Partners: NFI, EAB, AAU, AAP, ATO, EBY, ICC, LTU, TNO, NXW, NGE, OPL, TUD, TIM, UBW, UC3, WIN, TUK, SON

HEXA-X-II D3.5 Deliverable

Final architectural framework and analysis

Hexa-X-II

hexa-x-ii.eu

Purpose of this D3.5 slide-set

- The purpose of this D3.5 slide-set is to be an introduction to the D3.5 word deliverable

- The scope of this deliverable is the same as D3.5 with only high-level descriptions of the concepts and enablers.

Deliverable properties

Document number	**D3.5**
Document title	Final architectural framework and analysis
Document editor	NFI; EAB; WIN; EBY; APP; UC3
Authors	NFI: Ozgur Akgul, Hamed Hellaoui, Tero Lotjonen, David Navratil EAB: Mårten Ericson, Selim Ickin, Paola Iovanna, Stefan Wänstedt AAU: Pere Garau Burguera, Olav Tirkkonen APP: Panagiotis Botsinis, Milan Zivkovic, Sameh Eldessoki, Alperen Gundogan ATO: Ignacio Labrador Pavón, Enrique Lluesma Martí EBY: Merve Saimler ICC: Panagiotis Charatsaris, Maria Diamanti, Grigorios Kakkavas, Symeon Papavasliou LTU: Jaap van de Beek, Amirreza Moradi, Payal Gupta NXW: Giacomo Bernini, Erin Seder, Michael De Angelis NGE: Apostolos Kousaridas, Hasanin Harkous, Bahare M. Khorsandi, Ece Ozturk, Mohammad Soliman, Arled Papa, Umur Karabulut, Panagiotis Spapis, Gerald Kunzmann TIM: Antonio Varvara UBW: Luis Pedro Santos UC3: Antonio de la Oliva WIN: Sokratis Barmpounakis, Vasilis Tsekenis, Panagiotis Demestichas TUK: Mohammad Asif Habibi, Hans D. Schotten SON: Torgny Palenius, Vivek Sharma TNO: Toni Dimitrovski, Nassima Toumi OPL: Halina Tarasiuk, Marcin Ziółkowski, Karol Kuczyński, Jan Palimąka, Janusz Pieczerak

Table of Contents

Short executive summary

The Hexa-X-II project continues to develop a forward-looking 6G architecture that addresses the evolving needs of next-generation networks. The final architectural framework focus on three core areas:

1. **Novel Services**: The architecture enables innovative services that go beyond traditional communication. These include advanced data-driven capabilities that will facilitate new applications in various domains. The architecture is designed to support flexible, service-oriented deployments that can scale across different use cases.
2. **Flexible Topologies**: Flexible topologies ensure that 6G can dynamically adjust to a wide range of environments, offering reliable coverage in urban, rural, and remote areas. The architecture provides adaptable network configurations that allow for seamless integration of both terrestrial and non-terrestrial networks (NTNs). This approach supports the vision of a fully connected world, regardless of location.
3. **Transformed Architecture for 6G**: At the core of the framework is a modular, cloud-native architecture designed for 6G. This transformed architecture optimizes resource allocation, signaling, and network management. It supports distributed processing and enhances scalability, ensuring that the network can evolve to meet future demands while maintaining high performance and energy efficiency.

Novel Services

- DataOps
- MLOps
- AIaaS
- Integrated Sensing and Communication (ISAC)
- Compute offloading

Flexible Topologies

- Network of Networks
- Multi-Connectivity
- Context-aware management

Transformed Architecture for 6G

- Architectural aspects of Migration
- Design of a module
- Interactions between modules
- Modularization examples
- Control of a slice
- Multi-domain/Multi-cloud federation
- Orchestration of the cloud continuum
- Application-layer BCS optimisation
- Cloud transformation in 6G quantum architecture

Introduction

Introduction

The Hexa-X-II project is a flagship initiative bringing together key stakeholders in Europe for 6G research, continuing the Hexa-X project work. The stakeholders include key industry players in telecom and major research institutes; a combination capable of introducing new value chains for future connectivity solutions.

The Hexa-X-II project comprises several work packages (WPs) that span over important parts of the 6G ecosystem. In this report results from WP3 are presented, which deals with the 6G architecture design.

The overarching objective of WP3 is to develop a 6G architecture framework and innovative enablers for a data driven architecture capable of powering new services, such as beyond communications services, a modular cloud-native network for improved signalling as well as new access and flexible topologies for improved reliability.

This is the final deliverable from WP3, D3.5 "Final Architectural Framework and Analysis".

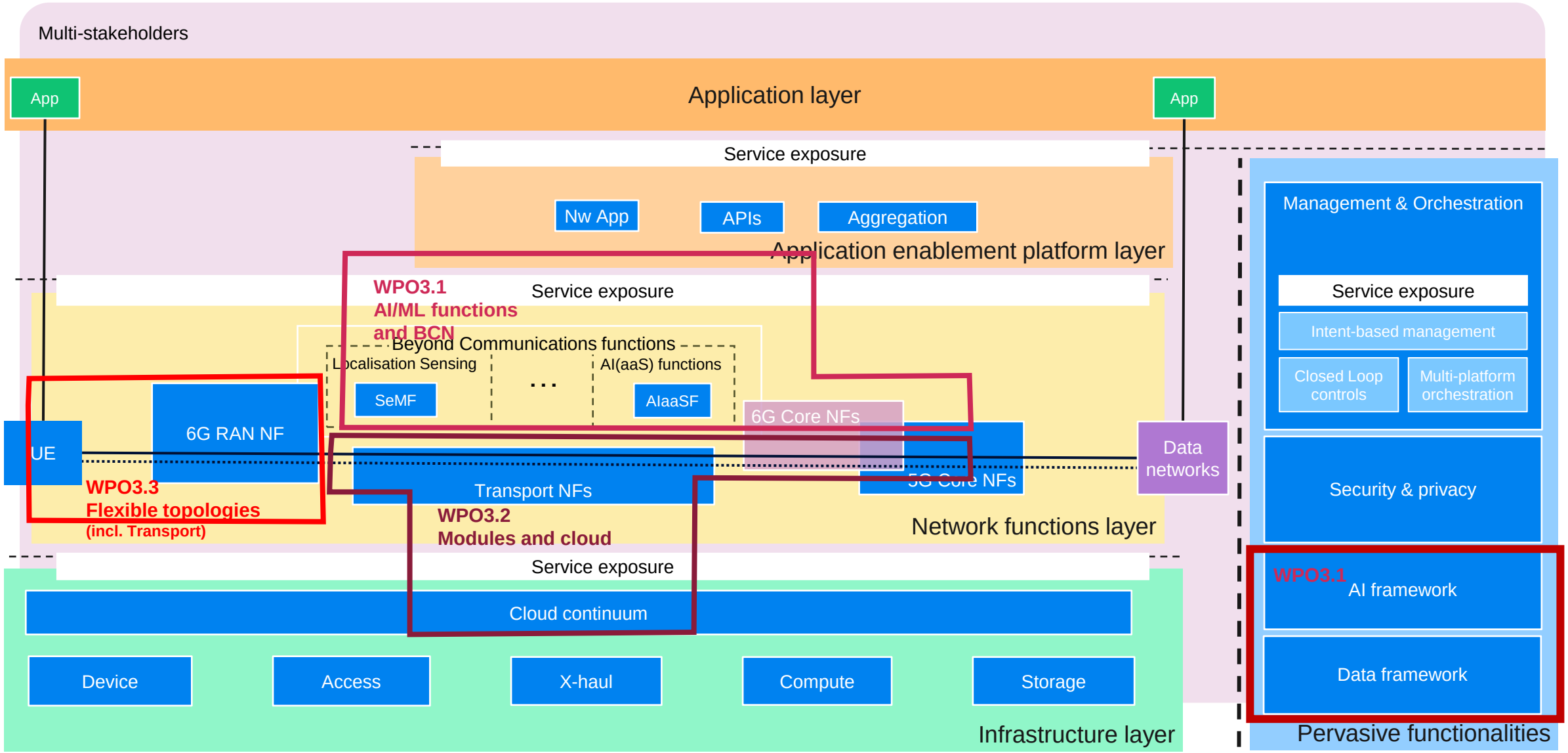
WP3 objectives

WP3 objectives

The objectives of WP3 are presented in the table below.

