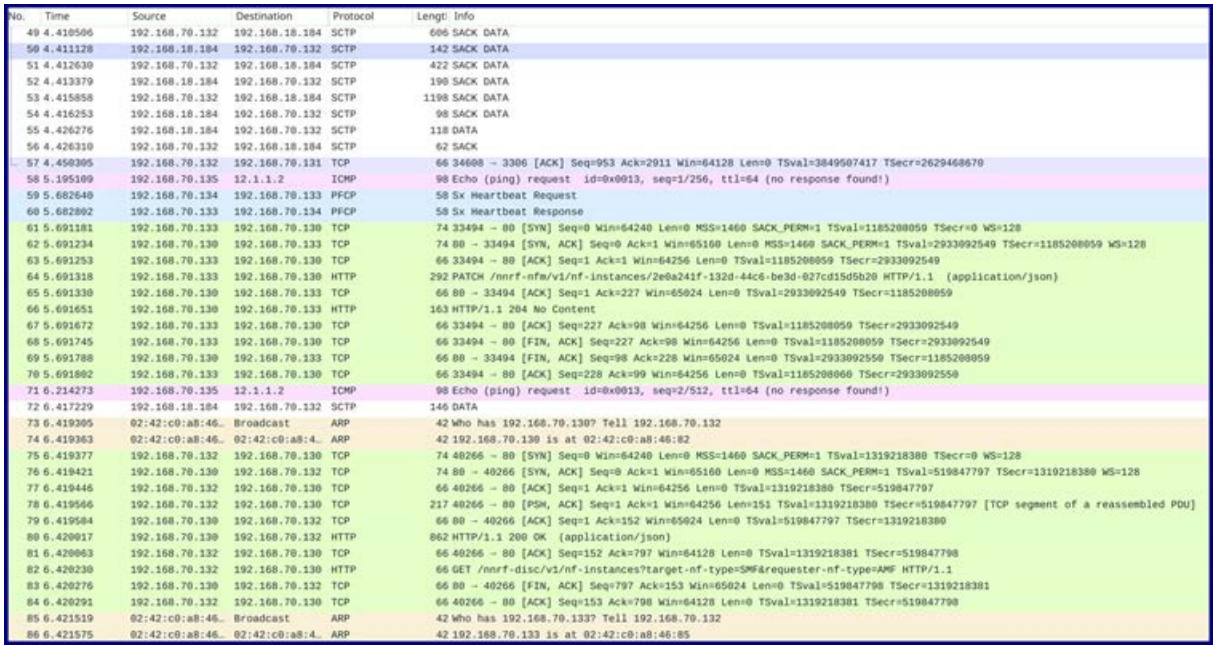
As data retrieval from the database must comply with all security requirements, 3GPP has introduced the TLS protocol in the 5G specification. Therefore, 5G kernel functions support innovative security protocols such as TLS 1.2 and 1.3 to protect communication at the transport level and the OAuth 2.0 framework at the application level to ensure that only authorized network functions are accessed, a service that offered by another function. The function of the TLSv1.2 protocol is shown in the communication between the database and the AMF, from package 33 onwards where the connection achieved and with the necessary encryption, and the access to the data is given.

From package 49, the connection to the access network and therefore to the user device starts, by sending Selective Acknowledgment (SACK) packets. SCTP applications submit data for transmission in messages (bytes groups) at the SCTP transfer level.



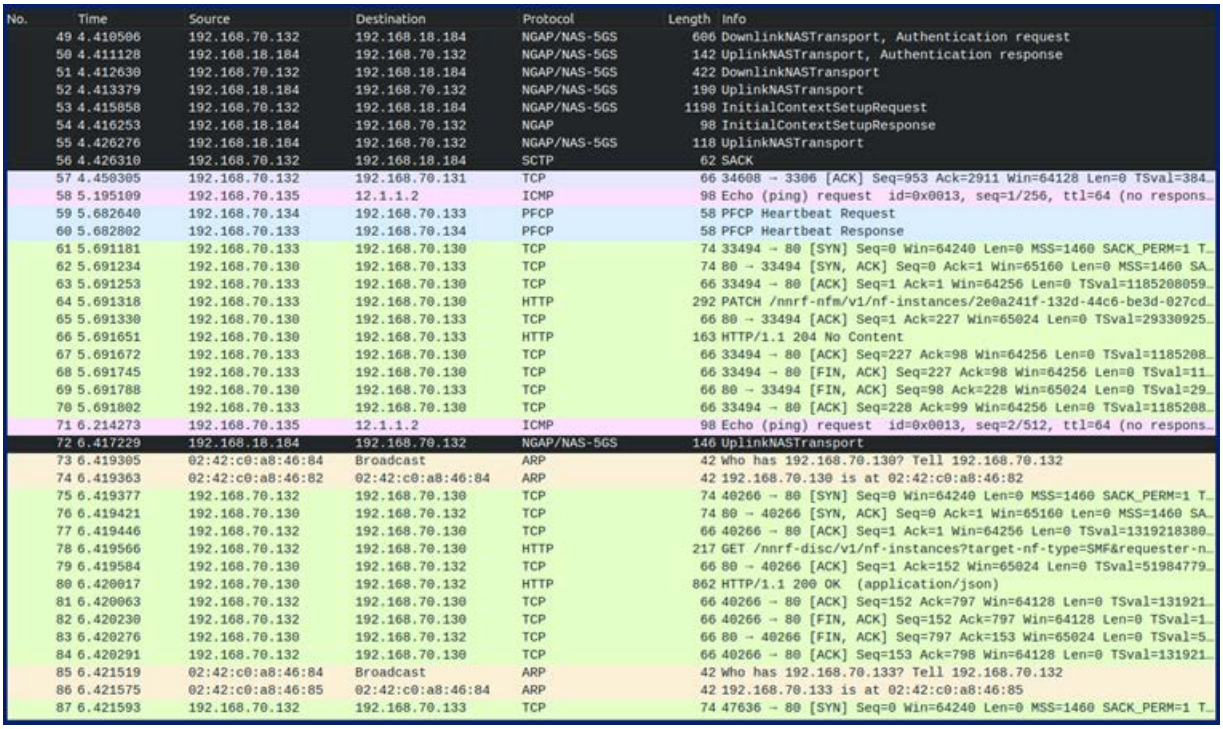
No.	Time	Source	Destination	Protocol	Length	Info
16	2.349636	192.168.70.132	192.168.18.184	SCTP	306	INIT_ACK
17	2.349757	192.168.18.184	192.168.70.132	SCTP	278	COOKIE_ECHO
18	2.349802	192.168.70.132	192.168.18.184	SCTP	50	COOKIE_ACK
19	2.360253	192.168.18.184	192.168.70.132	SCTP	162	DATA
20	2.360303	192.168.70.132	192.168.18.184	SCTP	62	SACK
21	2.363914	192.168.70.132	192.168.18.184	SCTP	574	DATA
22	2.364138	192.168.18.184	192.168.70.132	SCTP	62	SACK
23	4.392052	192.168.18.184	192.168.70.132	SCTP	130	DATA
24	4.397147	02:42:c0:a8:46…	Broadcast	ARP	42	Who has 192.168.70.131? Tell 192.168.70.132
25	4.397194	02:42:c0:a8:46…	02:42:c0:a8:4…	ARP	42	192.168.70.131 is at 02:42:c0:a8:46:83
26	4.397269	192.168.70.132	192.168.70.131	TCP	74	34608 → 3306 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 TSval=3849507364 TSecr=0 WS=128
27	4.397301	192.168.70.131	192.168.70.132	TCP	74	3306 → 34608 [SYN, ACK] Seq=0 Ack=1 Win=65160 Len=0 MSS=1460 SACK_PERM=1 TSval=2629468657 TSecr=3849507364 WS=128
28	4.397324	192.168.70.132	192.168.70.131	TCP	66	34608 → 3306 [ACK] Seq=1 Ack=1 Win=64256 Len=0 TSval=3849507364 TSecr=2629468657
29	4.397664	192.168.70.131	192.168.70.132	MySQL	144	Server Greeting proto=10 version=5.7.33
30	4.397700	192.168.70.132	192.168.70.131	TCP	66	34608 → 3306 [ACK] Seq=1 Ack=79 Win=64256 Len=0 TSval=3849507364 TSecr=2629468657
31	4.397900	192.168.70.132	192.168.70.131	MySQL	102	Login Request user=
32	4.397917	192.168.70.131	192.168.70.132	TCP	66	3306 → 34608 [ACK] Seq=79 Ack=37 Win=65280 Len=0 TSval=2629468658 TSecr=3849507365
33	4.399170	192.168.70.132	192.168.70.131	TLSv1.2	264	Client Hello
34	4.399184	192.168.70.131	192.168.70.132	TCP	66	3306 → 34608 [ACK] Seq=79 Ack=235 Win=65152 Len=0 TSval=2629468659 TSecr=3849507366
35	4.405605	192.168.70.131	192.168.70.132	TLSv1.2	2088	Server Hello, Certificate, Server Key Exchange, Certificate Request, Server Hello Done
36	4.405659	192.168.70.132	192.168.70.131	TCP	66	34608 → 3306 [ACK] Seq=235 Ack=2101 Win=64128 Len=0 TSval=3849507372 TSecr=2629468665
37	4.406858	192.168.70.132	192.168.70.131	TLSv1.2	171	Certificate, Client Key Exchange, Change Cipher Spec, Encrypted Handshake Message
38	4.407239	192.168.70.131	192.168.70.132	TLSv1.2	308	New Session Ticket, Change Cipher Spec, Encrypted Handshake Message
39	4.407423	192.168.70.132	192.168.70.131	TLSv1.2	266	Application Data
40	4.407544	192.168.70.131	192.168.70.132	TLSv1.2	117	Application Data
41	4.407595	192.168.70.132	192.168.70.131	TLSv1.2	102	Application Data
42	4.407649	192.168.70.131	192.168.70.132	TLSv1.2	106	Application Data
43	4.407704	192.168.70.132	192.168.70.131	TLSv1.2	184	Application Data
44	4.408151	192.168.70.131	192.168.70.132	TLSv1.2	381	Application Data
45	4.409282	192.168.70.132	192.168.70.131	TLSv1.2	215	Application Data
46	4.409584	192.168.70.131	192.168.70.132	TLSv1.2	147	Application Data
47	4.409675	192.168.70.132	192.168.70.131	TLSv1.2	176	Application Data
48	4.409848	192.168.70.131	192.168.70.132	TLSv1.2	147	Application Data

Figure 3-25 5G SA core network packets-registration an external network



No.	Time	Source	Destination	Protocol	Length	Info
49	4.410506	192.168.70.132	192.168.18.184	SCTP	606	SACK DATA
50	4.411128	192.168.18.184	192.168.70.132	SCTP	142	SACK DATA
51	4.412630	192.168.70.132	192.168.18.184	SCTP	422	SACK DATA
52	4.413379	192.168.18.184	192.168.70.132	SCTP	190	SACK DATA
53	4.415858	192.168.70.132	192.168.18.184	SCTP	1198	SACK DATA
54	4.416253	192.168.18.184	192.168.70.132	SCTP	98	SACK DATA
55	4.426276	192.168.18.184	192.168.70.132	SCTP	118	DATA
56	4.426310	192.168.70.132	192.168.18.184	SCTP	62	SACK
57	4.450305	192.168.70.132	192.168.70.131	TCP	66	34608 → 3306 [ACK] Seq=953 Ack=2911 Win=64128 Len=0 TSval=3849507417 TSecr=2629468670
58	5.195109	192.168.70.135	12.1.1.2	ICMP	98	Echo (ping) request id=0x0013, seq=1/256, ttl=64 (no response found!)
59	5.682640	192.168.70.134	192.168.70.133	PFCP	58	Sx Heartbeat Request
60	5.682802	192.168.70.133	192.168.70.134	PFCP	58	Sx Heartbeat Response
61	5.691181	192.168.70.133	192.168.70.130	TCP	74	33494 → 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 TSval=1185208059 TSecr=0 WS=128
62	5.691234	192.168.70.130	192.168.70.133	TCP	74	80 → 33494 [SYN, ACK] Seq=0 Ack=1 Win=65160 Len=0 MSS=1460 SACK_PERM=1 TSval=2933092549 TSecr=1185208059 WS=128
63	5.691253	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [ACK] Seq=1 Ack=1 Win=64256 Len=0 TSval=1185208059 TSecr=2933092549
64	5.691318	192.168.70.133	192.168.70.130	HTTP	292	PATCH /nnrf-nfm/v1/nf-instances/2e0a241f-132d-44c6-be3d-027cd15d5b20 HTTP/1.1 (application/json)
65	5.691330	192.168.70.130	192.168.70.133	TCP	66	80 → 33494 [ACK] Seq=1 Ack=227 Win=65024 Len=0 TSval=2933092549 TSecr=1185208059
66	5.691651	192.168.70.130	192.168.70.133	HTTP	163	HTTP/1.1 204 No Content
67	5.691672	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [ACK] Seq=227 Ack=98 Win=64256 Len=0 TSval=1185208059 TSecr=2933092549
68	5.691745	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [FIN, ACK] Seq=227 Ack=98 Win=64256 Len=0 TSval=1185208059 TSecr=2933092549
69	5.691788	192.168.70.130	192.168.70.133	TCP	66	80 → 33494 [FIN, ACK] Seq=98 Ack=228 Win=65024 Len=0 TSval=2933092550 TSecr=1185208059
70	5.691802	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [ACK] Seq=228 Ack=99 Win=64256 Len=0 TSval=1185208060 TSecr=2933092550
71	6.214273	192.168.70.135	12.1.1.2	ICMP	98	Echo (ping) request id=0x0013, seq=2/512, ttl=64 (no response found!)
72	6.417229	192.168.18.184	192.168.70.132	SCTP	146	DATA
73	6.419305	02:42:c0:a8:46…	Broadcast	ARP	42	Who has 192.168.70.130? Tell 192.168.70.132
74	6.419363	02:42:c0:a8:46…	02:42:c0:a8:4…	ARP	42	192.168.70.130 is at 02:42:c0:a8:46:82
75	6.419377	192.168.70.132	192.168.70.130	TCP	74	40266 → 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 TSval=1319218380 TSecr=0 WS=128
76	6.419421	192.168.70.130	192.168.70.132	TCP	74	80 → 40266 [SYN, ACK] Seq=0 Ack=1 Win=65160 Len=0 MSS=1460 SACK_PERM=1 TSval=519847797 TSecr=1319218380 WS=128
77	6.419446	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [ACK] Seq=1 Ack=1 Win=64256 Len=0 TSval=1319218380 TSecr=519847797
78	6.419566	192.168.70.132	192.168.70.130	TCP	217	40266 → 80 [PSH, ACK] Seq=1 Ack=1 Win=64256 Len=151 TSval=1319218380 TSecr=519847797 [TCP segment of a reassembled PDU]
79	6.419584	192.168.70.130	192.168.70.132	TCP	66	80 → 40266 [ACK] Seq=1 Ack=152 Win=65024 Len=0 TSval=519847797 TSecr=1319218380
80	6.420017	192.168.70.130	192.168.70.132	HTTP	862	HTTP/1.1 200 OK (application/json)
81	6.420063	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [ACK] Seq=152 Ack=797 Win=64128 Len=0 TSval=1319218381 TSecr=519847798
82	6.420230	192.168.70.132	192.168.70.130	HTTP	66	GET /nnrf-disc/v1/nf-instances?target-nf-type=SMF&requester-nf-type=AMF HTTP/1.1
83	6.420276	192.168.70.130	192.168.70.132	TCP	66	80 → 40266 [FIN, ACK] Seq=797 Ack=153 Win=65024 Len=0 TSval=519847798 TSecr=1319218381
84	6.420291	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [ACK] Seq=153 Ack=798 Win=64128 Len=0 TSval=1319218381 TSecr=519847798
85	6.421519	02:42:c0:a8:46…	Broadcast	ARP	42	Who has 192.168.70.133? Tell 192.168.70.132
86	6.421575	02:42:c0:a8:46…	02:42:c0:a8:4…	ARP	42	192.168.70.133 is at 02:42:c0:a8:46:85

Figure 3-26 5G SA core network packets-connection with the external network



No.	Time	Source	Destination	Protocol	Length	Info
49	4.410506	192.168.70.132	192.168.18.184	NGAP/NAS-5GS	606	DownlinkNASTransport, Authentication request
50	4.411128	192.168.18.184	192.168.70.132	NGAP/NAS-5GS	142	UplinkNASTransport, Authentication response
51	4.412630	192.168.70.132	192.168.18.184	NGAP/NAS-5GS	422	DownlinkNASTransport
52	4.413379	192.168.18.184	192.168.70.132	NGAP/NAS-5GS	190	UplinkNASTransport
53	4.415858	192.168.70.132	192.168.18.184	NGAP/NAS-5GS	1198	InitialContextSetupRequest
54	4.416253	192.168.18.184	192.168.70.132	NGAP	98	InitialContextSetupResponse
55	4.426276	192.168.18.184	192.168.70.132	NGAP/NAS-5GS	118	UplinkNASTransport
56	4.426310	192.168.70.132	192.168.18.184	SCTP	62	SACK
57	4.450305	192.168.70.132	192.168.70.131	TCP	66	34608 → 3306 [ACK] Seq=953 Ack=2911 Win=64128 Len=0 TSval=384…
58	5.195109	192.168.70.135	12.1.1.2	ICMP	98	Echo (ping) request id=0x0013, seq=1/256, ttl=64 (no respons…
59	5.682640	192.168.70.134	192.168.70.133	PFCP	58	PFCP Heartbeat Request
60	5.682802	192.168.70.133	192.168.70.134	PFCP	58	PFCP Heartbeat Response
61	5.691181	192.168.70.133	192.168.70.130	TCP	74	33494 → 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 T…
62	5.691234	192.168.70.130	192.168.70.133	TCP	74	80 → 33494 [SYN, ACK] Seq=0 Ack=1 Win=65160 Len=0 MSS=1460 SA…
63	5.691253	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [ACK] Seq=1 Ack=1 Win=64256 Len=0 TSval=1185208059…
64	5.691318	192.168.70.133	192.168.70.130	HTTP	292	PATCH /nnrf-nfm/v1/nf-instances/2e0a241f-132d-44c6-be3d-027cd…
65	5.691330	192.168.70.130	192.168.70.133	TCP	66	80 → 33494 [ACK] Seq=1 Ack=227 Win=65024 Len=0 TSval=29330925…
66	5.691651	192.168.70.130	192.168.70.133	HTTP	163	HTTP/1.1 204 No Content
67	5.691672	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [ACK] Seq=227 Ack=98 Win=64256 Len=0 TSval=1185208…
68	5.691745	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [FIN, ACK] Seq=227 Ack=98 Win=64256 Len=0 TSval=11…
69	5.691788	192.168.70.130	192.168.70.133	TCP	66	80 → 33494 [FIN, ACK] Seq=98 Ack=228 Win=65024 Len=0 TSval=29…
70	5.691802	192.168.70.133	192.168.70.130	TCP	66	33494 → 80 [ACK] Seq=228 Ack=99 Win=64256 Len=0 TSval=1185208…
71	6.214273	192.168.70.135	12.1.1.2	ICMP	98	Echo (ping) request id=0x0013, seq=2/512, ttl=64 (no respons…
72	6.417229	192.168.18.184	192.168.70.132	NGAP/NAS-5GS	146	UplinkNASTransport
73	6.419305	02:42:c0:a8:46:84	Broadcast	ARP	42	Who has 192.168.70.130? Tell 192.168.70.132
74	6.419363	02:42:c0:a8:46:82	02:42:c0:a8:46:84	ARP	42	192.168.70.130 is at 02:42:c0:a8:46:82
75	6.419377	192.168.70.132	192.168.70.130	TCP	74	40266 → 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 T…
76	6.419421	192.168.70.130	192.168.70.132	TCP	74	80 → 40266 [SYN, ACK] Seq=0 Ack=1 Win=65160 Len=0 MSS=1460 SA…
77	6.419446	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [ACK] Seq=1 Ack=1 Win=64256 Len=0 TSval=1319218380…
78	6.419566	192.168.70.132	192.168.70.130	HTTP	217	GET /nnrf-disc/v1/nf-instances?target-nf-type=SMF&requester-n…
79	6.419584	192.168.70.130	192.168.70.132	TCP	66	80 → 40266 [ACK] Seq=1 Ack=152 Win=65024 Len=0 TSval=51984779…
80	6.420017	192.168.70.130	192.168.70.132	HTTP	862	HTTP/1.1 200 OK (application/json)
81	6.420063	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [ACK] Seq=152 Ack=797 Win=64128 Len=0 TSval=131921…
82	6.420230	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [FIN, ACK] Seq=152 Ack=797 Win=64128 Len=0 TSval=1…
83	6.420276	192.168.70.130	192.168.70.132	TCP	66	80 → 40266 [FIN, ACK] Seq=797 Ack=153 Win=65024 Len=0 TSval=5…
84	6.420291	192.168.70.132	192.168.70.130	TCP	66	40266 → 80 [ACK] Seq=153 Ack=798 Win=64128 Len=0 TSval=131921…
85	6.421519	02:42:c0:a8:46:84	Broadcast	ARP	42	Who has 192.168.70.133? Tell 192.168.70.132
86	6.421575	02:42:c0:a8:46:85	02:42:c0:a8:46:84	ARP	42	192.168.70.133 is at 02:42:c0:a8:46:85
87	6.421593	192.168.70.132	192.168.70.133	TCP	74	47636 → 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 SACK_PERM=1 T…

Figure 3-27 Communication with NAS message

SCTP places control messages and information into separate chunks, each identified by a track header. The protocol can fragment a message into multiple pieces of data, but each piece of data contains data from a single user message. SCTP groups tracks into SCTP packets.

- **PDU session Establishment**

Once a connection between AMF and database has been reached, the user device communicates with NAS messages with the core network. Thus, we have an AMF and UE messages for direct certification through the NAS protocol and especially in the case of a 5GSM protocol to manage PDU and QoS sessions for the user

plane. From the packet 59 and then, NRF can be communicated with SMF. SMF receives UE subscription data (usually from UDM and PCF) and formulates a PFCP (Packet Forward Control Packet) installation request to schedule UPF to create a session management environment (ie, PDU Session) for UE.

Figure 3-28 shows that SMF uses PFCP via the N4 interface to create a session management (SM) element over UPF for the UE PDU Session. The PFCP Session Establishment Request Message Pack includes the Information Elements (IEs) to Sort UE, Tail, Programming and mapping/monitoring. After deployment the SM context environment in UPF, a similar session environment must also be deploying at gNB and UE respectively to configure the UE PDU Session and the default QoS Flow (i.e., one end-to-end PDU Session, from UE, gNB to UPF). So it must also contain the messages N1 and N2 to set the SM frames for the UE and gNB respectively. Figure 3-29shows the N2 message from SMF to AMF with AMF formulating the QoS profile in gNB. From packets analysis, it is obvious that the SMF advises the gNB through the interface N2 HTTP1.1 with all the necessary QoS information elements for the installation of the UE PDU session. This is designed to give the gNB an independent QoS decision to set up wireless transmission to extend the UE N3 GTP-U tunnel to the gNB. This separation of QoS control between access and core networks allows the 5GC to support different wired and wireless access networks with very different QoS capabilities and features.

In Figure 3-29 it is observed that UL and DL UE-AMBR (aggregate maximum bit-rate) are 20Mbps and 22Mbps. That is, gNB reduces any AMBR traffic other than GBR for UE above 20Mbps UL, and 22Mbps DL. An in-depth analysis of the packages reveals further information such as in which IP is GTP-Tunnel in UPF, as well as GTP-TEID for UE. This information is important for gNB to promote UL UE traffic to UPF for data network (DN). There is also QFI information that recognizes the default QoS Flow from UE to DN. Additionally describes the QoS attribute (e.g. best effort without GBR) of the default QoS stream in the UE's PDU session.

