6GSNS

Co-funded by the European Union

ORIGAMI

ORIGAMI

Optimized resource integration and global architecture for mobile infrastructure for 6G

Deliverable D2.1:

Initial report on requirements and definition of KVIs and KPIs

Date: 28/06/2024

Version: 1.0

DISCLAIMER

This document contains information, which is proprietary to the ORIGAMI ("Optimized resource integration and global architecture for mobile infrastructure for 6G") Consortium that is subject to the rights and obligations and to the terms and conditions applicable to the Grant Agreement number: 101139270. The action of the ORIGAMI Consortium is funded by the European Commission.

Neither this document nor the information contained herein shall be used, copied, duplicated, reproduced, modified, or communicated by any means to any third party, in whole or in parts, except with prior written consent of the ORIGAMI Consortium. In such a case, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced. In the event of infringement, the consortium reserves the right to take any legal action it deems appropriate.

This document reflects only the authors' view and does not necessarily reflect the view of the European Commission. Neither the ORIGAMI Consortium as a whole, nor a certain party of the ORIGAMI Consortium warrant that the information contained in this document is suitable for use, nor that the use of the information is accurate or free from risk and accepts no liability for loss or damage suffered by any person using this information.

The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Grant Agreement	101139270
Document number	D2.1
Document title	Initial report on requirements and definition of KVIs and KPIs
Lead Beneficiary	CMC
Editor(s)	Mika Skarp (CMC)
Author(s)	Simone Bizzarri (TIM) Alexis Duque (NETAI) Dimitris Tsolkas (FOGUS) Katerina Giannopoulou (FOGUS) Marco Gramaglia, Albert Banchs (UC3M) Mika Skarp (CMC) George Iosifidis (TUD) Fatih Aslan (TUD) Andres Garcia-Saavedra (NEC) Jose Antonio Ayala-Romero (NEC) Andra Lutu (TID) Stefan Geißler (JMU/EMN) Arifur Rahman (ISRD) Marco Fiore (IMDEA) Esteban Municio (i2CAT)
Dissemination level	Public
Contractual date of delivery	30.06.2024
Status	Final
File name	ORIGAMI_D2.1_V1.0.pdf

Revision History

Version	
0.1	Initial draft ToC
0.2	First stable draft
0.3	Ready for first round of review
0.4	Ready for second round of review
0.5	Ready for quality inspection
1.0	Final version following the Quality Check

ABBREVIATIONS

Abbreviations/ Acronym	Description
3GPP	3rd Generation Partnership Project
5GC	5G Core
5QI	5 QoS Identifier
AAL	Acceleration Abstraction Layer
ADRF	Analytics Data Repository Function
AF	Application Function
AI	Artificial Inteligent
AMF	Access and Mobility Management Function
AnLF	Analytics Logical Function
API	Application Programming Interface
ASIC	Application-specific integrated circuit
CB	Code Blocks
CCL	Compute Continuum Layer
CFA	Compute- and Fairness-aware resource Allocation
CPU	Central Processing Units
CRUD	Create, Read, Update and Delete
CSP	Communication Service Provider
CU	Central Unit
DCCF	Data Collection Coordination Function
DCH	Data Clearing Houses DCH
DICE	Dynamic Interconnections for the Cellular Ecosystem
DLP	Data loss prevention
DLT	Distributed Ledger Technology
DN	Data Network
DNN	Data Network Name
DU	Distributed Unit
E2SM	E2 service model key performance measurement
eMMB	Enhanced Mobile Broadband
eSIM	Embedded SIM
FCAPS	Fault, Configuration, Accounting, Performance, and Security
FEC	Forward Error Correction
FIFO	First In First Out
FPGA	Field Programmable Gate Arrays
FR	Functional requirements
GPU	Graphics processing unit
GSBA	Global SBA
HA	Hardware Accelerator
HMNO	Home Mobile Network Operator
IDS	Intrusion detection systems
IPX	IP Exchange
KPI	Key Performance Indicator
KPM	Key Performance Measurement
KVI	Key Value Indicator
LCM	Life Cycle Management
LLR	Log-Likelihood Ratios
LPU	Logical Processing Units
M2M	Machine to Machine
MAC	Medium Access Control
MAPE-K	Monitor-Analyze-Plan-Execute over a shared Knowledge

MCS	Mission Critical Services
MCS	Modulation and Coding Scheme
MDA	Management Data Analytics
MDAF	Management Data Analytics Function
MFAF	Message Framework Adaptation Function
ML	Machine Learning
Mmtc	Massive Machine Type Communications
MNA	Mobile Network Aggregator
MNO	Mobile Network Operator
MnS	Management Service
MOS	Mean Opinion Score
MTLF	Model Training Logical Function
MVNO	Mobile Virtual Network Operator
NF	Network Function
NI	Network Intelligence
NIF	Network Intelligence Function
NIF-C	Network Intelligence Function Component
NIS	Network Intelligence Services
NIO	Network Intelligence Orchestration
NPN	Non-Public Network
NR	New Radio
NRF	Network Repository Function
NWDAF	Network Data Analytics Function
OAM	Operations Administration and Maintenance
O-RAN	Open Radio Access Network
OTT	Over The Top
PCF	Policy Control Function
PDU	Packet Data Unit
PHY	Physical Layer
PLMN	Public Land Mobile Network
PoC	Proof-of-Concept
PPU	Physical Processing Units
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RB	Resource Block
RC	RAN Control
RIC	Radio Intelligent Controller
RIS	Reconfigurable Intelligent Surfaces
RRM	Radio Resource Management
RU	Radio Unit
SaaS	Software as a Service
SBA	Service-Based Architecture
SBI	Service-Based Interfaces
SLA	Service Level Agreement
SMF	Session Management Function
SMO	Service and Management Orchestrator
SMS	Short Messaging Service
SNR	Signal-to-Noise Ratio (SNR)
TCO	Total Cost of Ownership
TRL	Technology rediness levels
TTI	Transmission Time Interval

EXECUTIVE SUMMARY

ORIGAMI's target is to provide comprehensive architectural options for RAN, Transport, and Core functions and their interfaces, to be considered when standardizing 6G architecture in 3GPP. Deliverable 2.1 centers on a detailed analysis of the eight barriers identified by ORIGAMI, while also incorporating recognized architectural challenges in 6G to overcome such barriers. This document describes the precise requirements that ORIGAMI's innovative architectural solutions must address. Furthermore, it establishes key performance and value indicators to assess the effectiveness of ORIGAMI's solutions in various project use cases. By doing so, the document sets an important landmark towards ensuring that ORIGAMI not only meets its internal objectives but also aligns with the broader goals and constraints posed by interconnected projects. The deliverable will serve as a foundational reference, guiding the development and implementation of ORIGAMI's architecture to guarantee it is robust, adaptable, and capable of delivering measurable improvements. Through this approach, Deliverable 2.1 forms the architectural blueprint for further project deliverables and lays the groundwork for successful project outcomes and strategic advancements. Overall, the structured approach adopted in the document facilitates the positioning of ORIGAMI's architectural solutions towards significant contributions to the evolution of 6G standards.

KEYWORDS

6G architecture, mobile network barrier, 6G KVI, 6G KPI

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In Deliverable 2.1, ORIGAMI introduces the comprehensive analysis of the well-identified eight barriers, each of which presents significant challenges to the successful implementation of 6G architecture. These barriers are as follows:

Barrier #1: Unsustainable RAN Virtualization

The challenge here lies in ensuring that RAN virtualization is both sustainable and scalable. Current virtualization techniques are not equipped to handle the demands of future 6G networks, necessitating innovative solutions.

Barrier #2: Poor Interoperability of RAN Components

The lack of interoperability among RAN components hinders seamless network operation and integration. This barrier must be addressed to achieve a cohesive and efficient network infrastructure.

Barrier #3: High Latency and Unreliable Network Intelligence (NI)

High latency and unreliable NI impede the processing of complex 6G network problems. Enhancing the reliability and reducing the latency of NI are critical to managing 6G network operations effectively.

Barrier #4: Under-utilized Modern Programmable Transport

Modern programmable transport technologies are not being fully utilized. Leveraging these technologies is essential to optimize network performance and flexibility.

Barrier #5: Lack of Global Service Application Programming Interfaces (API)

The absence of standardized global service APIs limits the ability to offer seamless services across different networks and geographies. Establishing these APIs is vital for global interoperability.

Barrier #6: Obsolete Trust Model Hinders Performance

An outdated trust model can significantly hinder network performance and security. Updating this model is necessary to meet the security and performance needs of 6G networks.

Barrier #7: Inadequate Networking Data Representation

Current methods of data representation in networking are inadequate for the complex demands of 6G. Improving data representation techniques is crucial for effective network management.

Barrier #8: High Volume of Control Plane Signaling

The high volume of control plane signaling can overwhelm network resources, leading to inefficiencies. Reducing this volume is essential for maintaining network performance.

By examining these barriers, ORIGAMI has produced detailed architectural requirements specifications and guidelines. These will encompass both the anticipated functionalities and the objective technical indicators necessary for ORIGAMI's architectural models. The proposed ORIGAMI architecture is showcased in ten different use cases, each demonstrating how these models can be applied in real-world scenarios.

Links between barriers and use cases are shown in the Table 1.

Barrier	Use case	ID	Section
Unsustainable RAN Virtualization	Data-driven task offloading for reliable vRAN acceleration	SRV	6.1
	Compute- and Fariness-Aware Radio Resource Allocation Algorithms in Virtualized RANs	CFA	6.26.5
Poor Interoperability of RAN Components	Conflict Mitigation of xApps and Interoperability of O-RAN component	PIOR	6.2
	Enhancing Management and Stability in the 6G Architecture	EMSA	6.3
High Latency and Unreliable Network Intelligence (NI)	Interoperable Machine Learning Models Improving RAN Energy Efficiency	IMLE	6.4
	Compute- and Fairness-Aware Radio Resource Allocation Algorithms in Virtualized RANs	CFA	6.5
Under-utilized Modern Programmable Transport	Effective, distributed and streamlined access to u-plane computing capabilities	EAUC	6.6
Lack of Global Service Application Programming Interfaces (API)	Enabling the Global Operator Model	GMNO	6.7
Obsolete Trust Model Hinders Performance	Enabling the Global Operator Model	GMNO	6.7
	Limited Trust Network Analytics	LTNA	6.8
Inadequate Networking Data Representation	Anomaly Detection	KR	6.9
High Volume of Control Plane Signaling	Network Core traffic analysis and optimization	NCAM	6.10

Table 1: Links between barriers and use cases

D2.1 defines and describes the requirement structure associated with the use cases above in detail, hence setting the development framework needed for the implementation of ORIGAMI's architectural models. For each use case, an initial set of Key Value Indicators (KVI) and Key Performance Indicators (KPI) are identified to ensure that the solutions are both effective and measurable. The ultimate goal is to ensure that ORIGAMI's architectural solutions are robust, efficient, and capable of meeting the evolving needs of the project. This deliverable sets the foundation for achieving this goal by providing a clear and structured approach to identifying and addressing the key barriers and requirements.

Through this comprehensive approach, ORIGAMI aims to lay the groundwork for successful project outcomes and strategic advancements. Ensuring that ORIGAMI not only meets its internal objectives but also aligns with the current efforts in the SNS and in standardization for a is one of the guidelines followed in this document. The document is structured as follows: in Section 2 we introduce the Glossary used in this document and in the project in General. We discuss relevant state of the art architectures in Section 3 before discussing the Barriers in Section 4. We detail the ORIGAMI architectural innovations in Section 5 and the attached use cases in Section 6. Finally, we discuss our KVI framework in Section 7 before concluding in Section 8.

2 GLOSSARY

In this section, a common terminology framework to be used throughout ORIGAMI is presented. As often naming is slightly different according to the context, this section should always be considered as the main reference for the project activities and documents generated.

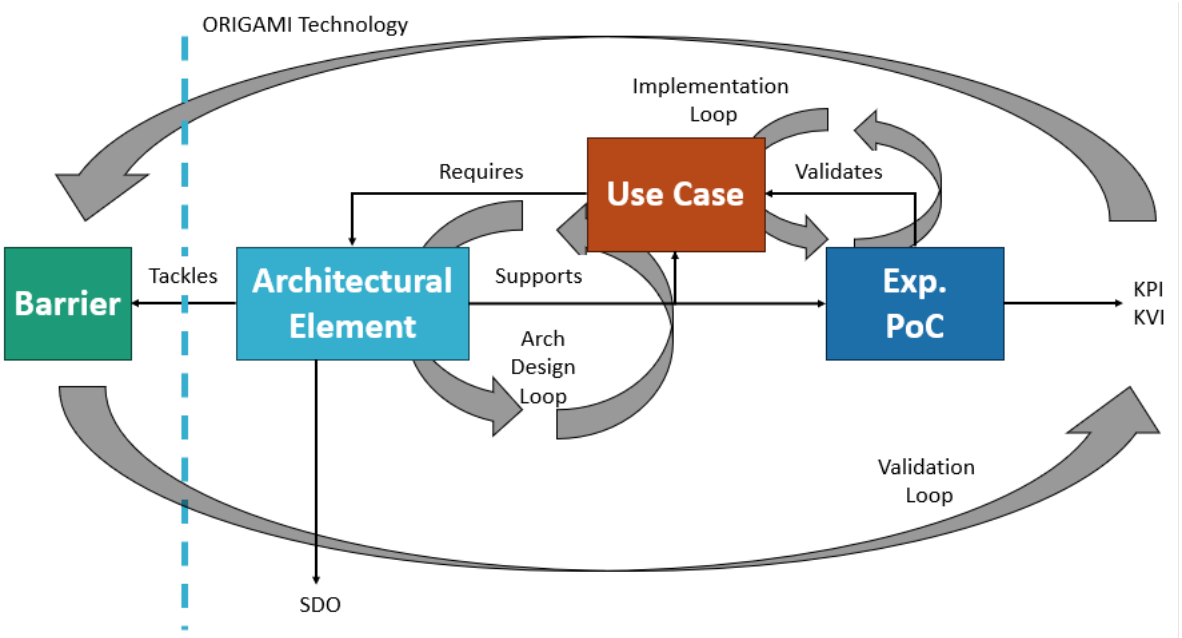

Figure 1: ORIGAMI project research methodology

2.1 BARRIER

In ORIGAMI, a Barrier is a technological limitation that is currently hindering a smooth transition towards the next generation (the 6th) of the mobile network architecture. The ORIGAMI consortium is working towards removing such barriers (described in Section 1) through the introduction of three architectural innovations (described in Section 5) that enable new use cases (discussed in Section 6).

2.2 USE CASE

In ORIGAMI, a use case is a new application or service enabled by tearing down one or more barriers, using at least one architectural element introduced by the project. use cases, analogously to Work Item and Study Items in 3GPP, introduce a novel functionality that classifies as 6th generation.

2.2.1 USE CASE REQUIREMENTS

In ORIGAMI two types of requirements are considered.

- Functional requirements (FR): They specify how the ORIGAMI technology should behave. In particular, ORIGAMI attaches functional requirements to use cases and Architectural modules. Functional requirements detail the capabilities the system must provide.
- Non-functional requirements (NFR): They define how a system performs in terms of KPI, constraining the operation of the Network Intelligence algorithms empowering a Use cases.

While functional requirements address the specific actions and responses of the system, non-functional requirements set the standards and constraints that ensure the operation under various conditions.

For Functional Requirements ORIGAMI further specifies the stages. Stage 1 Functional Requirement introduces Functional Requirements on the architectural elements that support a use case, Stage 2 also introduces such requirements on the NI functionality that empowers them.

2.3 USED TERMINOLOGY DEFINITIONS

2.3.1 TECHNOLOGY READINESS LEVEL

The TRL definition defined by the EC [1] is used in ORIGAMI as reference.

2.3.2 PROOF-OF-CONCEPT

In the context of ORIGAMI, *"a Proof of Concept (PoC) refers to a preliminary implementation designed to demonstrate the feasibility and potential of ORIGAMI's proposed architectural solutions in addressing the identified barriers within 6G networks"*.

Each PoC aims to validate the theoretical models and concepts by showcasing their practical applicability in real-world scenarios. Most of the Use cases introduced in ORIGAMI will lead to a Proof-of-Concept (PoC) that allows quantifying the Non-Functional Requirements set for the relevant use case. The target TRL for them is 3-4 at the end of the project. By successfully implementing a PoC, ORIGAMI can demonstrate the practical feasibility of its innovative architectural solutions, thereby laying a strong foundation for their adoption in the standardization of 6G architecture in 3GPP.

2.3.3 PILOT

In the context of ORIGAMI, *"a pilot refers to a small-scale implementation of the proposed architectural solutions in a real-world environment to test their functionality, performance, and feasibility after PoC deployment"*.

The ambition of the project is to impact industrial processes with the developed technology. Hence, especially the industrial partners of the project aim to provide pilots for the project technology (TRL 5-6). By successfully conducting a pilot, ORIGAMI could demonstrate the practicality and effectiveness of its architectural solutions in a real-world setting, ensuring they are robust, efficient, and ready for broader deployment. This step is crucial for building confidence and ensuring that the solutions can meet the demands of future 6G networks.

2.3.4 DEMONSTRATION

In the context of ORIGAMI, *"a demonstrator refers to a practical implementation or prototype of the proposed architectural solutions designed to showcase their capabilities and effectiveness".*

Unlike a PoC or a pilot, a demonstrator should be showcased at large events and congresses, yielding thus to a demonstrator. A demonstrator is used to exhibit the technology to stakeholders, including potential users, partners, and standardization bodies, to illustrate how the solutions work in a controlled but realistic environment. By providing a tangible and interactive representation of ORIGAMI's architectural solutions, a demonstrator should help stakeholders visualize the potential impacts and benefits, thereby facilitating better understanding, acceptance, and support for the project.

2.4 NETWORK INTELLIGENCE

The Use cases developed in the project are empowered by Network Intelligence Solutions (i.e., AI/ML Solutions applied to network functions as well as other solutions based on different autonomous algorithms) and require specifical features from the architectural modules, as captured by the Functional Requirements.

2.5 Network Domain

A network domain is a set of functions or infrastructural components that are specific to a given part of the network. For instance, the network function domain (encompassing core, transport, and access), the orchestration domain, the management domain, the infrastructure domain, and the service provider (including the Global Operator) domain.

2.6 Layer / Plane / Stratum

A Layer is an architectural element that connects two different domains. For instance, the CCL bridges the network function and infrastructure domains, the ZTL bridges the network function domain (of a given operator) to the others, creating for instance the Global Operator Model.

A subset of functions that are devoted to a specific task in a domain is defined as plane: for instance, the network function domain can be split into control plane and user plane functions.

Stratum is a subset of functions that span across multiple domains. For instance, the Network Intelligence Stratum integrates all the AI/ML functionality running in the different domains in the network. Thus, layers should support Stratums.

2.7 Mobile Network Aggregator (MNA)

Network operator model that upgrades the Mobile Virtual Network Operator (MVNO) approach to leverage existing infrastructure from multiple base operators in different countries, thus providing (close-to) global mobile connectivity. The MVNO is a virtual operator that does not run a full mobile network infrastructure to offer services to the end-users, and instead rents infrastructure from a single base operator, via contractual agreement.

3 State of the art Architectures

The architectural work of the ORIGAMI project does not propose a clean-slate approach. Instead, it builds upon the architectural foundations laid by ongoing activities. Below, ORIGAMI presents the most relevant ongoing and research activities and frameworks for the network domains that are targeted by the project: Access, Core, and Management for network services.

3.1 O-RAN

The O-RAN Alliance is a major carrier-led effort to define an open RAN architecture. The architectural model currently proposed by the O-RAN Alliance is depicted in Figure 2.

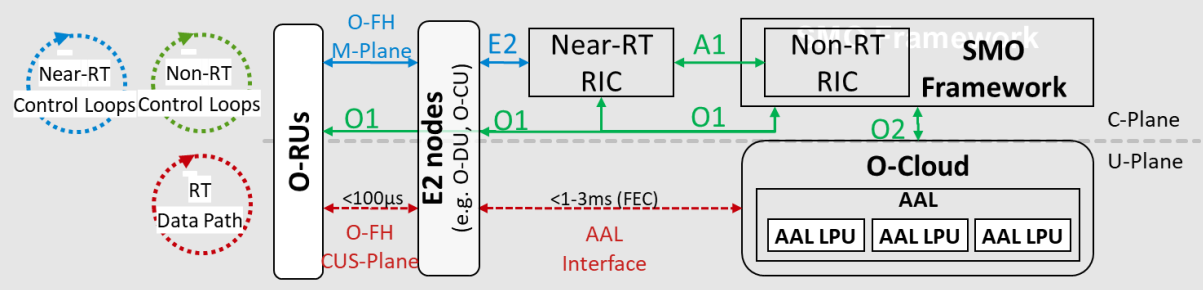

Figure 2: O-RAN architecture. The diagram illustrates the main components and their interfaces. The former include: Service Management and Orchestration (SMO) Framework, RAN Intelligent Controllers (RIC), Acceleration Abstraction Layer (AAL), Logical Processing Units (LPUs) Distributed Unit (DU), Central Unit (CU), Radio Unit (RU), and Open Fronthaul (O-FH)

The control plane includes two components, the Non-Real-Time RAN Intelligent Controller (Non-RT RIC) and the Near-RT RIC, that use A1 and E2 interfaces to manage network functions (NFs) such as 5G Central Units (CUs), 5G Distributed Units (DUs), 5G Radio Units (RUs), or 4G eNBs at, respectively, >1 s and >10 ms timescales.

On the one hand, the Near-RT RIC is a logical function that facilitates near-real-time optimization, control, and data monitoring of CU and DU nodes. It operates on a timescale between 10 milliseconds and 1 second. The Near-RT RIC receives guidance from the Non-RT RIC in the form of policies and machine learning models. While its primary focus is radio resource management (RRM), the Near-RT RIC also seamlessly supports third-party applications known as xApps. On the other hand, the Non-RT RIC is a part of the SMO and offers the A1 interface to the Near-RT RIC. It optimizes the RAN over longer timescales (seconds or minutes), creates policies, manages ML models (including training), and handles other radio resource management functions. Additionally, it adapts data management requests for the O1/O2 interface and shares contextual information with the Near-RT RIC via A1.

On top of that, the Service and Management Orchestrator (SMO) functions as a central hub for network orchestration and management, consolidating a range of services. It can potentially handle tasks beyond RAN management, such as 3GPP (NG-) core management and network slicing. In the O-RAN context, the SMO's primary tasks are FCAPS (fault, configuration, accounting, performance, and security) interfacing with O-RAN network functions, long-term RAN optimization, and managing O-Cloud resources (including scaling, software updates, and CRUD operations) through the O2 interface.

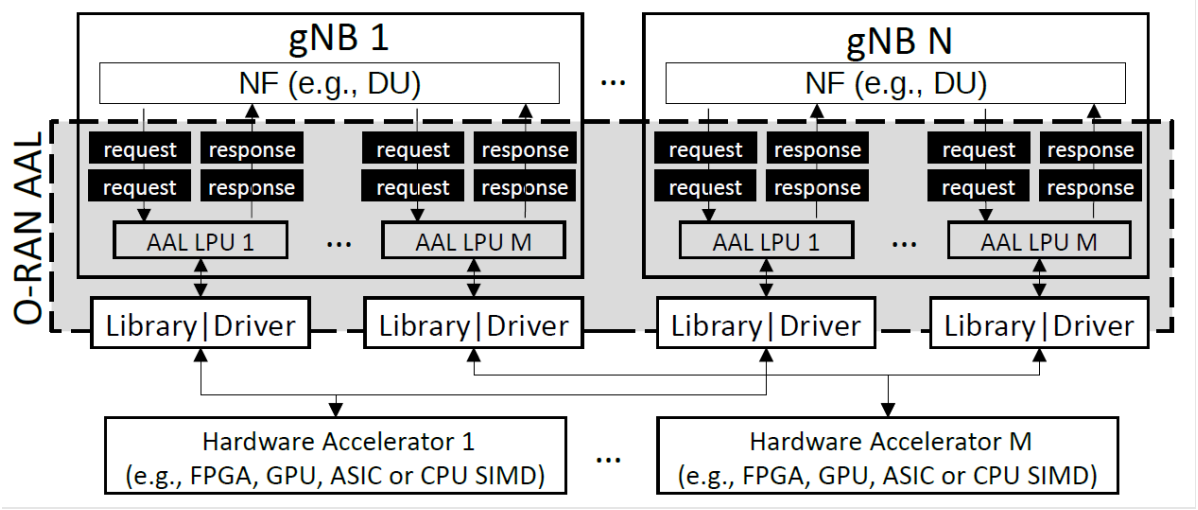

Figure 3: O-Cloud high-level architecture

The data plane has the O-Cloud, which provides computing resources, including CPUs and Hardware Accelerators (HAs), to NFs through an Acceleration Abstraction Layer (AAL) to support RAN NF virtualization.