Objective	Objective description	Section
WPO3.1: 6G architecture for AI and beyond communications	**WPO3.1**: Develop and analyse a 6G architecture framework and new innovative enablers for the beyond communications and data driven architecture, identify requirements a data-driven architecture will have on protocols, interfaces, data, and network nodes.	Novel Services
WPO3.2: Combine the cloud technology for a modular, scalable and extendable architecture	**WPO3.2**: Define and analyse solutions that combine cloud technology flexibility with distributed processing nodes into self-contained modules with minimum dependency that can be used to extend and scale the network deployments in stepwise manner.	Transformed architecture for 6G
WPO3.3: Architecture for flexible topologies	**WPO3.3**: Develop and analyse new access for flexible topologies and local communications, including different types of multi-connectivity, node roles and node coordination, as well as design control and management solutions for programmable and context-aware transport.	Flexible topologies

WP3 objectives mapping on system blueprint



WP3 enablers – Novel Services

DataOps

DataOps: Descriptions

- **DataOps concept in general**
  - DataOps is an approach that integrates practices, processes, and technologies to streamline data management by blending a process-driven view of data with automation and agile software development principles.
- **Kubernetes-based failure detection data sourcing for 5G Core**
  - Reliability requirements in considered use cases constitute the necessity for failure detection and prediction
  - Limited data available in open-source 5G Core implementations creates the need to utilize new Monitoring and Observability tools
  - Data collected for failure detection is stored in the Prometheus time series database
  - Exposure of collected data with the Prometheus API, accessible via PromQL language

- **Privacy-preserving data collection and learning description**
  - DataOps assumes methods and protocols for ensuring privacy-preserving UE data collection and analysis.
  - This allows NW to learn about the statistics of aggregated UE data, without revealing individual UE private data.
- **Sensing data exposure**
  - Exposure possibilities for sensing; both the control plane (CP) and user/data planes (UP/DP) can be leveraged (see ISAC slides for more information)
  - Data can be exposed both via the UP/DP and via the CP.
  - UP/DP exposure can be suitable when the data size is large while CP can be used for smaller outputs (such as Booleans, location etc.)
  - The 6G exposure solution may rely upon existing 5G network exposure function (NEF) framework, including Common API Framework (CAPIF) and/or Service Enabler Architecture Layer (SEAL).
  - The choice of application programming interfaces (APIs) and other interfaces may depend on use case.

Protocol and APIs Privacy-preserving data collection and learning

- To enable the privacy-preserving data collection and learning from potentially private user data at the network, we have previously proposed the Prio-based aggregation architecture [HEX224-D33].
- For 6G networks, this architecture can be mapped to the 5G framework for data collection and data analytics, initially introduced in Rel-17 [TS 23.288].
- An example of placement of different Prio entities in the cellular network and possible architectural and signaling adaptations are shown in Figure 1.
  - Leader and helper Aggregators are implemented as application functions (AFs), owned respectively by NW and device/app vendors.
  - Collector, implemented in NWDAF produces aggregated statistics
  - NF (e.g., network coordination function)consumes aggregated statistics
- In this system, NW learns distribution/aggregation of the data collected from multiple users, without revealing individual UE private data.

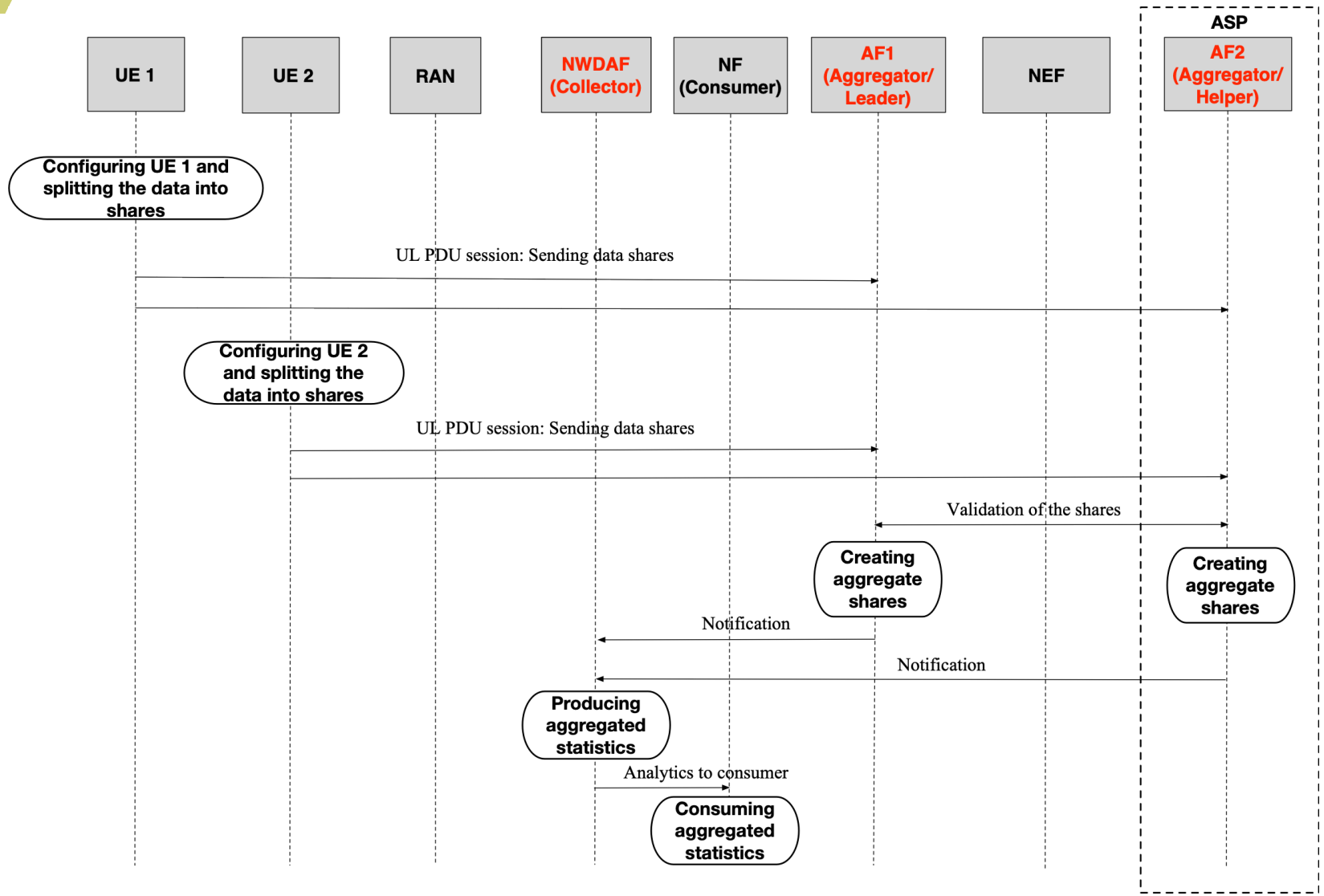


Fig 1: Privacy-preserving data aggregation procedure from UE: an example.

Data exposure, Data collection, DataOps - Description
Kubernetes-based failure detection data sourcing for 5G Core

- **Objectives**
  - To evaluate 5G Core network ability to perform a procedure based on data provided by the 5G Core Observability Platform (called 5GCOP here), Fig.2
  - Collect, process and expose data sourced by the Observability and Monitoring tools
- **Motivation**
  - To provide a minimum set of metrics for a 5G procedure*, based on which we can detect or predict a failure (risk of losing service continuity)