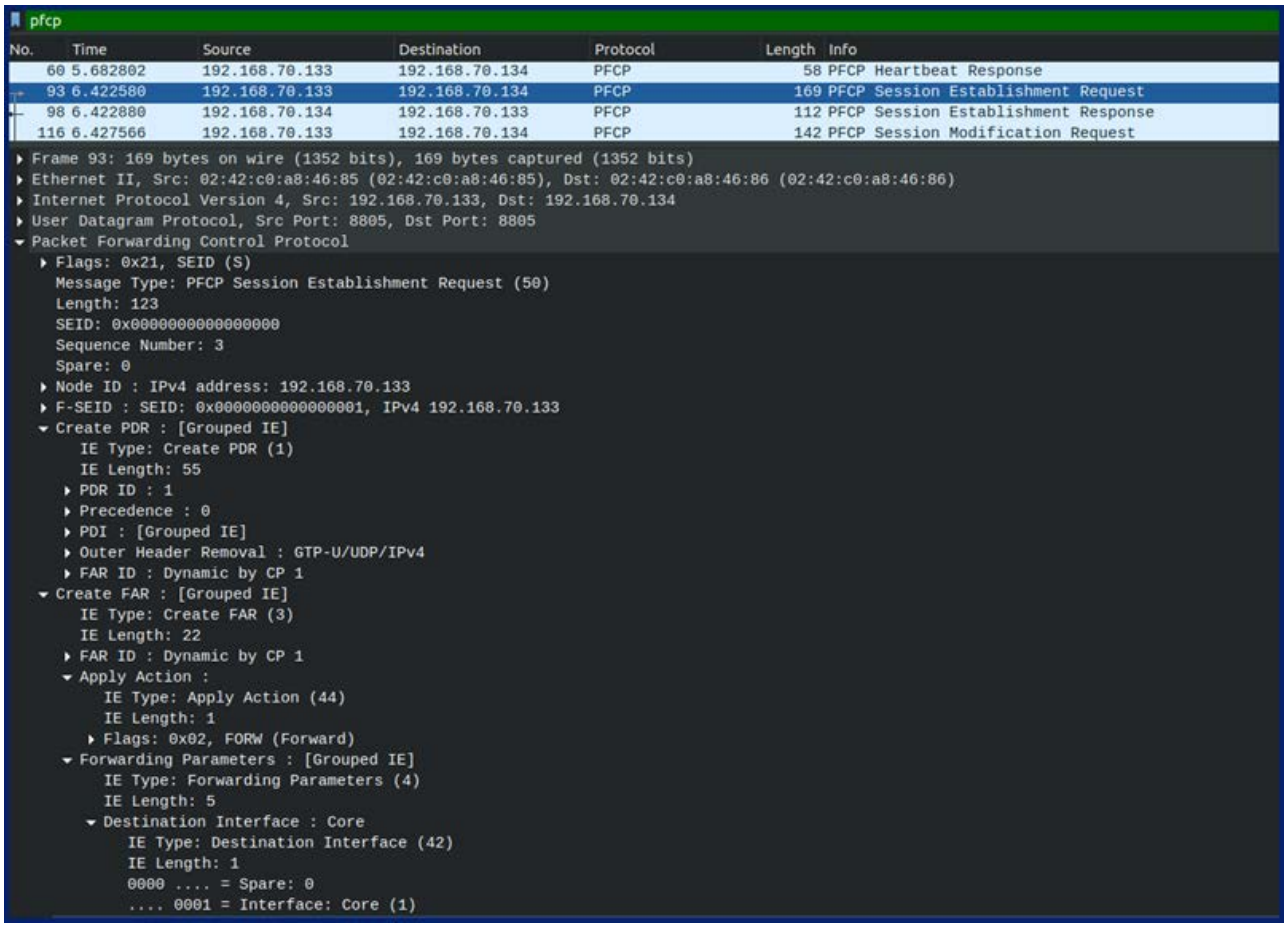


Figure 3-28 SMF to UPF-PFCP session establishment

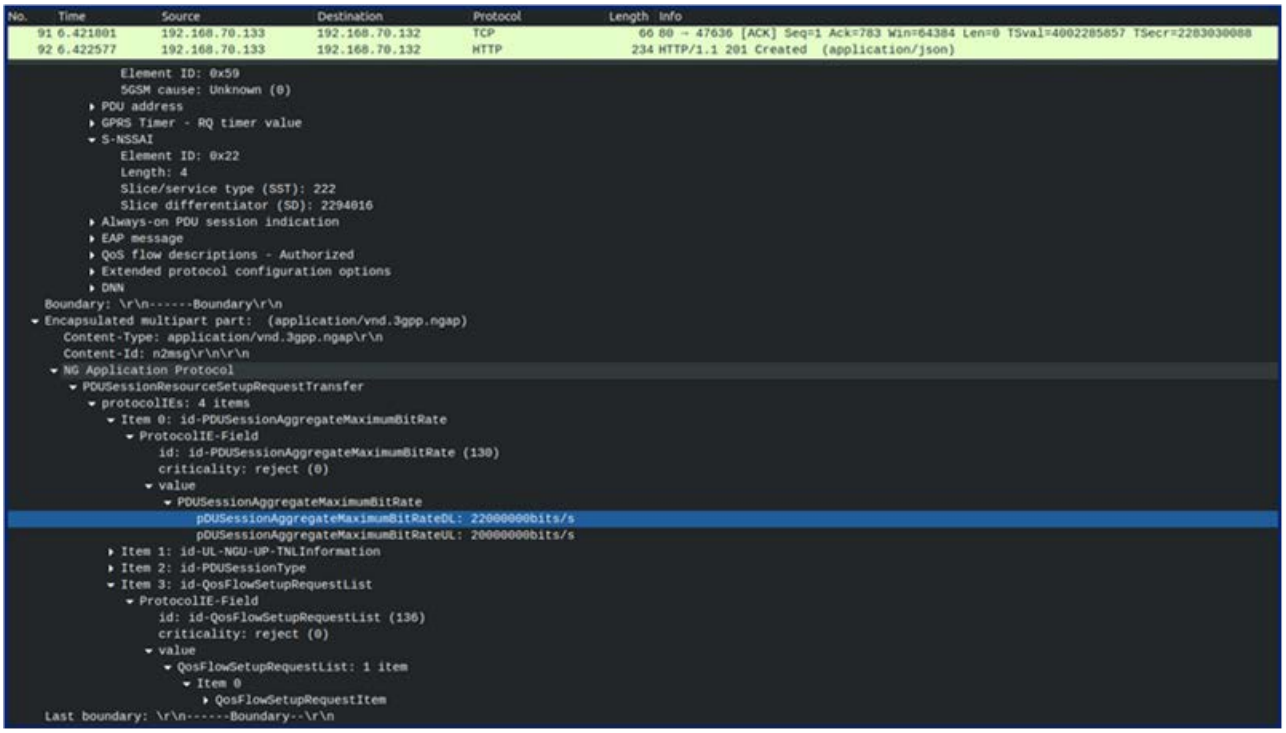


Figure 3-29 N2 message from SMF to AMF for the gNB

There is additional information for the selection of Network Slicing, followed by an integer that numbers the packet.

Following the above, the UE can now send its first UL packet to the DN because:

- UE has a default QoS stream to match the UL traffic in DRB (Data Radio Bearer) specified by QFI to gNB,
- The gNB knows the IP address of the UPF and the TEID of the UE from the Service/Slice Template message to forward the UL traffic to the UPF and DN.

When connecting, we used the ping terminal tool that sends ICMP Echo Request messages to another terminal or computer in general, to confirm the existence, connection, and time it takes for the message to reach that computer.

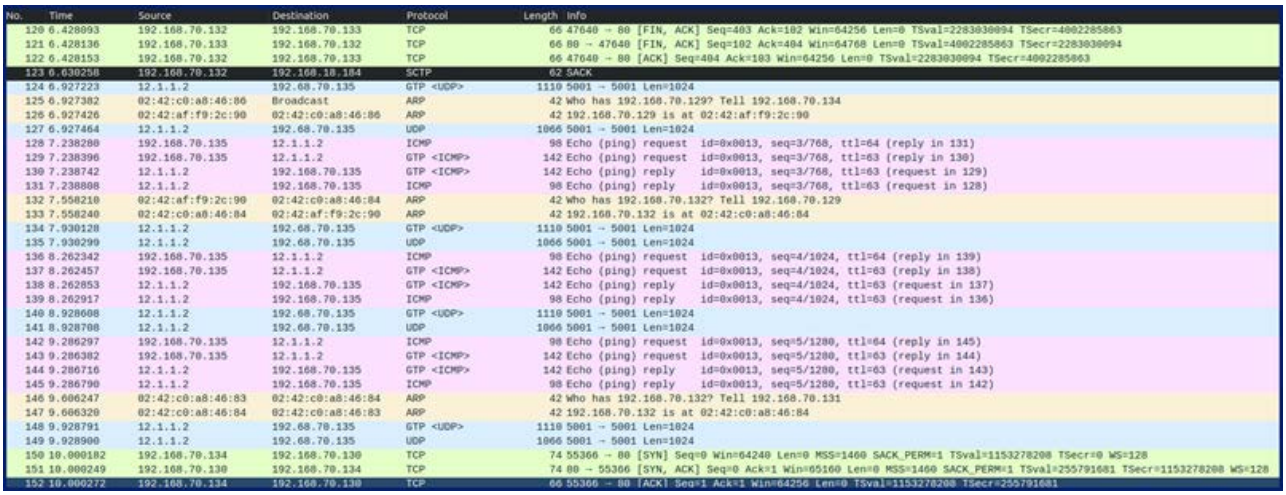


Figure 3-30 Using ping tool for connection validation

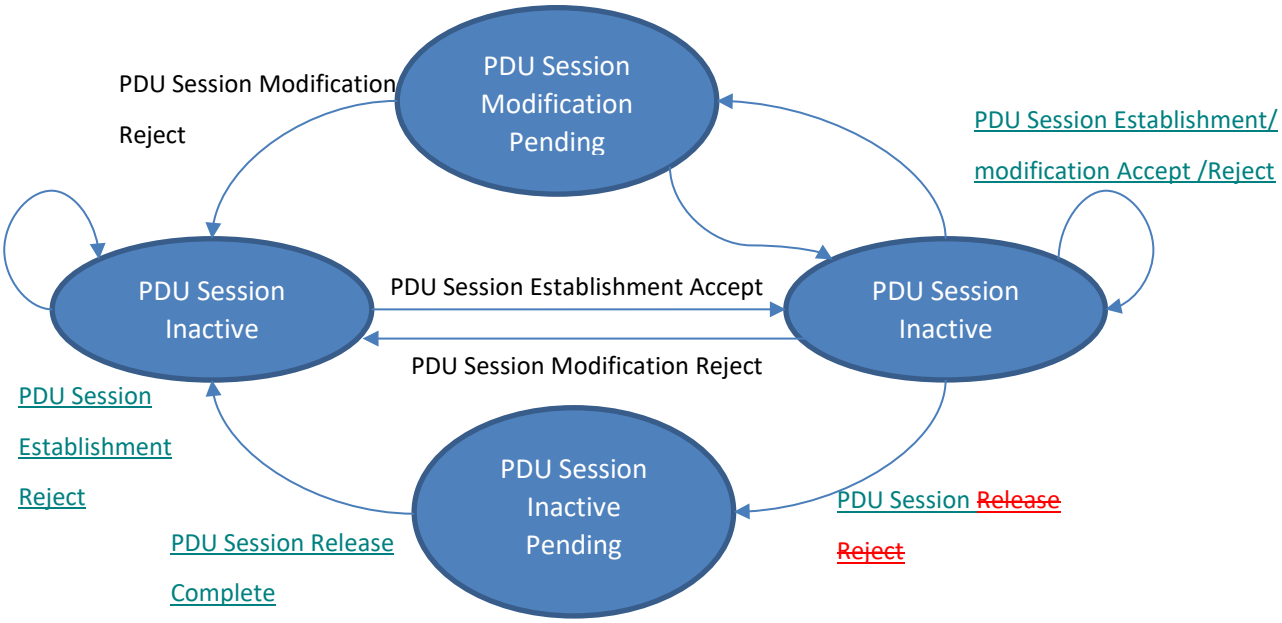


Figure 3-31 States of the PDU session establishment process

From package 124 onwards, the GTP protocol undertakes the data handling. GTP-U uses a tunneling mechanism to transmit user data traffic and traverses UDP or ICMP transport as appropriate and is identified and recognized at each node with an IP address and corresponding port.

In 5GS, GTP-U has been reused to transfer UP data via N3 and N9 (and N4) interfaces, since, as mentioned before, tunnel ID management and other controls used HTTP1.1 and NGAP.

GTP-U tunnels are deployed by providing GTP-U TEIDs and IP addresses between (R)AN and SMF. This signal is transmitted by HTTP1.1 between SMF and AMF and by NGAP between AMF and (R)AN. Therefore, there is no use of GTP-C in 5GC to manage GTP-U tunnels. Finally, we observe that a path is used for the GTP <ICMP> and GTP <UDP> tunnels with the TEID in the GTP-U header indicating to which tunnel a particular payload belongs each time.

Once the individual procedures of the control plane processes have been determined, the performance of the overall system can be modeled using Petri Nets. An example for the PDU session establishment process is given in Figure 3-31.

3.2.5 User Plane modelling

We consider a converged wired/wired infrastructure hosting specific 3GPP slices as shown in Figure 3-32 [1]. Each 3GPP slice can be modelled as a network of queues. In order to mathematically formulate this network, the Physical Infrastructure (PI) is modeled as an open queuing network, in which its node $n \in \mathcal{N}^p$ has m_n service modules (in the wireless access domain, m_n corresponds to the number of input queues at an eNodeB, while in the optical domain it corresponds to the number of receiver/transmitter queues in the edge node) with service rate μ_n. Due to the uncertainties introduced in these environments, we consider the general case where the inter-arrival times of the demands are not necessarily exponentially distributed. Assuming that:

- the external arrival process of the demands is any renewal process with mean inter-arrival time $1/\lambda_i$ and coefficient of variation σ_{Ai}. Both parameters are estimated based on the Autoregressive moving average model,

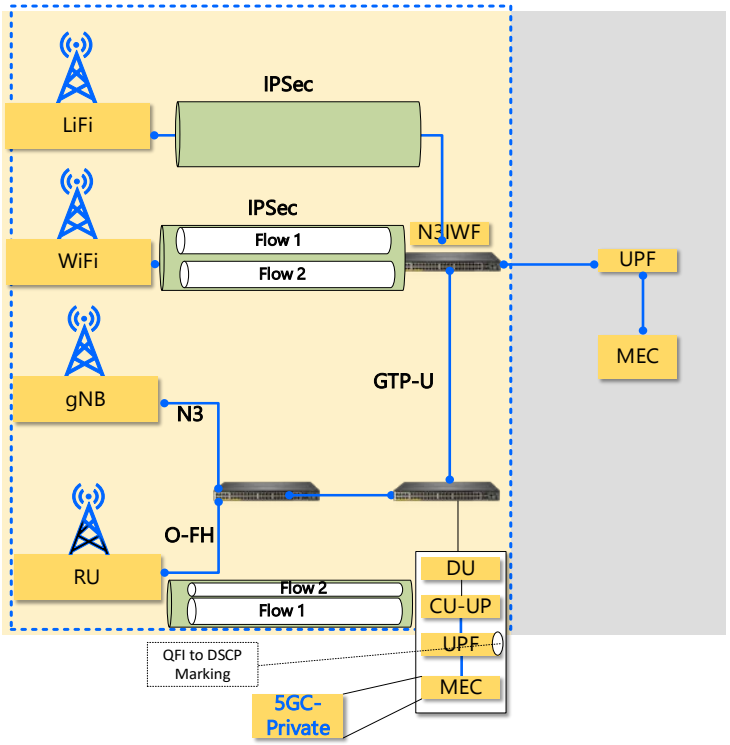


Figure 3-32 Example of a physical network topology

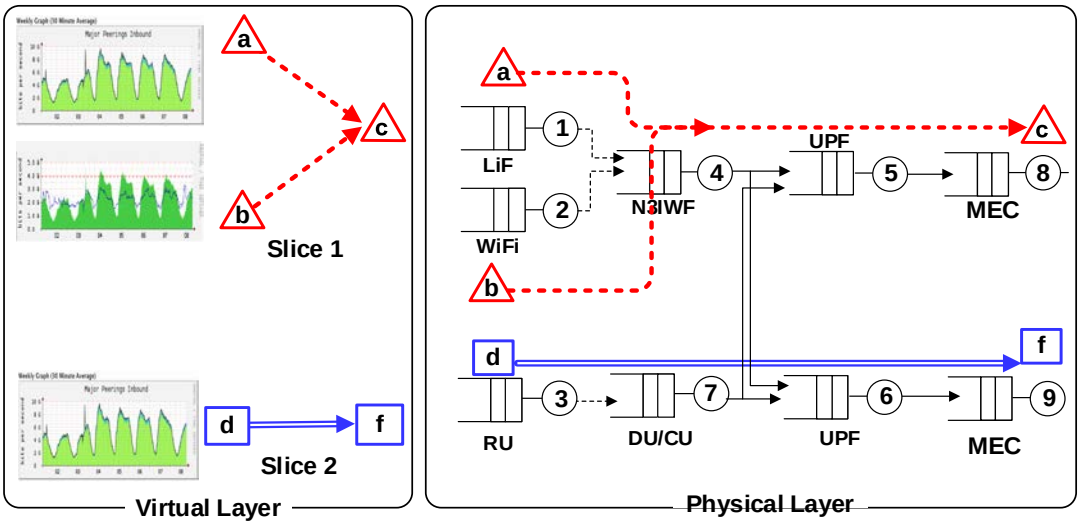


Figure 3-33 Toy Prediction of the network traffic and mapping of the requested service slice resources onto the multi-queuing model of the converged architecture presented in Figure 3-32

- the service times at the nth node of the physical infrastructure can follow any distribution with mean service time $\frac{1}{\mu_n}$ and coefficient of variation σ_{Bn} and,
- the demands are served according to the FIFO policy.

A closed form approximation for the end-to-end delay for the services that are provided by each virtual infrastructure (VI) can be extracted after applying the *method of decomposition* [32][33][34].

This method is based on the following steps [33] :

1. The arrival rate, λ_{ni}, and the utilization, ρ_{ni}, for the demands of VI_i at the nth node of the PI are

calculated

2. Once the λ_{ni}, ρ_{ni} have been determined, the coefficient of variation of the interarrival times at each node n, namely σ_{Ani}, using an iterative process that consists of three phases:
   a. *Merging Phase*: Traffic requests that arrive at each node are merged into a single arrival process.

      σ_{Ani} and λ_{ni} can be estimated using a plethora of approximation formulas. In the present work, the Decomposition of Pujolle [33]has been adopted.
   b. *Flow Phase*: The coefficient of variation for the inter-departure times at each node are estimated using as input the coefficient of variation of the interarrival times σ_{Ani} as well as the coefficient of variation for the service times.
   c. *Splitting Phase*: In this phase, the served demands are forwarded to the subsequent nodes for processing.
3. In the final step, using as input the parameters σ_{Ani} and λ_{ni}, the mean queue length and, consequently, the average waiting time per node can be evaluated using the well-known formulas for $\frac{GI}{\frac{G}{m}}$ e.g. [33]:

$$\bar{W}_{ni} = \frac{\frac{P_{m_{ni}}}{\mu_{ni}}}{1-\rho_{ni}} \frac{\sigma_{Ani}^2 + \sigma_{Bni}^2}{2m_{ni}}$$

where

$$P_{m_{ni}} = \frac{(m_{ni}\rho_{ni})^{m_{ni}}}{m_{ni}!\,(1-\rho_{ni})}\pi_0$$

Based on these results, end to end latency and throughput can be derived. The relevant analysis can be conducted under various network settings with emphasis on resilience and mobility management. To handle resilience, architectures offering 1:1 protection for the RAN and the transport can be considered.

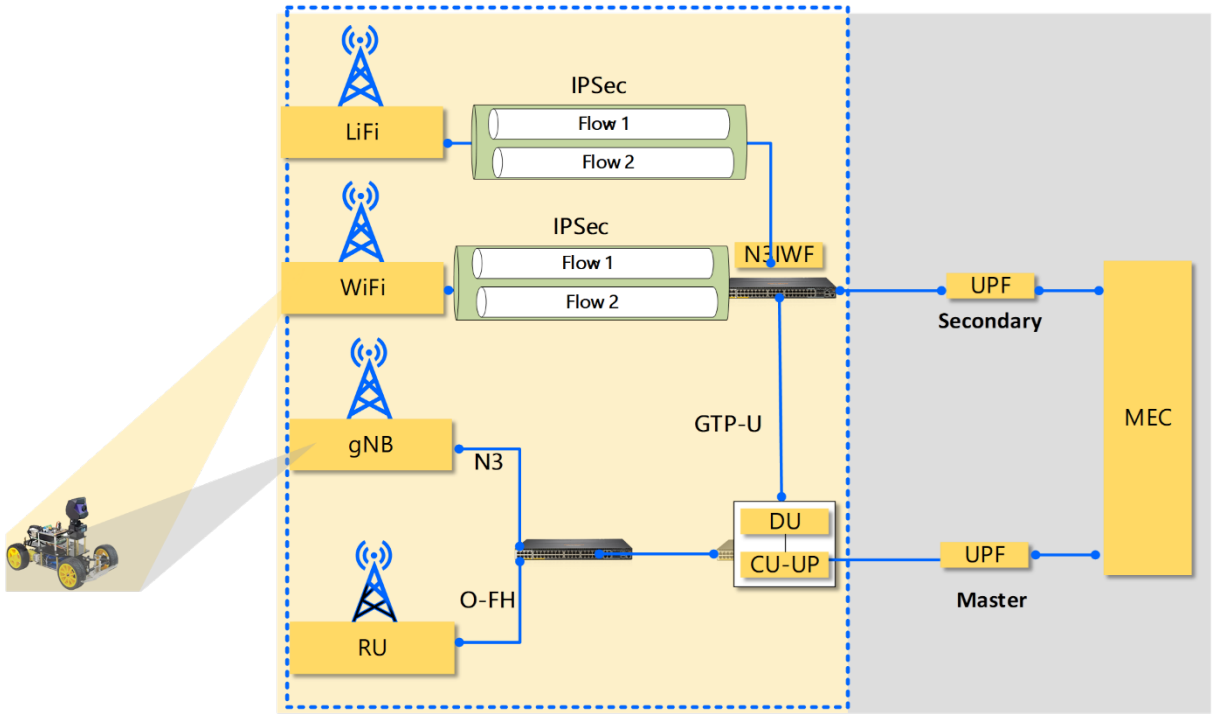


Figure 3-34 Example scenario for E2E redundant User Plane paths using dual connectivity

Figure 3-34 illustrates an example user plane resource configuration of dual PDU sessions when redundancy is applied. One PDU Session spans from the UE via gNB to the master UPF1 acting as the PDU Session Anchor, and the other PDU Session spans from the UE via Secondary Non-3GPP access to a secondary UPF2 acting as the PDU Session Anchor.

Based on these two PDU Sessions, two independent user plane paths are set up. Master and secondary UPFs connect to the same DN, even though the traffic via the two UPFs may be routed via different user plane nodes within the DN. The E2E analysis will be carried out under different availability levels with the objective to determine the tradeoffs between resource efficiency and availability.

For this case, the emphasis will be given on the evaluation of the ATSSS feature of 5G-CLARITY. The ATSSS feature enables a multi-access PDU Connectivity Service, which can exchange PDUs between the UE and a data network by simultaneously using one 3GPP access network and one non-3GPP access network and two independent N3/N9 tunnels between the PSA and RAN/AN. The multi-access PDU Connectivity Service is realized by establishing a multi-access PDU (MA PDU) Session, i.e., a PDU Session that may have user-plane resources on two access networks.

The UE may request a MA PDU Session when the UE is registered via both 3GPP and non-3GPP accesses, or when the UE is registered via one access only. After the establishment of a MA PDU Session, and when there are user-plane resources on both access networks, the UE applies network-provided policy (i.e., ATSSS rules) and considers local conditions (such as network interface availability, signal loss conditions, user preferences, etc.) for deciding how to distribute the uplink traffic across the two access networks.

Similarly, the UPF anchor of the MA PDU Session applies network-provided policy (i.e., N4 rules) and feedback information received from the UE via the user-plane (such as access network Unavailability or Availability) for deciding how to distribute the DL traffic across the two N3/N9 tunnels and two access networks. When there are user-plane resources on only one access network, the UE applies the ATSSS rules and considers local conditions for triggering the establishment or activation of the user plane resources over another access.

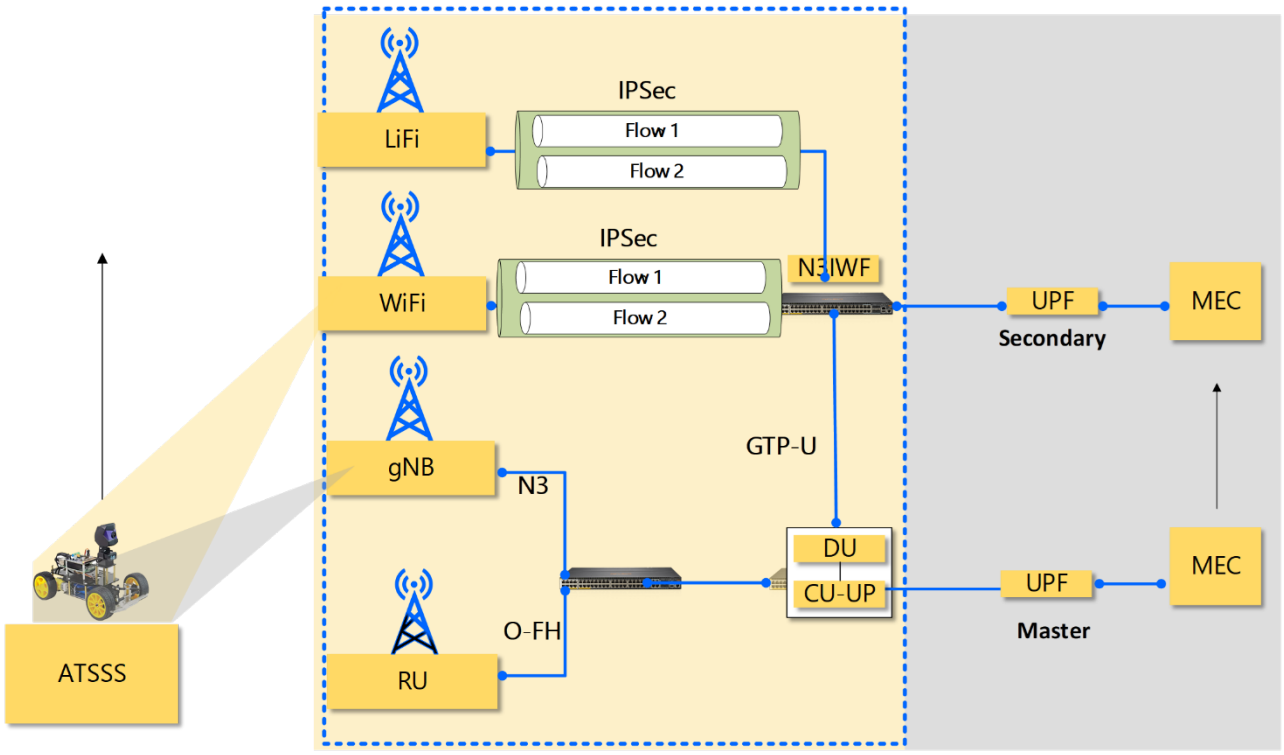


Figure 3-35 Example of non-roaming and roaming with local breakout architecture for ATSSS support

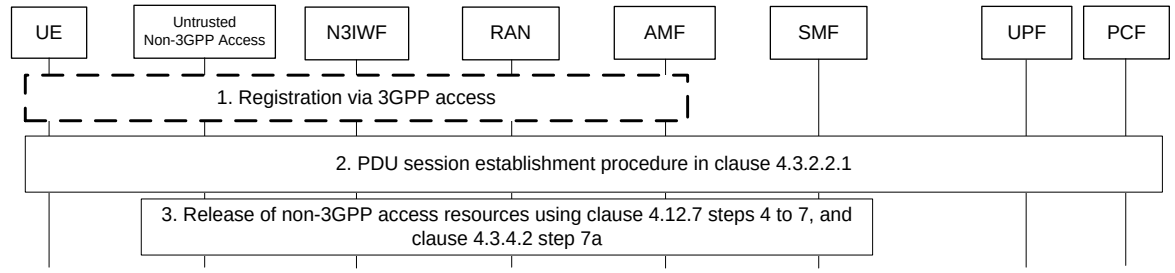


Figure 3-36 Handover of a PDU Session procedure from untrusted non-3GPP access to 3GPP access (non-roaming and roaming with local breakout)

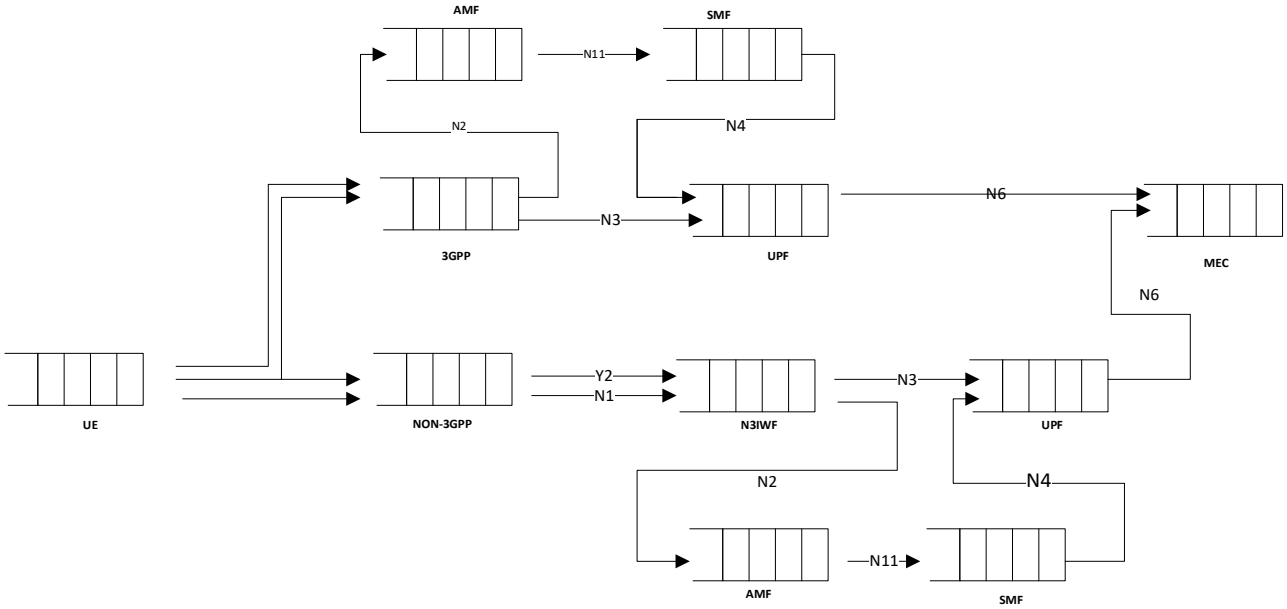


Figure 3-37 Network of queues for the hybrid 3gpp-non-3gpp system

In addition to this option, the 5G-CLARITY system architecture will be evaluated under other protection schemes using redundant transmission on N3/N9 interfaces. The second set of evaluation results will be focused mobility management considering the 5G system architecture shown in Figure 3-35.