There are two HA models, typically implemented with ASICs, FPGAs or GPUs: in-line, which processes PHY operations as wireless symbols arrive, without software intervention; and look-aside, which operates on data managed by a software controller to perform selected tasks like LDPC decoding. Traditionally, in-line HAs offer lower latency than lookaside HAs because the former does not require software mediation. However, in-line HAs tied the complete pipeline of Fig. 2 to the choice of HA, thus limiting the advantages of virtualization. Moreover, the performance gap between the two models is quickly closing, suggesting that look-aside HAs may become predominant.

The AAL abstracts O-Cloud resources (CPUs or HAs) as Logical Processing Units (LPUs). As shown in Figure 3, each LPU is dedicated to one NF via individual FIFO queues. Consequently, though a physical processor (CPU or HA) can be shared among NFs, the state of each LPU (e.g., its queue occupancy) is not shared. The O-Cloud is governed by the Service & Management Orchestrator (SMO) through the O2 interface but operates on several-second timescales. Moreover, the Near-RT RIC lacks O-Cloud visibility, hindering real-time compute-aware radio policies and DU coordination.

3.2 3GPP SA2

The Network Automation Framework [2] in 5G systems is a comprehensive approach designed to support the efficient operation of multi-service and multi-tenant networks through automation. This framework is pivotal for the deployment and management of 5G networks, allowing for enhanced responsiveness, scalability, and flexibility.

At the core of this framework is the Network Data Analytics Function (NWDAF), which plays a central role in the automation ecosystem. NWDAF is responsible for collecting data from different network functions (NFs) and other sources, processing this data to generate insights and analytics, and disseminating these insights to other NFs to facilitate informed decision-making.

The framework categorizes the network into three main domains, each with specific roles and interactions with NWDAF:

- 5G Core (5GC): Within this domain, NWDAF interacts with other core network functions to provide analytics that can help optimize network operations. For instance, analytics reports generated by NWDAF can inform network slicing, resource allocation, and other core network functions, enabling a more efficient and dynamic network operation.

- Operations Administration and Maintenance (OAM): In this domain, NWDAF plays a crucial role in enhancing network management and maintenance. It provides analytics that can aid in network monitoring, fault detection, and the automation of routine maintenance tasks. By leveraging real-time data, OAM can proactively address issues and optimize network performance.

- Service Domain: Here, the Application Function (AF) can interact with NWDAF to gain insights into network performance and user experience. These insights can help service providers tailor their services to better meet user needs and optimize service delivery over the 5G network.

The Network Automation Framework supports various services and analytical IDs, enabling NFs, OAM, and AFs to subscribe to or receive specific analytics reports. This capability allows for a wide range of automated functions, from predictive maintenance to dynamic resource allocation, enhancing the overall efficiency and effectiveness of the network.

Data gathering is a critical component of this framework, where NWDAF collects information from multiple sources within the network. This data is then used to generate analytics reports, providing valuable insights into network performance, user behavior, and potential issues.

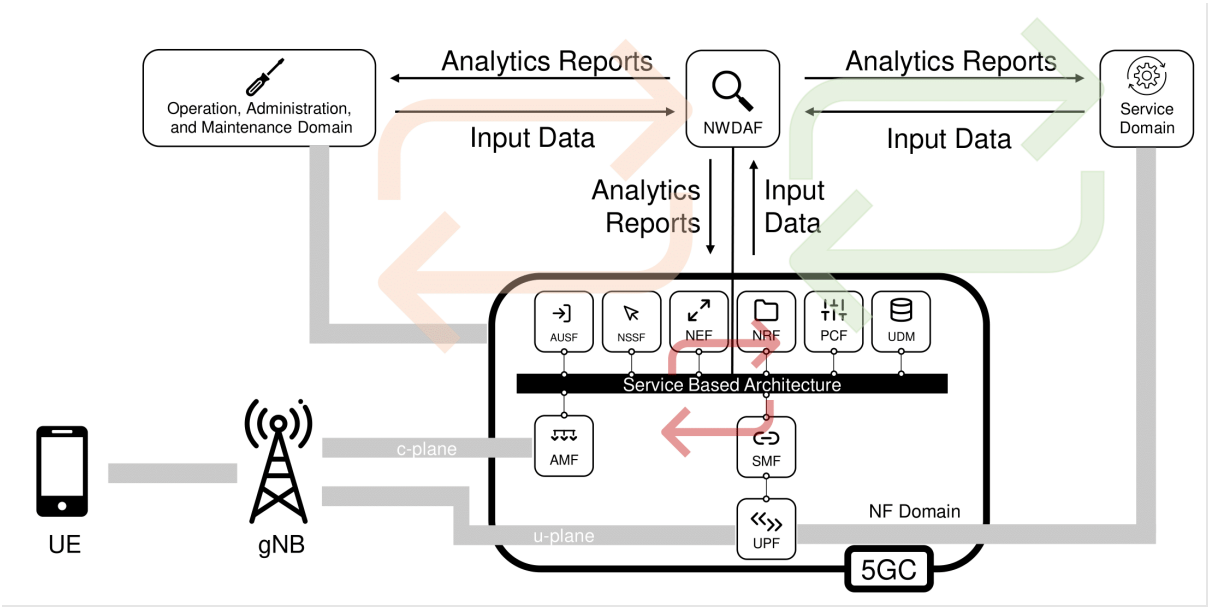

Figure 4: Feedback loops enabled by the NWDAF analytics [3]

3.2.1 ANALYTICS

NWDAF offers two primary service-based interfaces (SBIs):

- Nnwdaf_AnalyticsInfo: This service facilitates one-shot requests and responses, useful when a consumer needs a single analytic report from NWDAF. It allows for immediate retrieval of data-driven insights without establishing a long-term subscription.

- Nnwdaf_AnalyticsSubscription: Unlike the one-shot service, this subscription service supports ongoing interactions, allowing consumers to subscribe to and receive analytics reports periodically. This is instrumental for consumers needing continuous data monitoring or predictive analytics.

For generating analytics, NWDAF relies on comprehensive data gathering from multiple network sources:

- 5G Core Network Functions (NFs) and Application Functions (AFs): NWDAF collects data related to network operations, user behaviors, and service performance. This data could include metrics on network usage, service quality, and user experience.
- Operations Administration and Maintenance (OAM): NWDAF may also receive data from OAM systems, including performance metrics and maintenance data, enriching the context for its analytics.

The process of requesting analytics involves specifying an Analytics ID, which defines the type of report or insight needed, and a Target, detailing the specific network element, user group, or service to be analyzed. The analytics request can include filters to focus on relevant data points and reporting information to customize the delivery of analytics reports.

By providing these services and utilizing detailed analytics requests, NWDAF enables a wide range of network optimization and automation scenarios. This includes enhancing network efficiency, improving service quality, predicting network issues, and customizing user experiences. The flexibility and depth of the analytics offered by NWDAF underscore its critical role in the adaptive and intelligent operation of 5G networks. The overall set of analytics is summarized in Table 2.

Analytics ID	Definition and Data Gathering	Use case
Slice Load Level	This analytics provides information on the overall load of a network slice or NS instance, gathering data related to UE registrations, PDU sessions, resource usage, and load levels from NFs like AMF and SMF.	Used for decisions on resource provisioning, to throttle UEs or PDU sessions to avoid exceeding available resources, or for network orchestration to scale infrastructure accordingly.
Observed Service Experience	Focuses on the service experience quality, collecting data from AF, AMF, SMF, and UPF on metrics like Mean Opinion Score (MOS), delay, loss rate, and throughput.	Enables PCF to modify 5QI of flows, or SMF to select UPFs based on high-QoS demands, ensuring the service quality aligns with user expectations and network capabilities.
Network Function Load	Provides load information on one or more NFs, collecting data from NRF and UPF about their resource usage and status.	Useful for capacity planning, allowing AMF or SMF to select less-loaded NFs, optimizing network performance and resource utilization.
Network Performance	Analyzes performance in a specific area, including RAN performance, using data from AMF, OAM, and UPF to assess aspects like successful PDU sessions, handovers, and UE locations.	Helps in understanding network health for a given area, supporting decisions related to network configuration or resource allocation to maintain or enhance service quality.
UE Mobility Analytics	Provides insights on UE or group of UEs' mobility patterns, gathering location and timestamp data from AMF and OAM.	Assists in optimizing registration areas, adjusting paging strategies, or planning resource allocation for expected UE movements, enhancing network efficiency and user experience.
UE Communication Analytics	Offers predictions or statistics on UE communication patterns, using data from AMF, SMF, UPF, and AF, focusing on aspects like data rate and traffic volume.	Supports mMTC services by optimizing network access and control channel usage, helping in setting appropriate session

		inactivity timers, or tailoring PDU session parameters.
Abnormal UE Behaviour Analytics	Targets IoT devices to monitor unusual behaviors, collecting data from AMF, SMF, and AF to identify exceptions like unexpected locations or data usage patterns.	Enables rapid identification and response to potential security threats or malfunctions in IoT devices, maintaining network integrity and service reliability.
User Data Congestion	Provides insights into data congestion, sourcing information from AMF, OAM, and UPF about control and user plane congestion levels and identifying contributing applications.	Helps in managing network traffic, identifying bottlenecks, and making informed decisions to alleviate congestion, ensuring consistent service delivery.
QoS Sustainability	Focuses on QoS changes within an area, using data from OAM about RAN metrics to predict or report on QoS sustainment for different flows and 5QIs.	Aids in adjusting QoS parameters dynamically based on real-time network conditions, supporting service continuity and adherence to SLA requirements.

Table 2: 3GPP analytics

3.2.2 RECENT ADVANCES

Release 17 of the Network Automation Framework introduces the Analytics Logical Function (AnLF) and the Model Training Logical Function (MTLF), splitting NWDAF's functions for enhanced specialization. New functions like the Data Collection Coordination Function (DCCF), Analytics Data Repository Function (ADRF), and Message Framework Adaptation Function (MFAF) are added. Additionally, five new types of analytics are introduced to expand the framework's analytical capabilities.

Release 18 enhances NWDAF's operations with features for computing the accuracy of machine learning models and supports federated learning. It introduces new functionalities for roaming scenarios, finer UE location granularity, and a standardized interface for data collection from UPF. Furthermore, five additional analytics types are introduced, enhancing the analytical depth of NWDAF.

3.3 3GPP SA5

SA5 is the 3GPP working group responsible for producing specifications related to OAM and orchestration functionalities for communication services and mobile network functions. The working group also defines specifications for charging aspects.

SA5's reference architecture, introduced for 5GS, is a Service-Based Management Architecture, with each functionality defined as a service following the SOA paradigm.

SA5 defines each management service as a set of three main components:

- Component A, which includes management operations and/or notifications that are agnostic with regard to the entities managed (e.g., CRUD functions);
- Component B, which refers to information models representing the managed entities (e.g a network function or a network slice information model);
- Component C, which represents performance and fault information of the managed entity.

The management architecture is structured as a set of Management Functions (see Figure 5. Each function produces management services and consumes management services produced by other functions.

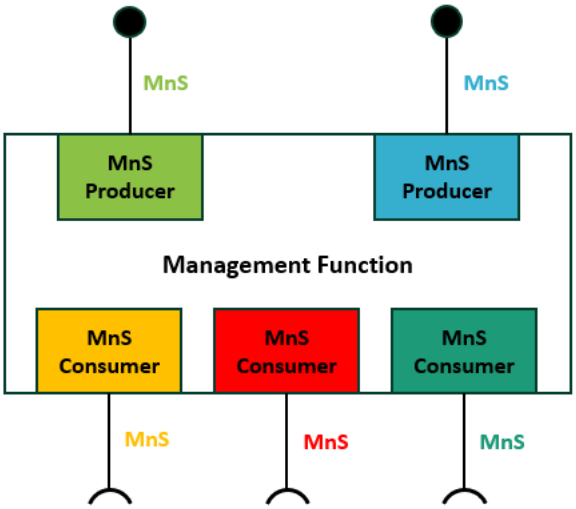

Figure 5: General Management Function structure

The management system is structured by using different management functions interacting with each other to define a completely open and flexible system (see Figure 6).

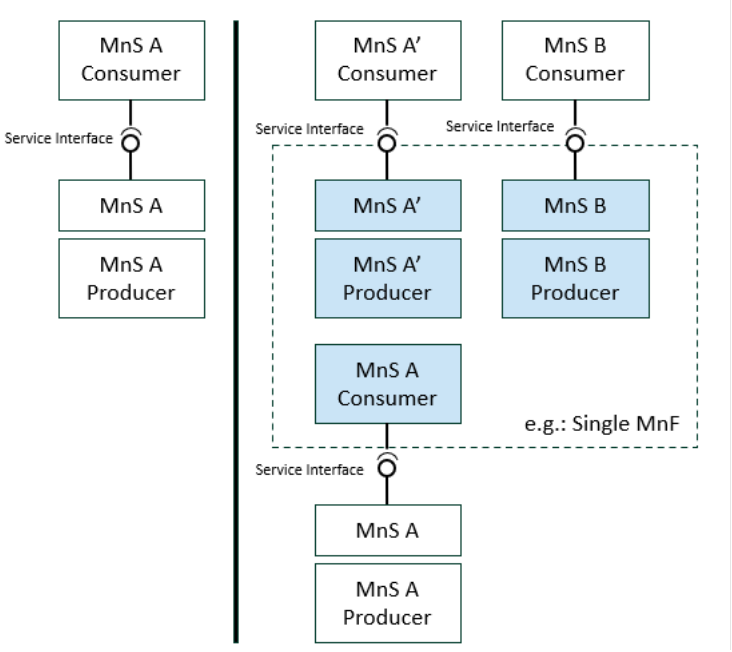

Figure 6: Reference architecture: producer, consumer and exposure concepts

There are two ways in which the management system and the network functions can interact:

The network function directly produces management services (see Figure 7 at right).

The Network Function Management Function module produces management services on behalf of the network function (see Figure 7 at left).

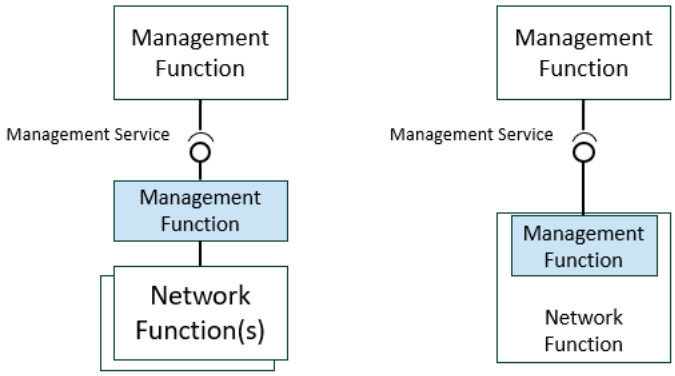

Figure 7: Interaction with the network

SA5 has defined a large number of management services. The most important ones are as follows:

- Configuration Management is a Management Service (MnS) that can be used to create, insert, modify and delete a Network Slice Instance (NSI) or a Network Function (NF);
- Performance Management is a Management Service for the definition of QoE, KPIs and performance measurement and all tasks for configuring, activating, collecting and deactivating performance measurements for NSIs and NFs;
- Fault Management deals with alarm presentation, management and reporting. An alarm is the management representation of a fault, a (detected) error or a failure that requires attention or a response from an operator or a machine.

The Security Management domain comprises all activities to establish, maintain and terminate the security aspects of a system. Examples of the features covered by the Security Management domain are:

- Management of security services;
- Installation of security mechanisms;
- Key management (management part);
- Establishment of identities, keys, access control information, etc.;
- Management of security audit trail and security alarms

In addition, SA5 also defines other management services, taking into account some recently advanced aspects such as:

- Self-organising networks: the ability to dynamically adapt network performance to changing resource requirements;
- Management Data Analytics (MDA) is considered a fundamental capability for the management and orchestration of mobile networks and services. It provides the ability to process and analyse data related to network and service events and status, such as performance metrics, KPIs, QoE reports, alarms, configuration data, network analysis data and service experience data, to provide analytical results, such as statistics or forecasts, root cause analysis questions, and may also include recommendations to enable necessary actions for network and service operations.
- Artificial Intelligence and Machine Learning Management Services to manage the full lifecycle (training, deployment and inference phases) of AI/ML-based algorithms (for optimisation, data analysis and event prediction).
- Intent-Driven Management Service: this MnS aims to provide a powerful interaction service between two or more actors with different roles (e.g. Communication Service Consumer, Communication Service Provider, Network Operator, etc.). In this context, an intent specifies the expectations including requirements, goals and constraints for a specific service or network management workflow; Management of cloud-native virtualised network functions; Energy efficiency management services and others.

3.4 3GPP SA6

The main objective of SA6 is to provide specifications for the application layer architecture of the 3GPP verticals, including architectural requirements, functional architecture, procedures, information flows, interworking with non-3GPP application layer solutions and deployment models [4].

SA6 defines the architecture for Mission Critical Services (MCSs). The architecture easily supports services for both public safety and general commercial applications, including utilities and railways.

5GS provides seamless access to the MCS service environment via the Data Network (DN) as defined in the 3GPP TS 23.501 standard. A Data Network Name (DNN) is an integral part of the 5GS user profile, allowing access to the Data Network with up to 8 connectivity sessions (PDU sessions), each with up to 64 communication flows (QoS flows). It is important to note that different data networks require different DNNs [5]MC services are independent of the type of network, which means that the available service options are identical in both public networks (PLMNs) and non-public networks (NPNs). An NPN can be deployed on premises defined by the organization, and 5G network services are provided to a specific set of users or organizations in accordance with [6].

In the figure below is reported the reference SA6 architecture.

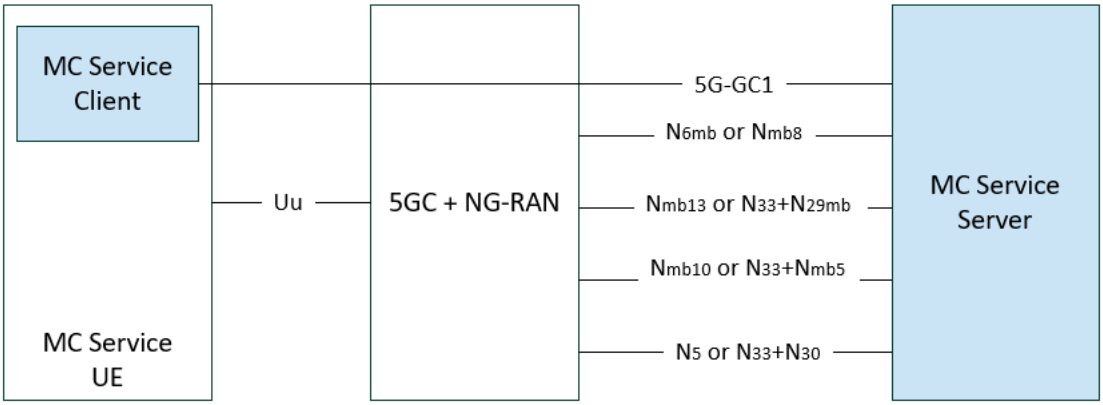

Figure 8: Architectural view of a mission critical system [7]

The next figure provides a view of the system architecture for the MC service UEs that support the delivery of mission-critical services via the MBS.

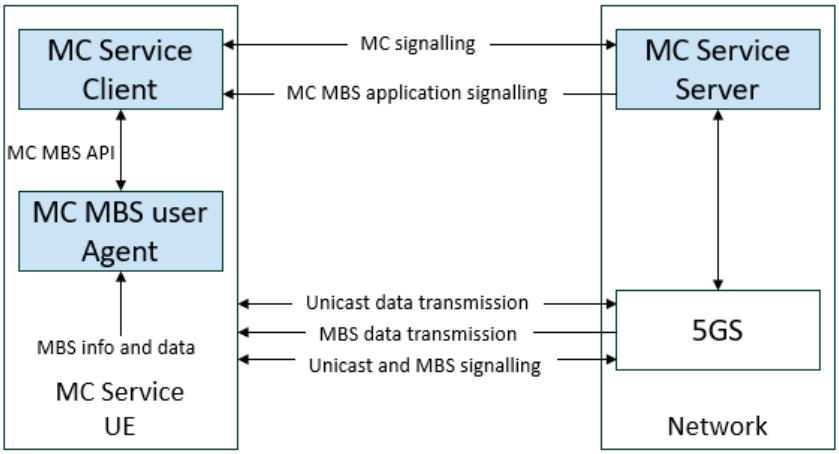

Figure 9: System architecture for MC MBS systems [7]

3.4.1 AI/ML ASPECTS

"SA1 Rel-18 identified requirements (in TS 22.261) for the support of AI/ML model distribution, transfer, training for various applications (e.g., video/speech recognition, robot control, automotive) and initiated AIML_MT_Ph2"

in Rel-19 for supporting Distributed AI training/inference based on direct device connection. Such use cases and requ*irements have application layer impacts*."[SP-231182]

SA6 has initiated a Release 19 study to:

- Identify key issues and develop corresponding architectural requirements at the application enablement layer, as well as potential enhancements to the application layer architecture for AI/ML model distribution, transfer, and training in Rel-18 and Rel-19.

- Examine in detail the architectural and functional implications of existing SA6 application facilities to support AI/ML lifecycle activities, covering all aspects of data collection, data preparation, and training/inference/federated learning for ML models to be used for analysis at the ADAE layer.
- Identify solutions, including information flows and developer-friendly APIs, to meet the architectural requirements and enhancements identified in bullets 1 and 2.
- Assess the potential impacts of application layer support for AI/ML services on different deployments and business models.

4 ORIGAMI BARRIERS

By identifying and addressing barriers, the ORIGAMI project aims to develop innovative architectural solutions that are robust, efficient, and capable of meeting the evolving needs of 6G networks. This comprehensive approach ensures that ORIGAMI not only achieves its internal goals but also contributes significantly to the standardization and advancement of 6G technology.

4.1 BARRIER #1: UNSUSTAINABLE RAN VIRTUALIZATION

Virtualized Radio Access Networks (vRANs) enable baseband processing on commercial off-the-shelf servers. This approach has many advantages over traditional hardwired RANs, such as mitigating vendor lock-in, streamlining upgrades, and enabling resource multiplexing. Led by the O-RAN Alliance [8], practically all the industry is building vRANs, breeding a new market with unprecedented business opportunities in an ossified RAN ecosystem. Analysts project that open vRANs may outgrow the traditional RAN market by 2028, with $29B in revenue.

5G base stations comprise a radio unit (RU), which performs basic radio operations such as signal sampling; a central unit (CU), which processes the highest layers; and a DU, which processes the radio link control layer, MAC layer and performs physical layer (PHY) functions including forward error correction (FEC). New Radio (NR) is 5G's PHY/MAC interface. The most common Frequency Range, which covers sub-6GHz bands, allows up to 100 MHz per carrier and have flexible numerology $\mu = \{0,1,2\}$.

The basic spectrum unit is the resource block (RB), which encompasses 12 subcarriers with $15 \cdot 2^{\mu}$ -KHz spacing. Time is divided into 1-ms subframes, each carrying 2^{μ} slots with, usually, 14 OFDM symbols lasting $66.7 \cdot 2^{-\mu}$ µs. Every Transmission Time Interval (TTI), often one slot, the DU's MAC schedules one TB for/from every active User Equipment (UE), which are signaled to UEs by grants. The TB size depends on the numerology, the amount of buffered data, the DU's RB scheduling policy, and the modulation and coding scheme (MCS), selected based on the signal-to-noise ratio (SNR).