5GCOP architecture and its mapping to the 6G E2E system blueprint

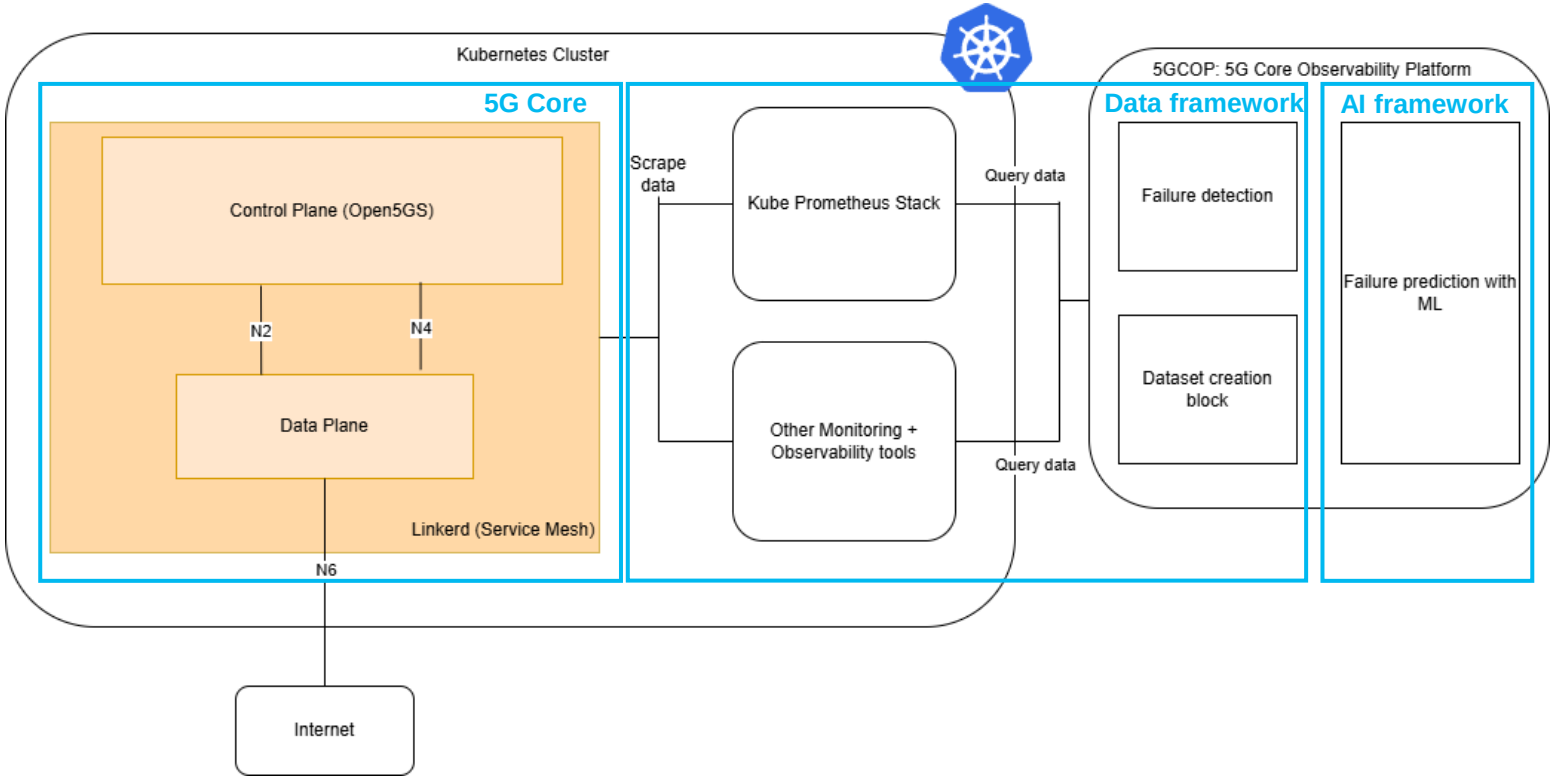


Fig. 2: 5GCOP (5G Core Observability Platform) architecture

* TS 123 502 - V15.2.0 - 5G; Procedures for the 5G System (3GPP TS 23.502 version 15.2.0 Release 15)

Data exposure, Data collection, DataOps - Evaluation Kubernetes-based failure detection data sourcing for 5G Core

- As said earlier, the purpose with the failure detection study is to predict failures in 5G Core network by the 5G Core Observability Platform. Therefore, several sources of data are available (Fig. 3) for 5GCOP, such as:
  - Logs
  - cAdvisor - 96 metrics
  - Kube-state-metrics – 253 metrics
  - Open5GS NF metrics – 44 metrics
  - Prometheus node exporter - 49 tools
  - Linkerd service mesh - 27 metrics
  - Kubernetes API
- KPI: reliability
  - To keep service continuity in cloud-native environment (to guarantee high reliability)
  - Use Cases for this study can be: #1 From robots to cobots, #3 Massive Twinning, #4 Immersive telepresence for enhanced interactions

```
root@s5-b:/home/mziolkowski/distributed-monitoring-demo/Open5GS_metrics# kubectl get po -A
NAMESPACE           NAME                                        READY   STATUS    RESTARTS      AGE
cert-manager        cert-manager-cainjector-5fd6444f95-6x9r2    1/1     Running   0             72d
cert-manager        cert-manager-d894bbbd4-pdbjx                1/1     Running   0             72d
cert-manager        cert-manager-webhook-869674f96f-4zkmf       1/1     Running   0             72d
chaos-mesh          chaos-controller-manager-5cf4f9c75d-dnrkw   1/1     Running   0             126d
chaos-mesh          chaos-controller-manager-5cf4f9c75d-jnck9   1/1     Running   0             126d
chaos-mesh          chaos-controller-manager-5cf4f9c75d-vzkhw   1/1     Running   0             126d
chaos-mesh          chaos-daemon-55g7h                          1/1     Running   0             126d
chaos-mesh          chaos-dashboard-85dbd44999-8m7mw            1/1     Running   0             126d
chaos-mesh          chaos-dns-server-6d777f574f-lgn6h           1/1     Running   0             126d
kube-flannel        kube-flannel-ds-xnjhr                       1/1     Running   0             223d
kube-system         coredns-5dd5756b68-b6zwg                    1/1     Running   0             223d
kube-system         coredns-5dd5756b68-k2ck7                    1/1     Running   0             223d
kube-system         etcd-s5-b                                   1/1     Running   1             223d
kube-system         kube-apiserver-s5-b                         1/1     Running   1             223d
kube-system         kube-controller-manager-s5-b                1/1     Running   0             223d
kube-system         kube-multus-ds-zsfn5                        1/1     Running   0             217d
kube-system         kube-proxy-bklss                            1/1     Running   0             223d
kube-system         kube-scheduler-s5-b                         1/1     Running   1             223d
kube-system         metrics-server-84989b68d9-tpfk7             1/1     Running   0             135d
linkerd-viz         metrics-api-675fb6bddc-87gnh                2/2     Running   0             44d
linkerd-viz         prometheus-5dcc5d9ff8-8227s                 2/2     Running   0             44d
linkerd-viz         tap-785d857fd-zw7bf                         2/2     Running   0             44d
linkerd-viz         tap-injector-6d47d5cc7-6mkdz                2/2     Running   0             44d
linkerd-viz         web-5b8d97994c-hckvf                        2/2     Running   0             44d
linkerd             linkerd-destination-66b849f5db-sqbrc        4/4     Running   0             69d
linkerd             linkerd-identity-97f5b5499-s8wqc            2/2     Running   0             69d
linkerd             linkerd-proxy-injector-7b4fcd46d5-9fjqb     2/2     Running   0             69d
local-path-storage  local-path-provisioner-6d9d9b57c9-8sn9c     1/1     Running   0             208d
open5gs             open5gs-amf-b77c54f6c-l6w8s                 2/2     Running   0             2m11s
open5gs             open5gs-ausf-6d855db87c-lj48g               2/2     Running   0             2m11s
open5gs             open5gs-bsf-755d967879-66jsw                2/2     Running   0             2m11s
open5gs             open5gs-mongodb-7d98bfc976-pk2lp            2/2     Running   0             2m11s
open5gs             open5gs-nrf-7c4d8d6855-vg7xh                2/2     Running   0             2m11s
open5gs             open5gs-nssf-6cbf5cfcd5-tlm5r               2/2     Running   0             2m11s
open5gs             open5gs-pcf-7fdbcd6d9c-v7c4q                2/2     Running   2 (2m1s ago)  2m11s
open5gs             open5gs-smf-659577598-8l2ln                 2/2     Running   0             2m11s
open5gs             open5gs-udm-8554b8d44c-4zb72                2/2     Running   0             2m11s
open5gs             open5gs-udr-98c95d864-cbr5c                 2/2     Running   2 (2m1s ago)  2m11s
open5gs             open5gs-upf-64dbff746c-twjmt                2/2     Running   0             2m11s
open5gs             open5gs-webui-7dcf577987-mmrdl              2/2     Running   0             2m11s
```

Fig. 3: Experimental setup Kubernetes cluster pods

Data exposure, Data collection, DataOps - Protocol and APIs
Kubernetes-based failure detection data sourcing for 5G Core