The set of results will evaluate the 5G-CLARITY system architecture in terms of E2E delay and throughput under mobility. In this case, handovers between 3GPP and untrusted non-3GPP access should be performed. The exact process described in Figure 3-36 is modelled as a stochastic process where during handover a set of PDU session establishment and resource release processes are instantiated.

The components involved in the handover processes are modeled as a network of queues (see Figure 3-37) where at each interface a specific traffic pattern (arrival/service process) is generated.

3.2.6 Modelling of the SDN controller northbound interface

The SDN applications view the state of resources in the data plane of HetNet shown in Figure 3-2 through the northbound interface of the control plane. They offer services via the centralized SDN controller by requesting to set up their forwarding rules in the data plane APs through the southbound Open-Flow protocol. However, the applications may renege on accessing the data plane, because of resource unavailability or controller processing capacity constraint. This reneging process influences the performance of network services provided throughout the HetNet data plane. Based on the HetNet system model shown in Figure 3-2, the SDN applications generate requests following a Poisson process with arrival rate, λ_n . The

M/M/1 queuing model describes the packets buffering and processing at the northbound interface of the controller. The M/M/1 retrial queueing system model with geometric loss and feedback [42] is adopted to model the requests processing for data plane access through the northbound interface, as shown in Figure 3-2. The state of the SDN-enabled HetNet is described by a pair $\left(\zeta(t), N(t)\right)$, where $\zeta(t)$ denotes the number of busy controllers and $N(t)$ denotes the number of requests in the retrial buffer at time t. A stochastic process $\left(\left(\zeta(t), N(t)\right): t \geq 0\right)$ is formed as a time-homogeneous Markov process with a state space (ζ, N) as a limiting variable of $\left(\zeta(t), N(t)\right)$. When the controller receives a small number of requests from the network applications to access the data plane, the average queue length of the controller is given as follows [39][42]:

$$E(N:\zeta=0) = C\ \frac{\alpha(\lambda_c+v)+\beta\mu_c}{v}\ F\left(\frac{\alpha(\lambda_c+2v)+\beta\mu_c}{\alpha v};\frac{\overline{\alpha}(\lambda_c+v)+\overline{\beta}\mu_c}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right) - \alpha F\left(\frac{\alpha(\lambda_c+v)+\beta\mu_c}{\alpha v};\frac{\overline{\alpha}(\lambda_c+v)+\overline{\beta}\mu_c}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right) - \frac{\alpha\lambda_c(\lambda_c+v)+\beta\lambda_c\mu_c}{v\left(\overline{\alpha}(\lambda_c+v)+\mu_{c\overline{\beta}}\right)}$$

$$\boldsymbol{F}\left(\frac{\alpha(\lambda_c+2v)+\beta\mu_c}{\alpha v};\frac{\overline{\alpha}(\lambda_c+2v)+\overline{\beta}\mu_c}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right) \quad (3)$$

where $F(a;b;w)$ is the Kummer's function; and C is a normalizing constant given by [39][42]:

$$\boldsymbol{C} = \left(\frac{\mu_{c\overline{\beta}}+\lambda_{c\overline{\alpha}}}{\lambda_c}\boldsymbol{F}\left(\frac{\alpha(\lambda_c+v)+\beta\mu_c}{\alpha v};\frac{\lambda_{c\overline{\alpha}}+\mu_{c\overline{\beta}}}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right)\right)^{-1} \quad (4)$$

However, when the number of requests generated from the SDN applications increases, the controller may become busy on its northbound and southbound interfaces. In this case, the average queue length of the northbound interface is expressed as follows [39][42]:

$$\boldsymbol{E(N:\zeta=1)} = \boldsymbol{C}\cdot\frac{\alpha\lambda_c(\lambda_c+v)+\beta\lambda_c\mu_c}{\left(\overline{\alpha}v(\lambda_c+v)+\mu_{c\overline{\beta}}v\right)}\boldsymbol{F}\left(\frac{\alpha(\lambda_c+2v)+\beta\mu_c}{\alpha v};\frac{\overline{\alpha}(\lambda_c+2v)+\overline{\beta}\mu_c}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right) \quad (5)$$

When the SDN controller runs beyond its capacity, the network is down or resources in the data plane are overbooked, the excess requests generated from the SDN applications are kept in the retrial queue or are dropped. In this case, the average length of the retrial queue is expressed as follows [42][39] :

$$E(N) = E(N:\zeta=0) + E(N:\zeta=1) = C\left[\frac{\alpha(\lambda_c+v)+\beta\mu_c}{v}\right] F\left(\frac{\alpha(\lambda_c+2v)+\beta\mu_c}{\alpha v};\frac{\overline{\alpha}(\lambda_c+2v)+\overline{\beta}\mu_c}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right) - \boldsymbol{\alpha F}\left(\frac{\alpha(\lambda_c+v)+\beta\mu_c}{\alpha v};\frac{\overline{\alpha}(\lambda_c+v)+\overline{\beta}\mu_c}{\overline{\alpha}v};\frac{\alpha\lambda_c}{\overline{\alpha}v}\right) \quad (6)$$

Analytical models have been developed in MATLAB, which evaluate (3), (5) and (6). They give indications regarding the relationship among the applications requests, queue length of the northbound interface, retrial queue length and controller processing rate. Based on (3), the request retrial probability, α, influences the retrial queue length more than the re-joining probability, β, as shown in Figure 3-38. This is attributed to the fact that not all the re-joining requests wait in the retrial queue length. The applications, which are allocated resources in the previous round, are quickly served by the controller. Based on (3), the controller service rate, μ_c , directly influences the retrial queue length. This even grows more rapidly with the increase of the requests' retrial probability, α, as shown in Figure 3-39. This also happens when the controller is busy or requested resources in the network data plane become unavailable during the access time of applications.

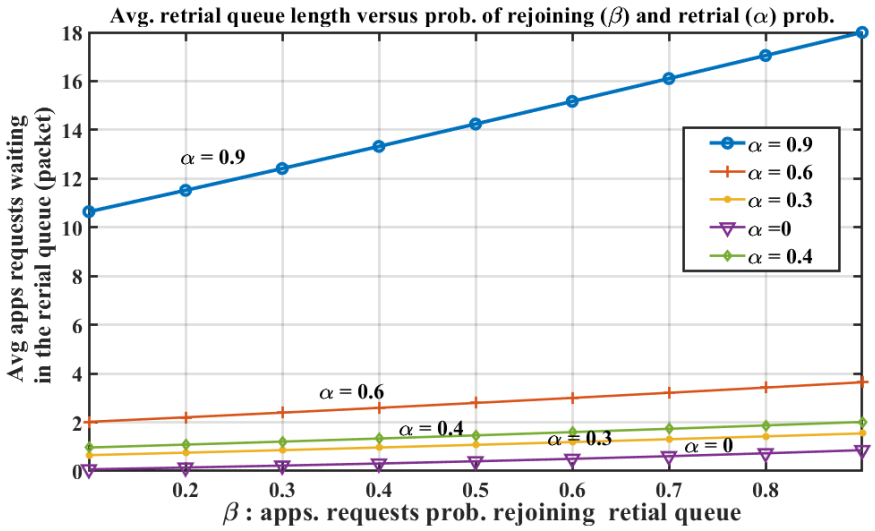


Figure 3-38 Average retrial queue length versus β

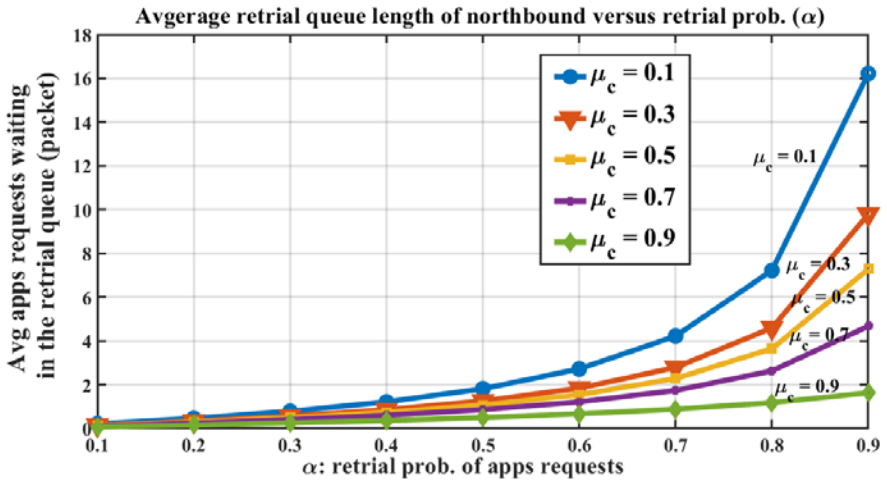


Figure 3-39 Average retrial queue length versus α

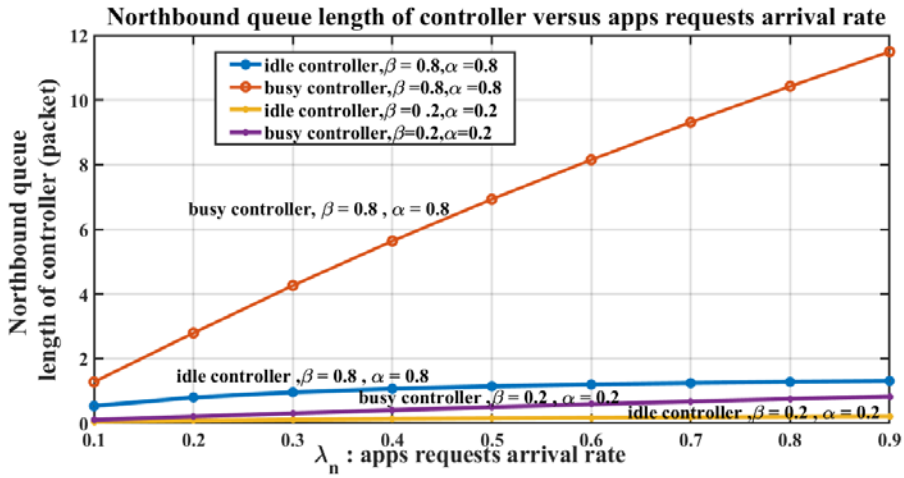


Figure 3-40 Average queue length of northbound controller versus λ_n

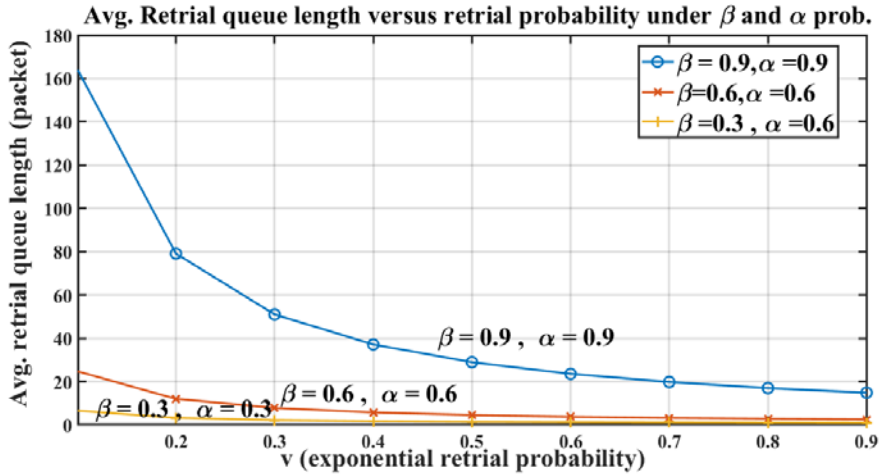


Figure 3-41 Average retrial queue length versus v

This emphasizes the importance of the controller processing rate and requested resource availability for applications to provide reliable services in the HetNet data plane. Based on (6), the retrial and re-joining probabilities significantly impact the queue length of the northbound interface, as shown in Figure 3-40.

When the exponential probability, v, of sending requests to the controller increases, it is obvious to see that the retrial queue length decreases. However, based on (5) and (6), the queue length of the northbound interface still depends on the controller processing rate, retrial, and re-joining probabilities, as shown in Figure 3-41. When the re-joining and retrial requests decrease, the average retrial queue length decreases as well and vice-versa. This case indicates that the controller can manage the allocation of resources that are requested by the different SDN applications. New flow rules can also be set in the APs based on the retrial and re-joining probabilities range.

3.2.7 Positioning system

The 5G-CLARITY localization system uses 4 different WATs for UE localization. These include:

- A sub-6 GHz proprietary localization system supporting DL/UL time difference of arrival (DL/UL-TDoA),
- mmWave localization system working in the 60 GHz ISM band and using two-way ranging (TWR),
- LiFi positioning supporting range-based localization as well as fingerprinting-based localization,
- Optical camera communication (OCC) – visible light positioning (VLP).

All of these WATs require installing of APs with known coordinates (called also anchor nodes) in order to enable the localization service as well as wireless data communications. Obtaining the optimal positions of the APs, as well as testing the system, depending on the 4 above mentioned WATs, requires modelling of each of this system separately. This is especially important in the initial phase, where not all WATs are available for real deployment, or they are not fully functional.

A simplified localization architecture developed within 5G-CLARITY is shown in Figure 3-42. All of the WATs, used for localization are connected to a localization server. The server collects all of the localization relevant data from each WAT, in order to perform a position estimate for the given UE. Not all of the WATs shown in this figure would be included in the final system. In 5G-CLARITY, a simulation environment for simulating the different WAT positioning technologies was developed. Each WAT positioning technology model architecture is shown in Figure 3-43. This is a general architecture and it is independent of the underlying WAT. These models were developed within 5G-CLARITY and are written in Python. They are used for simulation of the positioning technologies and for testing of the localization server. This is especially useful when the WAT

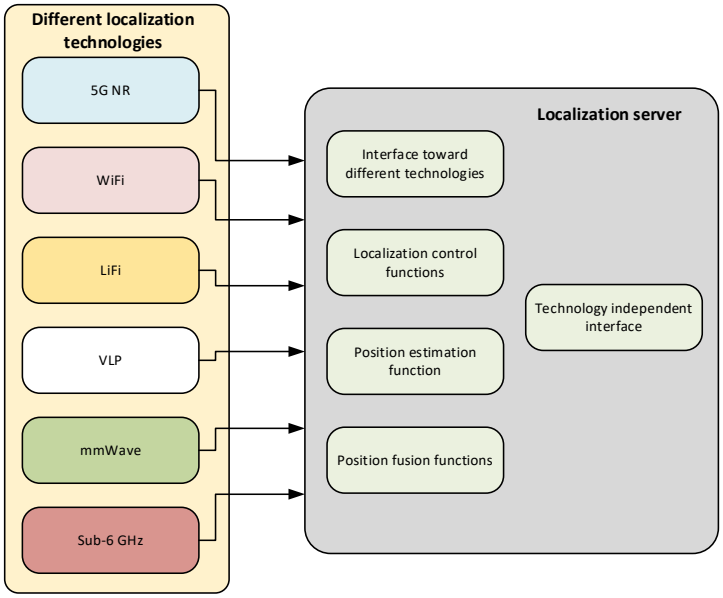


Figure 3-42 Simplified localization architecture

positioning technologies are not available on site. Additionally, they are used for simulations used for investigating optimal positions of the APs used for positioning.

The UE path is the path travelled by the UE of interest. It is supplied to the WAT positioning model as a timestamped 2D (or 3D) true positions of the UE. They are usually supplied in a text file. The anchor node positions are different for each of the available WATs. Each WAT has its own anchor nodes, placed in different positions. These positions are supplied to each WAT positioning model using a separate text file.

Each positioning technology has its own features and parameters, which must be supplied to the model in order to make the model more precise. These parameters include, for example in LiFi/sub-6 GHz/mmWave, the transmit power, receiver noise figure (NF), the temperature of the environment, in order to estimate the receiver thermal noise etc. All of these parameters are technology-specific and they are specified for each technology separately, in a separate text file. They enable tuning of the parameters for each technology, in order to find the optimal ones for a given scenario.

Finally, the positioning technology model shown in Figure 3-43 is supplied with all of these data and generates the necessary positioning parameters needed for the localization server in order to estimate the position of the UE.

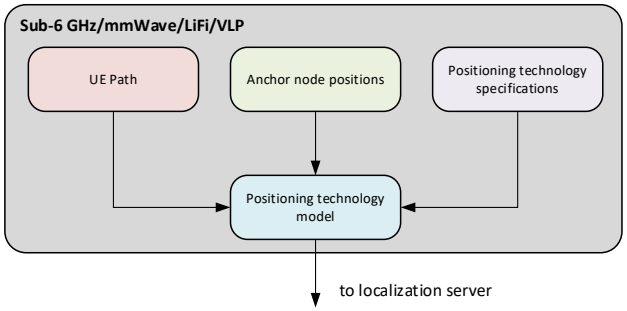


Figure 3-43 WAT positioning model

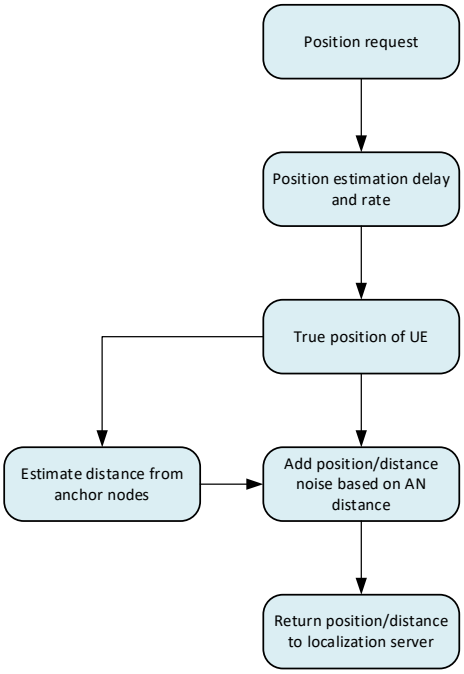


Figure 3-44 Positioning technology model

The positioning technology model details are given in Figure 3-44. In this model, after the position/distance or positioning parameter is requested form a localization server, the model first introduces a delay, based on the underlying WAT. This delay represents the delay needed by the used WAT to estimate the positioning parameters. Further, based on the timestamp, at which the position is requested, the true position of the UE is estimated.

The UE true positions are given with their coordinates and timestamps in a separate file. These positions are not given for each possible timestamp. Therefore, the positions between two timestamps are interpolated using a liner interpolation. This means that the UE is moving with a constant speed between two positions specified in the positions file. A more realistic interpolation model can be also used.

The interaction between the localization server, the UE and the WAT positioning model is shown in Figure 3-45. The position request is initiated using the UE or other entity in the network which requires the UE position. After the position request is initiated, the localization server contacts each WAT to obtain all of the necessary positioning relevant information in order to perform a position estimate.

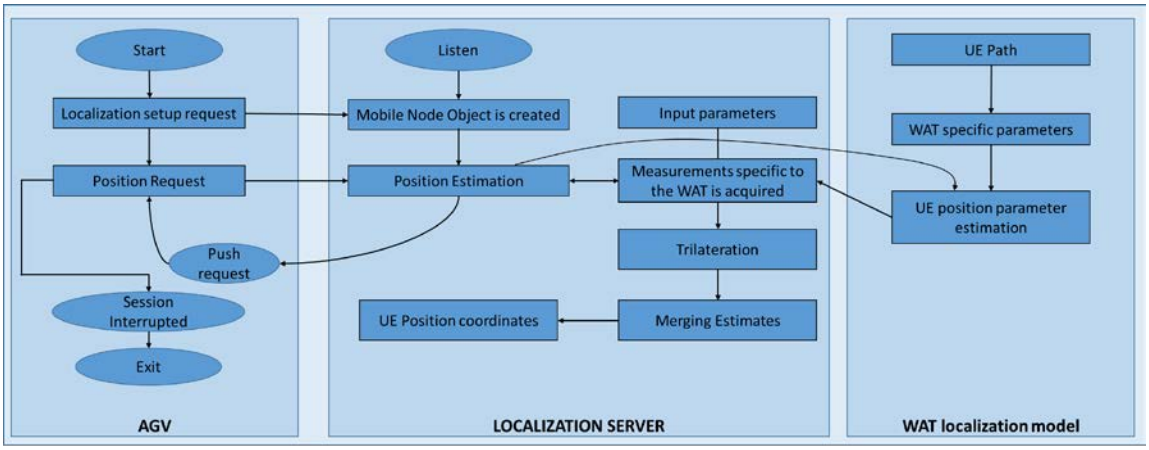


Figure 3-45 Interaction between the localization server and the WAT localization model

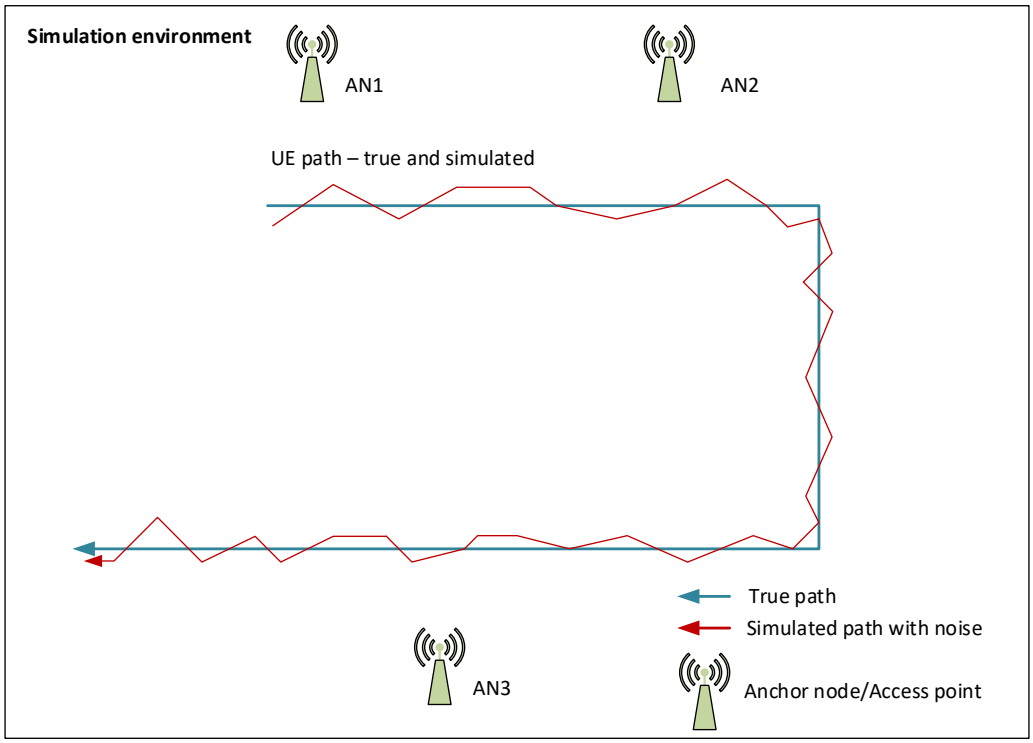


Figure 3-46 Position estimation simulation scenario

In Figure 3-46 a simulation scenario is shown. The simulation environment consists of 3 ANs and an UE moving along a given path (blue), being consisted of 3 straight lines. Using the WAT simulation model the localization server estimates the position of the UE based on the parameters of the used WAT.