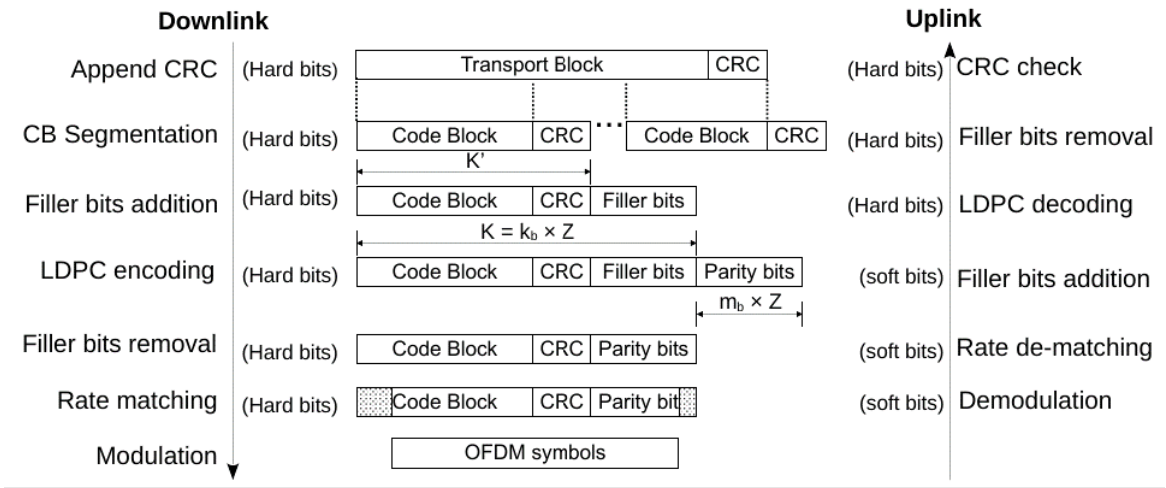

Figure 10: 5G DU data processing pipeline

Figure 10 shows the pipeline of DU operations required to process a TB. At the transmitter side, TBs are divided into code blocks (CBs) with individual CRC fields. Filler bits adapt the CB size to the requirements of the LDPC encoder used for FEC, which produces a codeword with parity bits. Finally, the codeword is aligned to the capacity of the allocated RBs (which depends on their MCS) via rate matching, by applying puncturing or repetition. At the receiver side, a soft-output detector computes the reliability of the data as log-likelihood ratios T (LLR) called soft bits. Then, an LDPC decoder maps soft bits into hard bits through an iterative belief propagation algorithm. The algorithm terminates

after a maximum number of iterations (usually 10), or earlier if CRC validates the codeword. The TB is reconstructed once all of its CBs are successfully decoded.

To adhere to 3GPP and O-RAN requirements [9], processing the heavier LDCP tasks has a deadline D = {1, . . . , 3} ms, depending on the base station, which must be met with 99.999% probability to reach the industry's 5-nines reliability target. These requirements make conventional virtualization approaches, which rely on software running in general-purpose CPUs, insufficient for industry-grade DUs: for instance, Figure 11 (left) shows that a state-of-the-art FEC LDPC decoding library in a CPU can take over 1-3 ms to process a large transport block (TB) compromising the latency budget.

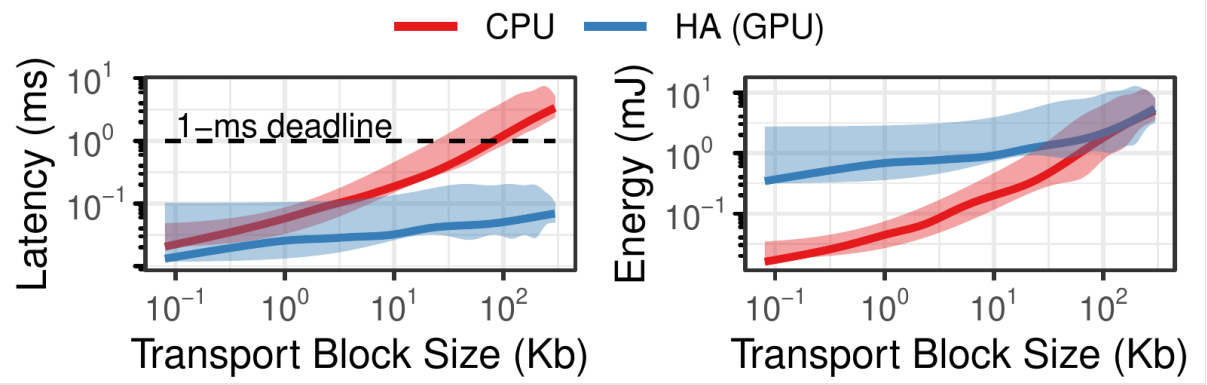

Figure 11: Mean (line) and max-min range (shaded area) latency and energy consumption to decode an LDPC-encoded transport block. Intel FlexRAN LDPC library on an Intel Xeon Gold 6240R CPU core @ 2.40GHz; and commercial driver on an NVIDIA V100 GPU

To address this, vRANs on the market today resort to offloading compute-intensive FEC tasks to dedicated hardware accelerators (HAs) that are co-located with every DU [20, 39]. HAs are ASICs [62], FPGAs [31], or GPUs [63] that, using in-line or look-aside models, can provide >10× latency gains over CPUs when processing large TBs, as shown in Figure 11 (left).

However, HAs are expensive, as exemplified in Table 1, and are energy-hungry, as shown in Figure 11 (right) for a GPU-based HA. More broadly, an Intel ACC100 ASIC and an NVIDIA V100 GPU consume up to 52W and 250W respectively, i.e., 20-82% of the overall consumption of a commodity server. In fact, the economic and energy costs of DU- dedicated HAs are so high that they have cast doubts in the industry about this approach, as implied by e.g., Nokia [11], Ericsson [12] or Mavenir [13].

	CPU core	**GPU**	**FPGA**	**ASIC**
Programmable	Software	Software	Yes	No
Time-to-Market (months) [14]	<2	<2	2	30
Non-Recurrent Engineering Cost ($) [14]	0	0	0	350K-1000K
Unit cost ($) *	110	8000	4000	3000

Table 3: Comparison of processors for 5G LDPC workload

*Intel Xeon 6240R CPU, NVIDIA V100 GPU, Intel PAC N3000 FPGA, and Intel ACC100 ASIC, as observed in Dec. 2023.

In light of the considerations above, ***it is of paramount importance to devise novel RAN virtualization solutions that provide sustainability in the RAN arena***.

4.2 Barrier #2: Poor inter-operability of RAN components

Achieving the envisioned 6G performance KPIs requires the seamless integration of 6G VNFs with the Radio Intelligent Controller (RIC) of O-RAN and with all the automatic functions and algorithms defined by 3GPP for optimizing network performances and resources. This, in turn, requires distributing network elements across edge and cloud networks while enabling real-time control of these RAN nodes (Central Unit, Distributed Unit, Radio Unit) within the RAN architecture. To address this need, one possible solution is to introduce a service-based architecture in the RAN architecture (or part of it). Thanks to this introduction, it is possible to have more flexibility for all network management processes (i.e., for configuration management, performance management and optimisation). It will be also possible to introduce a RAN bus for collecting key performance indicators (KPI) and performance measurements from nodes and delivering control decisions, allowing for controllability of the underlying RAN infrastructure. Furthermore, this RAN bus can manage the interactions among centralized and distributed algorithms. Achieving this goal demands providing multi-timescale control within the computing fabric, taking into account the system's multi-time scale controllability, conflict mitigation across different control loops, rigorous anomaly detection, and agile management of vast amounts of data to handle AI/ML pipelines.

Regarding O-RAN architecture, one of the main challenges in the current O-RAN RIC design is distributing network elements across the edge and cloud network. The RIC has limited insight into connections between subscribers (e.g., xApps, rApps, and E2 nodes). Therefore, to enhance controllability through the RIC and pave the way for the RAN bus, improvements to xApp and RIC architecture must be made, including xApp packages, conflict mitigation mechanisms, time-series handling, message queue support, scaling out methods, and enabling real-time AI technologies (online training, reinforcement learning, and federated learning) at the RIC. Another challenge is conflict mitigation, which requires a novel system design and interaction along with algorithmic solutions for policy-based conflict resolution [1]. Identifying conflicting actions on network configurations is crucial, given the distribution of tasks in an O-RAN architecture through various rApps and xApps from different providers [2]. The Open RAN architecture’s disaggregated nature relies on communication channels to facilitate effective communication between components, increasing the number of communication channels and potential anomaly surfaces. As a result, there may be a need for anomaly detection solutions and tools to automate AI-based anomaly detection systems, enabling detection, analysis, and action against anomalous behavior [3]. The O-RAN architecture incorporates intelligence into the network through AI/ML processes, leveraging vast amounts of data generated by RAN nodes and exposed through the O-RAN interface in the Near-RT RIC and SMO [4]. Despite available interface standardization, data access, pipelines, and validation cannot fully scale due to a lack of standardized network configuration and performance data exchange. During the xApp development and testing in the Near-RT RIC platform, several challenges appeared that are summarized in Table 4.

Aspects	Challenges
Available simulator	Available simulation framework to simulate the RAN does not provide complete functionality that specifically needed to test practical scenarios (e.g., network size, network operation duration, capabilities of the E2 nodes, etc.) The KPM generated in a simulation framework is not completely follows the real behaviors of the underlying RAN that can be further used to generate RAN Control (RC) decision from xApps running from the Near-RT RIC.
Intelligent conflict management	Absence of intelligent conflict management prevent to test the operation of multiple xApps running in the same platform subscribe same KPM and generating conflicting RC decision on underlying RAN On-boarding multiple xApps developed by third party providers working simultaneously in RICs with the utilization of the same E2 nodes lead to conflicts between control actions affecting the performance of the nodes Multiple policies from the xApps providers should be aware of the applied policy on the underlying RAN in case of any perspective conflicts
xApps API	Lack of standard compliance solutions: the E2 nodes or the simulation framework do not provide the parameters for the API that used to deploy and test 3rd party xApps in the Near-RT RIC platform Abstraction of the E2 messages of certain functionalities of the underlying RAN that simplify the xApp deployment and integration process
Interoperability of the components	Interoperability amongst the components, particularly the compatibility of the onboarding xApps with the Near-RT RIC platform The 3rd party xApps should be provided in such a way that it covers the functionalities of the platform as much as possible
Portability	Portability of the xApps that heavily involves manual integration work to deploy it as an xApp in the RIC platform Not having matured enough standardized E2 interface and xApp API to guide clear implementation
Automated testing of xApps	Lack of automation related to deploy and testing the xApps on the RIC platforms Unified way to smoothly introduce new/upgraded xApps to the Near-RT RIC platform, which consumes the resources of both platform providers and the 3rd party xApps providers

Table 4: Neat-RT RIC aspects & Challenges

The successful implementation of the RAN bus will address these challenges and pave the way for an efficient and interoperable RAN ecosystem. This process involves working with a specific RAN bus within the RIC platform. The RAN bus collects KPM samples from nodes and delivers RAN control (RC) decisions to enable controllability of the underlying RAN infrastructure. The RIC platform hosts several xApps that utilize the RAN bus to:

Collect performance measurements from RAN nodes, a functionality provided by the RAN bus service models, i.e., the E2 service model key performance measurements (E2SM-KPM) defined by the O-RAN Alliance.

Receive and modify Information Elements (IEs) in various types of signaling messages without requiring xApps to decode the entire network interface signaling message. This functionality is provided by the RAN node using E2SM RAN Control (E2SM-RC). The xApps in the RIC platform may simultaneously perform control tasks by modifying network interface signaling messages using E2SM-RC. Typically, xApps change one or more IEs to optimize a specific metric of RAN nodes.

xApps should perform tasks that do not conflict with each other. The proposed KPI collection and slice admission and congestion control xApps developed in ORIGAMI are designed to avoid conflicts. However, third-party xApps hosted in the RIC platform may configure IEs that result in conflicts within RAN nodes. Within each xApp, the following APIs can enable interoperability amongst 3rd party xApp with underlying RAN nodes.

The Shared Data Layer API allows to collect the information related to E2 nodes including their types, supported component interfaces and SMs.

The E2-related API supports the E2 Subscription, E2 Indication, E2 Control and E2 Guidance.

The A1-related API supports the A1 policy and A1 Enrichment Information.

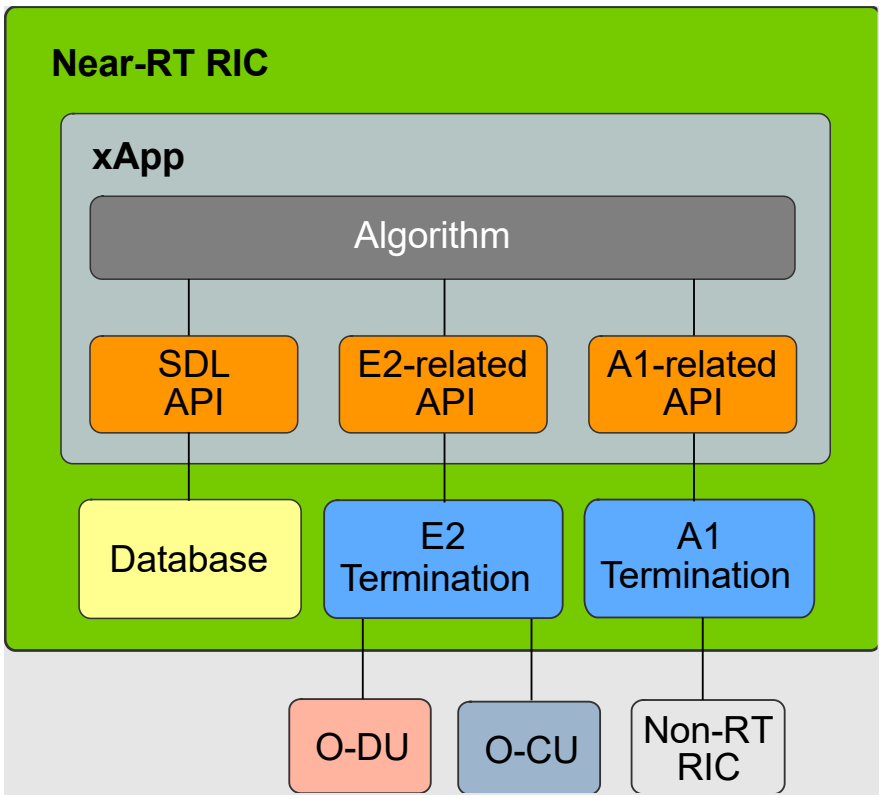

Figure 12: The Near-RT xApp API [15], [16], [17] and [18]

4.3 Barrier #3: High latency and unreliable Network Intelligence (NI) to process complex 6G network problems

The traditional approach used to develop Network Intelligence (NI) solutions frequently fails to leverage the unique capabilities and features of the underlying computing infrastructure. Instead, these solutions are often implemented directly as they are described in AI and ML literature, without considering the specific requirements and opportunities presented by the network environment. For instance, the limited execution times or the limit in the data gathering interfaces. This method can result in suboptimal network utilization, characterized by excessive data exchanges that not only escalate resource consumption but also potentially degrade system performance.

In contrast, a more tailored approach, which aligns the design and deployment of NI solutions with the specificities of the network infrastructure, can significantly enhance efficiency. By optimizing the interaction between the NI algorithms, the targeted network services, and the underlying hardware, it is possible to reduce unnecessary data transmissions, hence conserving bandwidth and computing

resources. Furthermore, focusing on learning from more compact, relevant datasets (rather than larger and bulky ones) can improve the speed and accuracy of the learning process.

Enhancing NI solutions in this way also opens the door to more advanced strategies, exploiting also edge computing facilities, where data processing occurs closer to the source of data generation. This can lead to faster insights and actions, as the latency associated with data transmission to centralized servers is minimized. Additionally, by leveraging the full potential of the infrastructure, it is possible to implement more complex and adaptive learning algorithms that can dynamically adjust to changing network conditions and demands, ensuring optimal performance at all times.

Ultimately, by moving away from a one-size-fits-all application of AI and ML models and toward a more integrated, infrastructure-aware approach, unlocking new levels of efficiency, responsiveness, and intelligence in network operations are possible. This will not only improve the performance and sustainability of the network but also pave the way for more innovative and effective NI applications such as the ones listed in Enhancing Management and Stability in the 6G Architecture 6.3

The overall characteristics of the solutions shall target the following aspects.

Objective	Description
Developing a comprehensive Network Intelligence (NI) toolbox for 6G network operations	Enhance reliability, explainability, unbiasedness, and long-term robustness.
Designing scalable first-order learning algorithms	Do not require training data and offer performance guarantees in various network conditions, including non-stationary and adversarial environments.
Addressing large-scale network problems (scheduling, routing, placement)	Solve without compromising optimality or deviating from the problem's timescale requirements.
Enabling real-time decision-making adaptation to network conditions, user demands, and system events	Allow for adaptive responses to changes in network conditions, user demands, and system events without dependency on training data.
Replacing outdated analytical tools used up to 5G with new scheduling and optimization tools	Suit new tools for 6G's dynamic conditions, moving beyond the limitations of previous generation tools.
Developing hybrid learning algorithms	Combine deep learning and online learning to ensure optimal performance under both stationary and non-stationary conditions.
Utilizing deep learning for real-time solutions to complex NP-hard network problems	Enhance efficiency and reduce solution times, closer to real-time, for large-scale complex network problems.
Introducing Bayesian learning algorithms using Gaussian Processes	Leverage Gaussian Processes for data-efficient runtime learning of performance and cost functions, aiding in system control.
Addressing the computational demands of Bayesian learning	Explore solutions like Restarting Gaussian Process (GPs), using shorter memories, and mixing models to integrate Bayesian learning into 6G platforms.
Integrating new computing paradigm natively into the architecture	Integrate next generation computing paradigm such as Quantum.

Table 5: AI objectives in 6G

4.4 Barrier #4: Under-utilized modern programmable transport

In the transport and core domains, user planes implemented with programmable switches, smartNICs and Network Processing Units (NPUs) offer novel compute resources co-located with the hardware that physically transfers bytes. These new resources provide opportunities to realize VNFs that can

operate, for the first time ever, on the network traffic at packet level and at line rate. Investigations about how programmable user planes can support VNFs for network monitoring and management operations are still in their infancy and quite limited in terms of what they can achieve and with which performance.

The models proposed to date for user-plane inference that can operate on industry-grade equipment have major limitations in terms of supported features and maximum achievable complexity. Solutions for inference with programmable switches, e.g., powered by Intel Tofino application-specific integrated circuit (ASIC), are limited to Decision Tree (DT) and Random Forest (RF) models with a small number of trees, e.g., below 5, and reduced depth, e.g., within 10 levels. Attempts at integrating neural network models onto programmable switches have failed, mainly due to the inherent limitations of the devices, and this even when considering software-only emulated equipment, which is known to be much less restrictive than real hardware. Moreover, many state-of-the-art solutions for in-switch inference only operate on a per-packet basis, i.e., cannot employ flow-level features to perform the target computational task, do not support hierarchical or distributed approaches, or are not viable in practice as they consume a too large fraction of memory resources that risks hampering the regular operation of the equipment for forwarding tasks.

SmartNICs and NPUs provide a slightly more flexible environment than switches, and, for instance, have been proved to be able to accommodate basic binarized neural network models, whose size is still limited to 3 layers with a few tens of neurons for layer. As a result, all the models for user-plane inference proposed to date have been only tested with straightforward classification tasks with typically less than 10 classes that can be effectively told apart using a few simple features.

Also, demonstrations of the feasibility, scalability and performance of in-band computation with demanding use cases that are closer to problems encountered in real-world networks remains elusive. The current state of in-band inference is a clearly missed opportunity. Indeed, exploiting computing models that are fully deployed into the programmable user plane has the potential to lead to 100x gains in latency and growth by one order of magnitude in throughput when compared to legacy control-plane decision models. Such dramatic advantages cannot be attained with other approaches, including hybrid strategies that split the inference process across user and control planes, making substantially more complex models possible but also losing the key line-rate capabilities of a pure in-band approach. Similarly, considering external components that must be attached to the network equipment and are dedicated to the inference task is hardly viable in production systems, since they are prohibitive in terms of capital expenditures and hardware complexity.

Finally, coding and deploying models for in-band computing is a largely manual process that requires in-depth expert knowledge of the target hardware equipment and a significant amount of time to tailor the solution to the specific task at hand. The lack of easy-to-use paradigms and interfaces that make the realization of these solutions almost a transparent process for the end user create a major obstacle to their adoption in production systems. Indirectly, this inhibits innovation in terms of original 6G VNFs and 6G services that could exploit the unique line-rate operation capabilities of complex user-plane traffic processing. In particular, there is a vast range of applications that could be enabled by a seamless implementation of intelligence directly into the user plane, controlled in a streamlined way by the MNO and possibly by the service providers as well.

The limitations above also curb the exploitation of user-plane computing capabilities to support new and compelling 6G applications, such as remote sensing. Programming the infrastructure to feed data that is relevant to applications in and beyond networking, such as in sociology, demography, or urbanism can unlock a completely new market for network infrastructure operators, drawing a flurry of additional classes of tenants into the ecosystem. Making in-network processing relate to the applications –and not only to the management of the network as in 5G systems– raises additional substantial challenges, including: (i) programming of probes that collect at ultra-low latency the necessary metadata to support the new classes of applications from traffic flowing at hundreds of Gbps; (ii) performing in-band data processing that provides by-design privacy guarantees already

when the collected information leaves the user plane; (iii) devising solutions of large (and possibly NP-hard) scheduling problems that can operate in real-time; (iv) designing decision-making policies under uncertainty for the performance of the employed ML libraries that new inference and processing tasks will use.

All the aspects above jointly prevent a streamlined development of original 6G VNFs into the user plane, hindering access to the compute capabilities offered by modern programmable network equipment and blocking the realization of innovative services that can build on top of such user-plane VNFs.

4.5 Barrier #5: Lack of global service APIs

In the past, different branches of standardization efforts were responsible for developing procedures related to network management, network orchestration, and network control. As a result, the functions responsible for carrying out these procedures, such as OSS functions, element managers, orchestrators, or radio and core NFs, were designed in a way that was specific to their respective domains. In some cases, these designs were even proprietary, and any optimizations made were limited to a "per domain" approach. This limited the interaction between these functions to peer-to-peer reference points within a domain, such as the N4 reference interface between SMF and UPF functions as defined in 3GPP 5G Core [19].

The adoption rate of Network Slicing technology within the telecommunications market has exhibited a gradual progression. Despite its inherent versatility and potential applicability across various sectors, the predominant utilization of network slices pertains to business-to-business (B2B) contexts. Specifically, multinational corporations and mobile networks operating within limited geographic regions stand to derive substantial benefits from the implementation of network slicing functionalities. Notably, the absence of standardized frameworks for external Application Programming Interfaces (APIs) within the industry poses a notable impediment to the widespread adoption and seamless integration of network slicing technologies across diverse organizational structures.

In setups that rely on reference points, optimizations can be open-loop, meaning there is no feedback among different modules in the system, or they may require costly human engineering procedures to close the loop. This is the case for example in video encoding where encoding is adapting network environment. While this approach has been considered valid in legacy networks due to their limited number of configurations such as 4G and the first releases of 5G, it is inadequate for what is expected for the 6G ecosystem. Legacy NFs typically have function-specific data acquisition and processing procedures or no procedures at all. Therefore, simple configurations or rulesets are usually enough to achieve optimization goals.

In the 5G era, network slicing needed a more modular design for NFs, enabling them to be shared and reused across slices in a more precise and targeted manner. For instance, a single Network Slice Subnet Instance (NSSI) and its constituent NFs may be used across multiple slices and services, such as common radio access NFs across slices [20]. As a result, there is a need to design interfaces dedicated to data acquisition and processing from NFs or for feeding and pushing data to specific AI modules. To enable automated loop closure by incorporating AI and big data solutions, new interfaces and functionalities are required. First, to allow effective communication between different network domains (such as RAN, Core, Management, and the Service Provider), a publish-subscribe methodology should be utilized to enable flexible data exchange. In addition to producing and consuming data, Network Functions (NFs) in all domains should provide ways to authorize and configure relevant parameters, with different levels of access based on the enforced resource provisioning scheme. This includes offering reliable and scalable configurations to authorized service providers, who may have different levels of configuration capabilities depending on their specific requirements. For instance, some providers may require full configuration capabilities, while others

may only need limited visibility of configurable parameters. Rising demands on existing cellular infrastructure are driving network densification, with nascent 5G networks requiring larger numbers of smaller cell sites to deliver their promised network capacity. Deploying dedicated radio infrastructure for each provider is capital-intensive and inefficient. In ORIGAMI, we recognize the need to facilitate low-friction infrastructure sharing, allowing any number of brokers to take advantage of telco infrastructure deployment.

While the pub/sub scheme has been recently introduced in the 3GPP standard, capability sharing across different domains is still far from being achieved. At present, state-of-the-art network architectures are often defined in a silo-based manner. While they may include some analytics features within a specific domain (such as the notable example of the NWDAF in the core), they lack the ability to openly exchange information among different domains.