Data sourcing and failure detection sequence diagrams

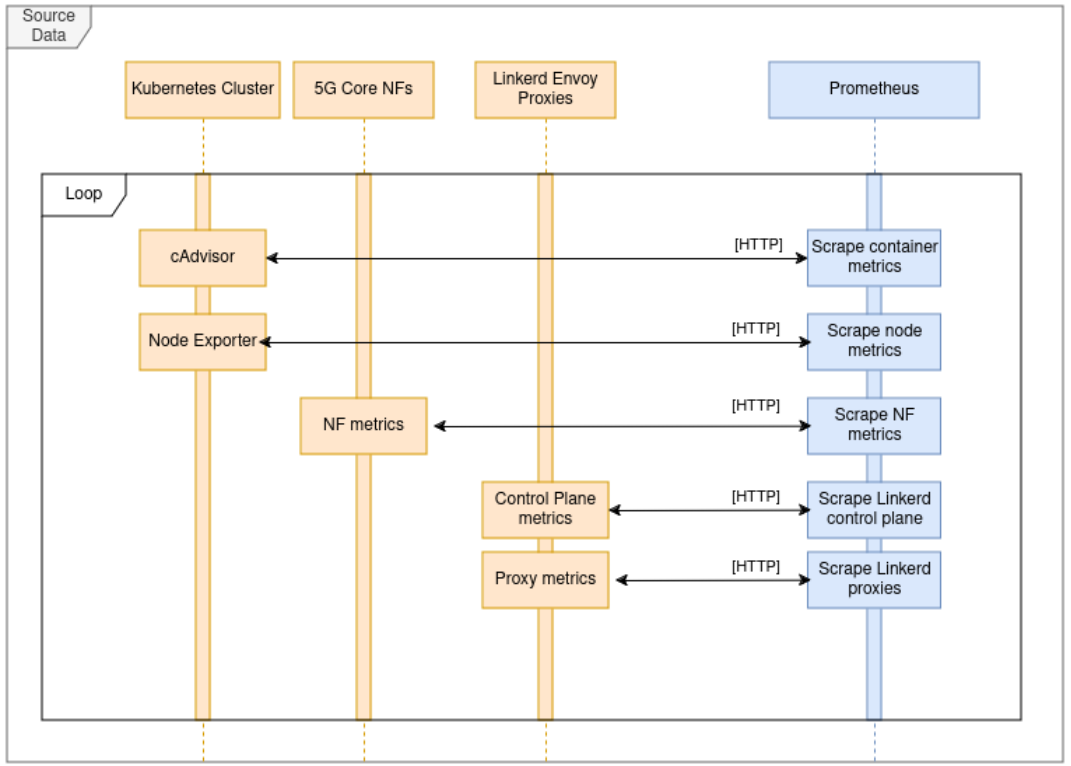


Fig. 4: Data sourcing sequence diagram

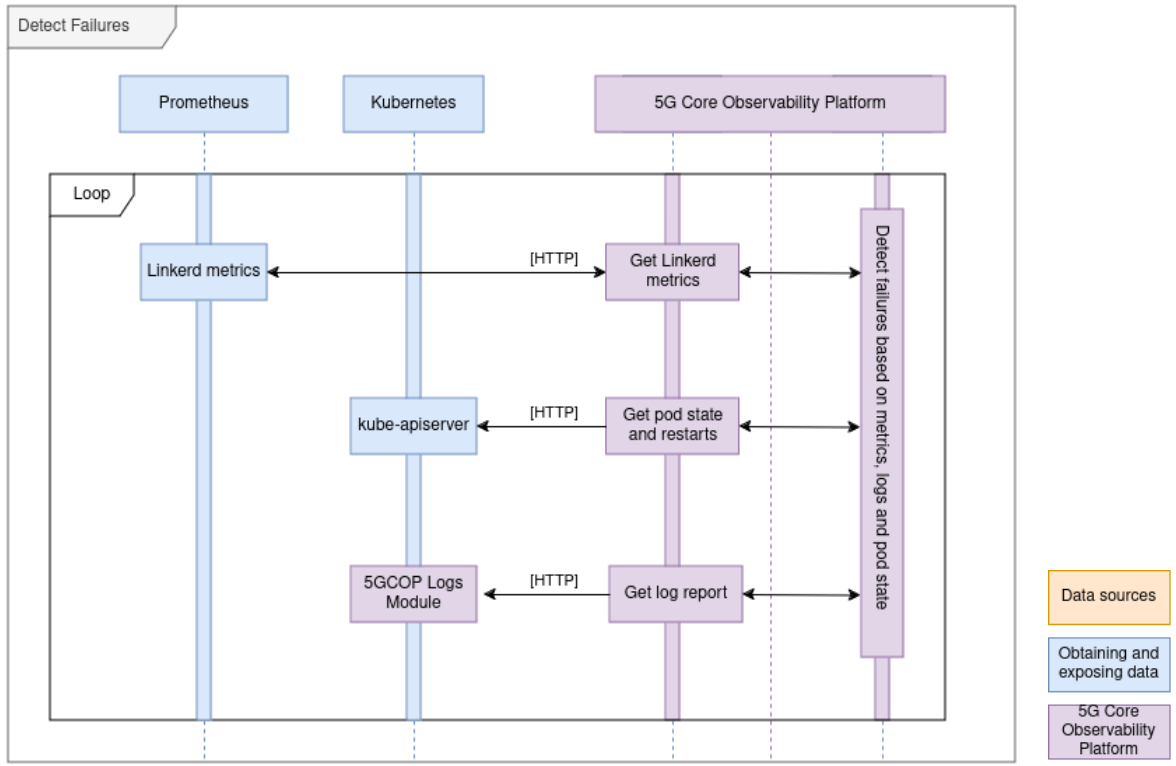


Fig. 5: Failure detection sequence diagram

Fig. 4 shows the integration of existing tools using their interfaces, Fig. 5 contains the elements implemented in 5GCOP that implement DataOps for failure detection (purple color).

Data exposure, Data collection, DataOps - Implementation
Kubernetes-based failure detection data sourcing for 5G Core

Exemplary failure detection results

Table 1: Failure detection by 5GCOP

Test case series	Experiment parameters	Information source for detection	Accuracy	Accuracy [%]
Latency introduced for a single network function	NetworkChaos* Delay 50ms	Service Mesh	70/70	100%
Jitter and Latency introduced for a single network function	NetworkChaos* Delay 50ms Jitter 20ms	Service Mesh	76/80	95%
Packet loss introduced for a single network function	NetworkChaos* Loss 30%	Service Mesh and kube-apiserver	54/70	77.14%
Pod failures and restarts	PodKill	kube-apiserver	13/13	100%
Error in network configuration (Error, Warning)	N/A	Logi	4/4	100%
Accuracy		**217/237**		**91.56%**

Table 2: Failure detection for Delay experiment

Network Function	Detected	Undetected	Accuracy [%]	NFs experiencing increased latency
AMF	16	4	80%	SMF
PCF	10	0	100%	SMF, PCF
AUSF	10	0	100%	SMF, AUSF, PCF
SMF	10	0	100%	SMF
UDM	10	0	100%	SMF, AUSF, UDM
UDR	10	0	100%	SMF, UDR, UDM, AUSF, PCF
UPF	10	0	100%	SMF
Accuracy			**95%**	

* NetworkChaos - Fault type name from Chaos Engineering tool, Chaos Mesh

- **Conclusion:**
  - Introducing a Service Mesh in a data-scarce environment has been proven to be a valuable data source, allowing us to achieve detection accuracy level of over 90%. At the same time, the increase in Procedure Completion Time created by the Service Mesh sidecar containers warrants a need to evaluate existing sidecarless Service Mesh tools as a replacement for Linkerd.
  - Three data sets (representing Idle 5G, 5G during the registration procedure, Failure in 5G during a procedure)** were created for Machine Learning models for training purposes for the Failure Prediction Model.

** Hexa-X-II MLOps Failure Prediction Datasets, Data sets

Data exposure, Data collection, DataOps
Key take-away

DataOps for failure detection in 5G Core

- Kubernetes-based failure detection
  - The search for early indicators of a possible failure in the 5G Core network needs to minimize the impact of introduced tools on the resource usage and performance of the 5G Core
- Take-aways from the study
  - While data from Service Mesh allows to achieve high failure detection accuracy, the additional resource cost and scalability concerns warrant the need for different data sources (sidecarless Service Mesh, new Monitoring and Observability tools)
  - Collected data is exposed and transformed to supply both Failure detection and prediction modules
  - The Failure Prediction Machine Learning module was implemented in order to improve the 5G Core reliability

Exposure and privacy preserving

- **Sensing exposure key-take aways** (see ISAC slides for more information)
  - **Streamlined Data Management**: DataOps integrates agile principles with automation to optimize the collection, processing, and exposure of data across 6G networks, leveraging both CP and UP/DP.
  - **Exposure Flexibility**:
    - **CP**: Suitable for small data (e.g., Boolean, location)
    - **UP/DP**: Ideal for large datasets
    - 6G exposure relies on the 5G NEF framework, including CAPIF and SEAL, adaptable to various use cases.
- **Privacy-Preserving Architecture**:
  - Adopting a **Prio-based** aggregation method ensures privacy in data collection and analysis by the network, where NWDAF acts as the **collector** and AFs manage aggregation.
- **Multi-source Data Integration for failure detection**:
  - Uses data from **multiple sources**, including **logs** and **metrics** from Kubernetes API and monitoring tools such as **Service Mesh proxies**, cAdvisor, and metrics from Open5GS NFs collected by Prometheus.
  - **Core Objectives**:
    - Ensure **service continuity** and reliability in cloud-native environments.
    - Provide real-time **monitoring** and data exposure to prevent or predict failures.
  - **Use Cases**: Examples include advanced robotics (cobots), massive digital twins, and immersive telepresence for enhanced interaction