The implementation of the WAT localization models is performed in Python. Each of the WAT positioning models are performed as classes containing the methods needed for initialization of the model, as well as the methods for obtaining position/range from the WAT model, as well as the other localization relevant parameters. The localization server communicates with the objects, created from these classes, using these methods. The objects for each WAT access the necessary UE position data as well as the necessary anchor node and technology specific data.

The described modelling framework would enable estimation of different parameters of the positioning system, as well as optimization of the system in advance, for a given scenario. For each WAT, the positioning error CDF can be estimated separately, as well as the CDF of the positioning error CDF, obtained from the localization server by merging the positions obtained from the different WATs. Additionally, different configurations and different positions of the APs can be tested in simulation, for a given scenario, in order to obtain their optimal position.

Finally, initial simulation for a sub-6 GHz system was performed in order to test the developed models before using them for a real scenario and multiple WATs. The scenario for this simulation is shown in Figure 3-47. In area of 10x10 meters, 4 anchor nodes are placed in the corners, marked with black dots. The UE is placed in 3 different positions, marked with red dots, and position estimates are performed, marked with blue dots. It can be noticed that the blue dots, i.e. estimated positions, are concentrated around the true positions. Due to the limited bandwidth and transmit power in this simulation, the estimated positions are widely spread around the true positions.

Additional KPIs can be also estimated using the tools developed. In Figure 3-48, the empirical CDF functions of the position error for the 3 simulated UE positions are shown.

For the given parameters of the scenario, it can be noticed that they differ slightly. This is normal, since the distance from the UE and the anchor points is different in the different positions. In this example, only a single technology was simulated, but the tool is not limited to the number of the technologies. For most of the RF technologies, the same model is used, only the parameters of the model are changed.

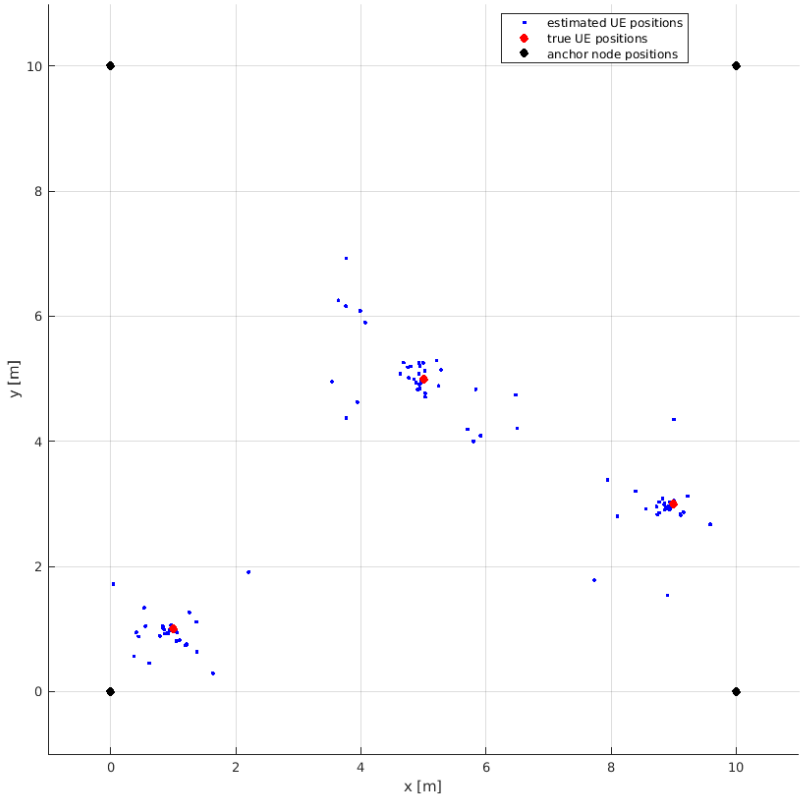


Figure 3-47 Simulation scenario

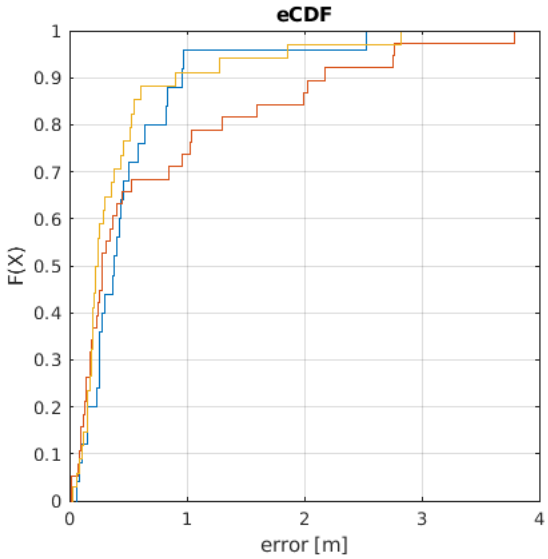


Figure 3-48 Empirical CDF of the position estimates for the 3 UE positions used in simulation

4 Scenario Description

This section provides a brief overview of the scenarios that are evaluated in the project including a private network supporting the operation of robots and an Industry 4.0 use case in a real factory. For the former scenario emphasis is given to the modelling and evaluation of dynamic decision-making algorithms. Special attention is given on the problem of MEC and UPF selection for end-to-end latency minimization. The Industrial scenario addresses several topics related to traffic offloading of mobile traffic from 3GPP to Wi-Fi network, multi-technology access and network selection, slicing for URLLC services and positioning.

4.1 Scenario 1: enhanced human-robot interaction

The main objective of the proposed use case is to evaluate the 5G-CLARITY system architecture through emulation and analytical modelling of infrastructure slices supporting uRLLC and eMBB services. To achieve this, we consider a MEC assisted private 5G network used to interconnect UGVs with onboard sensing devices and cameras with the application server hosting the AR/VR content delivery platform, the IoT platform, the teleoperation service etc as shown in Figure 4-1. To evaluate the overall architecture, each component will be modelled using the analysis described in Section 3.2 whereas the overall analysis will be conducted using the modelling tools described in Section 3.3. In the analysis, the UPF/packet gateway will mark the IP flows with the appropriate DSCP codes and apply the necessary forwarding rules. Specifically, data from uRLLC-related flows will be redirected at the local MEC whereas traffic flows with relaxed latency requirements will be sent to the central cloud platform. The MEC platform will be able to process the received data and apply specific AI schemes that will allow to detect possible anomalies and sent the necessary notifications to the devices and the operator. The overall architecture will be evaluated with its ability to co-host different slices.

This scenario aims at addressing all aspects of AGV communications and services including AGV-to-AGV and GV to ground.

Mobility management: Towards this direction a challenging topic that needs to be addressed is associated with service continuity for moving UGVs.

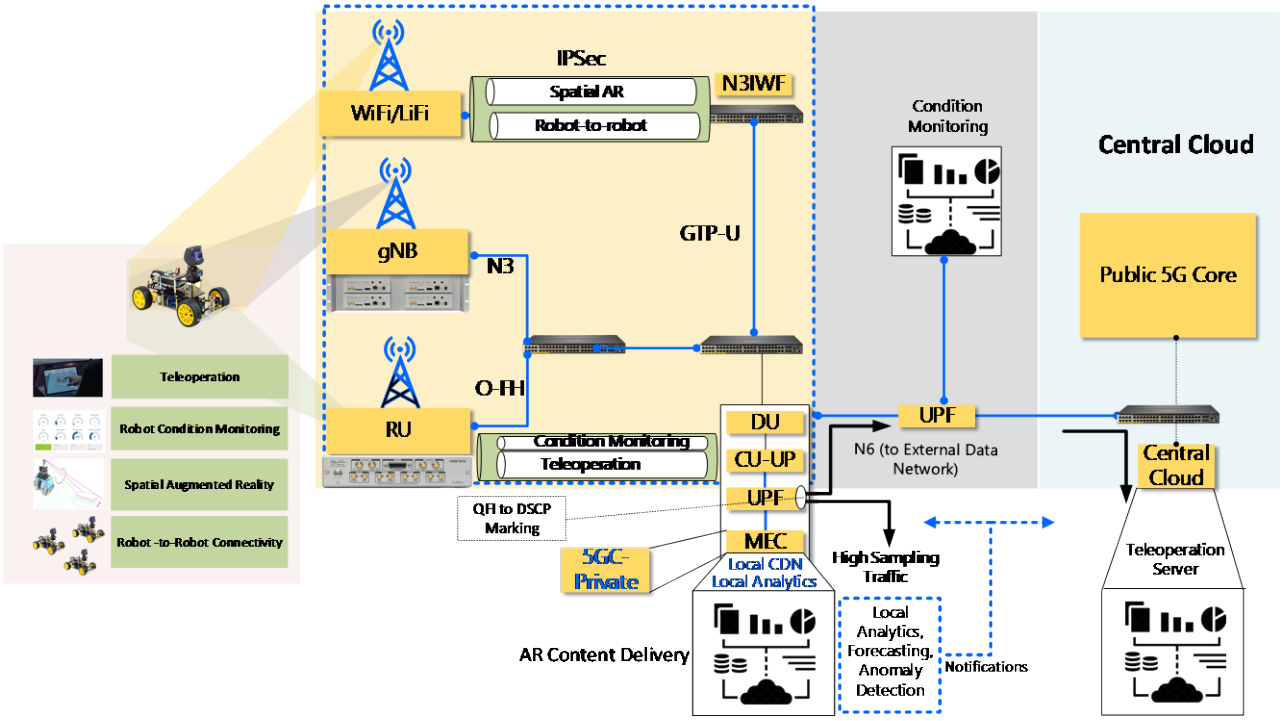


Figure 4-1 5G-CLARITY architecture supporting AGV operation

The required connectivity will be provided by the integrated 5GNR/Wi-Fi/LiFi network. The relevant network infrastructure is shown in Figure 4-1.Two scales of interoperability are required to manage the handover seamlessly and maintain the service continuity for UGV: (1) short time-scale (milli-second): at the level of 5G network among the neighbouring RAN and 5GC, and (2) long time-scale (second): at the level of 5G services across the neighbouring MEC platforms. The first issue will be solved through appropriate network and frequency planning to ensure the required level of cell overlapping as well as the tuning of handover triggering parameters among the neighbouring cells. The second issue will be addressed through user context and service migration among neighbouring MEC platforms. Considering the compatibility of 5G networks of different vendors, the only challenge is the appropriate network planning and configuration to maintain the service continuity with QoS. As for the 5G services, there are a number of implantation challenges to allow seamless transfer of the service and user context between the source and the target MEC host that are listed below:

- Support of application and service mobility among cross-border MEC platforms
- Synchronization 5GC SMF assignments across different MEC platform as well as service state

IoT slice: In addition to mobility management, this UC will demonstrate effective handling and decision making based on data, massively generated from sensing devices and transmitted over the deployed 5G network. The use case will take advantage and demonstrate the benefits of several key technologies adopted in 5G-CLARITY, for the establishment of multiple links from sensing devices, 5GNR for transmission of collected data and MEC for carrying out data intensive computational and authentication tasks. The following components will be deployed:

- An open wireless access communication system offering connectivity services for a massive number of low-powered sensing devices used for monitoring critical parameters of the UGV such as temperature, humidity, vibration, pressure on an end-to-end basis, power consumption, etc.,
- A data management platform (data lake plus data semantic fabric as defined in 5G-CLARITY D2.2) allowing scalable data collection, aggregation and processing of the collected information,
- A processing platform to facilitate optimal decision making.

CDN Slice: This scenario will evaluate the dynamic reconfiguration of the communication network in terms of slice provisioning/activating and migration towards achieving efficient utilization of resources and (virtual) service continuity, complemented by MEC infrastructure capabilities. The use case will use separate slices with guarantees over the wireless infrastructure used, prioritizing critical (i.e., teleoperation traffic) over non-critical traffic (critical slice) generated by the corresponding CDN server (i.e. AR/VR streaming). The infotainment applications will be migrated in a seamless manner to edge-resources across the edge.

To control and operate the 5G-CLARITY communications platform in a centralized manner, Service and Slice Provisioning subsystem (see 5G-CLARITY D2.2 [2]) is defined. This subsystem will be used to manage and orchestrate the envisioned services, incorporating MEC-enabled locations, and multi-domain functionality, across different network/infrastructure operators. MANO/RIC interactions will be also considered in order to evaluate the provisioning time of the optimal infrastructure over the multi-technology network.

Table 4-1 Scenario 1 Specifications

Services
URLLC: • Teleoperation, UGV-to-UGV connectivity eMBB: • CDN slice mMTC: • UGV monitoring data

Technical Challenges
• Integration of 3GPP with non-3GPP devices through the N3IWF • UPF processing capabilities • FH and BH traffic separation • Coordination of RIC with MANO
Architectural Features
• Multi-technology access • ORAN 5G-RAN with CU-DU separation • N3IWF • UPF • MEC platform • Data management platform hosted at the edge cluster • AMF/SMF for control
Configuration / Implementation Setup
• URLLC Slice #1: uRLLC UGV-RU- Transport- RAN (DU-CU)-UPF -Transport -N3IWF – LIFi/Wi-Fi-UGV • mMTC Slice #2: UGV-RAN (DU-CU)-UPF -MEC • eMBB Slice: #3: • Control information: UGV-AMF-SMF

4.2 Scenario 2: Wi-Fi offloading in an industrial scenario

Spectrum is an expensive and scarce resource. This is the reason why the availability of spectrum might be an entry barrier for private network owners as it is costly and limited, being not affordable by these private operators. On the contrary, the private network owner might sub lease the spectrum to a public operator upon stablishing an agreement to reduce the cost, but this is at the expense of losing control as the spectrum is managed by the public operator.

In this regard, the utilization of technologies like Wi-Fi for non-delay sensitive services like eMBB ones will be crucial for a private network operator, not only because of the management advantages of using unlicensed spectrum, but also because this technology cheapens the deployment costs.

The objective of this scenario evaluation is to assess the amount of 5G radio resources (licensed spectrum) that is released in an industrial scenario when Wi-Fi technology is used.

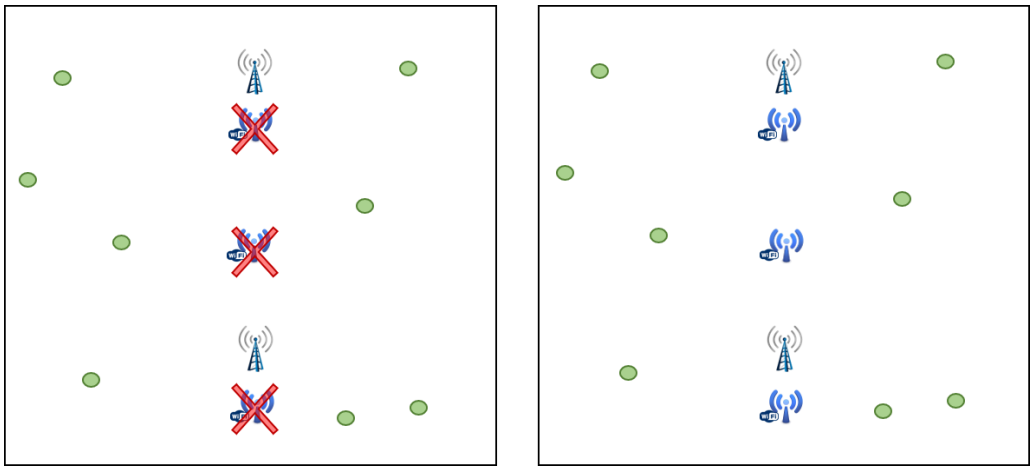


Figure 4-2 Wi-Fi eMBB offloading scenario (right) and baseline scenario without Wi-Fi (left)

Table 4-2 Scenario 2 Specifications

Services
Typical eMBB services in an industrial scenario (e.g., VR/ AR, video-streaming and mobile broadband access required by workers).
Requirements/KPIs
• Wi-Fi DL throughput: rate of data successfully delivered over the communication channel between the Wi-Fi access point and the user. • 5G DL throughput: rate of data successfully delivered over the communication channel between the gNB and the user.
Technical Challenges
• Ability of multi-WAT technology to provide higher data rates and capacity (relative to 5GNR) for eMBB traffic, and the integration of multi-WAT with slicing
Architectural Features
• Multi-WAT • Wi-Fi connectivity
Configuration / Implementation Setup
• Wi-Fi APs transmitting at 2.4 GHz. • 5G femtocells operating at 3.5 GHz and 100 MHz of bandwidth. • Industrial scenario layout including several production lines and the eMBB users randomly located through the factory floor.

Considering an industrial scenario similar to the one envisioned in UC2.1, we assume there are several eMBB users (e.g., AR-assisted workers) distributed over the geographical area (see Figure 5-3). Part of the eMBB users is served through Wi-Fi according to the SINR they perceive and the available Wi-Fi radio resources. The bandwidth consumed by eMBB users in this scenario will be compared with a baseline one in which there is no Wi-Fi access points deployed. In this way, the 5G radio resources freed up by Wi-Fi technology can be estimated for the industrial scenario, and hence indirectly approximate the cost saving associated with licensed spectrum acquisition.

4.3 Scenario 3: 5G-CLARITY slicing for URLLC services in an industrial scenario

One of the key features of the 5G-CLARITY architecture is an infrastructure-level slicing model to facilitate multi-tenancy in 5G non-public networks. 5G-CLARITY slicing concept extends the notion of network slicing to offer a higher degree of isolation among the network slices, becoming a highly suitable option for multi-tenancy support in 5G non-public networks. In this vein, the primary goal of this evaluation is to verify the effectiveness of the 5G-CLARITY architecture and slicing concept to ensure the full isolation among the network slices.

The scenario considered for the evaluation tries to resemble the 5G-CLARITY UC2.1 scenario in BOSCH factory, dubbed “Alternative Network to Exchange Production Data” (see Figure 5-3) [54]. In this scenario, we evaluate how the malfunctioning of a production line that results in the generation of a non-conformant traffic (traffic excess) affects the E2E mean response time of the rest of production lines for two configurations:

- A configuration in which there is a dedicated 5G-CLARITY slice to serve the traffic of each production line of the factory floor, thus providing isolation between production lines. Each 5G-CLARITY slice includes segregated resources for the different network domains, e.g., wireless, compute, and transport quotas, in order to serve the aggregated traffic generated by each production line.
- A baseline configuration with a single shared 5G-CLARITY slice to serve the aggregated traffic from

Table 4-3 Scenario 3 Specifications

Services
• Motion control.
Requirements/KPIs
• Packet loss ratio at the NR-Uu interface: The fraction of the packets that are lost at the radio interface. The target packet loss ratio at the radio interface is 10^{-4}. • DL E2E Latency: The time the network takes to transport a packet between the PSA UPF and the UE. The target E2E UP maximum delay is 1 ms as specified in 5G-CLARITY project [1].
Technical Challenges
• Ensuring the full Isolation of the 5G-CLARITY slices in an industrial scenario.
Architectural features
• Transport node: the two technologies considered in 5G-CLARITY project, namely, standard Ethernet and TSN. More precisely, for the latter, asynchronous TSN with non-preemptive traffic prioritization is considered. • gNB-RU user plane data transmission. • gNB-DU user plane data transmission. • gNB-CU-UP: 5GNR user plane data transmission. • UPF user plane data transmission. • 5G-CLARITY slice isolation for the different network domains.
Configuration / Implementation Setup
• URLLC traffic generated by each production line served by a 5G-CLARITY slice. Over the 5G-CLARITY slice is deployed a 5G system that includes dedicated virtualized UPF and gNB-CU instances to serve the traffic generated by the respective production line. There are also isolated radio and transport network resources destined for the slice. • The upper layers of the virtualized UPF and gNB-CU instances follow a FCFS discipline to serve the packets following a run-to-completion strategy. They are instantiated at the edge cluster and have dedicated physical CPU cores for this task (CPU pinning). • The gNB-DU and the radio unit are deployed as a physical network function (PNF) (small cell) operating at 3.5 GHz and 100 MHz of bandwidth. • The transport network interconnects the 5G components (gNB-DU, gNB-CU and UPF). We consider both standard Ethernet and an asynchronous TSN network. The constituent TSN bridges of the TSN network include an Asynchronous Traffic Shaper (ATS) at every egress port. Every ATS includes eight priority levels and sixteen shaped buffers. The transmission capacity for every link was set to 1 Gbps. • The primary delay bottlenecks considered for the DL at each slice are: UPF processing, gNB-CU processing, involved TSN bridge transmission at the transport network, gNB-DU processing, gNB-RU processing, and radio interface transmission.

all the production lines will be evaluated in order to demonstrate the benefits brought by 5G-CLARITY slicing in terms of isolation.

4.4 Scenario 4: mobility and traffic load management in Wi-Fi/LiFi integrated networks

A Wi-Fi AP can provide signal coverage for light blocked users moving around LiFi APs, while these can enhance the data transmission coverage of Wi-Fi APs. Indoor users suffer from low Wi-Fi signal coverage in most parts of their local private network (e.g., mall large commercial surfaces or private homes), while wireless service providers (WSPs) strive to provide diverse high data rate services and multimedia contents wherever their users located in indoor environments. An indoor user may use a LiFi AP to watch high definition (HD) Netflix films in his bedroom, while his father runs on move a video skype call at the back end

of their garden using the Wi-Fi AP meanwhile his sister plays online games using the LiFi AP or the Wi-Fi AP. These indoor users trigger the high-data rate requests for wireless data communications, where the WSPs should be able to engineer traffic delivery anywhere in their indoor places by using the integrated network shown in Figure 4-3. This keeps the indoor users connected anywhere in their places, while routing traffic to them through the LiFi or Wi-Fi AP according to the requested service type, location of indoor users, and integrated network conditions.

The objectives of this scenario are:

- Supporting vertical handover between LiFi and Wi-Fi wireless technologies with handover times < a specific time threshold.
- Design and validation of a multi-technology coexistence framework for private LiFi and Wi-Fi networks.
- A mobility management plane based on the principles of Software Defined Networking (SDN) for the LiFi/Wi-Fi joint networks.

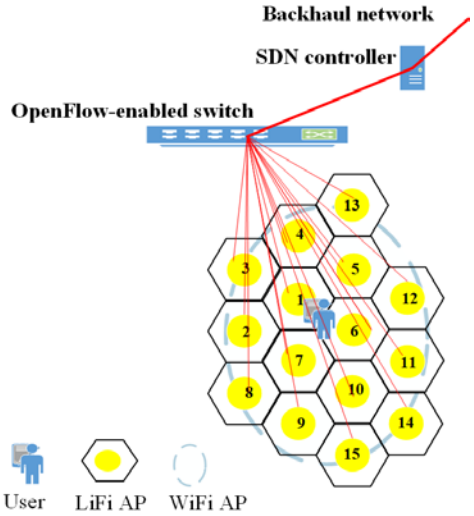


Figure 4-3 SDN-enabled Wi-Fi-LiFi joint networks

Table 4-4 Scenario 4 Specifications

Services
It supports two main services: traffic load balancing and aggregation, and user mobility management. Every measurement/simulation time interval (MTI), in each cell, the data rate of each user will be measured, and the number of users associated with each AP will be counted. Users will generate real and non-real time data traffic flows with various short packet sizes and transmission intervals. The interfaces of both LiFi and Wi-Fi have different queue sizes.
Requirements/KPIs
User requirements: • Keep users connected anywhere in their indoor places. • Provide and maintain high quality in provisioned services to indoor users. Wireless service provider requirements: • Associate users with the AP that can best support their service requests and traffic volume. • Provide users regular information about the current traffic volume and number of associated users per AP to make them aware of their network and service status. KPIs: • Blocking probability: the probability of denying a service for a user subject to the data rate or number of

users threshold set per AP.
- Delay (latency): the total average delay of packet experienced along its journey from the multimedia server collocated with the controller down to the user.
- Target cell throughput: the total data rate (throughput) of users associated with an AP over a time window.
- Target network throughput: the total data rate (throughput) of all users received services from APs over a time window.

Technical Challenges

- Integrating the SDN controller with the operations of LiFi and Wi-Fi APs using software agents.
- Developing software agents for association and disassociation of users with APs, particularly disassociating those with strong signal strengths and connecting them to another AP providing a sustainable lower signal strength.
- Combining the operations of the different software agents to provide on real-time E2E services.

Architectural features

- 5G-CLARITY Building blocks involved in the evaluation and their interconnection.
- SDN controller
- Traffic flows routing
- Traffic packets scheduling
- Users' association and dissociation
- Network monitoring dashboard

Configuration / Implementation Setup

Technologies:
- LiFi and Wi-Fi APs
- Open Day Light SDN controller
- Open flow enabled switch
- LiFi and Wi-Fi enabled user devices

Specifications:
- Users and network information (Global state) collection
- Generic software agents running on the APs
- Customised OpenDaylight software module
- Interlinking software agents
- Enforcement policies at the SDN controller and the APs

4.5 Scenario 5: joint synchronisation and localization using multi-wireless access technologies

The 5G-CLARITY network leverages multi-WATs to ensure reliable data communications as well as high precision positioning with a good coverage of the area of interest. A total of four different WATs are used for positioning in 5G-CLARITY: sub-6 GHz UL/DL-TDoA, mmWave, LiFi, VLP/OCC, as described in Section 3.

The main scenario that would be evaluated includes multi-WAT deployment, where different localization capable WATs are deployed in different areas. These areas do overlap partially. This enables evaluation of the localization precision in areas covered by multiple WATs as well as in areas covered with a single WAT.

The positioning framework would be tested in a real scenario in the BOSCH factory, use case UC2.2[54] . In 5G-CLARITY WP2, a simulation framework was developed to test the positioning precision and accuracy using the different available WATs. Additionally, the developed framework can be used for simulation of different deployments of the available WATs, in order to choose the most optimal configuration. This simulation environment also can use the data fusion approach developed in WP3 [D3.2]. This would enable simulation of the all WATs used for the simulation scenario and investigation of the contribution to each WAT towards the improvement of position estimation precision and accuracy.