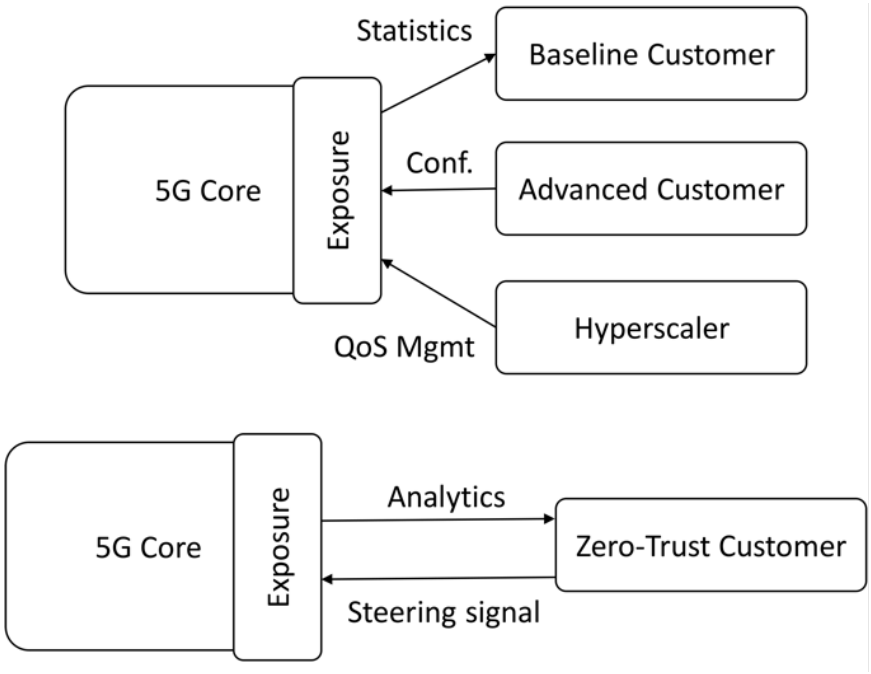

Figure 13: The current view as studied by 3GPP (top), the additional view as proposed in ORIGAMI

The advent of network softwarization in 5G and beyond 5G networks has enabled a diverse landscape of tenants, including industrial verticals, to play a more integral role in network operations compared to the legacy 4G networks with full over the top (OTT) service delivery models. With new configuration primitives, tenants and service providers can act on underlying network slices, leading to increased attention in standardization efforts, such as the NSaaS model, which allows tenants to manage network slices as managers via an exposed management interface. An example of this is the one brought by the 5G-ACIA industrial initiative, which also emphasizes the need for a 5G exposure reference point toward enterprise tenants to enable better integration between service providers and network operators. This framework proposes the exposure of selected functionalities, such as the 5GC control plane, 5GC capability, and 5GS management capability. However, 5G-ACIA does not mandate any specific solution for the exposure reference point, but rather emphasizes usability, simplicity, modularity, and extensibility. To support different business models, more flexible interfacing beyond the traditional operator model and NSaaS model may be beneficial. In addition to industrial consortia, research projects such as network applications are also promoting open APIs for tighter interaction between service providers and network operators, particularly for edge computing and vehicular use cases that require low latency. Efforts to integrate architectures such as O-RAN with MEC are still ongoing, but the lack of exposure functionality toward other domains such as the Core remains an important barrier.

4.6 BARRIER #6: OBSOLETE TRUST MODEL HINDERS PERFORMANCE

Due to the absence of a global horizontal exposure capability across telecommunications domains, the provision of international mobility for end-user devices via roaming remains a crucial aspect of the mobile industry's value proposition to global service aggregators (such as Google Fi, Twilio, or Truphone) and digital nomads. This is a critical success factor in the mobile industry's efforts to maximize its share of this revenue opportunity. Nevertheless, 4G and 5G architectures continue to regard global mobility as an exceptional and temporary event rather than a default mode of operation, which is the case for some IoT vertical applications that globally connect their devices by leveraging the roaming function.

In ORIGAMI, it is highlighted the necessity for globally connected devices to cross provider boundaries more frequently than the 5G architecture design currently anticipates. Presently, global infrastructure sharing via roaming is only achievable through pre-established contractual agreements, fostering trust between two entities that cooperate in authenticating and billing users (e.g., through inter-provider roaming protocols). Despite this, IoT manufacturers now widely adopt pre-provisioned global connectivity. This approach means that IoT devices come with built-in connectivity, allowing customers to avoid managing the connectivity of each device and directly benefit from smart services. For instance, when the power grid in Spain purchases smart meters, these meters come with pre-provisioned connectivity from the vendor through a managed connectivity provider. The power grid provider can then deploy them without running a mobile network or negotiating individual contracts for each smart meter. Instead, they rely on a single connectivity provider and use its international mobility agreements to deploy devices worldwide. This method is attractive for many global digital service providers as it (i) provides more stable connectivity/coverage, (ii) eliminates the cost of establishing technical and commercial relationships with operators in the countries where IoT verticals deploy their devices, and (iii) simplifies management since all SIMs have a single base MNO and home country.

However, the extensive cellular infrastructure that international carriers and providers have built over decades is not well-suited to the global operating model required by many global service providers. Although permanent roaming is not a new concept, its implications in the context of widespread global applications have become a critical issue for mobile operators to address within their native architecture. For example, permanent international mobility of IoT devices disrupts the fundamental business and technical assumptions related to international mobility of end-users, such as charging models, transparency, and fraud detection. Furthermore, permanent international mobility may expose operators to regulatory, tax, and corporate compliance challenges in some jurisdictions. The home operator plays a crucial role in charging end-users and paying roaming partners in the visited country, necessitating direct visibility of end-user activity due to a lack of trust between partnering Mobile Network Operators (MNOs). Consequently, the communication of IoT devices is routed back to the home provider's Core Network, resulting in significant performance penalties due to data paths traversing costly and congested international transit links.

Routing end-user traffic through the home operator enables zero-trust billing in the current cellular ecosystem but imposes severe performance penalties on end-user communication. **The ORIGAMI decentralized identity aims to allow end-users to interact directly with various entities within the cellular ecosystem, bypassing their service provider.** Local breakout enables end-users to access resources within the country they operate in, significantly altering the cellular ecosystem architecture. With local breakout, the end-to-end paths between the end-user device and the application server only traverse a small part of the cellular ecosystem and instead utilize Internet paths. The benefits of peering are well-known among ISPs and content delivery networks, especially with public peering via an IPX. Indeed, in the 6G era, the cellular ecosystem should evolve to adopt such practices to reduce latency, increase redundancy and capacity, and lower costs.

In this context, the home operator remains key in authenticating the end-user's identity and handling billing, including making cascade payments to partners along the data path. This setup requires that end-user traffic be visible and accountable to both the home operator and the operator in the country visited. By routing end-user traffic through the home network, both the visited and home networks gain visibility of user traffic, allowing consistent billing. This arrangement eliminates the need for trust between the visited and home networks, as both can account for the end-user traffic. The current setup relies on third-party services (Data Clearing Houses) to reconcile mismatches between the records of international partners. This approach has largely remained unchanged despite the significant evolution in cellular technology. Therefore, **MNOs must develop sustainable strategies for next-generation interconnections to remain relevant.**

4.7 Barrier #7: Inadequate networking data representation

Today's cellular architecture still operates on privacy and security assumptions that are outdated. In previous decades, mobile providers were heavily regulated and centralized, few users traveled internationally with their devices, and data broker ecosystems were undeveloped. Consequently, a closed platform for interconnecting mobile operators, such as an Internet Exchange Point (IXP) Network, seemed appropriate for offering international mobility to the limited number of users, promising guaranteed performance and privacy. However, the current ecosystem diverges significantly from these original goals, revealing vulnerabilities in roaming signaling protocols, such as the ability to read text messages, determine users' locations, and facilitate various types of fraud. Additionally, vulnerabilities in SIM cards, particularly in two major IoT device manufacturers, allow for the duplication of SIM cards, including the IMSI, authentication key, and payment information. These issues underscore the need for proactive monitoring to address anomalies and malicious activities. However, efforts are hampered by a scarcity of ground truth datasets and the lack of explainability in machine learning approaches for anomaly detection.

While AI techniques have proven effective in solving real networking problems and integrating into real products, there is still much progress needed before AI can be fully implemented across the entire cellular ecosystem. ORIGAMI emphasizes the importance of focusing on data representation and data governance, building a pipeline from global infrastructure providers to tenants that prioritizes "data" provisioning. This is crucial for enabling network intelligence (NI) functionalities that support inter-domain converged operations. High-quality data is essential for network intelligence, yet data has not been adequately prioritized in efforts to develop intelligent networking solutions.

The vision proposed by ORIGAMI aims to enable new methods for tenants to interact with network infrastructure and create opportunities for ultra-fine granular management of infrastructure, such as individual transmission blocks or virtualized tasks in the RAN, and individual packets in the transport plane. This offers unprecedented possibilities for service monitoring and performance improvement but also presents significant challenges in network automation. The volume of decisions and the speed at which they must be made (e.g., allocating resources to tenant requests or micro-managing network tasks) push current requirements to the extreme, making human involvement increasingly impractical. In this context, full network automation becomes essential. However, existing decision-making solutions based on statistical modeling, optimization, ML, or AI are tailored to current mobile network architectures and standards, whereas the ORIGAMI vision demands a higher level of performance. Therefore, designing NI instances that fit the global ORIGAMI ecosystem is a significant challenge that requires highly efficient models.

To address these barriers comprehensively, ORIGAMI should also focus on developing NI instances suitable for automating operations through novel architecture components.

4.8 BARRIER #8: HIGH VOLUME OF CONTROL PLANE SIGNALING

The Service-Based Architecture (SBA), a notable feature carried over from 5G to 6G, is designed to transform monolithic applications into a suite of independent services that can be developed, tested, deployed, and managed separately. This architectural approach is particularly beneficial in 6G, where it facilitates the definition of detailed network services and the integration of SBA into the user plane or Radio Access Network (RAN). The 5G SBA introduced flexible and unified interactions among Network Functions (NFs) of the network core through HTTP/2 calls, setting the stage for enhanced network functionality in 6G.

Network Functions (NFs) are key components within a network's infrastructure, defined by their external interfaces and behaviors. In 6G, NFs are essential for further atomization and distribution, enabling both centralized and distributed nodes to collaborate seamlessly. These functions are dynamically combined with services based on user needs, routing protocols, and security mechanisms. They are organized into separate entities for the control plane and data plane, capable of interconnecting through a unified API interface. The Network Repository Function (NRF) allows each NF to discover services offered by other NFs, enabling independent updates with minimal disruption, which supports network slicing and the overall flexibility of the network. This approach, along with the complete softwarization of functionality (provided as virtual or containerized network functions), has enabled network communication service providers (CSPs) to realize [24].

Despite these advancements, the 5G SBA faces significant challenges in signaling and control planes. Each NF must communicate directly with others, causing a rapid increase in signaling traffic, aggravated by the growing number of connected devices and sessions in the post-5G era. This surge in signaling traffic leads to higher energy and cost consumption and limits the scalability of the [25] . Additionally, the monolithic design of each NF binds various operations together, creating single points of failure and reducing resilience. Thus, optimizing signaling management is crucial for the progression of 6G networks, especially as mobile networks are increasingly integrated into IT infrastructures within private sectors, where external applications interface with mobile networks through APIs, further increasing the signaling load.

5 ORIGAMI ARCHITECTURAL INNOVATIONS

5.1 LAYERS AND LOOPS

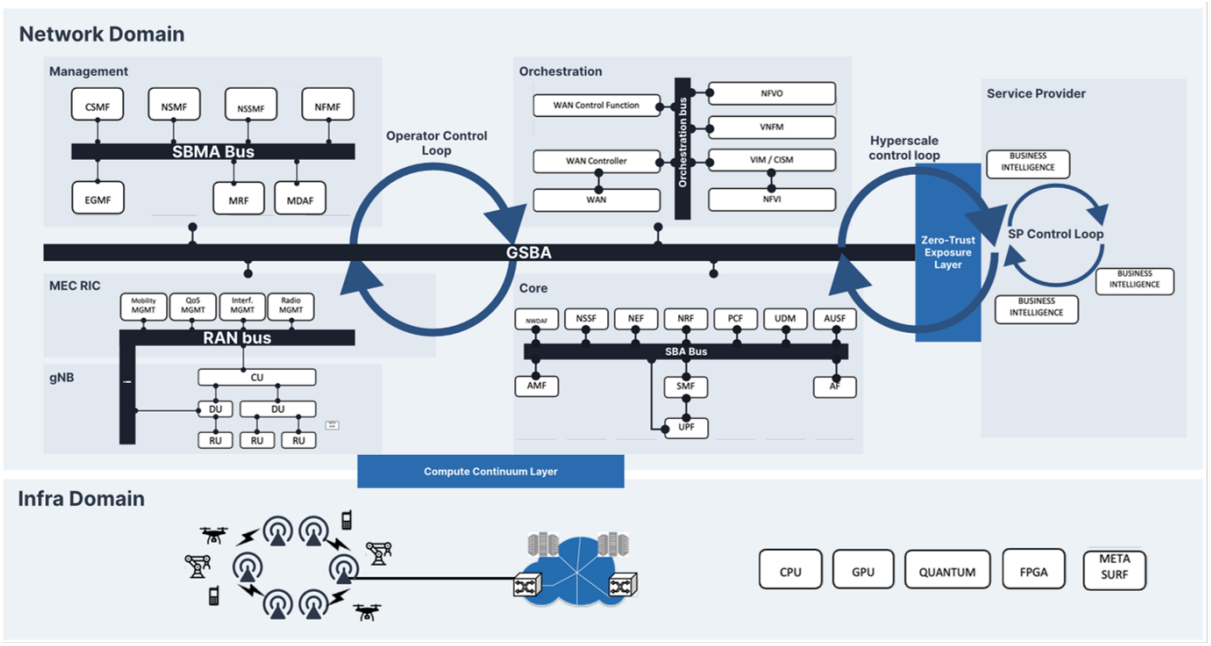

Figure 14: The ORIGAMI architecture innovations that enable next-generation global services and Network Intelligence (NI) functionalities

To overcome the barriers described in Section 4, ORIGAMI does propose a methodology based on three architectural innovations that enable three global control loops that will support different Network Intelligence (NI) functionalities and global services.

The three envisioned architectural enablers will evolve the network architecture beyond the current 5G SBA approach, effectively tearing down these barriers. This re-design, while maintaining the essence of the 5G architectural work, transitions into 6G by integrating three architectural evolutions, as depicted in Figure 14.

The **Operator Control Loop** expands upon existing control loops available per domain, such as those in the RAN domain between the RIC and other elements of the O-RAN architecture, or the loop created by the NWDAF in the core domain. This expansion aims to achieve a holistic view of network resources. This loop, enabled by the GSBA (see Section 5.4 will effectively integrate the currently scattered and siloed domains into a unified fabric, where all intelligent algorithms can coexist and cooperate for the optimal operation of the network, adhering to Network Operator policies. One example of interactions not currently feasible with the standard 5G architecture is the efficient global orchestration of RAN resources involving the RIC. This limitation exists because tight coordination between the RAN and other domains is necessary. For instance, the configuration of Reconfigurable Intelligent Surfaces (RIS) or smart reflectors in the vicinity need to be agreed upon with neighboring cells. Another example concerns the deployment of virtualized small cells in a shared edge data center, which must be coordinated with the deployment of edge services within the same infrastructure. [21]

In addition to enabling new network intelligence algorithms within the network domain, the enhanced GSBA will serve as a gateway for more flexible interactions between Service Providers and Network Operators, currently hindered by various factors, as explained in Section 4.5 enabling the **Hyperscale Control Loop**, offering new opportunities for both Service Providers and Network Operators. Service Providers can enjoy a broader range of exposed network capabilities, leading to seamless customization, while Network Operators can diversify their service portfolio based not only on service types (e.g., eMBB or URLLC) and associated KPIs, but also on the available customization capabilities.

By employing tiered pricing models used by Software-as-a-Service (SaaS) companies, Network Operators can monetize various levels of functionality and cater to a wider range of customers.

5.2 Zero Trust Exposure Layer (ZTL)

The ZTL enables the matching between the internal operation of the service provider's business logic and the network operator's continuous optimization. This cooperative control loop can unleash all the potential of the network deployment by i) having a more direct influence on the operation internals and ii) unlocking beyond data-transfer functionality such as remote sensing or digital twinning directly to the providers. Thanks to this view, ORIGAMI aims to provide verticals with network hyperscalers in a similar way computing hyperscalers (such as Amazon AWS, Microsoft Azure, and Google Cloud Services) are already providing for other purposes. Through this layer, ORIGAMI will tackle barriers #5, #6 and #7.

The ZTL developed by ORIGAMI will enable both vertical and horizontal exposure, as detailed next.

Vertical Exposure – Network Application Analytics Fusion. The data analytics framework proposed by 3GPP in TS 23.288 introduced a giant leap forward on how analytics are produced and managed within the network. The requirement for network automation has influenced the design of 3GPP system standardized in R15. Prior to this release, the generation of data and analytics involved communication between network elements and their managers through proprietary interfaces. However, with the subsequent consolidation in R16 and R17, the system architecture has been re-designed to natively support the collection of analytics towards automation loops. The Network Data Analytics Function (NWDAF) is the cornerstone of this system, which gathers data (i.e., metrics related to the current status of the network) from other network functions, computes analytics (i.e., refined statistics based on the gathered data), and shares them with other consumer functions in the network. A fundamental part of this exchange is the one with the management system, where the Management Data Analytics Function (MDAF) acts as a bridge towards other selected functions such as the radio related one.

In this framework, 3GPP defined several analytics that are usually backed by relevant use cases involving interactions with other network functions in the 5G Core Domain and with the Management system: for R18, 14 categories are listed, ranging from UE related analytics to the NF performance.

While the analytics system provides a framework to effectively perform self-optimization tasks in the 5G Core and Management domains, including federated learning, all these metrics are prominently related to network management. In a scenario, like the one described in Section 4.6 (Barrier #6), where the service provider is capable of directly customizing the network behavior in a zero-trust fashion, also the NWDAF-provided analytics shall reflect this view.

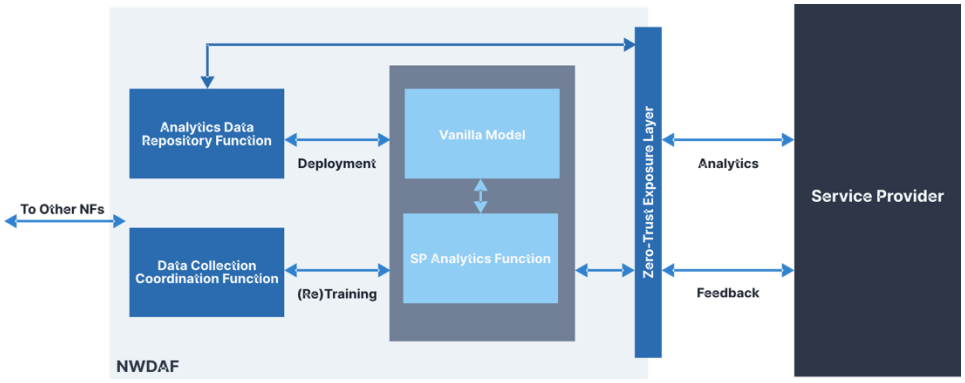

Figure 15: The enrichment of the NWDAF with Vertical Service Provider feedback

This problem is related with the traditional problem of the QoS to QoE mapping, which has been studied a lot in the past, also in the context of 5G systems. In general, the analytics framework proposed by 3GPP deals with QoS metrics that are related to the different Network Functions in the core. Still, Service Providers need to optimize their own metrics that are indeed related with the QoS perceived by users but may substantially differ because of e.g., business related aspects that are clearly unknown to the network operator.

Thus, ORIGAMI plans to enable ZTL with specific APIs that can be leveraged to provide customization of the analytics provided by the NWDAF, to also take into account feedback coming to the overarching service provider, as exemplified below.

Thanks to the Zero-Trust features enabled by the ZTL, as depicted in Figure 15, the provided analytics can be improved and matched to the specific QoE provider metrics without sharing the QoE metrics between these two parties, as they could contain either private or confidential features that shall not be directly exchanged among the parties. This requires the design of i) algorithms that can provide analytics in a parametrizable way and ii) the mechanisms that leverage the ZTL to expose the parametrization of the analytics. For the former we envision the usage of adaptable AI algorithms such as the one proposed in while for the latter, the usage of Reinforcement Learning schemes is appropriate for the task [22][23].

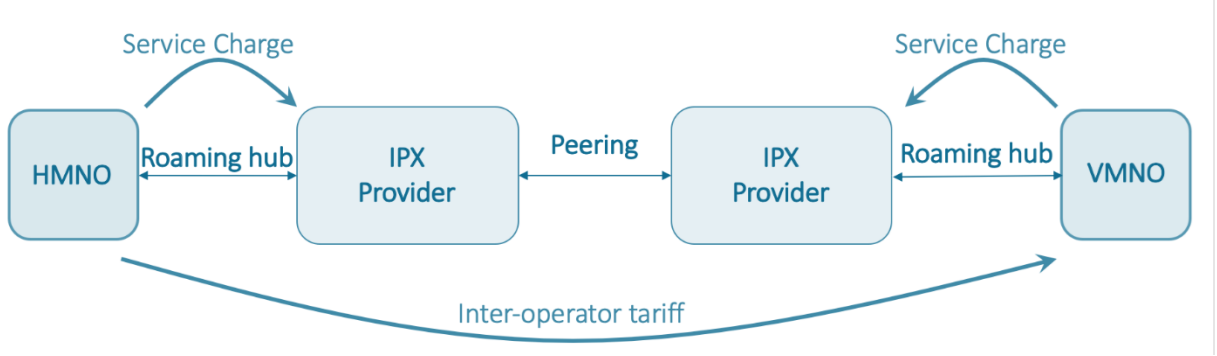

Figure 16: Business agreements to enable international roaming

Horizontal Exposure – ZTL is powering the Global MNO Model. Support for global operations is crucial for IoT verticals like connected cars, logistics, and wearables, driving the commercial success of IoT platforms. IoT device manufacturers need global connectivity solutions, an increasingly urgent requirement as IoT deployments accelerate. IoT verticals depend on providers ensuring reliable global data connectivity, such as M2M platforms relying on the cellular ecosystem. Therefore, international mobile roaming is essential for IoT verticals.

Roaming needs vary by use case (e.g., automotive, logistics tracking, smart meters). Logistics services prioritize international roaming for asset tracking, while payment services need reliable signal connections, selecting alternative networks if the primary fails.

Global operation of mobile devices involves the cellular ecosystem's complex infrastructure, connecting Mobile Network Operators (MNOs), business partnerships, and third-party Data Clearing Houses (DCHs) for billing. These rely on outdated practices, long financial clearing periods, complex billing models, and mechanisms for inter-MNO dispute resolution. To support roaming customers, Home Mobile Network Operators (HMNOs) and Visited Mobile Network Operators (VMNOs) must have commercial agreements. With technical solutions in place, commercial roaming allows MNO customers to use partner networks, generating roaming revenue based on data/voice/SMS usage by inbound roamers on the visited network. Roaming partners account for roaming activity and exchange records to claim revenue. The complex relationship between MNOs and other cellular ecosystem players, like IPX Providers, adds to the challenges. Figure 16 shows a schematic view of the business

interactions between entities in the cellular ecosystem to ensure that devices can currently enjoy global service.

Although radio technologies have evolved rapidly over recent decades, the logic of interconnection has largely stayed the same. Additionally, the platforms and systems involved are opaque, with minimal innovation in this area. Advancing the inter-provider charging system is particularly challenging because it demands standardization, followed by joint evolution and deployment across involved networks. This issue is becoming significant, as the current inter-provider charging system for global services results in penalties related to performance, operational costs, and business revenues. Meanwhile, the vision for next-generation cellular systems (6G and beyond) sets a very high standard for cellular networks. These networks are expected to deliver smart and global connectivity to a massive number of heterogeneous terminals operating in various environments worldwide. Achieving this ambitious goal requires increasing the already considerable complexity of the cellular ecosystem to instantly orchestrate physical resources and functions across different network domains, in alignment with time-varying user demand and multi-tenancy requirements, while providing a global service.

With ORIGAMI, ***the aim is to build an architecture that relies on a completely different trust model, which is a significant architectural change***. To achieve this, ORIGAMI proposes a new architecture where the identity of the end-user (IoT device) is no longer strictly bound to the home operator. The ORIGAMI architecture relies on the idea of decentralized identity and decouples the end-user authentication from the connectivity function those operators offer. With this, the ORIGAMI architecture aims to enable novel business models, and a dynamic approach for charging. For example, even with massive number of IoT devices operating under managed IoT connectivity model, retail charging is still being used by visited operators to charge their partner home operator. With ORIGAMI, the visited operator will be able to directly charge the (foreign) global end-user and give the home operator full visibility into these transactions.