MLOps

MLOps description

- **MLOps concept**
  - MLOps offers a set of tools and methods for efficient management of distributed AI functions in 6G networks.
  - These tools target to optimize the communication, computation, storage, and energy utilization across the distributed AI nodes [see Fig.1 on Slide 25 for distr. ML training and inference, see Figs. 1-2 on Slide 24 and Figs. 1-3 on Slide 25 for communication and computation optimization].
  - The studied methods include distributed ML model training (Federated Learning (FL), Split Learning (SL)), model generalization, model layer offloading, and optimized resource allocation [see Fig.1. on Slide 23].
- **Main goals and objectives**
  - Collaborative ML model training with network and application [see Fig.1 on Slide 23].
  - Distributed ML model generalization for multiple tasks (e.g., video bitrate or delay estimation), reducing the need to train and maintain separate models [see Fig.1 on Slide 23].
  - ML model layer offloading (e.g., neural network layers) between output consumer application and network functions, balancing ML model accuracy with device energy consumption [see Fig.1 on Slide 23].
  - Optimizing radio resource allocation (e.g., power control) across wireless links between communicating nodes during distributed ML model training [see Figs. 1-2 on Slide 24].
  - Optimizing computing resource allocation and pricing across distributed nodes belonging to different operators to balance profit and ML model accuracy [see Figs. 1-3 on Slide 25].
  - Design privacy-aware learning schemes for distributed ML model training [see Fig. 1 on Slide 27].

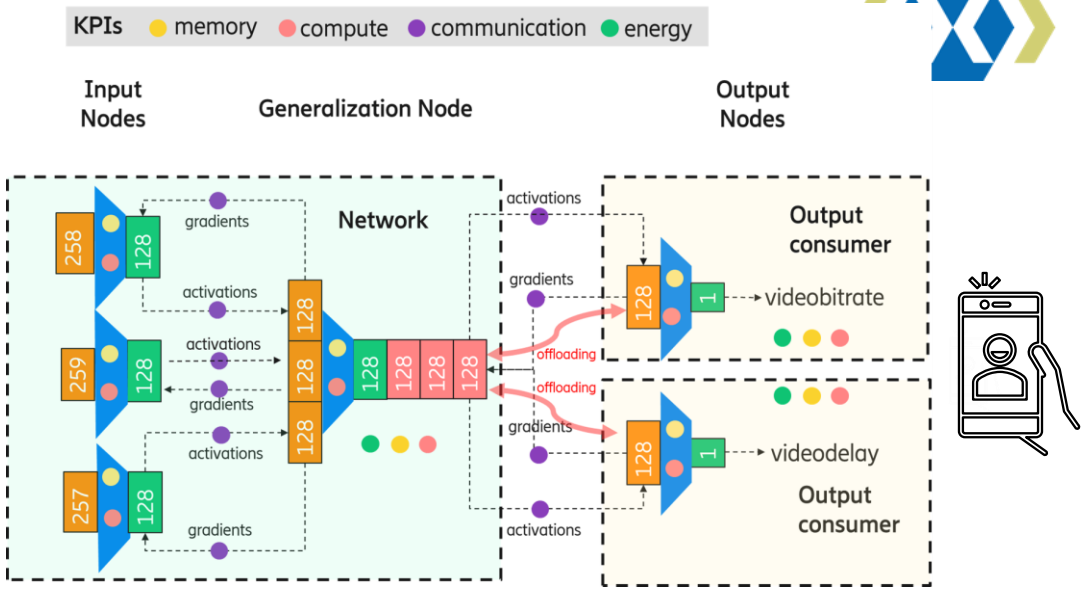


Fig. 1: Illustration of split learning setting and model layer offloading between generalization node that generalizes to two output nodes performing different tasks [see slide 23].

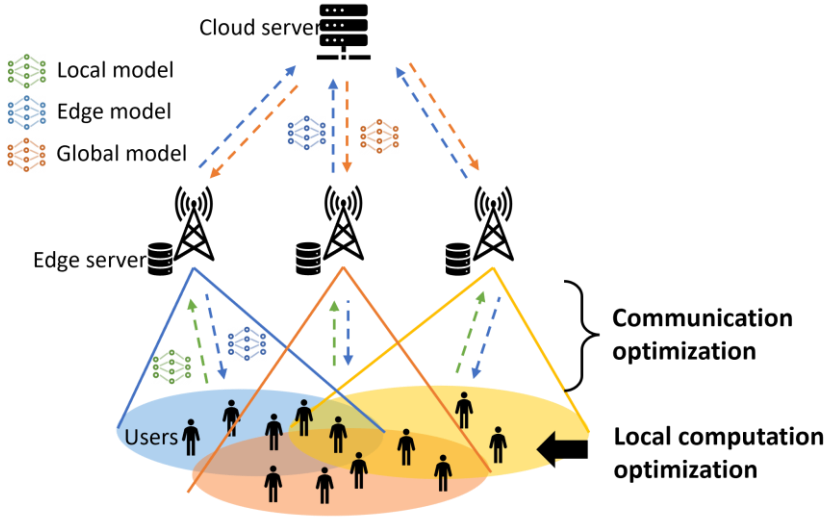


Fig. 2: Illustration of communication and computation optimization in wireless Federated Learning (FL) networks [see slides 24 and 25].

MLOps evaluations – benefits and KPI improvements

- **Main KPI improvements from MLOps**
  - Reduction in memory allocation(32KB->128B), ML model training time(68%) and energy consumption (73%) at the output consumer via offloading two ML model layers.[See Slide 23 – distr. ML training and inference]
  - Approximately 21% increase of global ML model accuracy via radio resource allocation in distributed ML model training compared to baseline/fixed allocation scenarios.
  - Reduction in communication time and energy consumption (by up to approximately 69%) via radio resource allocation in distributed ML model training compared to baseline/fixed allocation scenarios.
  - Improved computing resource utilization in distributed ML model training by splitting the ML model into distinct parts and effectively distributing computing workloads. [Slide 23 – distr. ML training and inference]
  - Reduced communication overhead in distributed ML model training architectures by minimizing the need to communicate large volumes of data to centralized server, and by dynamically allocating radio resources to optimize data transmission.
  - Improved data privacy in distributed ML model training, as the training and inference is performed without moving data in between entities.
- **Related Use Cases**
  - Cooperating mobile robots
  - Seamless immersive reality
- **Applicable Design Principles** (introduced in [HEX224-D33])
  - Principle 2:Full automation and optimization
  - Principle 3: Flexibility to different network scenarios
  - Principle 6: Persistent security and privacy
  - Principle 10: Minimizing environmental footprint and enabling sustainable networks

MLOps evaluations – Distributed model training and inference

Results from Component PoC #B.2

- Component PoC #B.2 addresses privacy, compute and energy overhead in split learning (SL) based vertical FL via:
  - model generalization in multi-task learning.
  - model layer offloading (focus in this deliverable).
- Reduction in forward (-58%) and backward propagation time (-78%), memory allocation (32KB →128B) and energy consumption (-73%) at the application via offloading 2 layers from application to network (see Fig. 1).
- In the PoC, the estimated energy consumption reduces with the offloading of model layers (22.7J → 5.9J) at the output node. Typically, the larger the models and offloaded layers are, the higher the energy saving (see Fig. 1).
  - Energy cost of transferring 2 NN model layers is negligible as compared to the transfer of activations, moreover offload is intended to be performed only on-demand in the whole training or inference phase.
  - Partial model offload does not necessitate offloading data in this vFL setting.
- The results indicate total energy saving when:
  - model layer offloading is triggered (focus in this deliverable),
  - generalization layer is used that serve multiple use cases simultaneously (shown in previous deliverable).
- Related WP3 enablers: compute offloading, DataOps.
- The potential role of Component PoC#B.2 in the System PoC:
  - Potentially assists exposing the localization data of the UE and application in a privacy preserving and communication efficient manner to the network functions in the network in the form of encodings (activations).
  - Necessitates deploying one side of the NN model at the UE (e.g., application) and the other side at the network, and having them communicate.