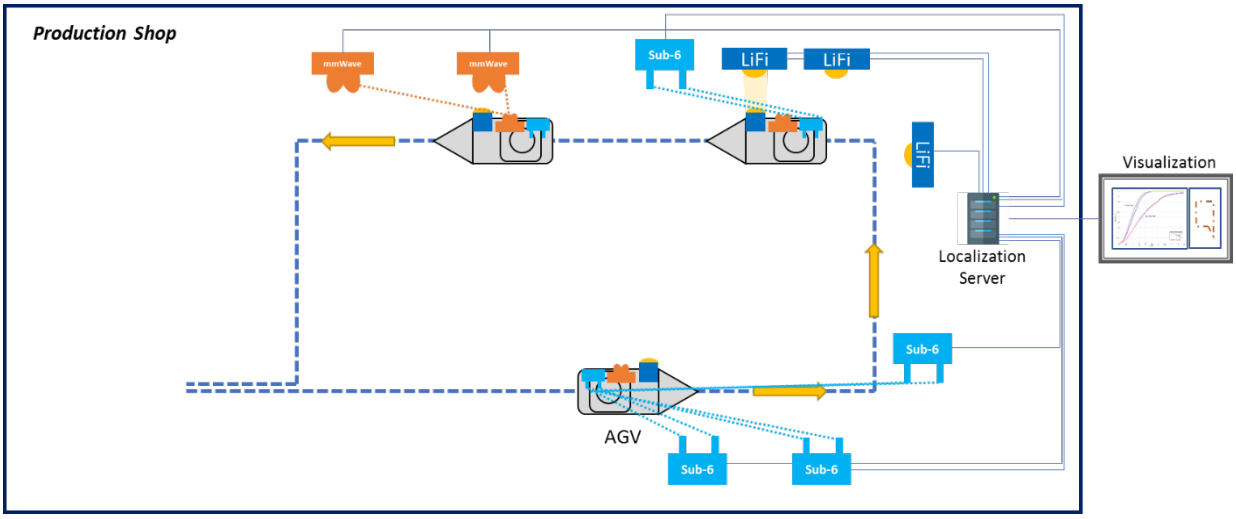


Figure 4-4 Positioning test scenario using multi-WATs

One typical simulation scenario is given in Figure 4-4. An AGV moved in areas covered by different WATs supporting localization. The number of the available WATs at a given moment differ from position to position. All of the available WATs are used in order to obtain the most precise position estimate.

Some of the positioning methods deployed in 5G-CLARITY are strongly depending on the synchronization accuracy and precision between the access points for the different WATs. Positioning and synchronization problems can be tackled independently, which is probably not the most optimal approach. Therefore, in this scenario, the both problems will be addressed and evaluated jointly, since they strongly overlap. Additionally, functions used for positioning can be also reused for synchronization.

The synchronization precision and accuracy and precision would also strongly affect the positioning precision. This will be also evaluated in this scenario in order to obtain quantitative measures which will be later used for optimizing of the positioning system architecture.

Table 4-5 Scenario 4 Specifications

Services
The main service supported is joint synchronization and localization. Different synchronization and localization methods and different WATs are supported and will be evaluated
Requirements/KPIs
• Performance of the joint synchronization and localization algorithm • Performance of joint synchronization and localization algorithm across time-stamping uncertainty
Technical Challenges
Precise synchronization of multi WAT APs with nanosecond precision, using different synchronization approaches.
Architectural features
• Multi WAT • Network wide synchronization
Configuration / Implementation Setup
• Network-wide synchronization • Pairwise synchronization • Hybrid Synchronization • Bayesian joint synchronization and localization

5 Scenario Evaluation

This section provides an overview of the main evaluation results for scenario presented in Section 4. Specifically, for Scenario 1 emphasis is given on mobility aspects and specifically, on the dynamic UPF selection problem. The main objective of this study is to identify the optimal UPF that can be used to server the mobile robot in order to minimize end-to-end latency. This is achieved through a reinforcement learning scheme that based on the location of the robot and the background can dynamically select the optimal location where PDU sessions can be terminated [17]. The second set of results focuses on the multi-wat offloading problem and it is demonstrated that by offloading traffic to Wi-Fi enhanced system performance in terms of throughput and packet loss rate can be achieved. Scenario 3 quantifies the benefits gained when TSN is integrated in the 5G-CLARITY solution. Through appropriate scheduling techniques TSN can be used create fully isolated infrastructure slices that can be allocated to URLLC services. It is shown that the performance of the proposed slices in terms of latency, is deterministic and not affected by background traffic. Scenario 4 evaluated the benefits gain by the integration of Wi-Fi/LiFi and 5G technologies whereas Scenario 5 evaluates a novel algorithm that has been developed allowing joint synchronization and positioning services to be offered over the 5G-CLARITY solution.

5.1 Evaluation of Scenario 1: enhanced human-robot interaction- dynamic UPF selection

A big part of the user plane functionality in 5G systems is handled by the UPF, which has to be designed to support challenging 5G services with very tight performance requirements. It connects with external IP networks hiding mobility related aspects from the external networks. Moreover, it performs different types of processing of the forwarded data, such as packet inspection, redirection of traffic and application of different data rate limitations. 5G-CUPS, supporting multiple UPFs, enables 5G edge capabilities, which is one of the key 5G advancements compared to 4G. The UPF related processing can be dynamically deployed and configured depending on the application needs. Overall, UPFs act as termination points for various interfaces and protocols and are also responsible to take several actions (rules) [9] including: mapping of traffic to the appropriate tunnels based on the QFI information, packet steering, packet counting, deep packet inspection and buffering and queuing for traffic service differentiation and assurance of E2E delays.

To perform these actions UPFs should support an extensive set of protocols such as, GTP-U, PFCP, IP and also assist in the operation of SDAP and PDCP through mapping of DSCP classified IP traffic coming from the external DN. It should be also capable of handling legacy and new protocols such as eCPRI/ORAN and Radio over Ethernet (RoE) at high data-rates. Towards this direction, *programmable edge nodes* can be effectively used to support transport network requirements as well as classify and steer the traffic. This is performed by adopting specific interfaces for control plane (N1/2, N4), user plane (N3, N6) and UPF handover (N9) communication. For example, the Network Interface Cards (NICs) can steer control plane protocol packets such as PFCP packets into the Session Management Function (SMF) or the control plane part of UPF and can steer UE sessions based on the PDU session, the flow, the QoS class, etc., through N3 and N6 interfaces. Programming can be also used to support extended header (EH) for 5G user plane traffic.

A high-level view of a 5G deployment option combining private and public 5GC is shown in Figure 5-1. In this figure the RU, the gNB-DU and the gNB-CU can be either collocated or located separately adopting either a MEC or a central cloud architectural approach. Based on the 5G-RAN deployment option and the type of service that needs to be provided, UPF nodes can be placed closer or further away from the 5G-RAN. In this context, as the network dimension grows, a larger number of rules is required to support policies, whereas network resources (e.g., memory in the switches) are limited. This may result in increased service delay as the number of flows requiring UPF processing increase.

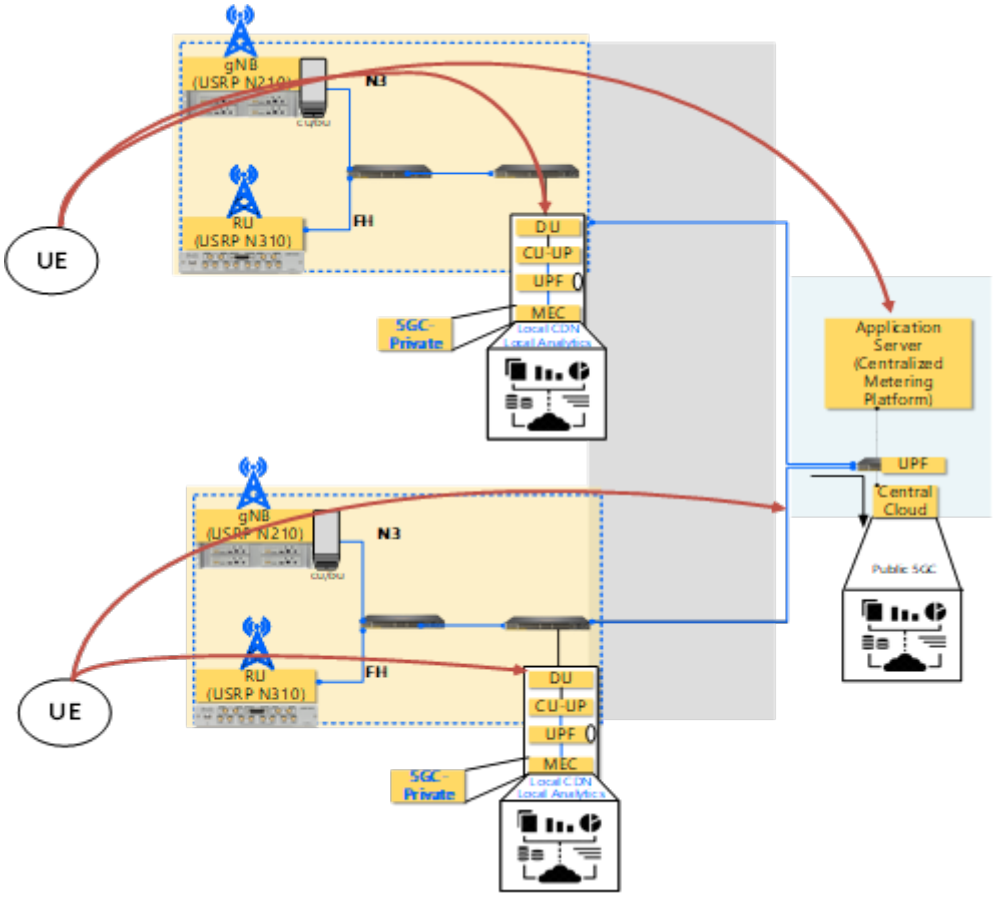


Figure 5-1 Hybrid private-public 5GC deployment

To address this problem, we apply Evolutionary Game Theory (EGT) to dynamically select the optimal UPF and MEC nodes where connection will be terminated. Specifically, the UL transmission of a 5G network shown in Figure 5-1 and discussed in Section 4 is considered. The UEs initiate PDU Session Establishment process by transmitting the relevant request to the AMF. The AMF contacts the SMF, which in turn checks whether the UE requests are compliant with the user subscription. Once subscription information is verified the SMF selects a UPF to serve the PDU session. This is a key decision to be taken as a UPF at close proximity to the RAN, that may be the optimal choice at first sight, since it should result in reduced latency.

However, if all UEs are associated with this UPF, congestion may arise resulting in increased latency. To address this challenge, a scheme that allows dynamic selection of the UPFs by the UEs is proposed. In this approach users try to optimize their own performance selfishly. The choice of UEs adaptation process can be formulated as an evolutionary game.

To formulate this problem, we consider a set of UEs each requesting a service of class $g \in G$ where G is the total number of available service classes. Let also $S^g = \left\{UPF_1^g, \dots, UPF_{N_g}^g\right\}$ be the set of available strategies in users belonging to the g-group. For each group, each UE tunnel needs to be terminated at a specific UPF. Assuming that N_g denotes the available UPFs for group g, then the population of the UEs in group g can be described at each time instance by vector $\boldsymbol{x}^g(t) = \left[x_1^g(t) \dots x_{N_g}^g(t)\right]$ where $x_i^g(t)$ is the proportion of UEs in group g that are currently being served by UPF_i. Each UE belonging to a specific group remains associated with a UPF for a time interval and reviews its choice periodically. When a revision occurs, the UE switches from UPF_i to another UPF_j according to a switching probability $p_{ij}^g(\boldsymbol{x})$ equal to the population probability distribution of strategies:

$$p_{ij}^{g}(\boldsymbol{x}) = x_j^g \qquad \text{(Eq. 5-1)}$$

where $\boldsymbol{x} = [\boldsymbol{x}^1(t) \dots \boldsymbol{x}^g(t)]$, is the population state of the system. If a switch occurs, the UE receives a payoff $u_j^g(\boldsymbol{x})$ that quantifies its satisfaction level associated with the selection of UPF_j. The obtained payoff affects the arrival rate of the revision opportunities. Assuming that the number of reviews of a UE that uses strategy i can be described by a Poisson process with arrival rate $r_i^g(\boldsymbol{x})$, and all UEs' Poisson processes are statistically independent, we can use the law of large numbers to approximate the adaptation process with the following deterministic dynamic model [11]:

$$\dot{x}_i^g(t) = \underbrace{\sum_{jeS^g} x_j^g(t) r_j^g(\boldsymbol{x}) p_{ji}^g(\boldsymbol{x})}_{inflow\ to\ strategy\ i} - \underbrace{x_i^g(t) r_i^g(\boldsymbol{x})}_{\substack{outflow \\ from\ strategy\ i}} \qquad \text{(Eq. 5-2)}$$

The UE updates its review rate, by linearly decreasing it to its current payoff. This means that the average review rate of a UE that uses strategy i is:

$$r_i^g(x) = a - \beta u_j^g(\boldsymbol{x}), \quad \beta > 0\ and \frac{\alpha}{\beta} > u_j^g(\boldsymbol{x}) \qquad (3)$$

This results in forcing UEs with higher payoffs to revise their UPF choice at lower rates than the rest, leading to the replicator dynamics:

$$\dot{x}_i^g(t) = \beta\left(u_i^g(\boldsymbol{x}) - \overline{u^g}(\boldsymbol{x})\right) x_i^g \qquad (4)$$

According to this equation, a selected strategy will either survive or be eliminated in the long run depending on whether its payoff is better or worse than the average payoff of all strategies. Since the objective of the UEs is to optimize their performance in terms of latency, greater payoffs correspond to lower delays. The observed latency can be decomposed into two main components. The first component is the propagation delay between the UE and the UPF and is proportional to the distance between the two entities. Assuming an underlaying optical transport network, the propagation delay due to the propagation time in the fiber links corresponds to 5 μs per kilometer (km) of fiber. The second component is the delay of processing inside the UPF and can be modeled by adding the processing and the transmission delay, that are constant, and the variable queuing delay. Mechanisms for bounding the processing delay within a network node can be found both in literature and in standardization. In this analysis, we assumed that the UPF, uses the bounded mechanisms described in [4].

Considering these assumptions, we formulate the payoff on a user of group g that selects action i, when the population state is $\boldsymbol{x}(t)$, as

$$u_i^g(\boldsymbol{x}) = \frac{1}{t_{prop}{}_i^g + t_{UPF_i}(\boldsymbol{x})} \qquad (5)$$

Where t_{prop} is the propagation delay and $t_{UPF_i}(\boldsymbol{x})$ the UPF_i delay that can be approximated by an exponential function:

$$t_{UPF_i}(\boldsymbol{x}) = e^{k_i \rho_{UE} \sum_{g=1}^{G} M_g x_i^g} \qquad (6)$$

Where ρ_{UE} is the traffic of one UE, M_g is the UE-population of group g and k_i is a variable related with UPF i and depends on the characteristics of the UPF node implementation including data rate, number of ports (fibres, wavelengths), buffering capability etc.

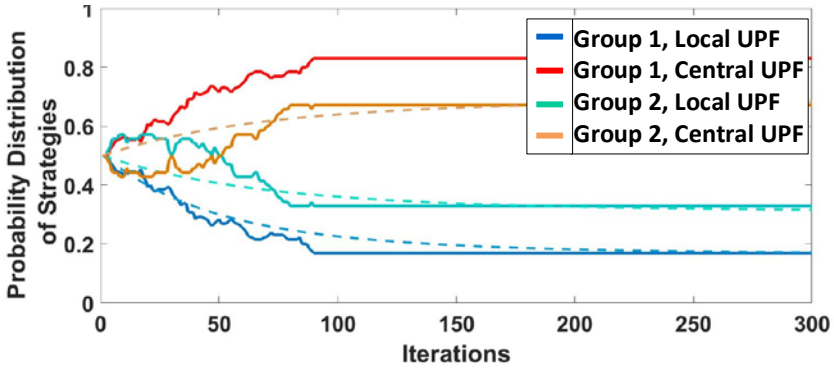
Based on the replicator dynamics of the EGT, we developed a scheme to attain the evolutionary equilibrium.

The following steps summarize the algorithm:

1. *Initialization*: Every UE in each group chooses a strategy at random and observes its payoff u. Then it calculates its review rate λ according to the formula $\lambda = a - \beta u$, where α, β are constants.
2. *Revision*: A revision opportunity may occur to each UE with probability equal to $p_{revision} = \lambda \cdot dt$, where dt is the time interval between two loops. If the revision occurs, the UE chooses to imitate, at random, one of the UEs of its group. Then it recalculates λ according to the obtained payoff. The same process is applied until the difference of each strategy's payoff compared with the average payoff of the population is lower that a limit ε.

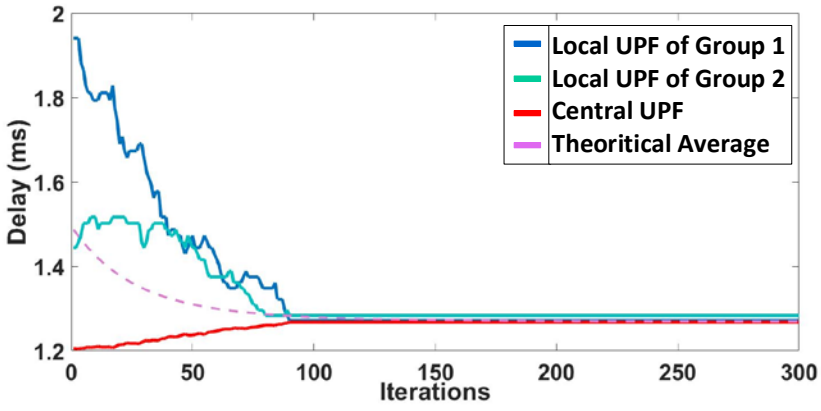
Note that the strategy adaptation process in the proposed EGT-based algorithm does not rely on the knowledge of the strategy selection of the other players. For the evolution, a UE requires a random matching with an opponent, a function that can be offered by a central controller (the SMF for example). Therefore, the amount of information exchange is reduced. The central controller will randomly match the UEs and stop the evolution process, if all payoffs are equal or differ by a small quantity.

The time interval (dt) between two repetitions must be higher than the communication time between the UE, the AMF, the SMF and the UPF that is going to carry the PDU session. dt is highly affected by the number of UPFs that are under the control of the SMF, since a large number of UPFs may result in increased processing delay for the SMF.



Full lines: simulation results, Doted lines: theoretical results

(a)



(b)

Figure 5-2 Trajectories of proportions of population and (b) convergence of the algorithm to the equilibrium (for $M_1 = 130, M_2 = 70, \frac{a}{b} = 1$). In the equilibrium 16% of group 1 UEs and 32% of group 2 UEs are served by their local UPFs, while the remaining are served by the central UPF

Taking into consideration the timing requirements of the network service ($t_{service}$), and the number of iterations of the algorithm (L), the number of UPFs (N) under the SMF's control can be evaluated so that the following relationship is true:

$$N < F^{-1}\left(\frac{t_{service}}{L}\right) \qquad (7)$$

where F^{-1} is the inverse function that relates dt with N.

The proposed theoretical model is evaluated using simulation. In the following we assumed a population of UEs that are organized into two groups. The UEs in each group can decide whether they want to use a local UPF at the edge of the network, that connects to a MEC, or to a UPF that connects to a central cloud as shown in Figure 5-1. The UPF in the central cloud can process a greater number of requests, compared to the local UPFs, and is shared by all groups in the UE population whereas the local UPF is dedicated to the population inside a group. The traffic generated by each UE is assumed to be $\rho_{\mathrm{UE}} = 100$ Mbps.

The limit ε of the algorithm is set to a payoff difference of 0.01. Figure 5-2 illustrates the simulation results (full lines) and the theoretical results derived through the model of the replicator dynamics (doted lines) demonstrating good agreement between theory and simulation. More specifically, Figure 5-2 (a) plots the evolution of strategy shares among the population of UEs. It can be observed that the system converges after some iterations to the equilibrium. In equilibrium all UEs achieve the same delay (Figure 5-2 (b)) indicating the fairness of the scheme. The number of total iterations of the algorithm is of vital importance for network planning. As it was discussed in the previous section, the number of iterations in combination with the time requirements of the service, can give an estimate (Eq. (7)) of the number of UPFs that the SMF can control, without compromising the stability of the system. Figure 5-2 shows that less than 100 iterations are needed for the system to converge.

5.2 Evaluation of Scenario 2: Wi-Fi offloading in an industrial scenario

In this section we intend to highlight the benefits of having multi-WAT in a private network environment. To that end, we focus on a private industrial network scenario. Specifically, we consider the scenario depicted in Figure 5-3. This scenario is inspired in the industrial network considered in 5G-CLARITY UC2.1 [54] . The considered private site occupies a geographical area with dimensions 100 m x 100 m, in which a multi-WAT RAN is deployed. Specifically, the RAN comprises two different WATs: 5GNR and Wi-Fi. In that way, the scenario includes four femtocells and five Wi-Fi access points, depicted as blue circles and green triangles, respectively. The eMBB UEs, represented as red squares, are uniformly distributed in the scenario. A total of 224 sensors considered as URLLC UEs are distributed in 4 production lines along the factory floor, being each of the production lines composed of two wings. The specific setup for the main parameters is included in Table 5-1.

The CDF of the SINR obtained from the scenario considered and described previously is depicted in Figure 5-4. Firstly, we begin assessing the Wi-Fi capacity for offloading eMBB traffic from 5GNR. This offloading experiment shows how Wi-Fi, which is a cheaper technology than 5GNR, can be leveraged to serve non-delay sensitive applications in order to release resources from 5G technology. The freed up 5G radio resources either will cheapen the private 5G network deployment and operation costs or might be allocated to URLLC type services to meet their stringent delay requisites.

Specifically, in order to show the benefits of having a multi-WAT we use a baseline scenario in which we only have 5GNR as the network radio access technology. In this baseline scenario we consider the setup configuration shown in Table 5-2.

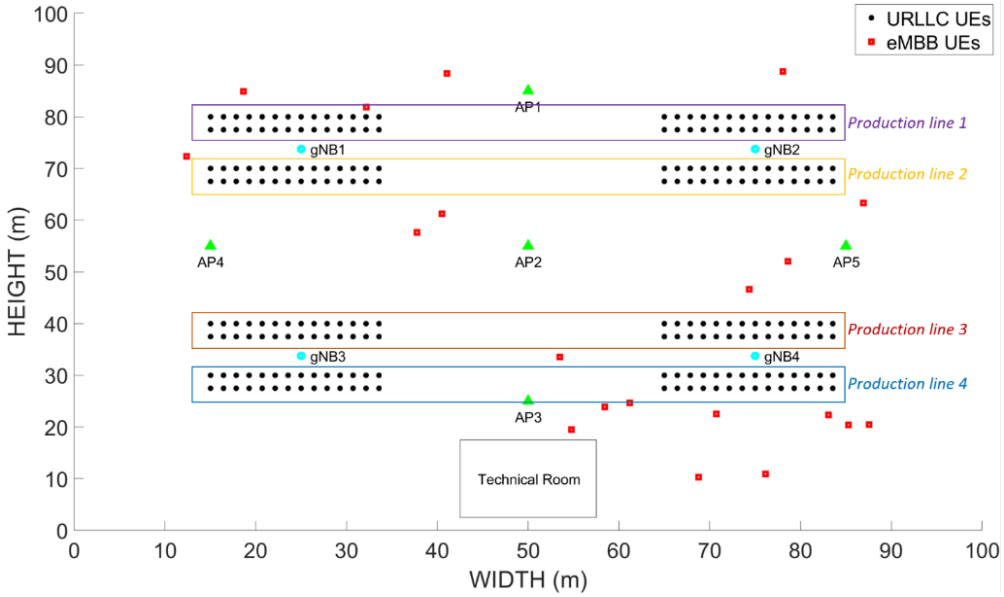


Figure 5-3 Industrial scenario layout of 5G-CLARITY UC2.1 [79]

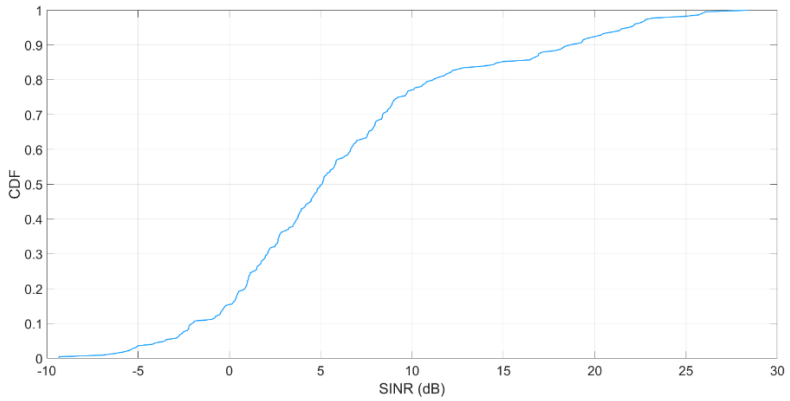


Figure 5-4 CDF of the URLLC UEs SINR obtained from the industrial scenario

Table 5-1 Simulation Parameters for Assessing the Wi-Fi Offloading Capacity

Parameter	Configuration
eMBB UEs guaranteed bitrate	5 Mbps
URLLC UEs bitrate	1.55 Mbps
URLLC delay requirement	1 ms
Number of eMBB UEs	50
Number of URLLC UEs	224 distributed in 4 production lines
Cell type	Femtocells and Wi-Fi cells
Number of femtocells	4
Number of Wi-Fi cells	5
Direction of transmission	DL
eMBB traffic distribution	Uniform
Path loss model for femtocells	Indoor Hotspot (InH)
Path loss model for Wi-Fi cells	Indoor Hotspot (InH)
Antenna height in femtocells	6 m

Antenna height of Wi-Fi APs	4 m
Transmission power in femtocells	30 dBm
Transmission power in Wi-Fi cells	20 dBm
UE height	1.5 m
UE thermal noise	-174 dBm/Hz
Noise figure	9 dB
Carrier frequency in femtocells	3.5 GHz
Carrier frequency in Wi-Fi APs	2.4 GHz
Bandwidth in femtocells	100 MHz
Bandwidth in Wi-Fi APs	40 MHz
Frequency reuse	1
URLLC packet size	80 bytes
Load sweep	From 1 to 28 URLLC UEs
Number of eMBB slices	1
Number of URLLC slices	4 (one per production line)

Table 5-2 Baseline Scenario Simulation Parameters

Parameter	Configuration
gNB bandwidth	100 MHz
Number of slices served per gNB	2 URLLC slices 1 eMBB slice
gNB bandwidth allocated to eMBB slice	40 MHz
gNB bandwidth allocated to URLLC slice	30 MHz

In this baseline scenario we measure the average throughput achieved by eMBB users in the network when no Wi-Fi APs are connected (i.e., both eMBB and URLLC users are served by 5G technology), and when Wi-Fi APs are available as a function of the amount of 5G bandwidth allocated to the eMBB slice. The models to obtain these metrics are described in Sections 3.1.2.6, 3.1.2.7, and 3.1.2.8. The results are shown in Figure 5-5.