The goal is to eventually tackle the barriers that are identified and enable a new global operating model (see Section 6.7), where roaming devices enjoy “first-class citizen” treatment. Specifically, ORIGAMI will integrate:

- Distributed ledger solution for archiving immutable records on the activity in the global federation, allowing for new interconnection models and business interactions.
- Network intelligence (NI) modules for anomaly detection (see Section 6.9).
- Mechanisms to ensure security, privacy, and confidentiality within the global federation.
- Novel models of global operations, relying on remote provisioning of eSIM-enabled devices.

5.3 Continuous Compute Layer (CCL)

ORIGAMI's Compute Continuum Layer (CCL), illustrated in Figure 17, is an innovative architectural component for 6G systems, which will enable network processing workloads to be executed in a heterogeneous computing fabric exploiting diversity opportunities. This architectural innovation holds the potential to streamline resource sharing, unlocking the full capabilities of diverse computing environments within a 6G system.

The CCL shall support a wide spectrum of computing resources – GPUs, TPUs, FPGAs, ASICs, NPUs, smartNICs, and even quantum computers to accelerate Virtual Network Functions (VNFs) and complement the work of traditional CPUs.

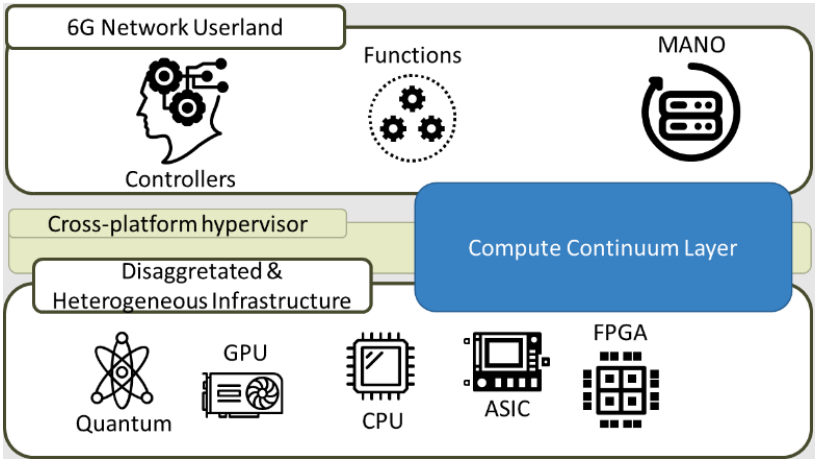

Figure 17: The ORIGAMI CCL

ORIGAMI's CCL shall pave the way for compute-aware network operations while maintaining the abstractions of a pure virtualization layer. This involves rethinking Network Functions (NFs) at both levels, enhancing efficiency and reducing resource usage by matching NFs to the CCL. Consequently, the CCL manages available resources in real-time, allowing the highest re-utilization factor in the edge-to-cloud continuum, but also imposes policies to NFs.

The CCL will allow a centralized and coordinated abstraction between NFs and specific hardware components within the underlying hardware substrate. This approach simplifies resource management and allocation, promoting efficient utilization and scalability. Furthermore, the CCL accommodates new sensing functionalities within the network, catering to tenants interested in metadata rather than data transport.

The computational resources assigned to specific network functions might impact both the overall system performance and their operational behavior, especially on virtualized RANs where the time constraints for some specific operations are tight. Therefore, ORIGAMI's CCL will enable the regulation of virtualized network functions by auditing and supporting decisions taken during their operation. In this way, it will ensure their correct operation and performance in real-time under a large variety of scenarios.

In summary, ORIGAMI's CCL shall provide a feature rich API that fulfills a twofold objective: i) it provides an abstraction of the heterogeneous (in terms of technology) and disaggregated (in terms of executions environments) computing infrastructure that can be used by the 6G Network Userland (which includes all the software components being executed in the network); and ii) at the same time exposes to the 6G Userland specific tools to exploit (if needed) the underlying infrastructure components, and also policies that bound their behavior. In this way, the underlying infrastructure can be efficiently pooled and effectively used just when the 6G Userland applications require it. This aspect is of particular importance to avoid vendor lock-in while guaranteeing the upmost performance in a challenging environment such as the 6G one.

By exploiting ORIGAMI's CCL, the project will tackle barriers #1 ("Unsustainable RAN virtualization"), #2 ("Poor inter-operability of RAN components"), #3 ("High latency to process complex 6G network problems"), #4 ("Under-utilized modern programmable transport"), and #7 ("Inadequate networking data representation"). Dependencies are shown in the Figure 18.

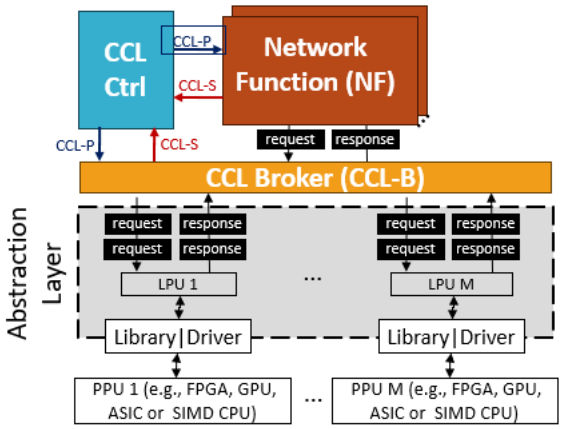

Figure 18: High-level CCL architecture

Physical Processing Units (PPUs) including hardware accelerators such as ASICs or FPGAs, GPUs, pools of CPU cores or even quantum computing functions (e.g., anneals), are accessed through an Abstraction Layer (CCL-AL). For instance, Forward Error Correction tasks, which may be accelerated by an ASIC or an FPGA; neural network execution for machine learning applications, which can be processed in a GPU; an optimization model, which can be expedited by a quantum annealer; or general-purpose functions, which require a regular pool of CPUs. The CCL-AL provides a common API to offload NF processing tasks into the CCL by describing heterogeneous PPUs as Logical Processing Units (LPUs), which homogenize access to the computing fabric.

Some of these tasks may require the deployment of a kernel into the PPU prior to real-time processing, e.g., for GPU, a PI for an FPGA, an embedding for a quantum annealer, or a library for a CPU. Such onboarding mechanisms will be designed within WP2. During such onboarding process, NFs also declare the KPIs that the CCL must satisfy.

During NF runtime, NFs *queue* processing requests, characterized by the kernel required and the inputs for the kernel (e.g., parameters of the optimization problem, input data of a neural network, or soft bits for a FEC decoder). The arrival process of NF requests may be regulated by *a policy (NF policy),* which NFs shall obey. For instance, CCL-C may bind the bandwidth eligible for radio scheduling in the MAC layer of a Distributed Unit to ensure the CCL-AL can process the resulting workload within strict deadlines. Conversely, the CCL Broker is in charge of routing requests (and the associated responses) between NFs and LPUs based on another *policy (CCL policy)*, e.g., prioritize low-energy LPUs as long as their queues are below a certain threshold, to ensure meeting processing deadlines.

In the control plane, the CCL Controller (CCL-C), mediates in between NFs and the CCL-AL, by computing policies at both ends (CCL and NF policies), using some algorithm, with the goal of minimizing cost and energy consumption while satisfying KPIs associated with the NF.

5.4 Global Services Based Architecture (GSBA)

Both ORIGAMI´s CCL and the ZTL (together with the other legacy domain buses, such as 3GPP SBA) rely upon the Global SBA (GSBA), which is meant to enable the declaration and management of services between different domains (e.g., network operators, and infrastructure providers). In ORIGAMI, a “domain” maps to an entity that owns different sub-domains, such as the radio access network, the core network, and the international carrier network.

Within this ecosystem, one of the major challenges to be tackled is the inherent lack of trust between the different entities, which makes resource sharing and the deployment of novel business models

difficult. While legacy domain buses such as 3GPP SBA shall assist the GSBA, there exist domains where new buses need to be developed.

5.4.1 Novel domain buses: RAN Bus

This process involves working with a specific RAN bus within the RAN Intelligent Controller (RIC) platform. The RAN bus collects key performance measurements (KPM) from nodes and delivers RAN control (RC) decisions to enable controllability of the underlying RAN infrastructure.

The RIC platform hosts several xApps that leverage the RAN bus to:

Functionality/Task	Provided by/Method	Details
Collect performance measurements	RAN bus service models (E2SM-KPM)	Measurements are collected from RAN nodes as defined by the O-RAN Alliance.
Receive and modify Information Elements (IEs)	RAN node using E2SM RAN Control (E2SM-RC)	xApps can modify IEs in signaling messages without needing to decode the entire message. This allows xApps on the RIC platform to perform simultaneous control tasks.
xApps task performance	RIC platform	xApps should perform tasks that do not conflict with one another. In ORIGAMI, developed xApps are designed to avoid conflicts. However, conflicts may still arise from third-party xApps.
Conflict types	RIC platform	Direct conflicts: Two or more xApps request different settings for the same IE. Indirect conflicts: Dependencies between IEs and resources are not directly observable but can be inferred. Implicit conflicts: Dependence between xApps is not obvious, and optimizing one metric may adversely affect another.
Conflict Mitigation	RIC platform	Direct conflicts: Resolved by the Conflict Mitigation component deciding on changes or the order of changes. Indirect conflicts: Resolved by observing the effects after actions are executed and deciding on any necessary corrections like rolling back actions. Implicit conflicts: Managed by ensuring xApps target different parameters or by establishing a generic approach to manage such conflicts. Mitigation strategies include avoiding conflicts or modeling them in schemes difficult to observe and manage.

Table 6: RIC platform xApps

In this way, third-party xApp providers may optimize RAN node performance. Each xApp should focus on optimizing a different metric. In case of conflicts, the Conflict Mitigation component will avoid or resolve them.

5.4.2 Novel domain buses: Network Intelligence Bus

As discussed in [26] the big impact of Artificial Intelligence in network management is to integrate a bus to manage the NI Stratum and to assist ORIGAMI's GSBA. The architecture, proposed in [27] is depicted in Figure 19.

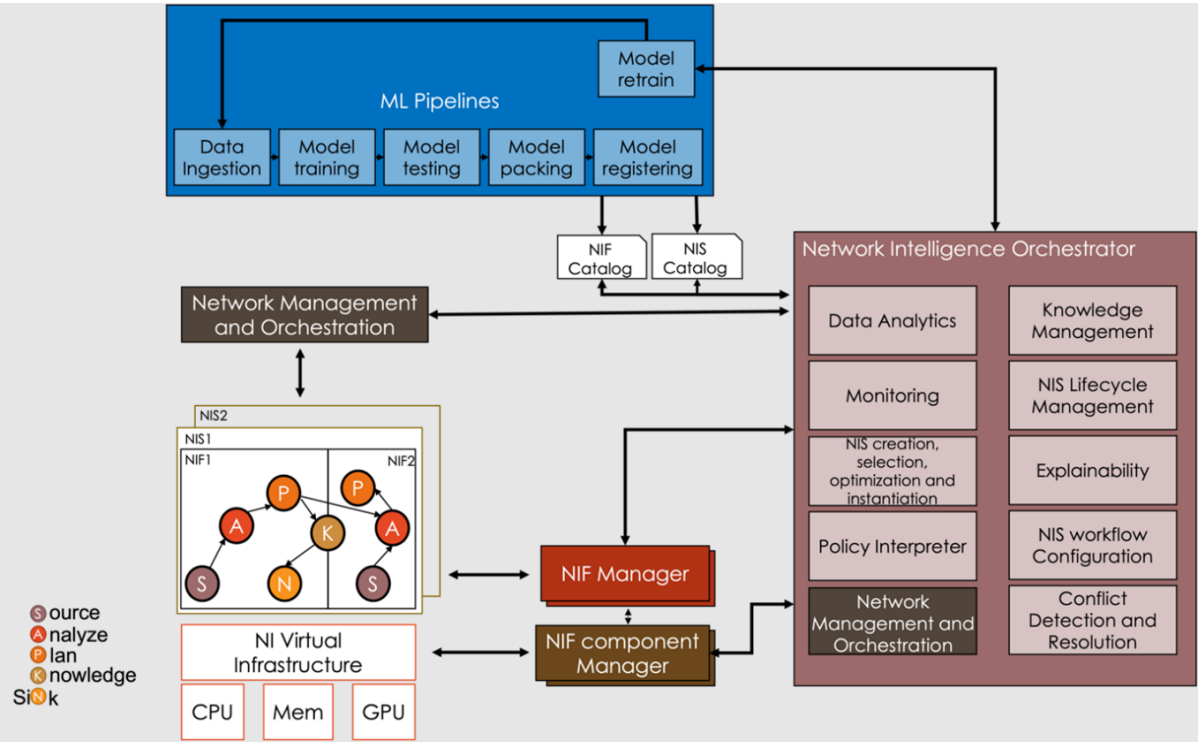

Figure 19: The Network Intelligence (NI) Stratum and the functional blocks of the Network Intelligence Orchestrator and ML pipelines

To describe the organization and operations of the Network Intelligence (NI) Stratum, a reference model has been introduced that organizes complex NI algorithms into a structured hierarchy consisting of Network Intelligence Services (NISs), Network Intelligence Functions (NIFs), and atomic NIF Components (NIF-Cs). This hierarchy helps to conceptualize and manage the layers of network intelligence.

A Network Intelligence Orchestration (NIO) system is responsible for managing the NISs and NIFs, using fundamental building blocks to structure its internal architecture. The NI Stratum incorporates the Monitor-Analyze-Plan-Execute over a shared Knowledge (MAPE-K) feedback loop, commonly employed in autonomous systems. This has been expanded into a Network MAPE-K (N-MAPE-K) model specifically adapted for the NI environment. This model encompasses an inference loop, a conventional supervised training loop, and an additional training loop for online learning, catering to the dynamic needs of network intelligence.

The N-MAPE-K model highlights six fundamental classes of atomic NIF-Cs:

- Sensor NIF-Cs: gather input measurement data.
- Monitors NIF-Cs: interact with Sensor NIF-Cs to gather raw data.
- Analyze NIF-Cs: preprocess, summarize, and prepare data for the specific NI algorithm.
- Plan NIF-Cs: implement the specific NI algorithm.
- Execute NIF-Cs: interact with the managed system and change configuration parameters.
- Effector NIF-Cs: update configuration parameters in the Network Function (NF) and specify APIs to be used.

This approach presents a unified framework that combines the operational hierarchy of NI components in the NIO and the N-MAPE-K representation of NIF-Cs, advancing the vision of a complete NI stratum. The framework demonstrates how multiple NIF-Cs can create NIFs, which can be combined to form NISs, such as a reliable virtualized RAN (vRAN) service.

5.4.3 GLOBAL BUS: HOLISTIC SMO FOR COST REDUCTION

Mobile Operators are increasingly focusing on optimizing network performance and reducing Total Cost of Ownership (TCO). The GSBA is pivotal in advancing Service Management and Orchestration (SMO) platforms, integrating existing tools like radio design planning and RAN configuration with O-RAN modules via standardized interfaces and open-source software. This integration facilitates the automation of Life Cycle Management (LCM) for apps and algorithms, enhancing network infrastructure awareness and efficiency through CI/CD and MLOps strategies. Additionally, transitioning to a microservices-based architecture with a service mesh paradigm and implementing an API manager under the 3GPP Common API framework are key steps to improve network management and interoperability. The GSBA's integration across domain buses further supports cohesive network resource management and the development of novel network control loops, aligning with Network Operator policies for better service provision and monetization opportunities.

Component	Role	Objective
GSBA	Integrates various network management tools and modules.	Streamlines mobile radio access management and orchestration.
Standardized Interfaces & Open Source	Facilitates interoperability with O-RAN modules and other network elements.	Ensures efficient process flows and integration across different network segments.
CI/CD and MLOps	Automates LCM for apps and algorithms.	Enhances network infrastructure awareness and efficiency at the service level.
Microservices and Service Mesh	Transforms core network architecture.	Improves performance and flexibility in network management.
3GPP Common API Framework	Standardizes API management across the network.	Enhances secure and interoperable API provisioning and consumption.
Domain Buses (RAN, NI, 3GPP)	Connects various network components.	Promotes efficient and cohesive management of network resources.
Novel Control Loops	Utilizes innovative global services and NI algorithms.	Allows for new capabilities and customization, facilitating better service provision and pricing.

Table 7: Global bus components

5.5 ORIGAMI OVERALL ARCHITECTURAL DESIGN

Building on the previously discussed high-level architectural description of the desired layers and loops, some initial gaps compared to the current state-of-the-art architecture are identified. This section details the overall structure of the ORIGAMI architecture.

5.5.1 EXTENSION OF THE SBA BUS TO THE RAN

There are currently many study activities in the 3GPP SDO (Release 19) related to the deployment and lifecycle management of AI/ML functionalities used for optimisation, analytics, etc.

These activities mainly concern the OAM, NR RAN and CN domains. While both the OAM and CN domains use a service-based paradigm, the NR RAN does not. This fact raises some issues on the following points related to the NR RAN domain:

- The different target location of the AI/ML algorithm requires a different solution for the data collection, data analysis, training, testing and validation, deployment and inference phases.

- different use cases based on AI/ML algorithms (i.e. energy efficiency, QoS, load optimisation, etc.) require a different solution in terms of the data required and the interaction between the NFs involved.

These issues do not characterize the equivalent functionalities in the OAM and core network domains.

For this reason, the possibility of extending the Service Base paradigm to the RAN (or a part of it) could simplify the possibility of introducing new AI/ML algorithms and the necessary services for their lifecycle. Furthermore, the extension of the Service Based paradigm to the RAN domain opens the possibility to introduce more flexibility into the network architecture for its deployment, provisioning and assurance phases.

5.5.2 O-RAN RT-RIC IN 3GPP WITH THE INTERFACES THROUGH THE GSBA

As is widely recognized, the O-RAN Alliance introduced the RAN Intelligent Controller (RIC) as a novel functionality designed to facilitate RAN resource control through programmable functions, namely rApps within the Non-RT RIC and xApps within the Near-RT RIC. Consequently, this standards development organization (SDO) has defined the following interfaces:

- E2, connecting the Near-RT RIC with the O-CU and O-DU network functions,
- A1, linking the Near-RT RIC with the Non-RT RIC,
- R1, serving as an internal interface within the Non-RT RIC to enable xApps to utilize the services provided by the O1, O2, and A1 interfaces.

The introduction of the RIC into the RAN and OAM domains presents both advantages and challenges. A significant benefit is the incorporation of programmable functionalities (xApps) that allow for flexible management and control of RAN resources. However, the solution outlined by the O-RAN Alliance also presents several issues:

- The RIC solution is not service-based; therefore, the interaction between the Near-RT RIC and the O-CU and O-DU occurs via a new interface (E2), adding complexity to the RAN.
- The Non-RT RIC within the OAM domain (SMO in O-RAN) does not provide services to other OAM functionalities. This contravenes the service-based paradigm adopted by 3GPP for OAM, creating an ambiguous separation between what lies inside and outside the Non-RT RIC. Additionally, the definition of the A1 interface may be redundant, as it overlaps with the O1 interface in providing management services.

A potential solution involves introducing an analogous concept of the RIC that adheres to the Service-Based paradigm within both the OAM and RAN domains. To achieve this, the Service-Based paradigm could be implemented in the RAN domain where feasible, considering the performance requirements of each network function. Subsequently, functionalities and services would be introduced to manage the lifecycle of programmable functionalities interacting with the RAN, thereby avoiding the need for new interfaces and specific solutions.

5.5.3 INITIAL ARCHITECTURAL STRUCTURE

The discussion detailed above motivates the need for a paradigm shift for the 6G Architectural Design. In particular, the aim is to integrate the two architectural components discussed above (CCL and ZTL) into the GSBA solution, while fulfilling the principles discussed above.

Starting from the access part, which will be extended to integrate the CCL and will incorporate the functionalities discussed in Section 5.4.1. the initial preliminary structure in Figure 20 is depicted.

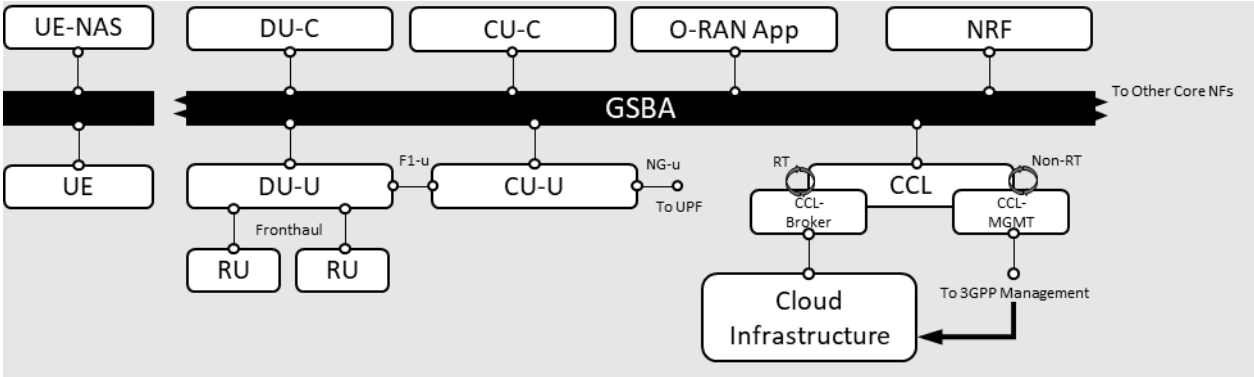

Figure 20: The initial definition of the GSBA extension to the RAN and the integration of CCL

The main design principles that will be followed in ORIGAMI are:

Split of the DU in two components. To allow a cleaner split of the control plane from the user plane function, ORIGAMI will extend the current approach of the CU split promoted by O-RAN also down to the DU, creating two subfunctions DU-U (managing the user plane issues) and DU-C. They will interact with each other through the GSBA, exposing API for the handling of relevant flows, in a similar way other Core functions interact with AMF and SMF (and from there down to the UPF). The very stringent requirements for fronthaul traffic make this approach unfeasible for the DU-U to RU interface, which is left out of this scope. CU-U finally connects to the UPF using the NG-u Interface as defined by O-RAN.

Integration of O-RAN apps. In this view, O-RAN xApps directly sit on the GSBA bus, which integrates part of the functionality of the O-RAN RT RIC. Over the GSBA, all the RAN elements interact using an API-based approach. Also, these elements will be fully integrated with other Core native functions, allowing an efficient exchange of data between this two-network domain. In particular, Network Repository Function (NRF) should be extended to cover also these RAN related functions.

CCL integration. As discussed in Section 5.3, one of the functionalities of the CCL is the fast pace dispatching of requests coming from Network Functions. As the CCL has a broad range of functionalities to be supported, this element is identified as CCL-Broker, which directly interfaces with the underlying Cloud Infrastructure to perform, e.g., FEC decoding. This is supporting part of the functionality that is performed AAL interface in O-RAN, with very fast timing constraints. Besides this, CCL also must perform slower tasks, in the order of seconds, which fall into the 3GPP management system and the integration with Cloud Orchestration mechanisms. ORIGAMI will study the integration of such modules.

UE Integration. Besides the RAN network functions, ORIGAMI will study the integration of the UE into the extended GSBA. In particular, while the SBA approach touches radio related aspects (due to very stringent timing requirements), there is the potential to integrate Non Access Stratum procedures directly into the GSBA.

The integration of ZTL may require less architectural changes to extend the GSBA towards the other parties in the network, being them either Application Service Provider, or other operators, to empower the Global Operator Model.

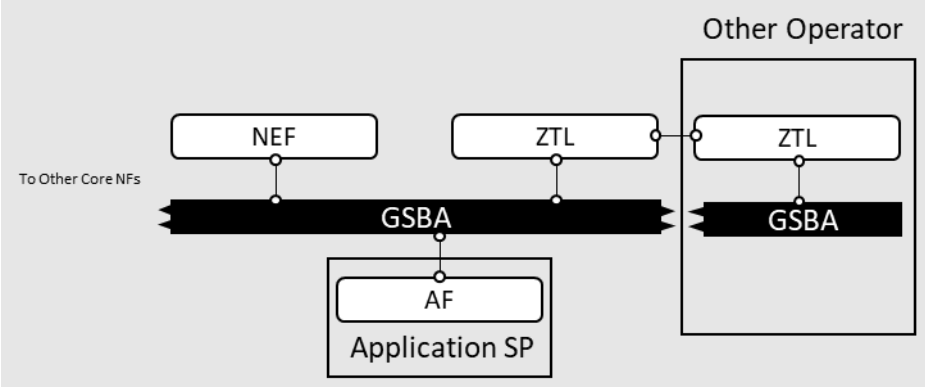

Figure 21: The initial definition of the GSBA extension to the RAN and the integration of CCL

As depicted in Figure 21, ZTL shall provide the required interfaces towards i) other ZTL from foreign operators, to ensure the horizontal exposure of relevant data between them and ii) to other Network Function in the core, especially the Application Function (AF), which interfaces towards the Application Service Provider for vertical exposure. Throughout the project, ORIGAMI will work in these areas to define the needed interfaces for this vision.