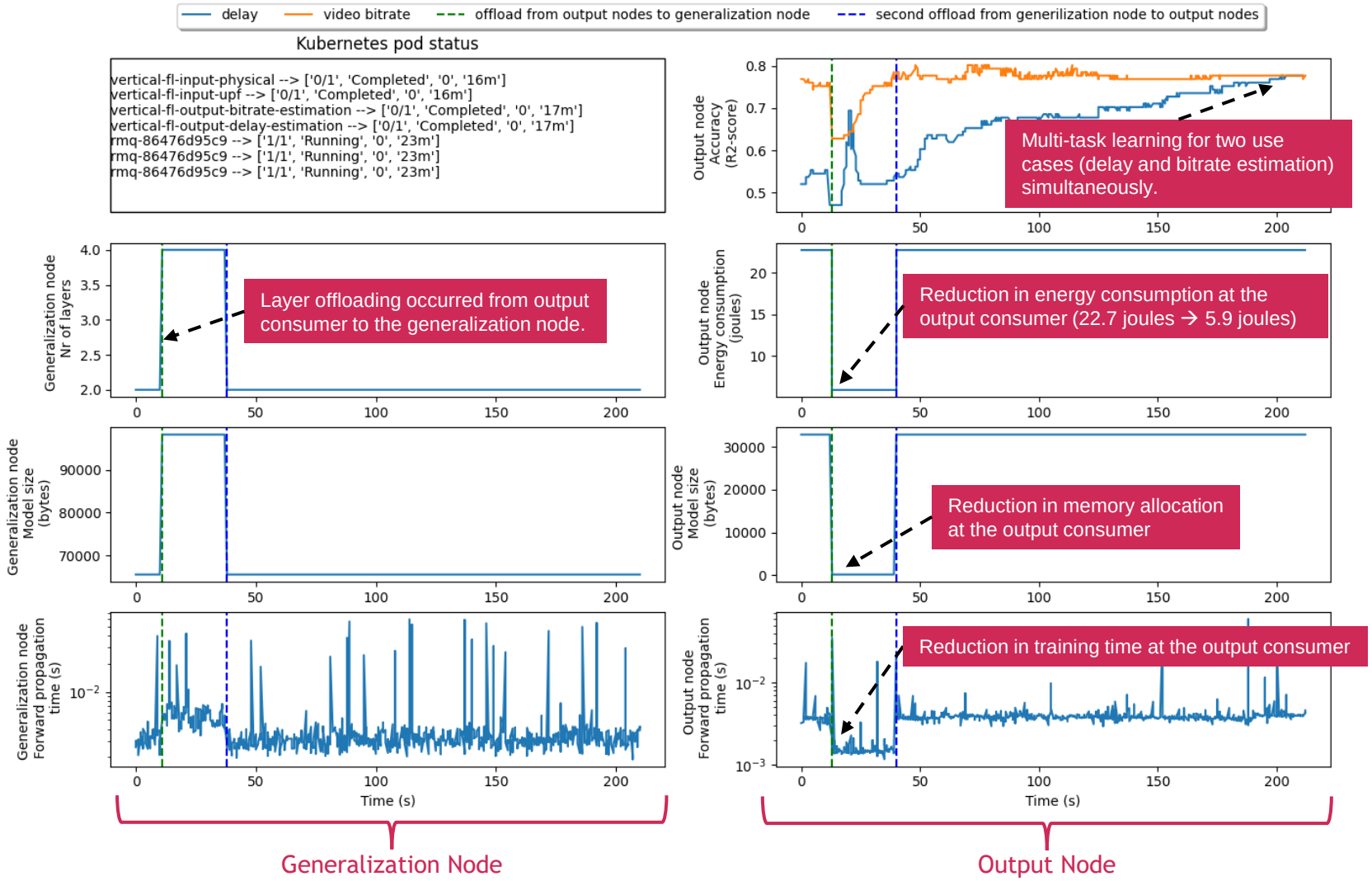


Fig. 1: Performance evaluation of split learning-based vertical FL via model layer offloading.

MLOps evaluations - On the accuracy-energy tradeoff in wireless hierarchical federated learning (HFL)

- The joint problem of association and uplink transmission power allocation of the users to the edge in a wireless HFL network using power-domain NOMA, is modeled as a Game in Satisfaction Form, such that each user pursues its tradeoff between local model accuracy and consumed time and energy overheads autonomously.
- The devised game may conclude different equilibria:
  - Satisfaction Equilibrium (SE) – Each user satisfies its targeted accuracy-time-energy tradeoff value.
  - Minimum Efficient Satisfaction Equilibrium (MESE) – Each user satisfies its targeted accuracy-time-energy tradeoff value with the minimum possible energy cost.
- Other association baseline comparative scenarios considered are the (a) Random user to edge server association and the (b) User association to the closest edge server, where users transmit with maximum power.
- The evaluation is performed over the MNIST dataset for handwritten digit classification of 6000 samples. The simulation parameters can be found in [CDP24].
- The proposed framework attains the highest global ML model accuracy (approximately 80%-90%) among all alternatives, under both types of equilibria (i.e., SE and MESE) (see Fig. 1).
- The increase in the number of users leads to higher congestion in the network and thus, more interference is sensed and caused between them (see Fig. 2). Consequently, lower data rates are achieved, leading to a longer time required for the transmission of their local model parameters to the edge, along with higher power consumption. The opposite behavior is observed when the number of edge servers increases.

Up to ~21% increase of global ML model accuracy compared to baseline association scenarios

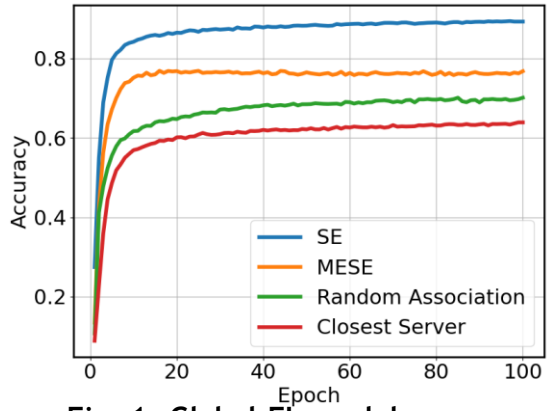


Fig. 1: Global FL model accuracy over FL epochs.

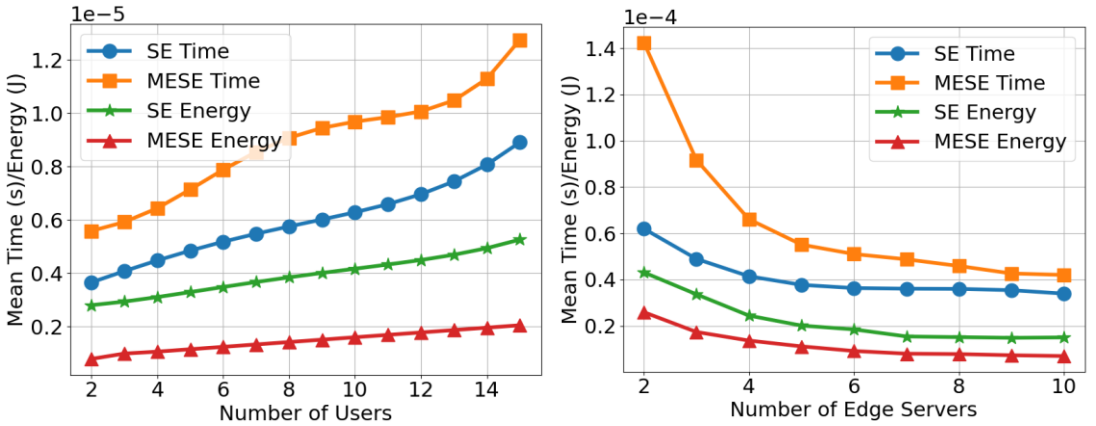


Fig. 2: Mean user transmission time and energy consumption during FL process, considering increasing number of users and edge servers in the network.

- At the MESE equilibrium point, lower energy consumption is concluded owing to the MESE's definition (see Fig. 2). SE results in better accuracy and time overhead values compared to the MESE point (see Fig. 1).

MLOps evaluations - Incentive mechanism design for wireless federated learning networks

- The joint problem of computing resource allocation and incentive provisioning in wireless Hierarchical Federated Learning (HFL) networks is tackled as a price-driven resource allocation problem based on the Market Equilibrium (ME) theory. The problem is modeled as a Fisher market where the edge servers are the buyers and the users' computing resources represent the goods to be purchased (see Fig. 1).
- At the ME point, the following two conditions are satisfied:
  - Given the equilibrium prices paid to the users, every edge server achieves its optimal resource allocation that maximizes its profit.
  - Either all computing resources are purchased by the edge servers, or their corresponding price is set to zero. This is known as the “market clearance” condition.
- For the performance evaluation consider 10 users with increasing computing capacity from 1 GCPU-cycles/sec to 10 GCPU-cycles/sec, indicated by the user ID, and 3 edge servers with increasing budget from 100 to 400 (unitless), indicated by S1, S2, S3, and S4.
- It is observed that edge servers with higher budget are allocated most of the computing capacity in the system (see Fig. 2), as shown in the vertical axis, which represents the percentage of each user’s computing resources. On the other hand, edge servers with smaller budget (e.g., S1 and S2) share the computing resources of users with higher computing capacity.
- Consider increasing the number of edge servers from the base scenario of three to two, three, four, and five times that amount. Then, the mean computing allocation that an edge server can expect decreases (see left side of Fig. 3). On the contrary, the prices paid to the users increase (see right side of Fig. 3).