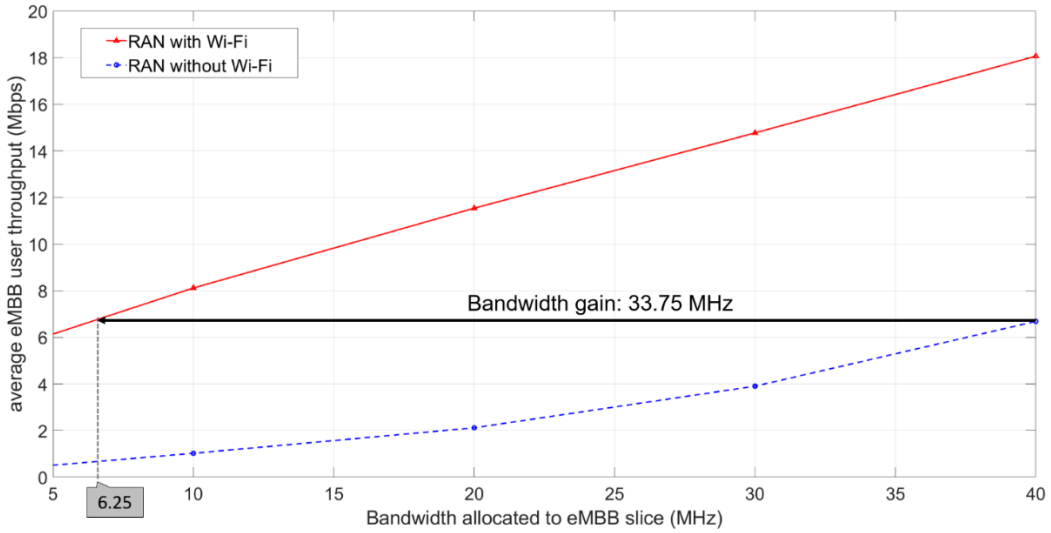


Figure 5-5 Average throughput achieved by eMBB users vs the 5G bandwidth allocated to eMBB slice

In this figure we can see in blue color the line that represents the mean throughput reached by an eMBB user in the network when Wi-Fi APs are disabled (i.e., in the baseline scenario), and in red color when Wi-Fi APs are active. As observed, the throughput achieved by eMBB users increases as the amount of bandwidth allocated to the eMBB slice is bigger. We can also see that the eMBB users throughput is significantly higher when Wi-Fi APs are deployed in the scenario, making evident one benefit of the multi-WATa feature. Additionally, in Figure 5-5 we can observe that when Wi-Fi technology is available in the scenario and some of the eMBB users can be offloaded to Wi-Fi, the eMBB users can achieve a mean throughput of roughly 6.5 Mbps with only 6.25 MHz of 5G bandwidth, while a bandwidth of 40 MHz is necessary to reach the same mean throughput when only 5G NR deployed in the scenario, given the setup described above. The implication of having a multi-WAT access network composed of both technologies 5G and Wi-Fi in an industrial scenario where different use cases coexist (eMBB and URLLC) is that the low priority traffic (i.e., eMBB traffic) can be steered through Wi-Fi technology. So that, large amount of 5G bandwidth can be saved in order to make it available for URLLC services that demand strict latency constraints. Particularly, our results show that this bandwidth saving is around 33.75 MHz for the given setup.

Next, we want to demonstrate how the 5G bandwidth saving can be profitable in a multi-WAT scenario. More precisely, we evaluate the packet loss ratio (PLR) at the radio interface for a URLLC slice with a delay constraint of 1 ms in our baseline scenario (without Wi-Fi APs) and in the multi-WAT scenario. To that end we use the analytical model described in Table 5 (Section 3.2.2). Given one of the gNBs deployed in the scenario (for instance gNB1), we measure the URLLC slice PLR for different traffic loads considering that each of the URLLC slices served by this gNB is allocated a bandwidth of 30 MHz (the remaining 40 MHz of bandwidth is allocated to the eMBB slice). The dashed blue line of Figure 5-6 represents this metric.

Then, when we introduce the multi-WAT functionality in our scenario being Wi-Fi technology integrated in the 5G-CLARITY RAN architecture some of the bandwidth will be released from 5G due to part of the eMBB traffic can be offloaded to Wi-Fi technology. The 33.75 MHz of bandwidth freed up from 5G (see Figure 5-5) can be now destined for URLLC services. In this way, the gNB has more bandwidth available to be allocated to the URLLC slices. As observed in Figure 5-6, the PLR of the URLLC services significantly decreases when the Wi-Fi technology is used. By way of illustration, given a PLR requisite for URLLCs of 10^{-4}, there is a gain of 11.5 Mbps of throughput to serve URLLCs. In other words, the radio interface can additionally withstand 11.5 Mbps for URLLCs guaranteeing the same PLR. In the end, it can be concluded that for this specific industrial scenario and thanks to the multi-technology functionality of the 5G-CLARITY architecture we achieve a reduction of up to 84.4 % of 5G bandwidth to reach the same throughput for eMBB users. Moreover, this reduction in the use of 5G resources for eMBB traffic can be translated into an increase of throughput for URLLC services of up to the double while the same packet loss ratio is ensured.

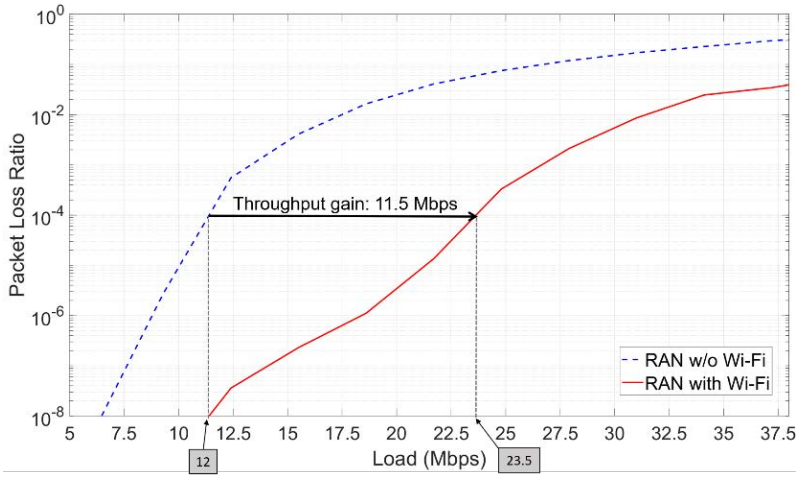


Figure 5-6 URLLC slice packet loss ratio vs the traffic load

5.3 Evaluation of Scenario 3: 5G-CLARITY slicing for URLLC services in an industrial scenario

This section includes the numerical results for showing the benefits brought by 5G-CLARITY slicing concept in terms of isolation. To that end, we rely on the E2E mean delay model detailed in Section 3.2.1. We consider the industrial scenario layout shown in Figure 5-3, while Figure 5-7 shows the specific substrate network infrastructure assumed together with the placement of the virtualized network functions (e.g., UPF and gNB-CU). For simplicity, we consider only the motion control (MC) service, characterized by the sustainable data rate and maximum burst size generated per device, for all the production lines. Specifically, we assume each production line has a fixed number of MC devices whose traffic has the same features. Figure 5-7 includes the paths followed by each slice in the midhaul network. For the sake of clarity, the path followed by the aggregated traffic from each cluster of servers to a given gNB or AP is specified all along the network, though there is a single full-duplex link interconnecting each bridges pair at most. For instance, the aggregated traffic from URLLC slices #2, #3, and #5 share the link between TSN switch #6 and TSN switch #7. Each URLLC slice generates the same amount of aggregated traffic for each gNB. In the same way, each eMBB slice generates the same amount of aggregated traffic for each AP. The main parameters for the simulations are included in Table 5-3.

The methodology used to demonstrate the effectiveness of 5G-CLARITY slicing in terms of isolation consists in comparing the E2E and per component mean delay of the following two configurations:

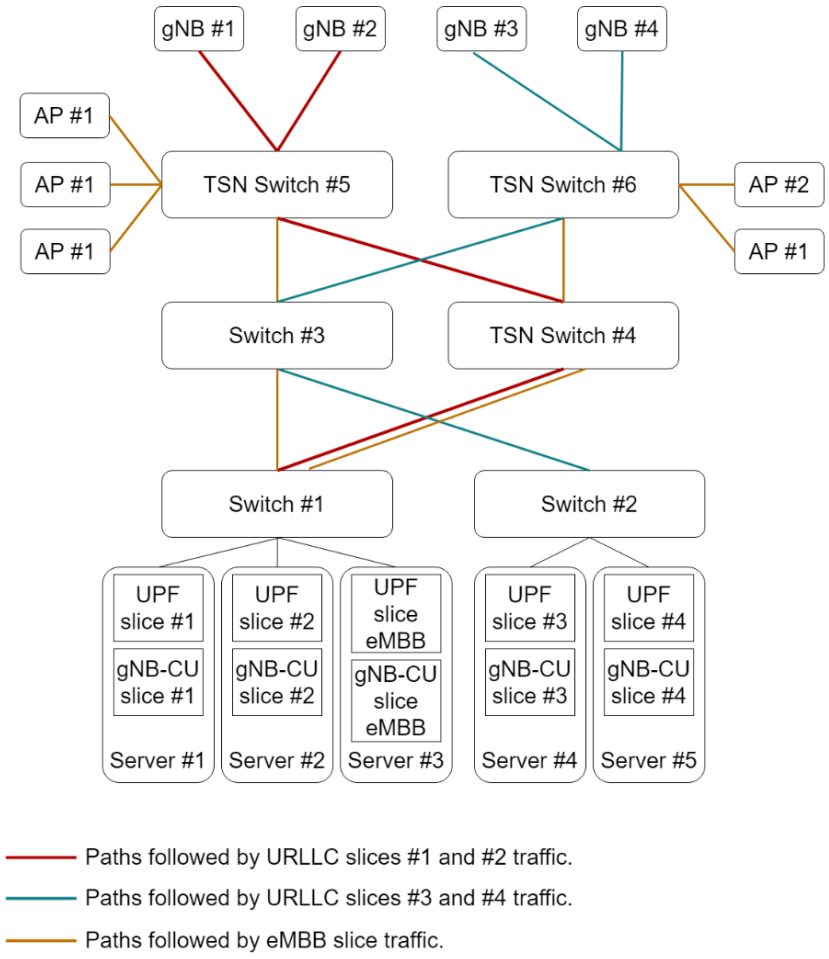


Figure 5-7 Infrastructure setup for the evaluation of the 5G-CLARITY degree of isolation

- Configuration 1: The URLLC traffic generated by each of the four production lines in the factory floor (see Figure 5-3) is served by a segregated 5G-CLARITY slice. The production line #1 generates an aggregated non-conformant traffic that does not meet the aggregated committed data rate due to a failure in its operation.
- Configuration 2: The URLLC traffic generated by all of the four production lines in the factory floor is served by a single 5G-CLARITY slice. The production line #1 generates non-conformant traffic due to a failure in its operation.

We also consider the following two variants for each scenario in order to compare the two 5G-CLARITY transport network technologies:

- Variant A: The midhaul network in Figure 5-7 is realized as a standard IEEE 802.1Q Ethernet network where there is no traffic prioritization.
- Variant B: The midhaul network in Figure 5-7 is implemented as an asynchronous TSN network, whose building block is the asynchronous traffic shaper (ATS). There is an ATS instance at every TSN bridge egress port. The ATS includes a per-flow traffic regulation through the interleaved shaping and traffic prioritization.

Table 5-3 Main Parameters for Assessing the Degree of Isolation for 5G-CLARITY Slicing

Paramaters	Value
Number of production lines	4
Number of URLLC flows per production line	56
URLLC service	Motion Control (MC)
Packet delay budget MC	1 ms
Packet length MC	80 bytes
Sustainable rate per MC flow	1.55 Mbps
Burstiness per MC Flow	2592 bits
eMBB traffic generated from server #3 to each Wi-Fi AP	AP#1: 100 Mbps, AP#2: 100 Mbps, AP#3: 100 Mbps, AP#4: 100 Mbps, AP#5: 100 Mbps
eMBB packet size	1500 bytes
UPF service rate per processing unit (CPU core)	357140 packets per second [30][31]
SCV of the UPF service time	0.65
gNB-CU service rate per processing unit (CPU core)	601340 packets per second
SCV of the gNB-CU service time	0.65
CPU core power (Intel Xeon Platinum 8180)	25.657 GOPS
gNB-DU service rate per processing unit (CPU core)	7545 packets per second [30]
SCV of the gNB-DU service time	1
gNB-RU service rate per processing unit	78530 packets per second [30]
SCV of the gNB-RU service time	1
Processing units allocated to each network component. * *Designed to ensure that the utilization of the computing resources for every component is lower than 75%.*	Configuration 1: UPF: 1 CPU core (Intel Xeon 8081) gNB-CU: 1 CPU core (Intel Xeon 8081) gNB-DU: 24 CPU cores (Intel SandyBridge i7-3930K @3.20Ghz) RU: 3 CPU cores (Intel SandyBridge i7-3930K @3.20Ghz) Configuration 2: UPF: 3 CPU core (Intel Xeon 8081) gNB-CU: 2 CPU core (Intel Xeon 8081)

	gNB-DU: 96 CPU cores (Intel SandyBridge i7-3930K @3.20Ghz) RU: 10 CPU cores (Intel SandyBridge i7-3930K @3.20Ghz)
Visit ratios of the UPF and gNB-CU	1
Visit ratios of the gNB-DU, gNB-RU and radio interface	0.5
TSN links capacities	All links have a capacity of 1 Gbps
MC traffic-to-priority level assignment at every TSN bridge output port	1 (1 is the highest priority level and 8 is the lowest)
eMBB traffic-to-priority level assignment at every TSN bridge output port	8
Radio interface time slot duration	142.8 μs
Number of PRBs dedicated for each URLLC slice per gNB	Configuration 1: Slice#1: gNB#1: 166, gNB#2: 166, gNB#3: 0, gNB#4: 0 Slice#2: gNB#1: 166, gNB#2: 166, gNB#3: 0, gNB#4: 0 Slice#3: gNB#1: 0, gNB#2: 0, gNB#3: 166, gNB#4: 166 Slice#4: gNB#1: 0, gNB#2: 0, gNB#3: 166, gNB#4: 166 Configuration 2: Slice#1: gNB#1: 333, gNB#2: 333, gNB#3: 333, gNB#4: 333
Mean number of PRBs required to transmit a URLLC packet at the radio interface	15.8
Average spectral efficiency per user	2.8173 bps/Hz (MCS index = 22)
Average SINR per user	3.5368 dB
External arrival process (to the UPF)	Poissonian

Figure 5-8 shows the results of the E2E mean packet delay for the configuration 1.A, i.e., the traffic of each production line is served by an independent slice and standard Ethernet is used in the midhaul segment. The X axis in the figure stands for the bandwidth excess generated by the production line #1. On the face of it, the results suggest that only the mean packet delay of the production line #1 is negatively affected by the non-conformant traffic, thus proving the effectiveness of 5G-CLARITY slicing for ensuring the isolation among slices. Figure 5-9 depicts a breakdown of latency by network component for the configuration 1.A. It is apparent from Figure 5-9 that the radio interface is the main bottleneck for configuration 1.A given our setup. The mean packet delay of the 5G components and radio interface for URLLC slices #2, #3, and #4 are not affected by the excess of traffic load from slice #1 as they have dedicated computing and radio resources. However, note that in a real deployment we might observe a degradation in the mean packet delay at the radio interface of the slices #2, #3, and #4 with the slice #1 traffic load excess depending on the per gNB radio resources to slices assignment. Here the increasing in the traffic from slice #1 would only negatively affect to slice #3 because of the interferences as both slices use the same radio channels at different gNBs. Nonetheless, in our analytical simulator setup, we always consider the worst-case scenario regarding the interference regardless of the traffic load and this is why the aforementioned effect is not captured in the results. As far as the midhaul network is concerned, the performance of the slice #2 is negatively affected by the traffic excess of slice #1 (observe that URLLC slices #1 and #2 share the same paths in the midhaul network in Figure 5-7). This is because the considered standard Ethernet network cannot provide per link traffic isolation, i.e., there are no means to reserve a segregated link capacity per 5G-CLARITY slice. Then, using bare Ethernet as transport network technology hinders the full isolation among 5G-CLARITY slices.

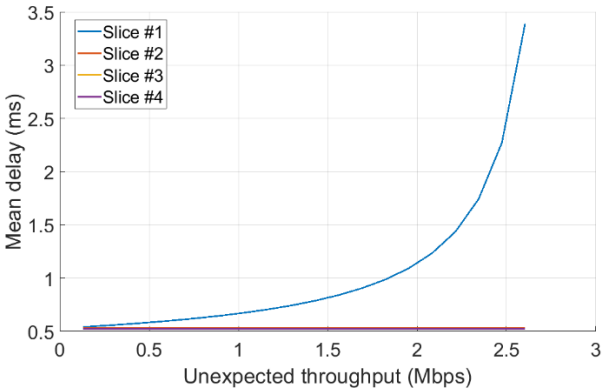


Figure 5-8 E2E mean packet delay per slice for the configuration 1.A

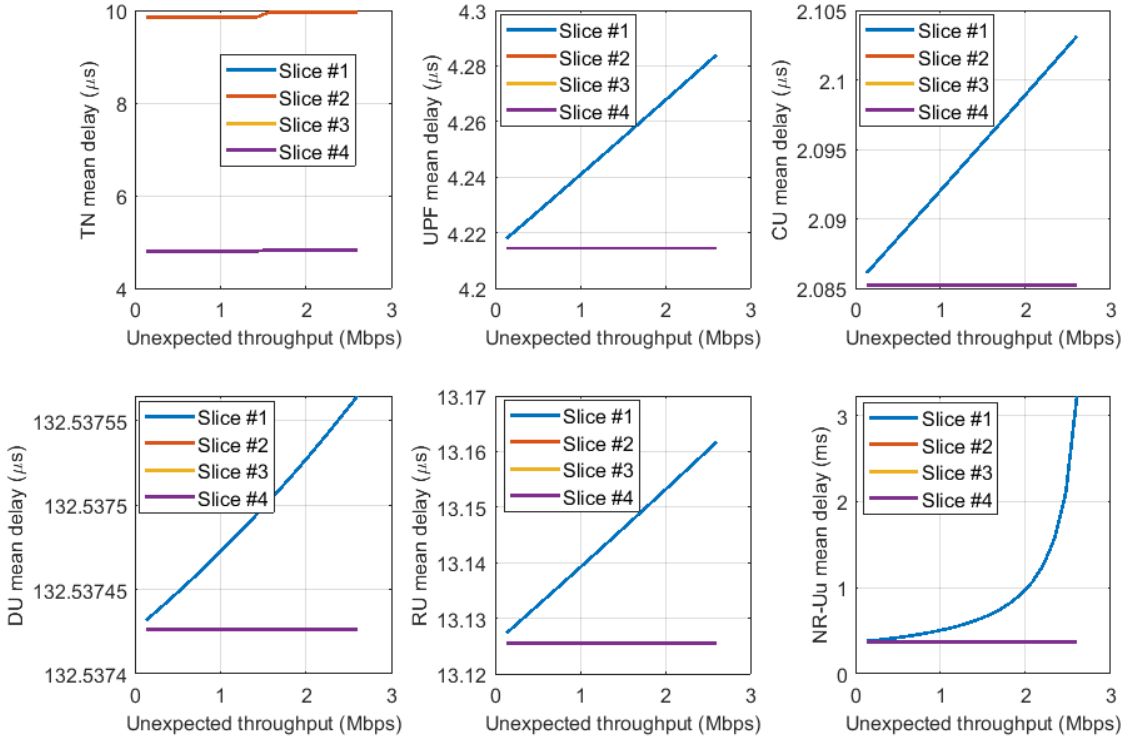


Figure 5-9 Mean packet delay per component and per slice for the configuration 1.A

Last, it is remarkable that slices #1 and #2 exhibit a higher mean packet delay at the midhaul network than slices #3 and #4 even for a low slice #1 traffic excess. That is due to the fact that slices #1 and #2 are sharing the link from switch #1 to switch #4 with eMBB traffic (see Figure 5-7) and there is no traffic prioritization.

Figure 5-10 includes the e2e mean packet delay per slice for configuration 1.B that has a similar setup as configuration 1.A previously discussed except for the midhaul network which is an asynchronous TSN network in this configuration. Again, we observe that the results suggest that only URLLC slice #1 is negatively affected by its non-conformant traffic. Nonetheless, it is remarkable that the throughput excess that eventually produces the e2e mean packet delay of the slice #1 shoots up is higher than for configuration 1.A. Observing the delay per component for the configuration 1.B in Figure 5-11, we realize that now the UPF becomes the main bottleneck of the network. Interestingly, the mean packet delays of the midhaul network, gNB-DU, gNB-RU, and radio interface for slice #1 do not depend on the traffic excess. This is due to the asynchronous TSN network performs a per flow traffic regulation at every TSN bridge egress port, thus

filtering the non-conformant traffic. Finally, we observe again that slices #1 and slices #2 have a higher mean packet delay at the midhaul network than slices #3 and #4 despite of asynchronous TSN network includes traffic prioritization. This can be explained by the fact that the traffic prioritization of the TSN network is non-preemptive. Then, since slices #3 and #4 do not share any link with eMBB traffic in the midhaul network (see Figure 5-7), their mean packet delays in this segment are lower.

Figure 5-12 shows the E2E mean packet delay for configuration 2.A. In this configuration, the URLLC traffic from all the production lines are served by the same 5G-CLARITY slice and standard Ethernet is used to implement the midhaul network. In contrast to configurations 1.A and 1.B, the non-conformant traffic from the production line #1 negatively impacts on the rest of production lines. These results further support the effectiveness of 5G-CLARITY slicing for providing isolation. As in configuration 1.A, the primary bottleneck is the radio interface as shown in Figure 5-13.

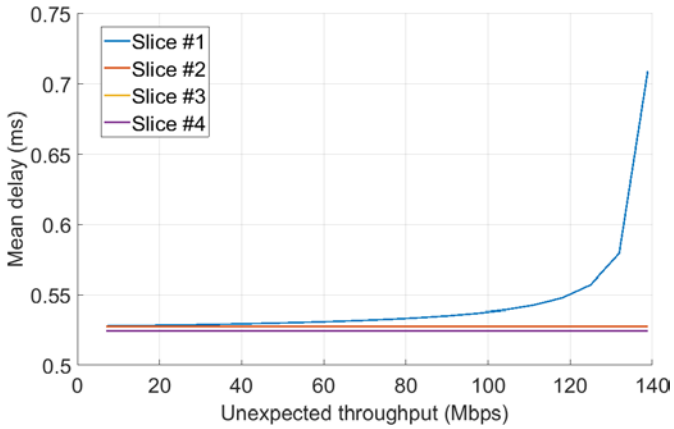


Figure 5-10 E2E mean packet delay per slice for the configuration 1.B

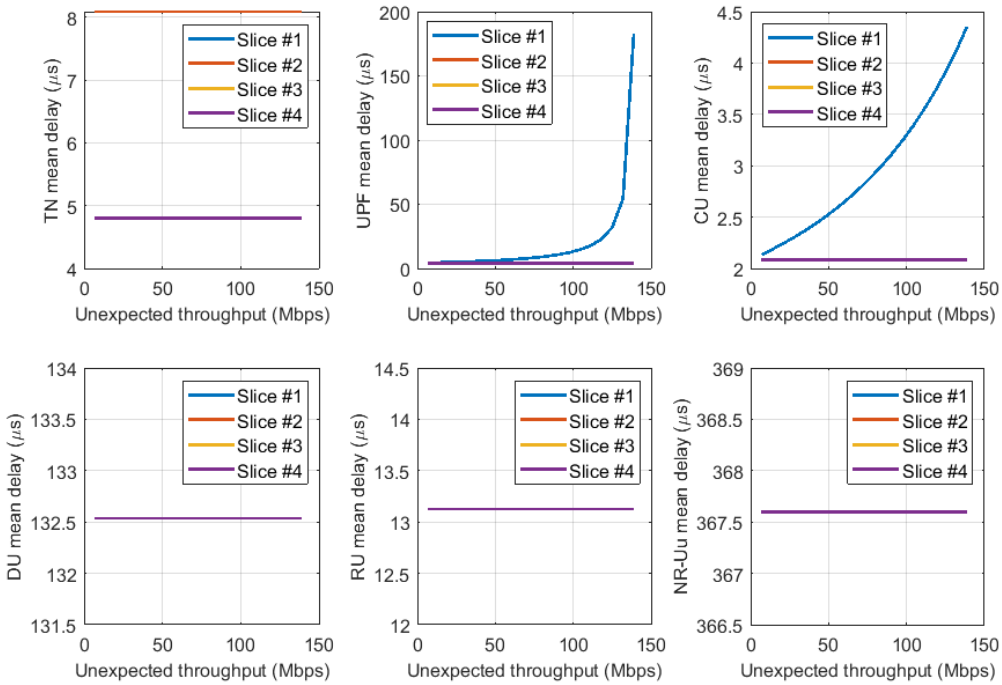


Figure 5-11 Mean packet delay per component and per slice for the configuration 1.B

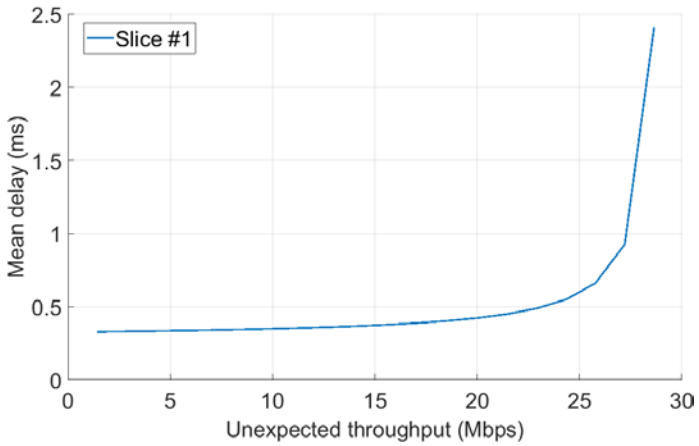


Figure 5-12 E2E mean packet delay per slice for the configuration 2.A

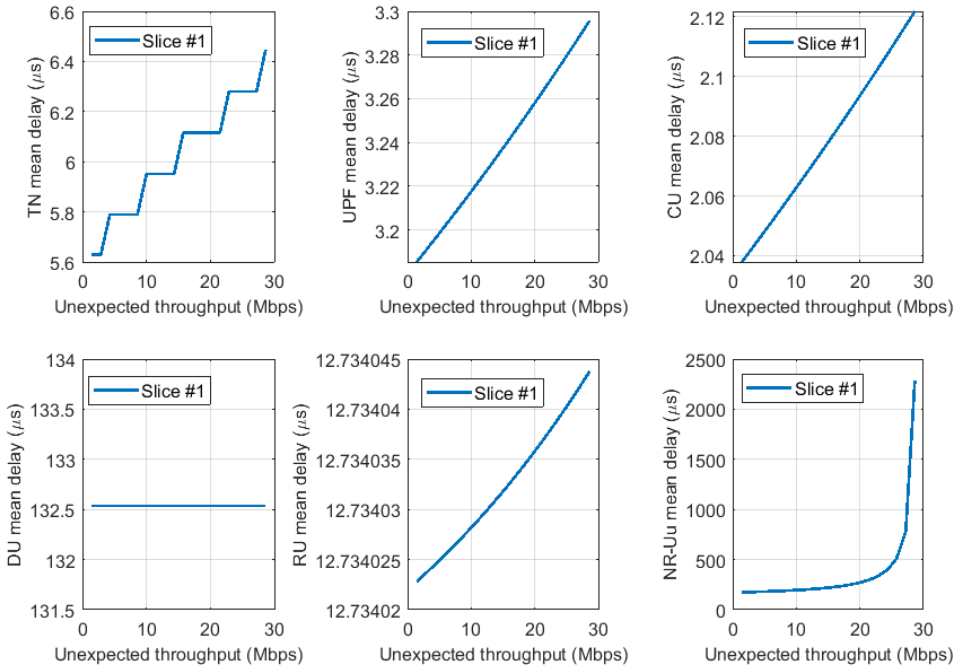


Figure 5-13 Mean packet delay per component and per slice for the configuration 2.A

Finally, Figure 5-14 includes the E2E mean packet delay for configuration 2.B which, in contrast to configuration 2.A, considers an asynchronous TSN as layer 2 technology in the midhaul network segment. As in configuration2.A, the non-conformant traffic from slice #1 negatively impacts on all the production lines. Nonetheless, the UPF becomes the primary bottleneck and the TSN midhaul network does not allow the traffic excess pass through as in configuration 1.B.