6 ORIGAMI USE CASES

As discussed in Section 2, ORIGAMI aims to overcome the barriers listed in Section 4 through the introduction of a set of 10 Use cases, summarized in Table 8 and thoroughly discussed in the rest of the section. For each of them a general description is provided as well as the discussion on the targeted barrier and the architectural components that will be used (together with a set of Non-Functional Requirements that need to be fulfilled) and the KPIs that are targeted.

Use case	Id	Architectural innovation	Barrier	KPI
Data-driven task offloading for reliable vRAN acceleration	SRV	Compute Continuum Layer	Unsustainable RAN Virtualization	K2, K3
Conflict Mitigation of xApps and Interoperability of O-RAN component	PIOR	Compute Continuum Layer	Poor Interoperability of RAN Components	K2
Enhancing Management and Stability in the 6G Architecture	EMSA	Compute Continuum Layer	Poor Interoperability of RAN Components	K2
Interoperable Machine Learning Models Improving RAN Energy Efficiency	IMLE		High Latency and Unreliable Network Intelligence (NI)	K3, K4, K5
Compute- and Fairness-Aware Radio Resource Allocation Algorithms in Virtualized RANs	CFA	Compute Continuum Layer	Unsustainable RAN Virtualization High Latency and Unreliable Network Intelligence (NI)	K1
Effective, distributed and streamlined access to u-plane computing capabilities	EAUC	Compute Continuum Layer Global SBA	Under-utilized Modern Programmable Transport	K4
Enabling the Global Operator Model	GMNO	Zero Trust Layer Global SBA	Lack of Global Service Application Programming Interfaces (API) Obsolete Trust Model Hinders Performance	K7, K9
Limited Trust Network Analytics	LTNA	Zero Trust Layer	Obsolete Trust Model Hinders Performance	K10
Anomaly Detection	KR	Global SBA	Inadequate Networking Data Representation	K10
Network Core traffic analysis and optimization	NCAM	Global SBA	High Volume of Control Plane Signaling	K12

Table 8: Use cases summary and Relevant KPIs

A description of the set of target KPIs is listed in the following table.

KPI	Description
K1	Energy efficiency (bits-per-joule)
K2	Cost efficiency (bps-per-$)
K3	Reliability (%)
K4	In-band ML model inference latency (ms/ µs)
K5	In-band ML model inference accuracy (%)

K6	In-band ML model inference throughput (Gbps)
K7	Network CAPEX ($)
K8	Network energy consumption (KWh)
K9	Control plane latency (ms)
K10	Anomaly detection recall and sensitivity
K11	OPEX gains ($)
K12	Control-plane efficiency (%)

Table 9: KPI Definition

Both the Use cases and the KPI lists are under development by the time the present deliverable is submitted. Further elements for both categories will be added in D2.2.

6.1 DATA-DRIVEN TASK OFFLOADING FOR RELIABLE VRAN ACCELERATION (SRV)

6.1.1 GENERAL DESCRIPTION

To address the unsustainability issues highlighted in Section 4.1, ORIGAMI proposes two strategies.

1. **Opportunistic HA offloading.** As hinted by the experimental results shown in Figure 11, CPUs alone may handle **some** 5G PHY workloads in Distributed Units (DUs) without the assistance of Hardware Accelerators (HAs) by exploiting SIMD programming and other optimizations. Additional information can be found in [29] [30]. However, CPUs alone cannot ensure 5-nines reliability for **all** workloads, as shown in Figure 11, and, consequently, they are usually shunned for this job in industry-grade RANs. Instead, ORIGAMI will show that CPUs can be a valuable complement to HAs in these tasks and that balancing DU workloads between CPUs and HAs can substantially improve the cost- and energy-efficiency of vRANs. The rationale is that minimizing processing latency brings no benefit as long as DU processing deadlines are met, hence CPUs may be occasionally exploited to alleviate the HAs' energy toll. As an example, Figure 11 (left) shows that a CPU core can decode within 1 ms TBs below 100 Kb (which correspond to a large portion of today's real-world TBs, as reported in the literature) consuming ~5.7× less energy than a GPU-based HA (right plot).

2. **Processor Pooling.** HAs co-located with (and thus exclusively used by) individual DUs suffer from low usage under real workloads. ORIGAMI will seize this opportunity to share HAs among multiple DUs, so as to amortize the cost of these expensive resources, and provide the needed acceleration at an affordable cost per DU. The concept of RAN pooling is not new, though. Indeed, 71% of US operators intend to realize RAN pooling solutions by 2025 [31], and some already implement it [32], but the traditional RAN centralization approaches *only* exploit long-term traffic variations, such as day-night ones, which is insufficient for cost-efficient RAN virtualization.

6.1.2 INVOLVED BARRIERS AND ARCHITECTURAL ELEMENTS

Implementing the above two strategies is technically challenging as pooling heterogeneous computing resources efficiently requires: (i) harnessing real-time multiplexing opportunities at sub-millisecond timescales where both PHY processing latencies and user loads fluctuate; and (ii) anticipatory operation that effectively copes with fluctuations in the future user demand. Note that, because resources are no longer over-dimensioned, rare peak loads risk violating PHY processing deadlines, which compromises reliability.

Moreover, the solution to these challenges goes well beyond the current capabilities of O-RAN. Although O-RAN provides convenient abstractions for heterogeneous processors, it falls short to support (i) real-time coordination among multiple DUs and (ii) radio scheduling policies that are *compute-aware*, two requirements that are essential to attain multiplexing gains while adhering to the stringent reliability of the industry (see FR-SRV-002 next).

Consequently, to break Barrier #1 above, ORIGAMI´s Compute Continuum Layer (CCL) shall meet the following functional requirements.

FR-SRV-001	
Description	ORIGAMI shall integrate NI solutions in vRAN systems
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not integrate NI solutions into vRAN systems, as ORIGAMI partners were already capable of integrating such kinds of solutions in Open Source vRAN environments.

Table 10: FR-SRV-001

FR-SRV-002	
Description	NI solutions integrated into vRAN systems shall comply with O-RAN specifications
Version	Y1M3
Stage	Network Intelligence
Risk	2/3
Risk Description	There is a mild risk that ORIGAMI may require extend O-RAN architecture to support novel vRAN technologies

Table 11: FR-SRV-002

FR-SRV-003	
Description	The inference time of NI solutions integrated into vRAN shall respect 3GPP and O-RAN latency budget requirements in the Distributed Unit
Version	Y1M3
Stage	Network Intelligence
Risk	1/3
Risk Description	Low inference times may be achieved by reducing the complexity of NI models.

Table 12: FR-SRV-003

FR-SRV-004	
Description	NI solutions integrated into vRAN systems may implement anticipatory operation
Version	Y1M3
Stage	Network Intelligence
Risk	2/3
Risk Description	Fluctuations in base station workload demands by real-world RANs may be unpredictable.

Table 13: FR-SRV-004

FR-SRV-005	
Description	ORIGAMI CCL shall provide interfaces to coordinate Network Functions sharing infrastructure
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	The risk of not designing such an interface is low, given the expertise of the consortium

Table 14: FR-SRV-005

FR-SRV-006	
Description	ORIGAMI CCL shall provide interfaces to coordinate with radio schedulers within Network Functions
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	The risk of not designing such an interface is low, given the expertise of the consortium

Table 15: FR-SRV-006

6.1.3 TARGET KPIS

To provide sustainability, novel RAN virtualization NI solutions shall provide energy- and cost-efficiency gains over current virtualization strategies based on dedicated hardware accelerators (ORIGAMI´s baseline). Consequently, NI solutions addressing Barrier #1 shall meet the following non-functional requirements.

NFR-SRV-001	
Description	Given pre-determined PHY processing deadlines, NI solutions integrated into vRAN systems shall ensure that such deadlines are met with at least 99.999% probability to attain reliability.
Version	Y1M3
Target KPIs	K3 (Reliability): This requirement directly establishes a target on KPI K3 of 99.999% reliability.
Risk	1/3
Risk Description	There is a low risk that the technologies developed within ORIGAMI will not comply with this requirement because baseline solutions, e.g., using dedicated hardware accelerators, already comply with this requirement.

Table 16: NFR-SRV-001

NFR-SRV-002	
Description	NI solutions integrated into vRAN servers shall increase the ratio of bits per unit of capital expenditure (cost-efficiency) over the baseline approach by at least 10 times with realistic workloads.
Version	Y1M3
Target KPIs	K2 (Cost-efficiency): This requirement directly establishes a target on KPI K2 of 10x increase on cost-efficiency.
Risk	2/3
Risk Description	There is a mild risk that pooling strategies intended to increase cost-efficiency cannot provide the required reliability.

Table 17: NFR-SRV-002

NFR-SRV-003	
Description	NI solutions integrated into vRAN servers shall increase the ratio of bits per unit of energy consumption (energy-efficiency) over the baseline approach by at least 2 times with realistic workloads.
Version	Y1M3
Target KPIs	K1 (energy-efficiency): This requirement establishes a directly target on KPI K1 of 2x increase on energy-efficiency.
Risk	2/3
Risk Description	There is a mild risk that opportunistic offloading strategies intended to increase energy-efficiency cannot provide the required reliability.

Table 18: NFR-SRV-003

6.2 Conflict Mitigation of xApps and Interoperability of O-RAN component (PIOR)

6.2.1 General Description

To address the poor inter-operability of RAN components issues highlighted in Section 4.2, ORIGAMI proposes two strategies.

- **Conflict mitigation of the 3rd party xApps in Near-RT RIC platform**: To solve the conflict amongst 3rd party xApps sharing same RAN, the conflict mitigation techniques will be designed and developed to ensure interoperability with RAN components. The direct conflicts will be resolved by the Conflict Mitigation component, which will make the final determination on whether any specific change is made, or the order in which changes are applied. The Indirect conflicts will be resolved by post-action verification. The actions are executed, and the effects on the target metric are observed. Based on these observations, the system must decide on potential corrections, such as rolling back one of the xApp actions. The implicit conflicts are the most challenging to mitigate, as these dependencies are difficult to observe and model in any mitigation scheme. Conflicts will either be avoided by ensuring xApps target different parameters or by establishing a generic approach to manage such conflicts.
- **Enabling interoperability of RAN nodes over E2 interface**: Enabling interoperability of Radio Access Network (RAN) nodes over the E2 interface is crucial for ensuring seamless communication and cooperation between different components within the O-RAN ALLIANCE Near-RT RIC architecture. In ORIGAMI, the E2 interface and the message exchange over the E2 interface will be configured in such a way that it will ensure that the RAN nodes can interpret and respond to messages exchanged over the E2 interface appropriately.

6.2.2 Involved Barriers and Architectural Elements

Implementing the above two strategies is technically challenging the expectation of the as the 3rd xApps running at the Near-RT RIC can trigger conflicting operation that can interrupt KPI on the underlying RAN. The interoperability and the effective management of the conflicts can be achieved through : (i) effective deployment of the RAN bus (i.e., the E2 interface and messaging infrastructure); and (ii) Interoperability of the Near-RT Radio Intelligent controller (Near-RT RIC) by onboarding 3rd party xApps in the Near-RT RIC platform.

Moreover, the solution to these challenges goes well beyond the current capabilities of O-RAN. Although O-RAN provides the E2 interface specifications it falls short to support (i) new E2 service model (E2SM) related to key performance measurement (KPM) subscription from underlying RAN and (ii) new messaging structure and RAN control (RC) message to trigger the same optimization decision towards underlying RAN, two requirements that are essential to overcome the barrier related to poor interoperability of the RAN components. Consequently, to break Barrier #2 above, ORIGAMI´s Compute Continuum Layer (CCL) shall meet the following functional requirements.

FR-PIOR-001	
Description	ORIGAMI shall effectively deploy the RAN bus (i.e., the E2 interface) to enable multi-time scale controllability
Version	Y1M3
Stage	Network Intelligence
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not deploy the E2 interface effectively in the evaluation framework, as ORIGAMI partners were already capable of deploying such solutions in several experimental platforms.

Table 19: FR-PIOR-001

FR-PIOR-002	
Description	ORIGAMI shall distribute the RAN components across edge while enabling real-time control of the E2 nodes (i.e., CU, DU, and RU) with the RAN architecture
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not be able to enable real-time control the E2 nodes (i.e., CU, DU, and RU), as ORIGAMI partners are implementing and integrating the E2 nodes with several such edge platforms to achieve the closed controlled loops that are specified in O-RAN ALLIANCE specifications.

Table 20: FR-PRIOR-002

FR-PIOR-003	
Description	ORIGAMI shall mitigate the conflicts of 3rd party xApps onboarding on the Near-RT RIC controlling the E2 nodes (i.e., CU, DU, and RU).
Version	Y1M3
Stage	Network Intelligence
Risk	1/3
Risk Description	There is a medium risk that ORIGAMI will not be able to resolve all the conflicts to ensure interoperability with RAN components. The Conflict Mitigation component of the RIC platform can address those conflicts that occurred on underlying RAN.

Table 21: FR-PRIOR-003

6.2.3 Target KPIs

To provide interoperability of RAN components over RAN bus (i.e., the E2 interface from O-RAN specifications) and conflict mitigation of xApps in the RAN, the CCL solutions shall provide multi-timescale control within the computing fabric, taking into account the system's multi-time scale controllability, conflict mitigation across different control loops, rigorous anomaly detection, and agile management of vast amounts of data to handle AI/ML pipelines. Consequently, CCL solutions addressing Barrier #2 shall meet the following non-functional requirements.

NFR-PIOR-001	
Description	The RAN components deployed as a containerized network function (CNF) in the servers with efficient utilization of the computing resources shall improve 2X cost-efficiency (bps/$) over the baseline deployment of the RAN.
Version	Y1M3
Target KPIs	K2 (Cost efficiency): This requirement directly establishes a target on KPI K2 of 10x higher than today's vRANs
Risk	2/3
Risk Description	There is a medium risk that optimal computing resources utilization to satisfy underlying users demand may increase the consumption of hardware resources of the server.

Table 22: NFR-PIOR-001

6.3 Enhancing Management and Stability in the 6G Architecture (EMSA)

6.3.1 General Description

To exploit the potential improvements and solve the issues discussed in Section 4.2, in ORIGAMI, a set of approaches are conceived and listed in the Table below. The list of functional requirements is discussed in more detail in Section 6.3.2.

Objective	Solution
Developing xApps for network management solutions in Open RAN architecture that interoperable with 3rd party RAN and Near-RT RIC vendors	The set of solutions addressing Barrier #2 by creating an xApps that provide scalable deployment of the xApp in the Near-RT RIC platform that subscribe KPM and provide RC decision by resolving direct conflict of the xApp.
Developing network management solutions (i.e., the radio resource quota optimization and multi-MNO slice admission and congestion control xApps) deployed in Near-RT RIC architecture	ORIGAMI addresses the challenge of solving network management problems, specifically within the context of O-RAN. It will be developed and integrate the network management solutions as an xApps with Near-RT RIC by considering Operator Core Networks (MOCN) architecture that allows a network operator to provide access to a single radio access network by other operators.
KPM extraction from the E2 nodes deployed in an Edge infrastructure	ORIGAMI will develop data collection methodologies that enhance the quality of the gathered information from the infrastructure that deploy the RAN components.

Table 23: Link between Objective and Solutions

6.3.2 Involved Barriers and Architectural Elements

Integrating the aforementioned solutions into the 6G ecosystem will require the extensive support of the ORIGAMI architectural models, as follows:

The CCL shall support the operation of the NI solutions in almost real time conditions, supporting hence the gathering of the data coming from underlying RAN that connect Near-RT RIC over E2 interface. This is also true for the support that address interoperability between Open RAN components aiming to create a seamless ecosystem that enables effective deployment of 6G RAN functions.

The scalable and generalizable network management focusing on solving several network management problems and leveraging ORIGAMI's architectural innovations to enhance data collection, ultimately improving interoperability and facilitating the deployment of a RAN bus and RIC in Open RAN architecture.

FR-EMSA-002	
Description	ORIGAMI shall develop network management solutions
Version	001M3
Stage	Network Intelligence
Risk	2/3
Risk Description	Although ORIGAMI partners are experienced in developing network management solutions, there is a medium risk of integrating the network management solutions with underlying RAN.

Table 24: FR-EMSA-002

6.3.3 Target KPIs

NFR-EMSA-001	
Description	ORIGAMI shall enhance the performance of throughput (i.e., Mbps or bit/s/Hz) by radio resource quota optimization solutions for multi-MNO configuration by ≥20% in the case of congestion of the resources per slice to satisfy the user demands.
Version	001M3

Target KPIs	K2 (Cost efficiency): This requirement directly establishes a target on KPI K2 of 10x higher than today's vRANs. This will further enhance the cost efficiency from the baseline deployment of RAN with MOCN configurations. Throughout will be measured in Mbps or bit/s/Hz.
Risk	2/3
Risk Description	There is a medium risk that ORIGAMI will not be able to provide network management to improve the throughput in the RAN in real-time when resources are limited on the specific slices while the demands for the users are higher than the availability.

Table 25: NFR-EMSA-001

6.4 Interoperable Machine Learning Models Improving RAN Energy Efficiency (IMLE)

6.4.1 General Description

To address the limitations and enhance the functionalities identified in Section 4.2, ORIGAMI proposes to develop interoperable ML models embedded into xApps deployed in the Near-RT RIC platform to provide efficient, scalable network management policies to improve energy efficiency in the RAN when 3rd-parties xApps and ML models coexist. To do so, ORIGAMI xApps will leverage the RIC conflict resolution mechanisms, subscribe to KPMs and data exposed in the RIC Shared Data Layer by 3rd parties xApps to develop enhanced, accurate and fine-tuned ML models and NI solutions.

6.4.2 Involved Barriers and Architectural Elements

While the proposed strategy (Interoperable ML models in xApps) seems promising, technical challenges arise. 3rd parties xApps running on the Near-RT RIC might expose inaccurate data, or with insufficient metadata. Developing and training ML models without prior knowledge of 3rd party xApps behavior might be challenging and thus, might results in inaccurate forecasts and inefficient energy reduction policies despite the ORIGAMI conflicts resolution solutions presented in section 6.2.

FR-IMLE-001	
Description	ORIGAMI shall develop advanced machine learning models for network management solutions improving network energy efficiency when conflicting policies coexist.
Version	001M3
Stage	Network Intelligence
Risk	2/3
Risk Description	Although ORIGAMI partners are experienced in developing ML model, there is a medium risk of effectively training the models due to lack of high-quality data related to RAN and without knowledge of 3rd-parties ML models and shared data.

Table 26: FR-IMLE-001

6.4.3 Target KPIs

NFR- IMLE -001	
Description	ORIGAMI shall reduce energy consumption in the RAN by ≥20%, maintaining network capacity ≥99.9% of the time when conflicting energy efficiency policies coexist, and running at a Sub-s timescale.
Version	001M3

Target KPIs	K3 (Reliability): This requirement directly establishes a target on KPI K3 of 99.999% reliability. K4 (In-band ML model inference latency): This requirement directly establishes that NI shall operate with latencies of hundreds of ms at most. K5 (In-band ML model inference accuracy): This requirement directly establishes a target on KPI K5 of above 95% accuracy.
Risk	2/3
Risk Description	There is a medium risk that ORIGAMI will not be able to provide ML driven network management to reduce energy consumption in the RAN in real-time when potentially conflicting energy efficiency policies coexist. The ML-driven solutions with an offline RAN related data can validate the accuracy of the forecasts driving energy efficient network management algorithms.

Table 27: NFR-IMLE-001

6.5 Compute- and Fairness-Aware Radio Resource Allocation Algorithms in Virtualized RANs (CFA)

6.5.1 General Description

To tackle the current network architecture limitations explained in Section 4.1 and 4.3, ORIGAMI proposes the following strategies:

- **Design of compute-aware policies for radio resource allocation.** As detailed in section 4.1, LDPC operations are computationally intensive limiting the adoption of network virtualization due to energy costs. Current hardware accelerators (HAs) such as GPUs or ASICs excel at parallelizing their operations, which is particularly beneficial for LDPC operations that are highly amenable to parallelization. Thus, by utilizing HAs, the processing time for tasks like FEC becomes nearly independent of transport block (TB) size, as previously shown experimentally [26]. Thus, **when processing larger TBs the usage of the HA is optimized and the energy consumption per bit is reduced**. This creates an opportunity to design radio resource allocation policies preventing transmission of small TBs. Nevertheless, such a policy will inevitably deteriorate the latency for users, as they might need to refrain from transmitting despite having non-empty (MAC-layer) buffers. It is therefore imperative to strike a balance between energy savings and transmission delays; and further, to disperse fairly these delays across the users so as to avoid excessive service deterioration for some of them.

- **Design of efficient tailored algorithms for fair resource allocation.** In contrast to the application of standard AI/ML approaches, ORIGAMI aims at designing algorithms that are tailored to the specific requirements of the problem. This allows us to design fast algorithms and tailor the objectives to the problem goals. For that purpose, ORIGAMI aims to develop new Online Convex Optimization (OCO) theory [33] to consider not only the minimization of the energy but also the fair distribution of the delay across users. In contrast to other vanilla approaches (e.g., black box models such as neural networks), the aim is to design algorithms with performance guarantees in terms of regret. Finally, ORIGAMI aims to enhance performance of the proposed algorithms with ideas from optimistic learning [34].

6.5.2 Involved Barriers and Architectural Elements

In order to integrate the previously mentioned solutions into the 6G ecosystem, the ORIGAMI architecture needs to play a major role. The CCL must allow near real-time operation of the network intelligence (NI) solutions. This ensures timely collection of data from network elements. To overcome barriers #1 and #3, ORIGAMI shall meet the following functional requirements.

FR-CFA-001	
Description	ORIGAMI shall integrate NI solutions into the RAN systems
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not integrate NI solutions into vRAN systems, as ORIGAMI partners were already capable of integrating such kinds of solutions in Open Source vRAN environments

Table 28: FR-CFA-001

FR-CFA-002	
Description	NI solutions integrated into vRAN systems shall comply with O-RAN specifications
Version	Y1M3
Stage	Network Intelligence
Risk	2/3
Risk Description	There is a small risk that ORIGAMI may require extend O-RAN architecture to support novel vRAN technologies

Table 29: FR-CFA-002

FR-CFA-003	
Description	ORIGAMI shall provide feedback loops in RAN systems to track the performance of the users in terms of QoS
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not integrate feedback loops into vRAN systems. ORIGAMI partners were already capable of integrating such kinds of solutions in Open Source vRAN environments

Table 30: FR-CFA-003

FR-CFA-004	
Description	ORIGAMI shall provide mechanisms to deploy radio scheduling policies in RAN systems
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not integrate mechanisms to deploy radio scheduling policies into vRAN systems. ORIGAMI partners were already capable of integrating such kinds of solutions in Open Source vRAN environments

Table 31: FR-CFA-004

6.5.3 TARGET KPIS

NI solutions addressing Barriers #1 and #3 shall meet the following non-functional requirements.

NFR- CFA-001	
Description	Provable sublinear regret under adversarial scenarios
Version	Y1M3
Target KPIs	K1 (energy-efficiency): This requirement extends on KPI K1 by imposing a fairness criterion to the energy-efficiency gains targeted by KPI K1
Risk	2/3

Risk Description	There is a mild risk that the proposed algorithms does not attain sublinear regret or the theory behind the proof is difficult to develop

Table 32: NFR-CFA-001

NFR- CFA-002	
Description	The algorithm shall balance between energy savings and average user delay.
Version	Y1M3
Target KPIs	K1 (energy-efficiency): This requirement extends on KPI K1 by imposing a fairness criterion to the energy-efficiency gains targeted by KPI K1
Risk	1/3
Risk Description	There is a low risk in achieving this goal as the partners have experience in developing algorithms managing this type of trade-offs

Table 33: NFR-CFA-002

NFR- CFA-003	
Description	The algorithm shall fairly distribute the delay among users independently on their context (e.g., channel quality or traffic load)
Version	Y1M3
Target KPIs	K1 (energy-efficiency): This requirement extends on KPI K1 by imposing a QoS criteria to the energy-efficiency gains targeted by KPI K1
Risk	2/3
Risk Description	There is a medium risk in achieving this goal as this requires fundamental innovations in OCO theory.