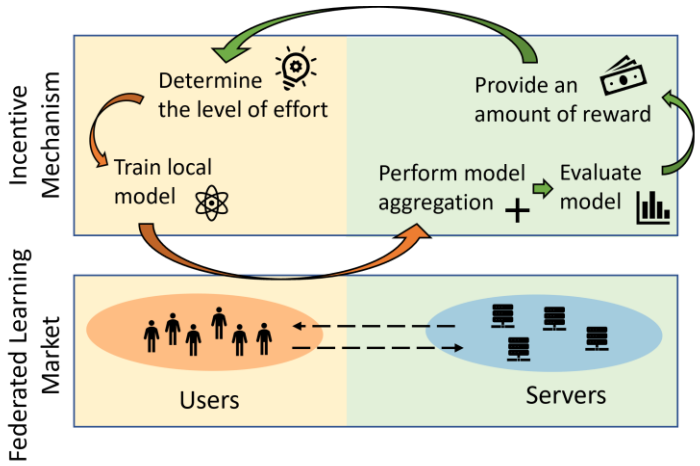


Fig. 1: Price-driven computing resource allocation in HFL networks as a Fisher market.

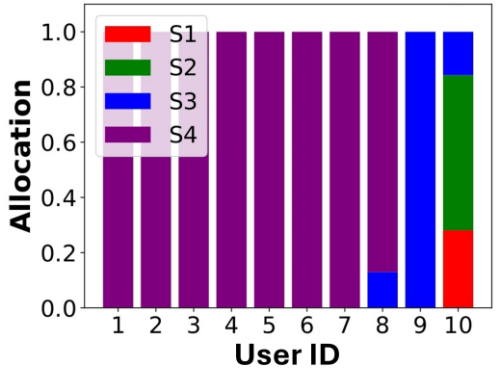


Fig. 2: Percentage of computing resource allocated by each user device to edge servers.

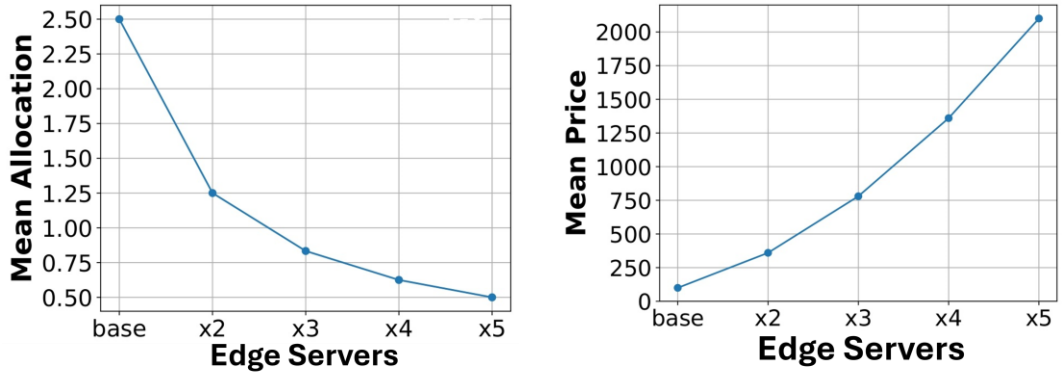


Fig. 3: Mean edge server computing allocation (left) and price (right), considering increasing number of edge servers in the network.

MLOps evaluations – Collaborative edge computing model

- Consider an integrated smart city system that facilitates inter-vertical communication and information sharing through edge nodes that host essential services for monitoring city verticals.
- Tapping into the computational capabilities of edge nodes, effectively offloading the computational load of ML model training towards the network while also taking advantage of the strategic placement of edge nodes for data collection.
- ML model training is performed in a distributed and secure way using a Federated Learning framework for contributions from edge nodes and user devices while protecting the privacy of the data collected.
- Workload offloading algorithms for heterogeneous devices and open-source solutions enabling federated learning among multiple devices and multiple tasks are implemented and tested.
- The Urban Platform is a digital tool that monitors and presents relevant city data on a single platform (see Fig. 1).
  - A Broker integrated on this platform that has the information of the devices and edge nodes connected along the federation.
  - Model aggregation is done on the cloud, minimizing the data processed and transferred towards the Urban Platform.
- The proposed implementation framework aligns with the goals and objectives of MLOps by (i) ensuring efficient computing resource allocation, (ii) enabling distributed ML model training, (iii) preserving data privacy, and (iv) supporting scalable and flexible model management across multiple devices and tasks.

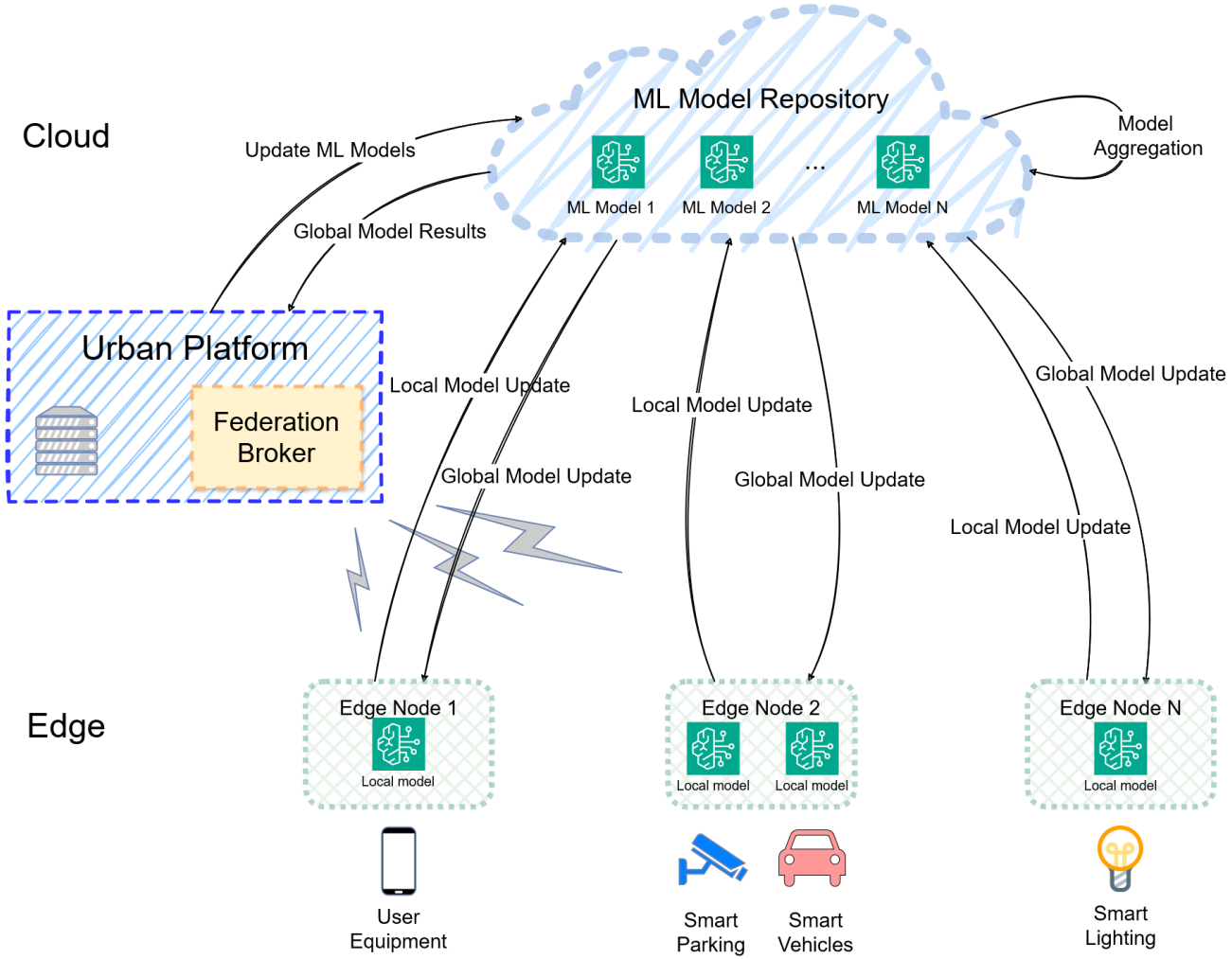


Fig. 1: Illustration of vertical FL implementation for smart city applications using the so-called Urban Platform.