As concluding remarks, the results above suggest the effectiveness of 5G-CLARITY slicing in ensuring a quite fair degree of isolation among segregated slices. Nonetheless, the use of standard Ethernet for realizing the transport network segments hinder the full isolation of the 5G-CLARITY slices. To overcome this issue, TSN technology might be used to enable the per link dedicated resources assignment to every slice. Furthermore, TSN enhances the performance of the transport network segments due to its traffic prioritization capability, thus reducing the negative effects of the interfering eMBB traffic.

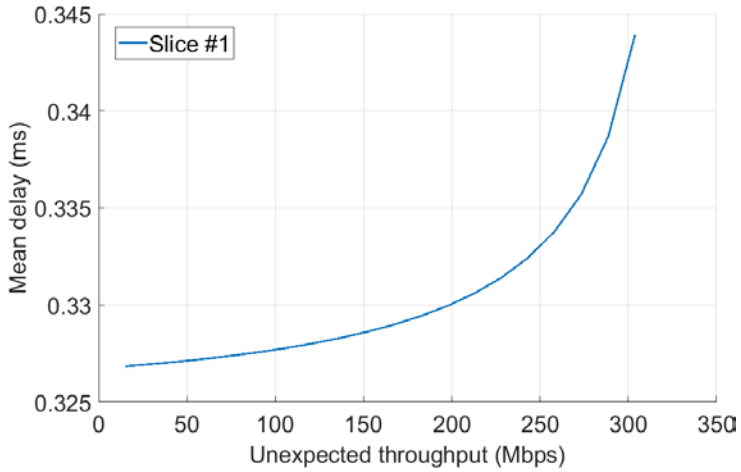


Figure 5-14 E2E mean packet delay per slice for the configuration 2.B

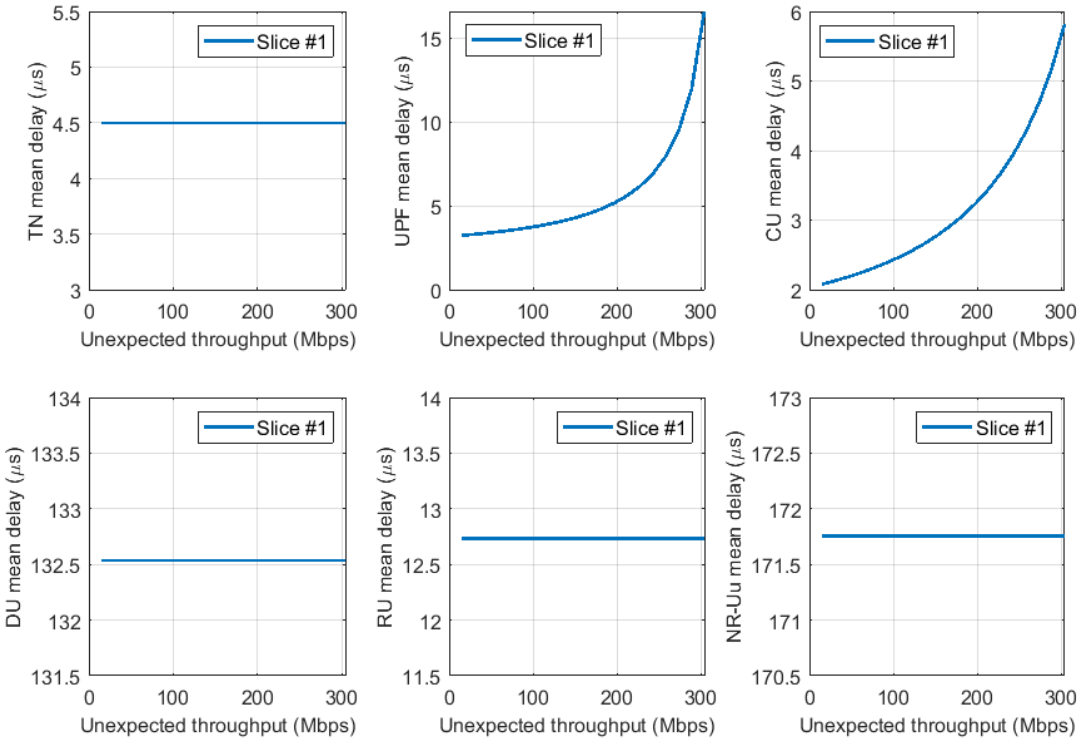


Figure 5-15 Mean packet delay per component and per slice for the configuration 2.B

5.4 Evaluation of Scenario 4: mobility and traffic load management in LiFi/Wi-Fi-integrated network

To facilitate the experimental validation of networking algorithms, such as handover, an SDN-enabled LiFi/Wi-Fi integration testbed platform was developed in the LiFi R&D centre (USTRATH) [49]. The testbed is composed of six LiFi attocells and a Wi-Fi AP. The APs are interconnected through a switch to a centralized OpenDaylight SDN controller wchich manages the SDN-enabled network via the southbound interface while supporting applications on its state transfer application program interface on the northbound.

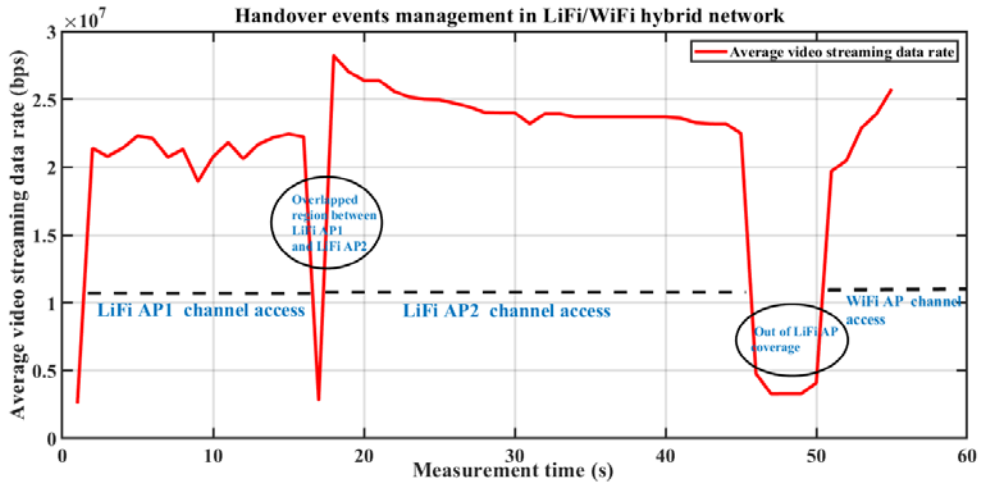


Figure 5-16 Measured average data rate during handover of user device from LiFi to LiFi and LiFi to Wi-Fi.

A LiFi access and traffic engineering application is running on top of the testbed, which supports network monitoring and management, user mobility, and network load balancing. The SDN controller has software agents running on the APs, which periodically send the state of APs to the controller. This expose, in turn, the collected network state to the developed application to support the mentioned services [41].

The testbed platform generates data relating to users, network, traffic flows, and supported services. As the testbed supports vertical handover between the heterogeneous LiFi and Wi-Fi networks, it is possible to trace the data flows of users during transitions from LiFi to LiFi and LiFi to Wi-Fi. An example of a horizontal and a vertical handover of a high-definition video service running on a mobile device is shown in Figure 5-16. The mobile user slowly moves from the centre of a LiFi AP to another LiFi AP, passing through the overlapping region. It then moves from the LiFi AP to the Wi-Fi AP. This result shows that the time for horizontal handover is shorter than the time for vertical handover, as shown in Figure 5-16. In both handover events the users experience short service disruption, which, however, is not noticeable as the service is running in a buffered mode.

A mobile connection is dropped out if a UE performs a handover into a cell, where there are no available resource units (channels). The handover dropping probability, P_{hd}, and forced termination probability, P_{ft}, are investigated as two important QoS parameters for evaluating the connection handover performance in network. The handover dropping probability is defined as the dropping probability of a mobile connection because of moving into a cell where there are no resource units (channels) can support the arrived connection [53]. The forced termination probability is defined as the probability of an in-progress connection termination due to handover dropping during its connection lifetime. The forced termination probability depends on the number of handovers during its connection lifetime and the handover dropping probability [53].

We consider the arrival of new connections in the system model follows a Poisson process with arrival rate λ per cell. The connection holding time is exponentially distributed with mean, $\frac{1}{\mu}$, and the connection dwell time in a cell is exponentially distributed with mean $\frac{1}{h}$, where the handover rate is h. It is assumed that each cell (AP) can support at most C connections. These are admitted to a cell according to a wireless connection admission control, which is modelled as an M/M/m/m queuing model. A new connection may be blocked either by a tree-based call admission control when the total number of connections in the tree exceeds a predetermined threshold N, or it can be blocked by a cell-based admission control where the available channels (i.e., resource units) are less than a predefined ratio of all channels $(\alpha C, \alpha \in [0,1])$. Let the probabilities of connection blocked by the tree-based admission control be denoted by P_t and those

blocked by the cell-based admission control be denoted by P_b. Since the two admission controls are independent, the total call blocking probability (P_{tb}) is given by $P_{tb} = 1 - \big((1 - P_t)(1 - P_b)\big)$.

A connection tree consists of seven cells. Thus, the new connection arrival rate in a tree is 7λ ; and the rate of handover connections into a tree is given by: $\lambda_{th} = 3h(1 - P_{hd})$. The $3h$ is considered as there are 18 edges in the mobile connection region and the handover rate to an edge of a cell is $\frac{h}{6}$. If the total number of connections in a tree exceeds the threshold N, then the tree admission control will deny new connections and only allow handover connections.

The effective Erlang load in any cell is given by $\frac{\lambda}{\mu}$, where $\frac{1}{\mu}$ is the average connection holding time in seconds. where $\lambda_b = \lambda(1 - P_t)$. This represents the new connection arrival rate, as the new connections should be admitted first to the tree-based admission control. The rate of handover connections into a cell is $\lambda_{bh} = h(1 - P_{hd})$. In the calculation process of λ_{bh} , P_{hd} was set initially set to 0.01.

New connections are blocked when the number of admitted connections in the cell-cluster exceeds N, where only the handover calls are admitted in the system. When the total number of in-progress connections in a cell, denoted as x, exceeds $m = \lfloor(1 - \alpha)\, C\, \rfloor$, then the cell rejects the new connections and accepts only the handover connections. So, to meet the QoS requirement, the tree-based admission control threshold and cell-based reserved fraction are used to control the handover dropping probability below a predetermined level N at all times. It is measured as follows.

$$P_{hd} = \pi_0 \left[\sum_{i=m+1}^{C} \frac{(\lambda_b + \lambda_{bh})^m \cdot \lambda_{bh}^{C-m}}{(\mu + h)^C \cdot C!} \right]$$

where

$$\pi_0 = \left[\sum_{i=0}^{m} \frac{(\lambda_b + \lambda_{bh})^i}{(\mu + h)^i \cdot i!} + \sum_{i=m+1}^{C} \frac{(\lambda_b + \lambda_{bh})^m \cdot \lambda_{bh}^{i-m}}{(\mu + h)^i \cdot i!} \right]$$

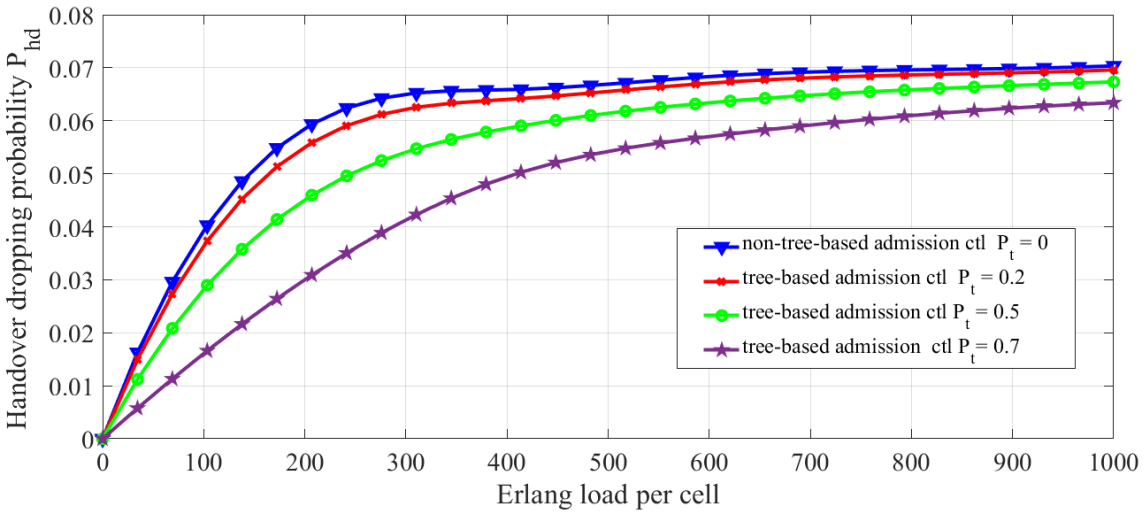


Figure 5-17 Handover dropping probability versus Erlang load under single and two-layer admission controls.

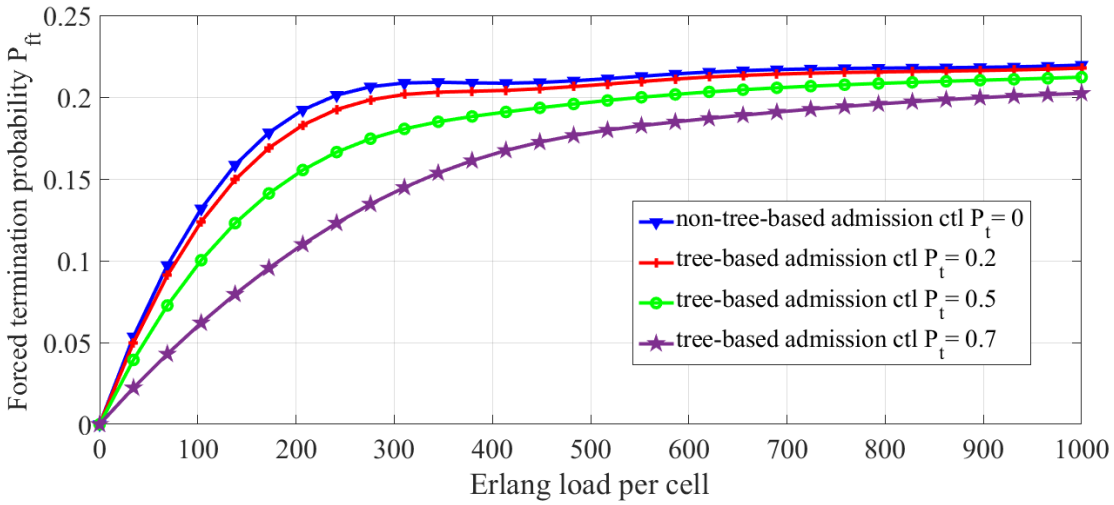


Figure 5-18 Forced termination probability versus Erlang load under single and two-layer admission controls.

From the user point of view, another QoS parameter, the forced connection termination probability, P_{ft}, is measured as follows:

$$P_{ft} = \frac{h \cdot P_{hd}}{\mu + h \cdot P_{hd}}$$

As a result, the blocking probability, hand-off dropping and forced termination probabilities are calculated as a function of Erlang load in the cell-cluster (tree) with 7 cells, each supporting up to C real-time connections. The average connection holding time $\mu = 0.1$; and the hand-off rate is $h = 0.5$ per unit of time, where the average dwell time is 2-time units.

The handover dropping probabilities are compared under the two-layer admission control and the single-layer admission control, as shown in

Figure 5-17. It is observed that as the Erlang load per cell increases, the handover dropping probability increases. In addition, the tree-based admission control ($P_t > 0$) can effectively reduce the handover dropping probability in comparison with no tree-based admission control $(P_t = 0)$. Increasing the new connection blocking probability (P_t) of the tree-based layer will reduce the effective new connection arrival rate, and thus have a smaller handover dropping probability.

Figure 5-18 shows the relation between Erlang load per cell and forced termination probability where four different tree-based new connection blocking probabilities are considered. It shows that the forced termination probability increases as the Erlang load increases. To guarantee the QoS, controlling the forced termination probability under a predefined level is an important criterion. From Figure 5-18 , a larger P_t can lower the forced termination probability. The values of P_t that the tree-based admission control should support to meet different P_{ft} requirements.

5.5 Evaluation of Scenario 5: joint synchronization and localization using multi-wireless access technologies

A wide range of high-accuracy localization techniques rely heavily on the synchronization between APs. In particular, to localize a Mobile User (MU), these techniques draw on the time measurements conducted among the APs and the MUs, requiring them to have a common time base [51]. Given that, it appears that the inter-AP synchronization, AP-MU synchronization, and MU localization problems are closely intertwined, suggesting that they may need to be tackled jointly [44].

In this section, we develop a joint synchronization and localization (sync&loc) algorithm for MUs, grounded on the hybrid synchronization algorithm developed in [47].

5.5.1 Hybrid network synchronization

To achieve network synchronization, there are in general two ways of tackling the problem: designing a network synchronization algorithm from scratch or, alternatively, expanding on the existing pairwise synchronization algorithms [46]. The former offers higher accuracy while the latter is capable of performing synchronization more frequently thanks to its low complexity. We, therefore, employ both algorithms in a hybrid manner to overcome their disadvantages and make the most of their advantages to achieve network synchronization. This will eventually lay the ground for MU joint sync&loc at the edge of the network.

The cornerstone of the above-mentioned approaches towards synchronization is time-stamp exchange. We employ the time-stamp exchange mechanism shown in Figure 5-19 and implemented by means of Precision Time Protocol (PTP). It functions as follows: node *j* transmits a sync message wherein the local time $c_j(t_1^k)$ is incorporated. Node *i* receives the packet and records the local reception time $c_i(t_2^k)$. After a certain time, the process repeats again with $c_j(t_3^k)$ and $c_i(t_4^k)$. Subsequently, at local time $c_i(t_5^k)$, node *i* sends back a sync message to node *j* with $c_i(t_2^k), c_i(t_4^k)$ and $c_i(t_5^k)$ incorporated. Upon reception, node *j* records the local time $c_j(t_6^k)$. Given that, the relation between local clocks can be written as:

$$\frac{1}{\gamma_i}(c_i(t_2^k) - \theta_i) = \frac{1}{\gamma_j}(c_j(t_1^k) - \theta_j) + d_{ij} + T_{ij}^{k,0}$$
$$\frac{1}{\gamma_i}(c_i(t_4^k) - \theta_i) = \frac{1}{\gamma_j}(c_j(t_3^k) - \theta_j) + d_{ij} + T_{ij}^{k,1}$$
$$\frac{1}{\gamma_i}(c_i(t_5^k) - \theta_i) = \frac{1}{\gamma_j}(c_j(t_6^k) - \theta_j) + d_{ij} + R_{ij}^{k}$$

where $(t_1^k, t_3^k)/t_6^k$ and $t_5^k/(t_2^k, t_4^k)$ are the time points where neighboring nodes *j* and *i* send/receive the sync messages, respectively. Variables θ_i and γ_i denote the clock offset and skew of node I, respectively. Variable d_{ij} represents the propagation time between nodes *i* and *j*. Moreover $T_{ij}^{k,0}, T_{ij}^{k,1}$, and R_{ij}^{k} denote the random variables due to the multiple hardware-related random independent processes and assumed to be i.i.d Gaussian random variables. For the sake of simplicity and without loss of generality, we assume same distribution $N(\mu_T, \sigma_T^2)$ for all of them. The time-stamps collected by this mechanism will be utilized in the synchronization algorithm to estimate the clock offset and skew estimation.

5.5.1.1 Network-wide synchronization

To achieve network-wide synchronization we draw on the statistical approach introduced in [43]. In particular, we assume a joint probability density for all clock parameters at all the nodes. The synchronization problem can then be formulated as computing the marginal distribution at each node.

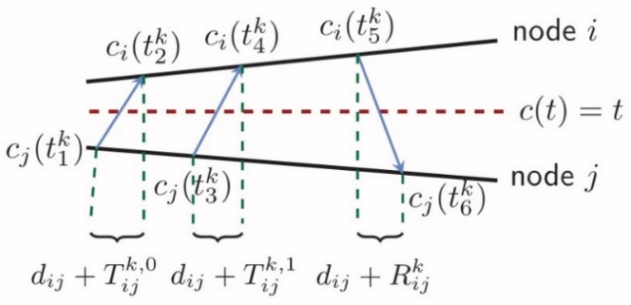


Figure 5-19 Time-stamp exchange mechanism implemented using PTP protocol [47].

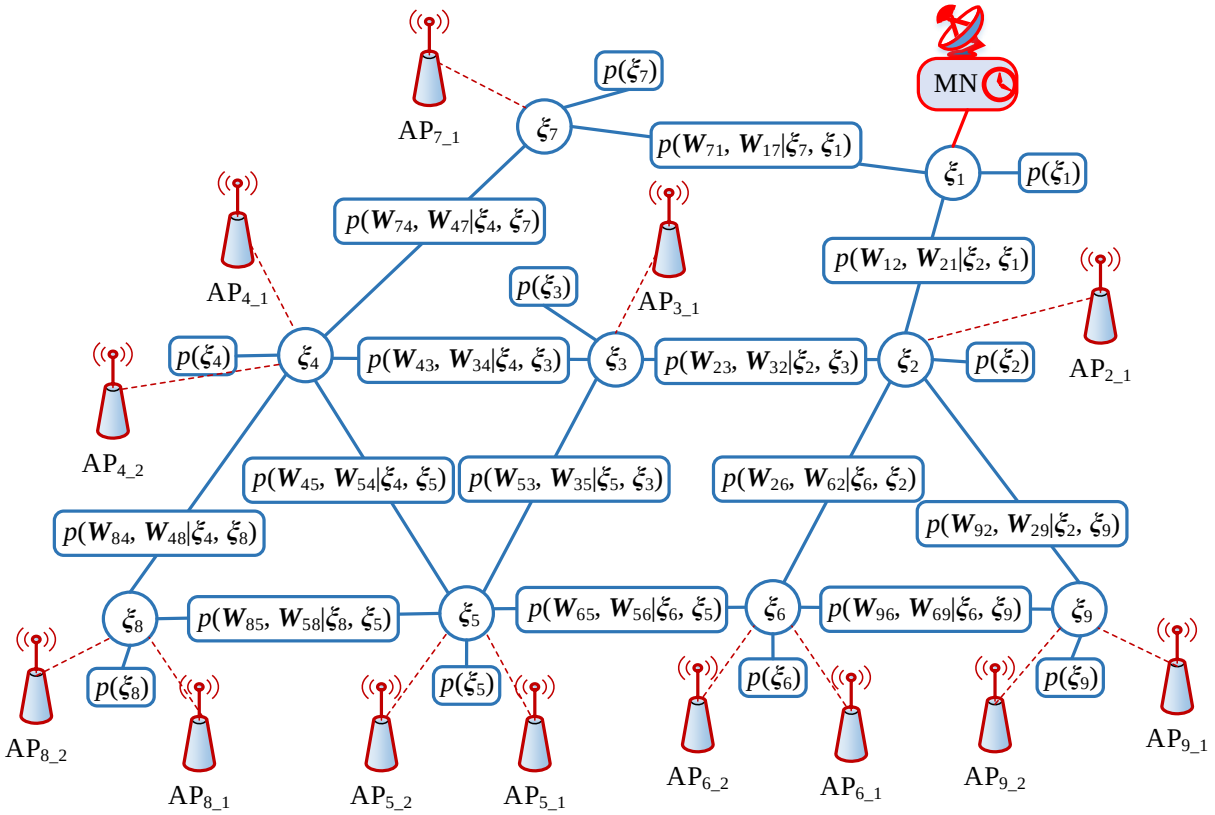


Figure 5-20 An exemplifying network where both network-wide and pairwise synchronization can be applied [48].