Table 34: NFR-CFA-003

6.6 Effective, distributed and streamlined access to u-plane computing capabilities (EAUC)

6.6.1 General Description

As a first step towards overcoming the limitations discussed in Section 4.4, this use case will focus on the design of practical solutions for the integration of ML models in distributed and heterogeneous programmable user planes of mobile networks. The solutions developed by ORIGAMI will hinge upon state-of-the-art ML models that have been very recently proposed in the literature and by the project partners, including hierarchical [35], flow-level [36], joint packet- and flow-level [37] or neural network-based [38] approaches. In particular, the principles underpinning this use case aim at filling gaps in such existing works and are as follows.

First, through the use case the plan is to improve the way ML models are integrated into resource-constrained programmable user-planes, like switch ASICs, network FPGAs and smartNICs. By considering the precise internal architecture of such programmable hardware, the goal is twofold:

- To develop ML models natively designed and trained for the specific platform and more efficient mappings of the ML models.
- To develop distributed solutions that operate across diverse user plane equipment in synergy to achieve a common inference goal. For example, it will combine limited but pervasive programmable switches with more capable smartNICs that are only deployed at specific network locations. The solutions developed by ORIGAMI shall be practical, i.e., apt to deployment in industry-grade programmable network hardware, where they shall operate at line rate in a scalable way and without disrupting legacy functionalities of the user plane.

6.6.2 Involved Barriers and Architectural Elements

To provide the aforementioned capabilities and break Barrier #4, the architectural elements CCL and GSBA are leveraged. In particular, the CCL will be involved in all five aspects mentioned in the previous subsection, while the GSBA will be harnessed to provide tenants with the interface for designing the user plane intelligence and performing metadata monitoring.

FR-EAUC-001	
Description	ORIGAMI shall improve the way ML models are integrated into industry-grade programmable user planes
Version	Y1M3
Stage	Network Intelligence
Risk	1/3
Risk Description	There may not be a significant margin of improvement in the integration of ML models into programmable user planes, as several efficient mapping paradigms have been proposed; yet little attention has been paid to optimizing the ML model design for user-plane deployment.

Table 35: FR-EAUC-001

FR-EAUC-002	
Description	ORIGAMI shall develop distributed solutions that operate at different user plane equipment in synergy to achieve a common inference goal
Version	Y1M3
Stage	Network Intelligence
Risk	1/3
Risk Description	Developing a distributed inference service requires thorough design of the distributed ML, planning of the inference elements placement and close coordination among multiple programmable user planes

Table 36: FR-EAUC-002

6.6.3 Target KPIs

NFR- EAUC -001	
Description	ORIGAMI shall achieve sub-μs inference latency in the transport layer so as to fully ensure line-rate operation
Version	Y1M3
Target KPIs	K4 (in-band inference latency): This requirement reflects the target that user-plane intelligence shall operate with latencies of hundreds of ns at most
Risk	1/3
Risk Description	There is a low risk not achieving this KPI as modern programmable switches operate below the microsecond delay, and ML models deployed in the user plane are expected to work at line rate

Table 37: NFR-EAUC-001

NFR-EAUC-002	
Description	ORIGAMI shall achieve in-band ML accuracy ≥ 95% in large traffic classification tasks with cardinality ≥ 20
Version	Y1M3
Target KPIs	K5 (in-band inference accuracy): This requirement reflects the target that user-plane intelligence shall solve relatively large network traffic classification problems with high accuracy
Risk	2/3
Risk Description	There is a mild risk that ORIGAMI will not achieve this KPI because solutions in the literature struggle to achieve 90% in complex tasks with high cardinality

Table 38: NFR-EAUC-002

NFR-EAUC-003	
Description	ORIGAMI shall allow up to 100 Gbps of ML-processed traffic in the user plane so as to align with high-end traffic forwarding capabilities of industry-grade network equipment
Version	Y1M3
Target KPIs	K6 (in-band inference throughput): This requirement reflects the target that user-plane intelligence shall operate on order-of-Gbps traffic data
Risk	3/3
Risk Description	There is a significant risk that ORIGAMI will not meet this KPI, as current solutions have been tested with a throughput in the order of Mbps, taking into account traffic in the order of tens of Gbps just as background only; hence scaling inference by 3-5 orders of magnitude may be complex

Table 39: NFR-EAUC-003

NFR-EAUC-004	
Description	ORIGAMI shall enable user-plane inference that does not consume more than 20% of the available key (e.g., memory) resources in the programmable hardware
Version	Y1M3
Target KPIs	K11 (OPEX reduction): This requirement contributes to the target that network intelligence shall reduce OPEX, in this case by limiting the number of resources consumed in the user plane to implement inference functions
Risk	2/3
Risk Description	There is a mild risk that ORIGAMI will not achieve this KPI because state-of-the-art approaches to user-plane inference available to date typically exceed the 20% memory usage threshold even for tasks that are simpler than those with high cardinality targeted by the project

Table 40: NFR-EAUC-004

6.7 Enabling the Global Operator Model (GMNO)

6.7.1 General Description

To tackle the network architecture shortcomings that are explained in Section 4.5 and Section 4.6, in ORIGAMI the aim is to propose a novel global model for cellular operators. The so-called **global mobile network operator (GMNO)** will explore two major ideas towards achieving frictionless global cellular access:

Global aggregation of cellular networks: global uninterrupted connectivity is inherent to the idea of cellular networks, and it currently relies of the international roaming function. In ORIGAMI, the aim is to first illuminate the existing business models of global operators that exploit the current implementation of the roaming function, such as the ones new virtual operators (e.g., Truphone, Airalo) already propose [40]. Our goal is to validate whether these models fall short from offering a native-like connectivity service in a visited location, as the one provided by a local operator. Then, with ORIGAMI, our goal is to propose a novel architecture that removes the existent barriers in realizing multi-PLMN access globally, for the same user. Specifically, one first promising idea to be explored is to move support for mobility from the network to the user device, so that a user can experience seamless mobility, even if she frequently switches between mobile providers [40]. Another potential avenue to enable global access to cellular networks is decoupling the identity of the end-user from the cellular infrastructure provider [41]. Both these paths require significant architectural innovations, which will be evaluated as part of this use-case.

Dynamic Interconnections for the Cellular Ecosystem (DICE): realizing the global operator model supposes significant changes in the billing models that are currently in place within the cellular ecosystem. The current business logic around roaming-based global MNO interworking has further implications in the communication performance. The intrinsic lack of trust between the Home Mobile Network Operator (HMNO) and the Visited Mobile Network Operator (VMNO), and the unwillingness of the former to expose to a foreign operator charging information for their users makes home-routed roaming (HR) roaming the default roaming configuration [42]. The purpose of DICE is to allow MNOs to exchange value easily, without the need of a third party to act as a trusted intermediary, to verify the interaction between the roaming partners. With DICE, MNOs can avoid the need for using third-party Data Clearing Houses (DCHs) and instead leverage the potential of Distributed Ledger Technology (DLT) and tokens to retrieve revenue from their roaming partners.

6.7.2 Involved Barriers and Architectural Elements

Implementing the global operator model assumes the evaluation of disruptive ideas that shift our current understanding of the cellular ecosystem, and break two very challenging barriers, namely **Barrier#5 and Barrier #6**. In this use, ORIGAMI explores innovation on both technical solutions for global access, and billing models that support these innovations. Specifically, DICE sets the very ambitious target of innovating the current billing models for international roaming, which have remained largely unchanged within the cellular network architecture until its latest 5G release.

For this reason, two main architectural innovations that are brought to ORIGAMI will be leveraged: the **ZTL** and the **GSBA**. In our effort to enable the GMNO, the aim is to achieve the following requirements:

FR-GMNO-001	
Description	ORIGAMI should design the global operator model to be compatible with 3GPP core network standard specifications
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not integrate NI solutions into 3GPP core systems, as prior work already leveraged 3GPP core network specification to design similar solutions

Table 41: FR-GMNO-001

FR-GMNO-002	
Description	ORIGAMI must design the global operator model to ease zero-trust interactions between the different entities involved in guaranteeing the connectivity service
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a low risk that ORIGAMI will not be able to propose a solution, given that the ZTL aims to rely on promising recent DLT advancements

Table 42: FR-GMNO-002

FR-GMNO-003	
Description	ORIGAMI must guarantee that any information considered private by the MNOs (e.g., billing information across roaming partners, or personal customer information) does not leak to unauthorized parties
Version	Y1M3
Stage	Architectural
Risk	2/3

Risk Description	There is a medium risk that ORIGAMI will not be able to guarantee the privacy requirements; the aim is to integrate state-of-the-art DLT solutions, which are currently thought to enable privacy-by-design

Table 43: FR-GMNO-003

FR-GMNO-004	
Description	ORIGAMI's global operator model should avoid the generation of billing disputes among partner MNOs
Version	Y1M3
Stage	Network Intelligence
Risk	2/3
Risk Description	The fundamental characteristics of distributed ledger technology guarantee the immutability and validity of the information shared between privileged parties. Thus, a medium risk that the billing disputes still arise in the ORIGAMI architecture

Table 44: FR-GMNO-004

6.7.3 TARGET KPIS

With the global operator model, ORIGAMI aims to tackle a fundamental performance-related issue that currently impacts the cellular ecosystem, namely, the impact home routed roaming has on the latency of data communications. Moreover, ORIGAMI should enable MNOs to bill a user and settle the roaming charges among them in almost real time. This would guarantee the user is able to consume the traffic she is allowed, avoiding incomplete and accumulated roaming records.

Consequently, the ORIGAMI solutions addressing Barrier #5 and Barrier #6 shall meet the following non-functional requirements (corresponding to KPIs K7 and K9, respectively).

NFR-GMNO-001	
Description	ORIGAMI's global operator model should decrease the operator CAPEX by 50% compared to the baseline by aggregating infrastructure from multiple providers, and by re-thinking functions within the core network
Version	Y1M3
Target KPIs	K7 (Network CAPEX ($): this requirement establishes a target of 50% to reduce the network capex related to deploying infrastructure to realize the global operator model
Risk	2/3
Risk Description	There is a mild risk that pooling strategies intended to increase cost-efficiency cannot provide the required reliability, that the trust model cannot be enabled or that it is susceptible to attacks from misbehaving actors within the ecosystem

Table 45: NFR-GMNO-001

NFR-GMNO-002	
Description	ORIGAMI's global operator model should decrease the operator control plane latency by 50% compared to current procedures. For example, in the case of billing, ORIGAMI shall enable close to real-time-billing (and depart from the monthly/yearly billing currently in place)
Version	Y1M3
Target KPIs	K9 (Control plane latency (ms)): this requirement establishes a target of 50% reduction in the billing time that currently entities within the cellular ecosystem experience when performing their financial clearing
Risk	3/3

Risk Description	There is a significant risk that selecting the most optimal RAN and network resources from different providers will not result in more efficient transport that will reduce latency

Table 46: NFR-GMNO-002

6.8 Limited Trust Network Analytics (LTNA)

6.8.1 General Description

The current approach of Service Provider (SP) deploying Over The Top (OTT) applications over Mobile Network Operator (MNO) is suboptimal for both MNOs and SPs. For the MNOs, it circumvents their billing systems and reduces them to “dumb pipes”; for the SPs, they cannot support traffic engineering via network re-configurations that would improve the performance of the service. With the arrival of 5G, the few supported interactions are limited to the exchanges of service templates for the deployment of network slices, and the situation is unlikely to change despite recent initiatives of open network (APIs) such as O-RAN [43][44] or [45].

The root cause of this decoupling of the operations of the SP and the MNO lies in the lack of trust among them, due to several reasons such as, e.g., the MNO and SP being market competitors (for instance, in the case of triple play services), or the MNO not trusting SP-driven re-configurations or even the installation of specific modules in its infrastructure. But this lack of trust results in capacity overprovisioning, a non- sustainable approach. In fact, recent initiatives call for tighter Network-Application Integration (NAI), to enable an information-driven management consistent with the changing network circumstances and rapid development of new technologies towards 6G.

To support a tighter integration while circumventing the above trust issues, solutions that enable limited-trust collaborations between stakeholders are needed. This type of collaboration is characterized by a restricted exchange of information between the parties, to prevent the leakage or inference of sensitive information (confidential, strategic, etc.). For instance, in environments where AI is used to drive the autonomous operation of systems, a limited-trust collaboration forbids the exchange of e.g., raw data, labels, or even gradients from the training models, while allowing the exchange of aggregated or less critical information.

An example of the operation of a limited-trust collaboration a network analytics service, which serves to illustrate how the MNO and the SP can collaborate to align their interests without disclosing critical information: on the one hand, the MNO provides a qualitative classification of flows (good vs. bad performance) without revealing the sensitive metrics used to compute this specific information.

Here a limited trust operation by enriching the interaction between an MNO and the SP to support a QoE-driven operation of networks is exemplified. As mentioned above, both parties share only aggregated or limited information for confidentiality or privacy issues, with the common goal of optimizing the accuracy of the analytics. This is aligned with the current trends in architectural design, which envision more direct interactions between different players in the 5G ecosystem [46]. For instance, the 3GPP study items reported in TR 28.824 [47]. define who, what, and how management services can be exposed to third parties, effectively enforcing three levels of access, which range from baseline (i.e., consumer access) to hyperscalers (more advanced control such as Quality of Service (QoS) Management).

One of the new interactions between the MNO and SP may revolve around Network Analytics. The Network Data Analytics Framework introduced by 3GPP [48] allows different network functions, including those from the SP, to access analytics that could be used to optimize performance. Since the default analytics provided by the MNO is oblivious of SP-specific QoE information, cooperation is required to learn the best approach to provide a tailored analytics service.

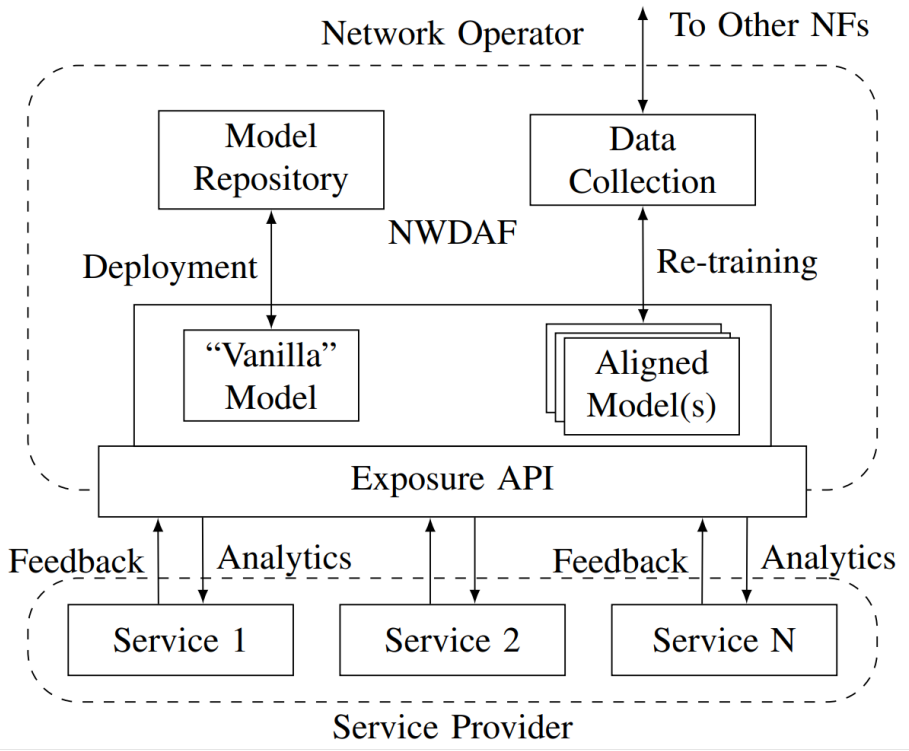

Figure 22: Overview of the MNO-SP Loop. Please notice the (s) to identify possible multiple models, one for each service

The objective is for the MNO to provide an analytics service that is tailored to the QoE of the SP. This service could both support a QoE-driven operation of the network and constitute a new revenue stream for the operator. Following the 5G architecture (although it could apply to other architectures), it is envisioned that it can be implemented at the Network Data Analytics Function (NWDAF), with a catalogue of models tailored to each specific service and provider, as illustrated in Figure 22. There, a scenario which is composed of a set of different services running on top of the same network is depicted. The network operator supports the different services through the Analytics Framework, initially with a vanilla model, which is then refined to a set of different aligned models that provide an optimized version of the analytics for the specific service under consideration. To support this vision, two related but conflicting challenges need to be addressed.

Usually, the mapping between QoS and QoE follows non-linear and multivariable behavior that requires tailored approaches (i.e., there is no one-size-fits- all solution) and calls for the use of data-driven and therefore time- and resource-consuming approaches, such as the one proposed in [49].

The transmitted QoE and QoS information is costly and sensitive. Neither the MNO nor the SP are likely to share such type of raw data, especially with possible competitors

6.8.2 Involved Barriers and Architectural Elements

The integration of such limited trust analytics frameworks requires acting on the ZTL, between SP and MNOs. Due to the complexity of QoE forecasting from current network QoS statistics the usage of machine learning is required. The most straightforward approach would be to train a model using QoS and QoE raw information, but this breaks the limited trust principle discussed above. Following this, each party reveals only the strictly needed information: instead of reporting per-flow or per-user information, one side may convey only aggregated metrics (e.g., summary statistics) that obfuscate detailed information. For instance, the MNO does not report per-flow fine-grained QoS statistics, but only a coarser estimation (e.g., good vs. bad) of their performance.

Since no entity has access to the raw data to perform the training, a single intelligent component, either centralized or spread across the MNO and the SP cannot be expected. Hence, there is a requirement of two separate intelligent entities that cooperatively learn to achieve a common goal. This entails a set of requirements, as discussed below.

FR-LTNA-001	
Description	ORIGAMI shall allow limited trust Network Analytics through the ZTL
Version	Y1M3
Stage	Architectural
Risk	2/3
Risk Description	There is mild risk that ORIGAMI cannot provide this functionality as in previous works, partners already demonstrated this possibility

Table 47: FR-LTNA-001

FR-LTNA-002	
Description	The ZTL should allow the alignment of the SP model with the MNO model with a reduced information exchange providing Limited Trust Network Analytics.
Version	Y1M3
Stage	Architectural
Risk	2/3
Risk Description	There is a mild risk, as in seminal work partners already proved the capability of collaboratively learning among different parties in the network

Table 48: FR-LTNA-002

6.8.3 TARGET KPIS

The provisioning of Limited Trust Network Analytics will be validated upon the achievement of certain Non-Functional Requirements, to effectively tackle Barrier #5.

NFR-LTNA-001	
Description	Limited Trust Network Analytics shall improve the performance in terms of Accuracy and Precision with respect to solutions that do not involve the interaction through the ZTL
Version	Y1M3
Target KPIs	K10, as the plan is to have a comparative analysis with baseline component.
Risk	1/3
Risk Description	When the Limited Trust Network Analytics can be provided through the ZTL, their effect will be beneficial for the specific analytics provided by the MNO.

Table 49: NFR-LTNA-001

NFR-LTNA-002	
Description	The performance improvement of NFR-LTNA-001 should be achieved for multiple services and multiple QoE thresholds
Version	Y1M3
Target KPIs	K10 we plan to improve it for at least 4 services
Risk	1/3
Risk Description	Although the effect of the Limited Trust Network Analytics is dependent on the specific configurations of MNOs and SPs, the Limited Trust Network Analytics will improve the solution without the ZTL.

Table 50: NFR-LTNA-002

6.9 Anomaly Detection (KR)

6.9.1 General Description

Challenges that are explained in Section 4.7 are tackled in this section (i.e., the lack of inadequate network data representation) with a specific and relevant use-case: the detection of anomalies in the connectivity of IoT devices that depend on global cellular services (through international roaming). A major hurdle in solving the anomaly detection issue for cellular IoT devices is the fact that the end-to-end path supporting the corresponding IoT application depends on multiple entities, making the root-case detection an uphill battle. This complexity also makes it challenging to identify a dataset that can capture the connectivity status of the IoT devices.

For building a feasible anomaly detection approach, the feasibility of different large datasets from two global providers is investigated. Specifically, the international roaming signaling behavior of IoT devices that serve different applications (e.g., connected cars, shipment containers, elevators, etc.) is investigated. Monitoring signaling traffic allows for a much less intrusive view than monitoring application traffic. In mobile networks, complex and diverse protocols let devices connect to the radio network first, and then establish the data communication channel over which application traffic is carried in an encrypted fashion. Visibility is thus much more limited compared to passive measurements in the traditional Internet.

With this use-case, ORIGAMI plans to make several contributions, as follows:

- Use-case specific **knowledge representation and reasoning approaches** will be investigated, which would allow us to build significant features that capture the connectivity status of the IoT devices. For this, a first check for IoT signaling traffic patterns that will allow us to build a generalizable network data representation will be done. These resulting representations (or engineered features) will be used towards the design of network intelligence (NI) solutions for the IoT devices.
- The tradeoff between different machine learning approaches, or the optimal design for deep learning **solutions for anomaly detection** will also be investigated. For being able to select a feasible solution, we will work towards building extensive ground truth information about anomalies that affected different types of IoT devices.

6.9.2 Involved Barriers and Architectural Elements

The aim of this use-case is to build effective knowledge representations that allow us to detect anomalies at the device level -- a very changeling problem within the cellular ecosystem. By working on this problem from the point of view of two separate global connectivity providers, a generalizable approach to network data representation is proposed, and thus break **Barrier #7**.

FR-KR-001	
Description	ORIGAMI should design network data representations that optimally capture expert knowledge, and that enable an entity to reason about the global cellular ecosystem
Version	Y1M3
Stage	Architectural
Risk	2/3
Risk Description	To enable this requirement, the aim is to collaborate with expert users within the global cellular ecosystem, whose input is paramount towards achieving an efficient knowledge representation

Table 51: FR-KR-001

FR-KR-002	
Description	ORIGAMI should propose meaningful network data representations for NI anomaly detection applications in the global cellular ecosystem
Version	Y1M3
Stage	Network Intelligence
Risk	1/3
Risk Description	The risk that ORIGAMI will fail to propose a meaningful data representation is very low, given that the consortium includes two expert teams providing their extensive operational knowledge to drive this effort

Table 52: FR-KR-002

FR-KR-003	
Description	NI for anomaly detection should leverage ground truth information on anomalies in the cellular system of interest.
Version	Y1M3
Stage	Network Intelligence
Risk	2/3
Risk Description	There is a mild risk that the ground truth datasets might not be adequate. To build a high-quality dataset of known anomalies, ORIGAMI relies on historical ticketing information from the operators of large systems.

Table 53: FR-KR-003

6.9.3 TARGET KPIS

The ORIGAMI solutions addressing Barrier #7 shall meet the following non-functional requirements (corresponding to KPIs K10 and K11, respectively):

NFR-KR-001	
Description	ORIGAMI's NI for anomaly detection should enable the identification of anomalies with a high performance, achieving a recall and sensitivity on the ground truth of above 85%
Version	Y1M3
Target KPIs	K10 (Anomaly detection recall and sensitivity): This requirement directly relates to the target KVI of developing a NI solution to detect anomalies with high accuracy
Risk	3/3
Risk Description	There is a significant risk that the performance of the anomaly detection approach might be low, given that anomalies change within the system. The aim is to study here topics related with normality drift detection and adapt our anomaly detection approaches accordingly

Table 54: NFR-KR-001

NFR-KR-001	
Description	ORIGAMI's NI for anomaly detection should enable a reduction in the OPEX of 30%
Version	Y1M3
Target KPIs	K11 (Open gains ($)): This requirement is directly related to efficiency gains realized through successful anomaly detection and the related reduction in signaling traffic, in turn reducing OPEX
Risk	2/3
Risk Description	There is a low risk that anomaly detection would not achieve this target. Focusing the effort towards detecting specific type of anomalies that are known to significantly impact operations would help achieve this target

Table 55: NFR-KR-001

6.10 Network Core traffic analysis and optimization (NCAM)

6.10.1 General Description

In the road towards 6G, significant challenges in signaling and control planes, chiefly due to the necessity for direct communication between NFs, leading to increased signaling traffic. This challenge intensifies with the connected devices and sessions across multiple NFs, resulting in elevated energy consumption, costs, and scalability constraints. Additionally, the monolithic design of NFs introduces resilience issues, creating potential single points of failure within the network architecture. Therefore, optimizing signaling management emerges as a crucial endeavor, given the substantial implications for the evolution of 6G networks.

Given that, towards the 6G era of communications, the current organization and implementation paradigm of the network core functionality is being revisited. Already, in 3GPP an enhanced SBA is foreseen where a Service Communication Proxy (SCP) is introduced to deliver the full capability of the Service Based Architecture [50]. The SCP is a service communication proxy used for indirect communication among NFs and other SCPs within the PLMN. The SCP can also interact with the Security Edge Protection Proxy (SEPP), i.e., the function that provides interconnection with other PLMNs. Practically, when a new NF is introduced, it needs only to be connected through the SCPs to the NRFs in the core, rather than directly to all other NFs that it may (or may not) use. More precisely, the SCP acts as hub for i) fundamental processes such as the registration and discovery of a NF to the NRF, and ii) service provisioning processes among registered NFs. This approach is the so called, signaling Model D in 3GPP, and it reflects the approach targeted from most of the service providers so far; while, already, Model D SCP implementations have recently emerged from all the major vendors and open-source network core projects (e.g., open5Gs).

The recently emerged concept of indirect signaling through a SCP reshapes the SBA and brings new, yet unexplored, capabilities for the CSPs. ORIGAMI focuses on those recently emerged changes in network core, and moves one step ahead, targeting the research study and implementation of this SCP-enabled SBA through a service mesh approach[1]. The concept of service meshes has been introduced recently in the cloud domain to efficiently solve critical problems, including scaling, interworking, and fault isolation. In this direction, a cloud native network core that is based on service mesh principles, brings a new and fresh approach toward 6G era of communications. The CSPs are expected to apply unified load balancing and overload control and operate a simple and extensible service-based scheme.

Recent report, and developments from the major vendors (Ericsson[2], Nokia[3], Huawei[4], Oracle[5]), converge on the statement that the SCP becomes an essential mechanism for network core scalability, and it also accelerates service delivery with reduced operational cost.

Regarding service mesh implementation of such an approach, there are multiple references that indicate the potential benefits, but only a few of them have implementation or tangible results, since the concept has only been introduced in 3GPP Release 18. In this direction, the research ambition is to investigate the potential benefits and drawbacks of a SCP-enabled 6G network core which is operated as a fully service-mesh.

6.10.2 Involved Barriers and Architectural Elements

As highlighted above ORIGAMI adopts the enhanced SBA foreseen in 3GPP (from Release 16 and onwards) where a Service Communication Proxy (SCP) is introduced. ORIGAMI's research ambition focuses on exploring the potential benefits and drawbacks of an SCP-enabled 6G network core operating as a fully service-mesh. In scope is the Network Core Analysis and Management (NCAM) using analysis tools such as ML techniques and graph theory. A key assumption for the study is that network functions (NF) are implemented and operate as microservices. The advantages of this approach have been evident since the early release of the SBA architecture, leading to ongoing

enhancements in network core signaling load by the research community. Recent studies on signaling load without SCP serve as benchmarks for a conceptual investigation.

Within the project framework, the opportunity to scale the network core based on statistical analysis of traffic loads among NFs will be examined. It will also be studied forecasting using ML techniques, moving beyond clustering methods, and capitalizing on the architectural changes brought about by SCP introduction. The approach involves: i) modelling network core signaling, considering topology and interface changes introduced by the enhanced SBA; ii) conducting statistical analysis of traffic flows at both NF-to-NF and global (network core) levels; and iii) providing decision-making schemes based on statistical analysis to inform the operation of the recently introduced Management Data Analytics Function (MDAF) in the network core.

Overall, the target is to address and remove Barrier #8, “High control-plane signaling overhead”, by investigating the potential benefits and drawbacks of an SCP-enabled 6G network core as a fully service-mesh, utilizing ML techniques and graph theory to model, analyse, and optimize traffic flows and network core signaling. This will contribute towards the architectural evolution foreseen in the ORIGAMI (Section 5), specifically, the global service-based architecture (GSBA).

FR-NCAM-001	
Description	Monitoring and telemetry of network core traffic flows should be available
Version	Y1M3
Stage	Architectural
Risk	1/3
Risk Description	There is a very low risk that the tools available in the literature require more time than expected to be configured

Table 56: FR-NCAM-001

FR-NCAM-002	
Description	The analysis of the traffic should lead to efficient deployment and restructuring of the network core functions without violating other performance metrics (e.g., delay, CPU consumption etc.)
Version	Y1M3
Stage	Network Intelligence
Risk	2/3
Risk Description	Applying changes to the network core structure and deployment schemes could lead to side effects not captured if the measurement only targets the optimization metric (e.g., add of delay due to buffering/overloading of the SCP)

Table 57: FR-NCAM-002

6.10.3 Target KPIs

The primary Key Performance Indicator (KPI) concerns the efficiency of the control plane (K12), particularly within the core of the 6G network. Specifically, implementing indirect communications among Network Functions (NFs) and Network Resource Functions (NRFs) in the network core can fully exploit the advantages of the service-based approach. The introduction of a Service Communication Proxy (SCP) can enhance the efficiency of the network core, improving scalability and accelerating service delivery while reducing operational costs. A key contributing factor to this efficiency is the decreased signaling overhead facilitated using SCPs. This streamlined signaling overhead is estimated to reduce by approximately 25% compared to a network core without SCP support, leading to potential reductions in operational costs, energy consumption, and environmental impact such as CO2 emissions and footprint size.

Further KPIs may be established in subsequent phases based on the initial Key Value Indicator (KVI) analysis outlined in subsection X. It is crucial to identify and define indicators that quantitatively and qualitatively describe the economic and environmental sustainability impacts of this solution and approach. Ultimately, SCP emerges as a vital mechanism not only for delivering efficient real-time 6G performance and scalability but also for expediting service delivery, reducing operational costs, optimizing link utilization, and enhancing End-to-End (E2E) service revenues, which may prompt the identification of additional KPIs.

NFR-NCAM-001	
Description	Monitoring and telemetry of network core provide adequate data for analysis.
Version	Y1M3
Target KPIs	K12 (Control-plane efficiency (%)): This requirement directly establishes a target on KPI K12 of 25% lower signaling overhead compared to network core without SCP
Risk	2/3
Risk Description	There is a mild risk as the traffic patterns of the signaling are unknown (since signaling load is not linearly related to the data plane traffic) and multiple scenarios should be checked in order to extract the adequate amount of data to enable a reliable analysis

Table 58: NFR-NCAM-001

NFR-NCAM-002	
Description	The NF topology and resource optimization increases the signaling efficiency by more than 10%
Version	Y1M3
Target KPIs	K12 (Control-plane efficiency (%)): This requirement directly establish a target on KPI K12 of 25% lower signaling overhead compared to network core without SCP
Risk	1/3
Risk Description	Given the current state of the art solutions there is very low risk the total gain to be below the threshold of 10%

Table 59: NFR-NCAM-002

7 KEY VALUES FRAMEWORK

A technology proves its societal value through its facilitation of Key Values (KVs), with Key Value Indicators (KVIs) serving as crucial metrics to demonstrate this worth. Following the methodology outlined in the relevant 6G-IA white paper [51], ORIGAMI analyzed it to define and pinpoint an initial set of Key Values and KVIs pertinent to the project's specific use cases.

The ORIGAMI project tackles the complex relationship scenario between the enabled business models and societal benefits, which in turn influences technology acceptance models by taking into consideration environmental impacts as well. The use cases identified and described in this document indeed have the potential to have socio-economic and environmental effects. ORIGAMI aims to establish a strong connection between technology and its positive impact on society, environment, and economy, hence the project analyzes values according to these three categories: **Societal, Environmental, and Economic** [51].

Among the project's objectives there is to develop an assessment framework that enables the evaluation of such use case dynamics for societal and environmental acceptance, specifically in the context of the 6G future roadmaps. Therefore, the concept of KV will be analyzed across the UCs in the project. This increased visibility not only benefits the industrial processes towards standardization, but also aids non-technical adopters, such as users in the public, commercial, or environmental sectors, in understanding the advantages.

According to the 6G-IA white paper, the utilization of KVIs in the development of 6G serves two main purposes: first, to demonstrate and validate that 6G can effectively address societal needs, and second, to steer technology development towards directions that yield value-driven benefits.

In the following, ORIGAMI exemplifies initial steps and questions to address to properly define the link between KVs, KVIs and KPIs.

Category	Definition
Key Value (KV)	What values are important to us? Which values hold the most significance?
Key Value Indicator (KVI)	What are the key indicators of these values? How can we measure or assess them?
Enablers	What factors contribute to the promotion of these values? What makes these values achievable? For instance, 6G features, low latency, reliability, etc.
KPIs	What are the technical impacts of these values? For example, coverage, capacity, energy efficiency, device access density, and localization accuracy.

Table 60: KVI Definition

With this definition, in ORIGAMI a list of Key Values that are relevant to the project's activities is defined as well as an initial mapping of the project's activities into this framework. The ORIGAMI project focuses on addressing KVI and Key KVs with an emphasis on sustainability. It identifies KVs relevant to ORIGAMI, including environmental sustainability, economical sustainability and innovation, and digital inclusion. These KVs can be either "use-case specific" or "architectural."

The project defines and conceptualizes KVIs, ensuring that each KV has at least one associated KVI, ideally more. The definition of a KVI includes a detailed description, the metric employed to assess performance or progress, the target value or goal, the percentage of improvement, the timeframe within which the target should be achieved, and whether the metric indicates an increase or decrease.

Specific KPIs are identified to help achieve these target values. For example, in the context of trustworthiness, which can be categorized as an architectural KV, the KVI is defined as architecture

resilience. The metric for this KVI is the ratio of computing resources successfully provisioned versus experienced failures, with a target value set at 99.999%. Several KPIs contribute to achieving this target. In the following table the KV list from [51] is presented as well as the introduction.

Key Value	Definition	Related Use cases
Environmental Sustainability	KV related to SDGs #6, 13, 14, 15	Unsustainable RAN virtualization ones directly target energy efficiency as KPI to demonstrate their successful achievement.
Societal Sustainability	KV related to SDGs #1, 2, 3, 4, 5, 7, 11, 16	Indirectly, all Use cases are related to this value, as the introduction of 6G technologies empowered by ORIGAMI will allow a sustainable society.
Economic Sustainability and Innovation	KV related to SDGs #8, 9, 10, 12	The new use cases that introduce the Global Operator Model will introduce new business models that can improve the economic sustainability.
Democracy	KV related to SDGs #5, 10, 16, as well as linked to securing "Political equality in a pluralistic, liberal society" and to "Protecting EU democracy from external interference"	Indirectly, ORIGAMI use cases and the new architectural innovations are linked to this value. Indeed, 6G technologies will not only support more advanced e-governance solutions (making it easier for citizens to interact with their governments and access services online) but will also facilitate more transparent governance by enabling real-time data collection and dissemination.
Cultural Connection	KV related to SDG #10, 11, 16, linked to fostering production and access to cultural products (e.g., art - movies, music, literature -, history, trends/new culture domains, e.g., games)	Indirectly, ORIGAMI solutions can foster a wider adoption of the 6G mobile network which can be used to improve the cultural connection.
Knowledge	KV related to SDGs #1, 4, 5, 8, 10, 17 especially referring to access to quality education systems and equal educational opportunities	Also in this case, a wider adoption of the 6G mobile network can improve the quality of the education system.
Privacy and Confidentiality	KV related to SDG #16; as privacy is an institutionally protected value related to the claim of individuals or institutions to decide on if, when, how, and to what extent information about them is communicated to others. and at the same time "the appropriate use of data relating an individual to a context"	Use cases related to Barriers 5,6, and 7 not only allow to introduce new business models, but they also allow to improve the operation of the network by enforcing privacy preserving operation of the networks through zero trust exposure.
Simplified Life	KV reflecting UN SDGs #3 (primarily), #9, #11	ORIGAMI solution will improve the overall 6G adoption, thus enabling a simplified life as envisioned by those goals.
Digital Inclusion	KV reflecting partly UN SDG #10, in people being part of the digital world.	Besides improving the performance of the Network, the improved

		sustainability of the network will increase the inclusion opportunities for the network.
Personal Freedom	KV referring to a positive freedom of an individual to control and impact their own life	Not addressed by ORIGAMI.
Personal Health and Protection from Harm	KV related to SDGs #2, 3, 6, 13	Not addressed by ORIGAMI.

Table 61: Key Values

In the following, an initial analysis is conducted for possible indicators associated with the use case described in Section 6.10. Understanding the challenge posed by the high volume of control plane signaling and its resolution using a Service Communication Proxy (SCP), it becomes imperative to assess the real-world impact of this use case and identify the areas poised for improvement and benefit from this solution.

Of particular importance is the examination of how this solution influences both environmental and energy factors. Given the escalating environmental impact of human decisions on our planet, prioritizing environmental parameters is paramount. Hence, the primary objective of the proposed study is to minimize signal exchanges among Network Functions (NFs), since fewer signals translate to reduced power and energy consumption across the network.

Economic sustainability and operational costs are another critical factor, particularly for Mobile Network Operators (MNOs). Employing an SCP for NF communication means that servers initially dedicated to NFs will experience reduced usage, thereby lowering energy consumption and mitigating potential damage incurred during frequent, high-capacity utilization. This outcome correlates with the SCP's caching functionality, which stores requested resources, delivering them directly to clients without server involvement, thereby reducing network latency and enhancing responsiveness. This caching mechanism not only enhances user experience but also alleviates server load, enabling efficient handling of increased request volumes.

Another vital consideration pertains to privacy and confidentiality. It is highly beneficial to incorporate mechanisms and security policies into proposed solutions and technology components. These policies play a crucial role in safeguarding the confidentiality of user-sensitive data and ensuring security. The SCP facilitates client authentication, regulates access to designated resources, and encrypts data transmission, thereby fortifying protection against unauthorized access and interception of sensitive information. Moreover, SCP seamlessly integrates with other security measures such as intrusion detection systems (IDS) and data loss prevention (DLP) systems.

Equally important is digital inclusion, a goal underscored repeatedly by the EU in the realm of technology. In scenarios of high demand, SCP efficiently distributes incoming requests across multiple servers to maintain resource equilibrium, thereby averting server overload. This intelligent routing enhances scalability and reliability, ensuring uninterrupted service even in the event of server failures by automatically redirecting traffic to alternative servers, thus minimizing downtime. By incorporating this mechanism for load-balancing network traffic, more users can access and utilize a reliable network, fostering their inclusion in technological advancements and societal progress.

8 CONCLUSION AND NEXT STEPS

This deliverable has meticulously analyzed the barriers identified by ORIGAMI and expanded upon them by incorporating potential challenges from interactions with other related 6G projects. Through this process, a comprehensive set of KPIs and KVIs that ORIGAMI's novel architectural solutions must support has been also established.

By reviewing eight barriers, specific Key Performance Indicator (KPI) requirements that will be facilitated through the implementation of ORIGAMI's architectural models have also been defined. The outcome includes detailed requirements specifications and guidelines, encompassing both expected functionalities and objective technical indicators. Additionally, design roadmaps that will guide the development and implementation of these models have been outlined.

The work conducted in this deliverable ensures that ORIGAMI's architectural solutions are robust, efficient, and capable of meeting the project's evolving needs. The structured approach taken to identify and address key barriers and requirements lays a solid foundation for the successful realization of ORIGAMI's objectives. Moving forward, these insights and guidelines will be instrumental in driving the project towards achieving its goals, ensuring that the architectural models developed are both innovative and practical in their application.

D2.1 will serve as architectural input for WP3 and WP4. Work under WP2 will continue to produce D2.2 and D2.3. In D2.2, ORIGAMI will analyze potential business models using both analytical and experimental approaches to quantify the benefits of joint decision-making on 6G KPIs, particularly for issues currently managed across different administrative domains. Technical advancements, such as optimizing the O-RAN bus, will facilitate new virtualization solutions like fine-grained cloudification. These advancements will necessitate regulatory policies to prevent inefficiencies and exploitation.

9 REFERENCES

[1] TRL definition https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf
[2] 3GPP, Technical Specification (TS) 23.288, March 2023, version 18.1.0.
[3] M. A. Garcia-Martin, M. Gramaglia and P. Serrano, "Network Automation and Data Analytics in 3GPP 5G Systems," in IEEE Network, doi: 10.1109/MNET.2023.3321524.
[4] https://www.3gpp.org/3gpp-groups/service-system-aspects-sa/sa-wg6
[5] TS23.289
[6] 3GPP TS 23.501
[7] 3GPP, TS 23.289
[8] Andres Garcia-Saavedra and Xavier Costa-Pérez. 2021. "O-RAN: Disrupting the Virtualized RAN Ecosystem". IEEE Communications Standards Magazine 5, 4 (2021), 96–103.
[9] 3rd Generation Partnership Project (3GPP). 2022. 3GPP TR 38.913; Technical Specification Group Radio Access Network; Study on Scenarios and Requirements for Next Generation Access Technologies
[10] O-RAN Alliance. 2022. Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN (O-RAN.WG6.CADS-v04.00). Technical Report.
[11] https://rethinkresearch.biz/articles/nokia-says-its-fpga-strategy-hit-5g-margins-but-other-factors-are-at-work-too/
[12] https://rethinkresearch.biz/articles/is-general-purpose-silicon-too-slow-and-expensive-for-the-vran/
[13] https://www.lightreading.com/open-ran/mavenir-unhappy-about-chip-prices-for-smaller-open-ran-players/d/d-id/781327
[14] Diksha Moolchandani, Anshul Kumar, and Smruti R. Sarangi. 2021. "Accelerating CNN Inference on ASICs: A Survey". Journal of Systems Architecture 113 (2021), 101887.
[15] T. Salem, G. Iosifidis, G. Neglia, "Enabling long-term Fairness in Dynamic Resource Allocation", ACM Sigmetrics 2022
[16] Polese, M., Bonati, L., D'Oro, S., Basagni, S., & Melodia, T. "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges" IEEE Communications Surveys & Tutorials.2023
[17] RAMEZANPOUR, K; JAGANNATH, J "Intelligent zero trust architecture for 5G/6G networks: Principles, challenges, and the role of machine learning in the context of O-RAN". Computer Networks 2022
[18] H. Lee, J. Cha, D. Kwon, M. Jeong and I. Park "Hosting AI/ML Workflows on O-RAN RIC Platform" p1-6 2020
[19] https://www.3gpp.org/technologies/5g-system-overview
[20] https://www.3gpp.org/technologies/slice-management
[21] 23.288, 3GPP Technical Specification (TS) Architecture enhancements for 5G System (5GS) to support network data analytics services, 3rd Generation Partnership Project (3GPP), December 2021
[22] al., F. Z. Yousaf et "Network slicing with flexible mobility and QoS/QoE support for 5G Networks" 2017
[23] Dario Bega, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez.DeepCog: "Cognitive Network Management in Sliced 5G Networks with Deep Learning". In IEEE INFOCOM 2019

[24]28.530, 3GPP TS Technical Specification Group Services and System Aspects, Management and orchestration, Concepts, use cases and requirements, Release 17, V17.4.0

[25]23.501, 3GPP TS Technical Specification Group Services and System Aspects, System architecture for the 5G System (5GS), Stage 2, Release 18, V18.0.0 2022

[26]Ayala-Romero, Jose A., et al. "Mean-Field Multi-Agent Contextual Bandit for Energy-Efficient Resource Allocation in vRANs." *IEEE International Conference on Computer Communications. 2024.*

[27]M. Gramaglia, M. Camelo, L. Fuentes, J. Ballesteros, G. Baldoni, L. Cominardi, A. Garcia-Saavedra, and M. Fiore, "Network Intelligence for Virtualized RAN Orchestration: The DAEMON Approach," in 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2022, pp. 482–487

[28]M. Gramaglia, M. Camelo, L. Fuentes, J. Ballesteros, G. Baldoni, L. Cominardi, A. Garcia-Saavedra, and M. Fiore, "Network Intelligence for Virtualized RAN Orchestration: The DAEMON Approach," in 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2022, pp. 482–487

[29]Jian Ding, Rahman Doost-Mohammady, Anuj Kalia, and Lin Zhong. 2020. "Agora: Real-time massive MIMO baseband processing in software". In Proceedings of ACM CoNEXT '20. ACM.

[30]Junzhi Gong, Anuj Kalia, and Minlan Yu. 2023. Scalable DistributedMassive MIMO Baseband Processing. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23). 405–417

[31]Heavy Reading. 2022. "5G Transport: A 2021 Heavy Reading Survey. White Paper" (Feb. 2022).

[32]NTT Docomo. 2016. Base-station Equipment with the Aim of Introducing 3.5-GHz band TD-LTE. NTT Docomo Technical Journal (2016).

[33]Hazan, Elad. "Introduction to online convex optimization." *Foundations and Trends® in Optimization 2.3-4 (2016): 157-325.*

[34]Mohri, Mehryar, and Scott Yang. "Accelerating online convex optimization via adaptive prediction." *Artificial Intelligence and Statistics. PMLR, 2016.*

[35]A.T.-J. Akem, B. Bütün, M. Gucciardo, M. Fiore, Henna: "Hierarchical Machine Learning Inference in Programmable Switches", NativeNI 2022, Rome, Italy, Dec 2022.

[36]A.T.-J. Akem, M. Gucciardo, M. Fiore, "Flowrest: Practical Flow-Level Inference in Programmable Switches with Random Forests" IEEE INFOCOM 2023, New York, USA, May 2023.

[37]A.T.-J. Akem, B. Bütün, M. Gucciardo, M. Fiore, "Jewel: Resource-Efficient Joint Packet and Flow Level Inference in Programmable Switches" IEEE INFOCOM 2024, Vancouver, Canada, May 2024.

[38]Z.Zhao, Z. Li, Z. Song, F. Zhang, B. Chen, "RIDS: Towards Advanced IDS via RNN Model and Programmable Switches Co-Designed Approaches", IEEE INFOCOM 2024, Vancouver, Canada, May 2024.

[39]Alcalá-Marín, Sergi, Aravindh Raman, Weili Wu, Andra Lutu, Marcelo Bagnulo, Ozgu Alay, and Fabián Bustamante. "Global mobile network aggregators: Taxonomy, roaming performance and optimization." In Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, pp. 183-195. 2022.

[40]Luo, Zhihong, Silvery Fu, Mark Theis, Shaddi Hasan, Sylvia Ratnasamy, and Scott Shenker. "Democratizing cellular access with CellBricks." In Proceedings of the 2021 ACM SIGCOMM 2021 Conference, pp. 626-640. 2021.

[41]Schmitt, Paul, and Barath Raghavan. "Pretty good phone privacy." In 30th USENIX Security Symposium (USENIX Security 21), pp. 1737-1754. 2021.

[42]Mandalari, A. M., Lutu, A., Custura, A., Khatouni, A. S., Alay, Ö., Bagnulo, M., ... & Fairhurst, G. (2021). "Measuring roaming in Europe: Infrastructure and implications on users' QoE". IEEE Transactions on Mobile Computing, 21(10), 3687-3699.

[43]A. Garcia-Saavedra and X. Costa-Perez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," IEEE Communications Standards Magazine, vol. 5, no. 4, pp. 96–103, 2021.

[44]"5G-ACIA, 5G Alliance for Connected Industries , a Working Party of ZVEI (German Electrical and Electronic Manufacturers' Association)." https://www.5g-acia.org/
[45]Linux Foundation, "Camara Project." https://camaraproject.org/
[46]M. Milani, D. Bega, M. Gramaglia, and C. Mannweiler, "Optimizing predictive analytics in 5g networks through zero-trust operator- customer cooperation," in 2023 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), pp. 123– 128, 2023
[47]3GPP, "Study on network slice management capability exposure," Technical Report (TR) 28.824, 3rd Generation Partnership Project (3GPP), July 2023. Version 18.0.1
[48]3GPP, "Architecture enhancements for 5G System (5GS) to support network data analytics services," Technical Specification (TS) 23.288, 3rd Generation Partnership Project (3GPP), March 2023. Version 18.1.0.
[49]A. Collet, A. Bazco-Nogueras, A. Banchs, M. Fiore, *et al.*, "Au tomanager: a meta-learning model for network management from intertwined forecasts," in *IEEE International Conference on Computer Communications*, 2023
[50]3GPP TS 23.502, Procedures for the 5G System, Stage 2, Release 18, V18.0.0 (2022-12).
[51]https://5g-ppp.eu/wp-content/uploads/2022/05/What-societal-values-will-6G-address-White-Paper-v1.0-final.pdf