Mathematically this can be expressed as:

$$p(\boldsymbol{\xi}_i) = \int p\left(\xi_1, \cdots, \xi_M \middle| \{\boldsymbol{W}_{ij}, \boldsymbol{W}_{ji}\}_{i=1:M, j\in ne(i)}\right) d\xi_1 \cdots d\xi_{i-1} d\xi_{i+1} \cdots d\xi_M,$$

where $\boldsymbol{\xi}_i = \left[\frac{1}{\gamma_i}, \frac{\theta_i}{\gamma_i}\right]^T$ denote the transformed clock parameters vector. Matrixes $\boldsymbol{W}_{ij}$ and $\boldsymbol{W}_{ji}$ are constructed by means of the collected time-stamps. Nevertheless, calculating the marginal in the above equation is computationally NP hard. As done in [45], we approximate the conditional term in the integral with the multiplication of the pairwise conditional probabilities and the prior knowledge on the clock parameters. It turns out that such an approximation can be well depicted by Factor Graphs (FGs). Figure 5-20 depicts an exemplifying network where the backhauling network is shown in the form of an FG. Each node is indicated by a variable node, whose parameters is related to its adjacent nodes by the factor nodes, whose behaviour is characterised by the pairwise conditional probabilities.

To compute the pairwise conditional probabilities each node exchanges time-stamps with all of its neighbouring nodes based on the protocol shown in Figure 5-19. Furthermore, to facilitate the computation process, BP is employed to obtain the marginal at each node in a distributed manner. To this end, each node *i* receives BP messages from its neighbouring nodes *j*, updates its own belief, and propagates it back to the neighbouring nodes. The details of this iterative process can be found in [47].

5.5.1.2 Pairwise synchronization

In pairwise synchronization one node plays the role of Master Node (MN) with which the other node synchronizes itself. Assuming node *j* to be the MN, the pairwise synchronization problem can be written as estimating the clock parameters of node *i* in the *k*-th round of time-stamp exchange. This is given by

$$p(\boldsymbol{\xi}_i^k | \boldsymbol{C}_{ij}) = \int p(\boldsymbol{\xi}_i^0, \cdots, \boldsymbol{\xi}_i^k | \boldsymbol{C}_{ij}) d\boldsymbol{\xi}_i^0 \cdots d\boldsymbol{\xi}_i^{k-1},$$

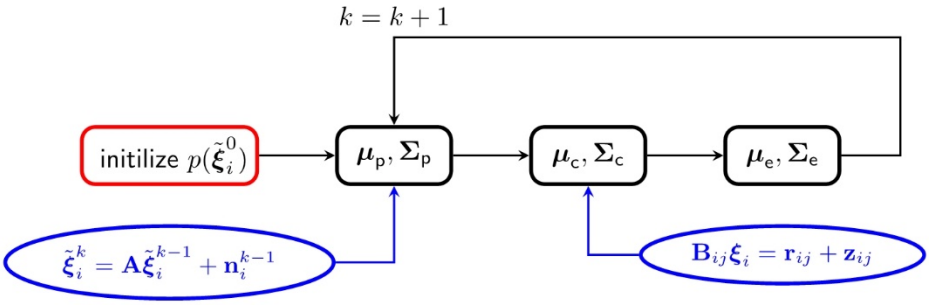


Figure 5-21 Recursive clock parameter derivation process

where $\boldsymbol{C_{ij}} = [\boldsymbol{c}_{ij}^1, \cdots, \boldsymbol{c}_{ij}^k]$ with $c_{ij}^k = [c_j(t_1^k), c_i(t_2^k), c_j(t_3^k), c_i(t_4^k), c_i(t_5^k), c_j(t_6^k)]$. A few mathematical manipulations can turn above-mentioned equation to a recursive estimation process known as Bayesian Recursive Filtering (BRF). The process comprises prediction, measurement, and estimation steps. In the prediction step, we predict the offset under the assumption of constant clock skew.

In the measurement step, we rely on the time-stamps to obtain the clock parameters. Both steps are combined in the final step to estimate the clock parameters. Such a recursive process is depicted in the Figure 5-21. The variables with subscript "p" indicate the vector parameters of prediction step, the ones with "c" indicate that of the measurement step, and finally the ones with “e” show that of the estimation steps. The blue ellipse connected to the prediction step contains the equation for predicting the clock parameters, while the one connected to the measurement step is constructed using the time-stamp exchange.

5.5.1.3 Hybrid synchronization

To ensure a low E2E synchronization error at the global level, BP can be run over the backhaul network. At the same time, we can employ the BRF algorithm to perform synchronization between the backhaul nodes and the APs at the edge of the network where fast and frequent synchronization is required to keep the relative time error small. This is, crucial to a number of applications such as localization. This synchronization scheme serves as the basis for the joint synchronization and localization in the next section.

5.5.2 Bayesian joint synchronization and localization

Grounded on the network synchronization achieved by the previously proposed algorithms, we develop a joint synch&loc technique based on time-stamp exchange between an MU and multiple APs. Our focus in this section is the edge of the network, where the APs, on one hand, are synchronized in a hybrid manner or fully using BP and, in the other hand, they perform joint synch&loc with the MUs. The latter is accomplished through exchanging time-stamps with MUs to which they have LoS.

5.5.2.1 Joint sync&loc algorithm

The principles of Bayesian joint sync&loc are similar to those described for the pairwise synchronization. However, $\boldsymbol{\xi}_i$ needs now to be updated to also account for the location-related parameters. These parameters can be uncovered when expanding the variable $d_{ij} \times v_c = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, where x_j and y_j represent the known position of *j*-th AP on the *x* and *y* axes, respectively. Variable v_c denotes the speed of light. The updated $\boldsymbol{\xi}_i$ can be then given by

$$\boldsymbol{\xi}_i = \left[\frac{1}{\gamma_i}, \frac{\theta_i}{\gamma_i}, \mathrm{x_i}, \mathrm{y_i}, \mathrm{v_{x_i}}, \mathrm{v_{y_i}}\right],$$

where v_{x_i} and v_{y_i} represent the velocity of the *i*-th MU on the *x* and *y* axes, respectively. Moreover, each AP is assumed to be able to perform Angle of Arrival (AoA) estimation given by

$$\arctan\frac{y_i-y_j}{x_i-x_j} = \phi_{ij} + n_\phi,$$

where ϕ_{ij} represents the true AoA and $n_\phi \sim N(0, \sigma_\phi^2)$ denotes the zero mean Gaussian noise resulted from the AoA estimation algorithm.

It is straightforward to see that the relation between the measurements, i.e. time-stamps, and the location parameters is nonlinear. To deal with this nonlinearity we draw on the Taylor expansion, details of which are thoroughly explained in [46][47].

5.5.2.2 Performance analysis

We perform our simulations for the pedestrian scenario shown in Figure 5-22. An MU moves with a constant velocity of 2 m/s and takes the turns randomly until it exits the map. During its journey, it is assumed that the MU exchanges time-stamps with two APs, each of which also perform AoA estimation. To analyse the impact of synchronization on the joint sync&loc, we consider two cases: a) APs synchronize to the backhauling network using only BP, and b) APs are synchronized to the backhauling nodes using the hybrid approach. Apart from that, we evaluate the performance of joint sync&loc algorithm across the uncertainty in time-stamping for both above-mentioned scenarios.

Figure 5-23 presents the Cumulative Distribution Function (CDF) of AP-MU offset estimation error as well as position estimation error. As can be seen, the offset estimation error is slightly high for the hybrid synchronization. This is in fact the cost that we bear to achieve a faster local synchronization at the edge of the network among the adjacent APs. We note that, without hybrid synchronization, when we only synchronize the whole network using only BP, the APs must wait *n* iterations (depending on the number of layers between each AP and the MN) to be synchronized to each other while in hybrid synchronization this is achieved immediately. Moreover, as can be observed, the performance deterioration of position estimation is negligible when compared to the BP synchronization.

Figure 5-24 shows the Root Mean Squared Error (RMSE) of the position/offset estimation error across the uncertainty in time-stamping for multiple number of BRF iterations. As can be seen, both RMSEs increase with the growth of σ_T. This is expected as the position estimation as well offset estimation is highly dependent on the accuracy of time measurements between the APs and the MUs. The slope of growth, however, turns out to be small, especially for scenario (a).
For scenario (b), the performance can be boosted by running more iterations of the BRF algorithm. This is in particularly straightforward as these iterations, in contrast to BP iterations, are fast and computationally inexpensive.

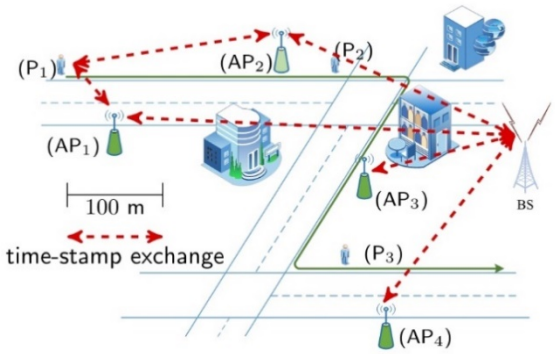


Figure 5-22 An example where MU joint sync&loc is conducted. At each point P1, P2, and P3 the MU is exchanging time-stamps with the two APs

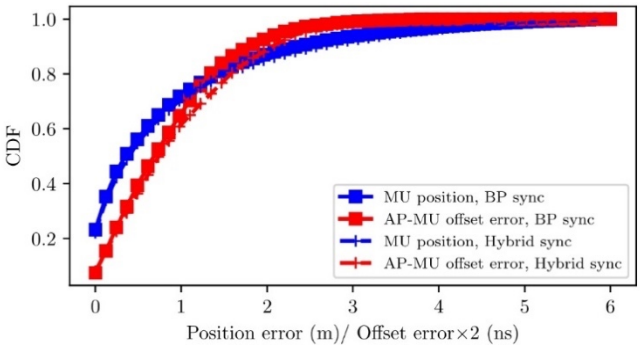


Figure 5-23 Performance of the joint sync&loc algorithm

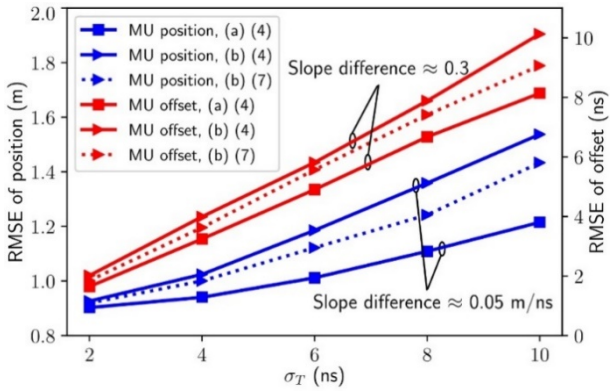


Figure 5-24 Performance of joint sync&loc algorithm across time-stamping uncertainty

6 Conclusions

5G-CLARITY aims at developing a heterogeneous 5G and beyond 5G (B5G) system integrating together a variety of wireless access technologies including 5G, Wi-Fi and LiFi suitable for private networks. This infrastructure will be operated through AI-based autonomic networking. Taking into consideration current standardisation activities and the requirements of the services and use cases that the project aims to support documented in 5G-CLARITY D2.1 [1] the project has defined a proposed architecture reported in 5G-CLARITY D2.2 [2]. The proposed architecture is structured in four strata: the Infrastructure stratum, the Network and application function stratum, the Management and Orchestration stratum, and the Intelligence stratum.

In this context, this deliverable focuses on a first evaluation of the 5G-CLARITY system that aims at quantifying the benefits of the proposed architectural features and technologies the project proposes. This analysis also aims at offering some benchmarking with respect to alternative architectural and technology approaches with similar functionality.

The work done in this deliverable has involved the following activities:

a) A review of the 5G-CLARITY high-level functional requirements and associated network capabilities, derived by the description of services expected to be supported and their associated KPIs.

b) Modelling of the 5G-CLARITY architectural functional elements organised in accordance to the overall project architectural structure described above. The reported models rely on the development of both theoretical and simulation tools describing the performance of the corresponding elements as well as experimental profiling of specific architectural elements where this has been feasible. The main functional elements that have been modelled include: the gNB nodes comprising both the DU and CU elements, the Wi-Fi and LiFi access points, TSN and standard Ethernet transport nodes, main elements of the 5G-CORE including the UPF, the SMF and the AMF, the elements providing the data management functionalities (data lake) and control plane elements including the SDN controllers used to manage the multi-wat access.

c) Development of E2E modelling capabilities exploiting the functional element models developed and integrating these in generic tools that can be used for the evaluation of the overall 5G-CLARITY architecture and infrastructure taking a system perspective. This includes functions providing convergence of the multi-technology access networks, functionalities for the provisioning of infrastructure slices, functionalities for synchronization and positioning, traffic offloading from 3g-ppp to non-3GPP infrastructures and orchestration of end-to-end resources.

d) Use Cased-based overall architectural evaluation, where the developed E2E modelling tools are fed with inputs and parameter values from the requirements derived by the use cases defined. This allows assessment of the overall 5G-CLARITY solution, indicating clear benefits with respect to the relevant state-of-the-art as well as associated trade-offs.

5G-CLARITY D2.4 will build on top of these modelling results evaluating more complex scenarios taking into account of all elements of the 5G-CLARITY solution involved in the service provisioning process. In addition to this, the theoretical models will be refined from the solutions that are currently under development taking into account more realistic constraints imposed by the hardware/software used in 5G-CLARITY.

Bibliography

[1] 5G-CLARITY D2.1, "Use Cases and Requirements2," march 2020. [Online] Available at: https://www.5gclarity.com/index.php/deliverables/

[2] 5G-CLARITY D2.2, "Primary System Architecture", October 2020. [Online] Available at: https://www.5gclarity.com/wp-content/uploads/2020/12/5G-CLARITY_D22.pdf

[3] 5G-CLARITY Deliverable D3.2: "Design Refinements and Initial Evaluation of the Coexistence, Multi-Connectivity, Resource Management and Positioning Frameworks", May 2021. [Online]. Available at: https://www.5gclarity.com/wp-content/uploads/2021/06/5GC-CLARITY_D32.pdf

[4] Hao Yu, Francesco Musumeci, Jiawei Zhang, Yuming Xiao, Massimo Tornatore, and Yuefeng Ji, "DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks," Optical Network Design and Modeling, Springer International Publishing, pp. 82-93, 2020.

[5] S. Navaratnarajah, M. Dianati and M. A. Imran, "A Novel Load-Balancing Scheme for Cellular-WLAN Heterogeneous Systems With a Cell-Breathing Technique," in IEEE Systems Journal, vol. 12, no. 3, pp. 2094-2105, Sept. 2018, doi: 10.1109/JSYST.2017.2733446.

[6] L. Simić, M. Petrova and P. Mähönen, "Wi-Fi, but not on Steroids: Performance analysis of a Wi-Fi-like Network operating in TVWS under realistic conditions," 2012 IEEE International Conference on Communications (ICC), 2012.

[7] N. Heiba, M. Alghoniemy, M. Elwekeil, A. Mokhtar and M. R. M. Rizk, "An adaptive algorithm for channel assignment and load balancing in elastic IEEE 802.11 WLANs," 2017 13th International Computer Engineering Conference (ICENCO), Cairo, Egypt, 2017, pp. 343-347, doi: 10.1109/ICENCO.2017.8289811.

[8] M. P. Anastasopoulos *et al.*, "Planning of dynamic mobile optical virtual network infrastructures supporting cloud services," *2014 European Conference on Networks and Communications (EuCNC)*, Bologna, Italy, 2014, pp. 1-5, doi: 10.1109/EuCNC.2014.6882679.

[9] 5G; System architecture for the 5G System (5GS), (3GPP TS 23.501 version 16.6.0 Release 16)

[10] I. Leyva-Pupo et al.,. Dynamic Scheduling and Optimal Reconfiguration of UPF Placement in 5G Networks. In Proc. of MSWiM '20, 2020 NY, USA, 103–111

[11] Jorgen W. Weibull, 1997. "Evolutionary Game Theory," MIT Press Books, The MIT Press, edition 1, volume 1, number 0262731215,.

[12] A. Tzanakaki et al., "Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services," IEEE Comms. Mag, vol. 55, no. 10, pp. 184-192,.2017

[13] Intel VTune Applifier, [Online]. Available at https://software.intel.com/content/www/us/en/develop/articles/intel-vtune-amplifier-2018-release-notes.html

[14] 3GPP TS 28.552 version 16.8.0 Release 16, Management and orchestration; 5G performance measurements ETSI TS 128 552 V16.8.0, 2021.

[15] ETSI TS 123 501 V16.6.0 (2020-10). 5G;. System architecture for the 5G System (5GS). (3GPP TS 23.501 version 16.6.0 Release 16)

[16] Intel White Paper, Samsung Achieves 305 Gbps on 5G UPF Core Utilizing Intel® Architecture, 12/2020, [Online]. Available at: https://networkbuilders.intel.com/solutionslibrary/samsung-achieves-305-gbps-on-5g-upf-core-utilizing-intel-architecture

[17] V. Alevizaki, A. Manolopoulos, M. Anastasopoulos, A. Tzanakaki, Dynamic User Plane Function Allocation in 5G Networks enabled by Optical Network Nodes, *in proc. Of* ECOC 2021

[18] A. Karamyshev, E. Khorov, A. Krasilov, I.F. Akyildiz, "Fast and accurate analytical tools to estimate

network capacity for URLLC in 5G systems", Computer Networks, vol. 178, 2020, 107331, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2020.107331.

[19] 3GPP TR 38.803 V14.2.0 (2017), Study on new radio access technology: Radio Frequency (RF) and co-existence aspects (Release 14)

[20] W. Whitt, "The queueing network analyzer," Bell System Tech. J., vol. 62, no. 9, pp. 2779–2815, Nov. 1983

[21] J. Prados-Garzon, P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Andres-Maldonado and J. M. Lopez-Soler, "Performance Modeling of Softwarized Network Services Based on Queuing Theory With Experimental Validation," in *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1558-1573, 1 April 2021, doi: 10.1109/TMC.2019.2962488.

[22] DongJin Lee, JongHan Park, Chetan Hiremath, John Mangan, and Michael Lynch, "Towards Achieving High Performance in 5G Mobile Packet Core's User Plane Function," Intel and SK Telecom White Paper, 2018.

[23] V. Quintuna Rodriguez and F. Guillemin, "Cloud-RAN Modeling Based on Parallel Processing," in *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 457-468, March 2018, doi: 10.1109/JSAC.2018.2815378.

[24] "IEEE Draft Standard for Local and Metropolitan Area Networks–Bridges and Bridged Networks Amendment: Asynchronous Traffic Shaping," IEEE P802.1Qcr/D2.1, Feb. 2020, 2020, pp. 1–152.

[25] J. Specht and S. Samii, "Urgency-Based Scheduler for Time-Sensitive Switched Ethernet Networks," Proc. 2016 28th Euromicro Conf. on Real-Time Syst. (ECRTS), July 2016, pp. 75–85.

[26] J.-Y. Le Boudec, "A Theory of Traffic Regulators for Deterministic Networks with Application to Interleaved Regulators," IEEE/ACM Trans. Netw., vol. 26, no. 6, Dec. 2018, pp. 2721–33.

[27] J.-Y. Le Boudec, and P. Thiran, (2001). *Network calculus: a theory of deterministic queuing systems for the internet* (Vol. 2050). Springer Science & Business Media.

[28] Jonathan Prados-Garzon, Lorena Chinchilla-Romero, Pablo Ameigeiras, Pablo Munoz, and Juan M. Lopez-Soler, "Asynchronous Time-Sensitive Networking for Industrial Networks," submitted to IEEE EuCNC 2021.

[29] B. Debaillie, C. Desset and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), 2015, pp. 1-7, doi: 10.1109/VTCSpring.2015.7145603.

[30] Nikaein, Navid. "Processing radio access network functions in the cloud: Critical issues and modeling." Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services. 2015.

[31] C. P. Li, J. Jiang, W. Chen, T. Ji, J. Smee, "5g ultra-reliable and low-latency systems design," in Proc. of 2017 IEEE European Conference on Networks and Communications (EuCNC), 2017, pp. 1-5.

[32] A. Tzanakaki et al., "Wireless and wired network convergence in support of cloud and mobile cloud services: The CONTENT Approach," European Wireless 2014; 20th European Wireless Conference, Barcelona, Spain, 2014, pp. 1-7.

[33] P. Kuehn, "Approximate Analysis of General Queuing Networks by Decomposition," in IEEE Transactions on Communications, vol. 27, no. 1, pp. 113-126, January 1979, doi: 10.1109/TCOM.1979.1094270.

[34] M. P. Anastasopoulos et al., "Planning of dynamic mobile optical virtual network infrastructures supporting cloud services," 2014 European Conference on Networks and Communications (EuCNC), Bologna, Italy, 2014, pp. 1-5, doi: 10.1109/EuCNC.2014.6882679.

[35] B R Rofoee, G Zervas, Yan Yan, Markos Anastasopoulos, A Tzanakaki, S Peng, R Nejabati, and D Simeonidou, "Hardware Virtualized Flexible Network for Wireless Data-Center Optical Interconnects

[Invited]," J. Opt. Commun. Netw. 7, A526-A536 (2015)

[36] M. Anastasopoulos, A. Tzanakaki, and D Simeonidou, "Stochastic Planning of Dependable Virtual Infrastructures Over Optical Datacenter Networks," J. Opt. Commun. Netw. 5, 968-979 (2013)

[37] M. Anastasopoulos, A. Tzanakaki, A. F. Beldachi and D. Simeonidou, "Network coding enabling resilient 5G networks," 45th European Conference on Optical Communication (ECOC 2019), Dublin, Ireland, 2019, pp. 1-3, doi: 10.1049/cp.2019.0906.

[38] 3GPP TS 22.261 version 15.6.0 Release 15. Table 7.1-1 Performance requirements for high data rate and traffic density scenarios

[39] 5G-CLARITY D2.1, "Use Cases and Requirements2," march 2020. [Online] Available at: https://www.5gclarity.com/index.php/deliverables/

[40] Alshaer, H., & Haas, H. (2020, April). Software-Defined Networking-Enabled Heterogeneous Wireless Networks and Applications Convergence. IEEE Access, 66672--66692.

[41] Alshaer, H., Uniyal, N., Katsaros, K., Antonakoglou, K., Simpson, S., & others, a. (2017, September). The UK Programmable Fixed and Mobile Internet Infrastructure: Overview, capabilities and use cases deployment. IEEE Access, 8, 175398-175411.

[42] Choi, B. D., & Kim, Y. C. (1998). The M/M/c Retrial Queue with Geometric Loss and Feedback. Computers and Mathematics with Applications, 36(6), 41-52.

[43] Du, J., & Wu, Y.-C. (2013). Distributed Clock Skew and Offset Estimation in Wireless Sensor Networks: Asynchronous Algorithm and Convergence Analysis. IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, 12.

[44] Etzlinger, B., & Wymeersch, H. (2018). Synchronization and localization in wireless networks. Foundations and Trends in Signal Processing, 12, 1--106.

[45] Goodarzi, M., Cvetkovski, D., Maletic, N., Gutiérrez, J., & Grass, E. (2020). A Hybrid Bayesian Approach Towards Clock Offset and Skew Estimation in 5G Networks. London, UK.

[46] Goodarzi, M., Cvetkovski, D., Maletic, N., Gutiérrez, J., & Grass, E. (2020). Synchronization in 5G: a Bayesian Approach. Dubrovnik, Croatia.

[47] Goodarzi, M., Cvetkovski, D., Maletic, N., Gutiérrez, J., & Grass, E. (2021). Synchronization in 5G Networks: a Hybrid Bayesian Approach towards Clock Offset/Skew Estimation and Its Impact on Localization. EURASIP Journal on Wireless Communications and Networking .

[48] Goodarzi, M., Maletic, N., Gutiérrez, J., & Grass, E. (2020). Bayesian Joint Synchronization and Localization Based on Asymmetric Time-stamp Exchange. 2020 International Symposium on Networks, Computers and Communications (ISNCC). Montreal, Canada: IEEE.

[49] Haas, H., Yin, L., Chen, C., Videv, S., Parol, D., Poves, E., . . . Islim, M. (2020, February). Introduction to indoor networking concepts and challenges in LiFi. Journal of Optical Communications and Networking, A190-A202.

[50] Kleinrock, L. (1976). Queueing Systems. New York.

[51] Koivisto, M., Costa, M., Werner, J., Heiska, K., Talvitie, J., Leppanen, K., . . . Valkama, M. (2017). Joint Device Positioning and Clock Synchronization in 5G Ultra-Dense Networks. IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS,, 16.

[52] McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., . . . Turner, J. (2008). OpenFlow: Enabling Innovation in Campus Networks. ACM Computer Communication Review, 69-74.

[53] Naghshineh, M., & Acampora, A. (1994). Design and control of micro-cellular networks with QoS provisioning for real-time traffic. Proc of IEEE International Conference Universal Personal Communications, (pp. 376–381).

[54] 5G-CLARITY D5.1, "Specification of Use Cases and Demonstration Plan", Feb. 2021. [Online] Available at: https://www.5gclarity.com/wp-content/uploads/2021/02/5G-CLARITY_D51.pdf

[55] 5G-PPP White paper, AI and ML – Enablers for Beyond 5G Networks, DOI: http://doi.org/10.5281/zenodo.4299895