MLOps protocols and APIs - Privacy-aware learning

- Cooperative learning among network nodes with different data privacy sensitivity levels utilizes data collection, learning and inference based on data privacy levels (introduced in [HEX224-D33]).
- Potential cooperative learning scenario is shown in Fig. 1:
  - **Network (via NF)** or **UEs** may initiate/ask NWDAF for analytics assistance (data/model training) using UE/Network data.
  - **UEs** report available data types (privacy sensitivity levels), determining the cooperative learning variant.
  - If UE contains multiple data types with different privacy sensitivity levels, the variants can be combined.
  - **NWDAF** performs the training and inference and can provide analytics and trained models both to consumer NF and UEs.

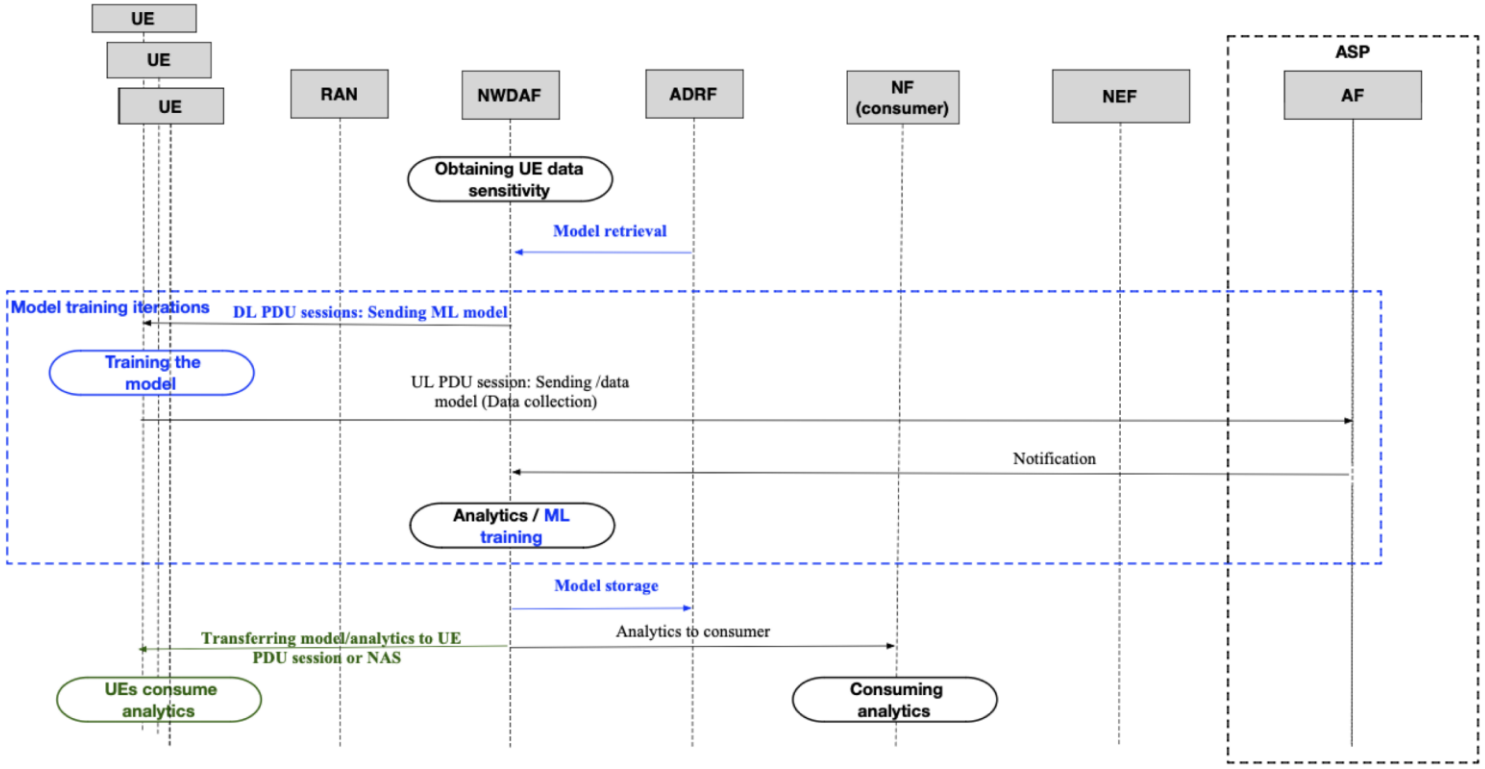


Fig. 1: Example of cooperative data collection and learning for privacy level 1/2 data.

MLOps Protocol and APIs - Federation process protocol

- Standard Federation process following ETSI MEC specifications in [ETSI-MEC-040], handling the registry of new devices on the so called Urban Platform, as illustrated in Fig. 1.
- UE resources are exposed when requesting to join the federation to more accurately assign models to train with their local data.
- UE ledger monitored by the Federation Broker is updated each time a new node joins or leaves the network.
- Global ML models are updated every round of tests in the Federation Learning process until an accuracy threshold is reached (or several updates are finished).
- ML Repo uploads the models to Authorized UEs for local training.
    - This assignment is done considering node information collected by the broker during federation process.

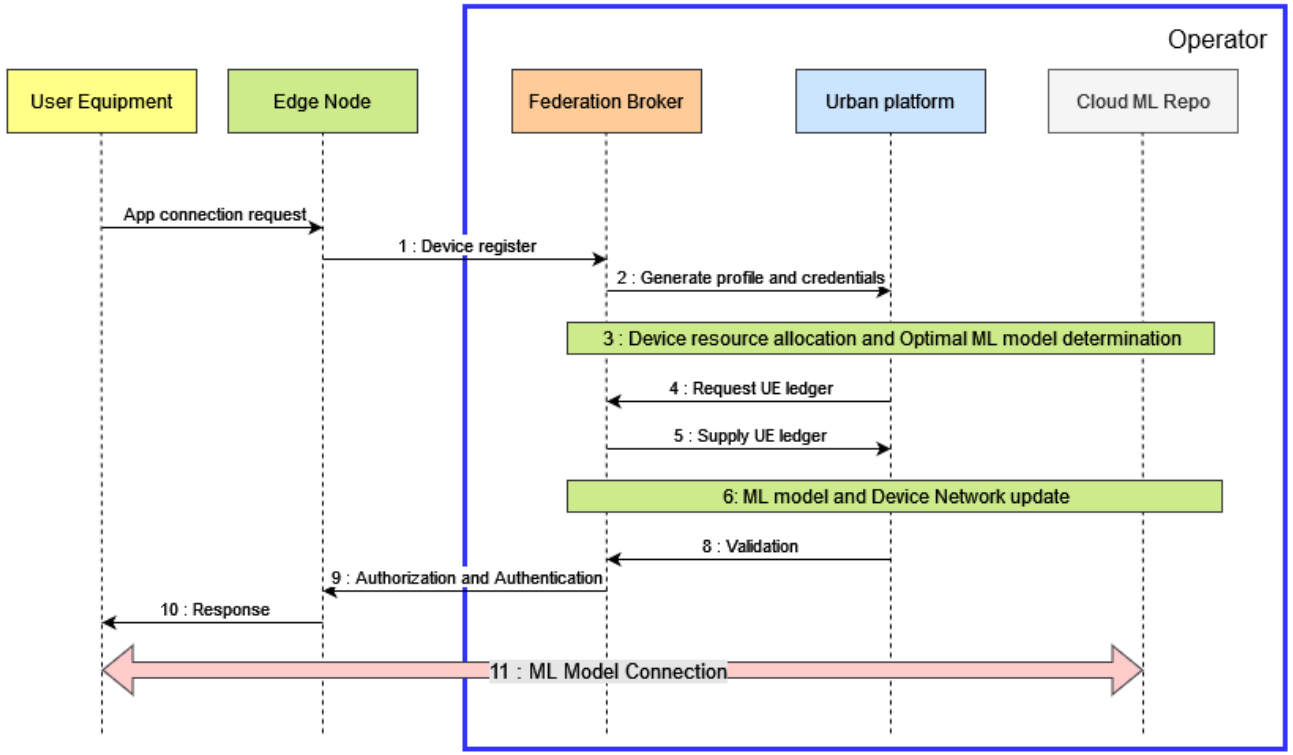


Fig. 1: Example of federation joining and the federated learning process flow.

MLOps evaluations – Failure prediction for 5G Core network

- With the emergence of new use cases of 6G networks (such as from robots to cobots, massive twinning, immersive telepresence for enhanced interactions), reliability and high uptime remain paramount.
- For this reason, we proposed and implemented Failure Prediction for 5G Core networks with Machine Learning (Fig. 1), capable of:
  - Collecting raw data from multiple sources, and then transforming it into data sets focused on TTFB latency between network functions
  - Creating data sets were focused on a variety of disadvantageous networking conditions and their impact on the core network
  - Using data sets to train and compare four ML models (Tree regressor, SVR, MLP regressor, LSTM), capable of providing the M&O with failure predictions
- With the MLOps designed to work in-real-time, we defined and measured four steps of the DataOps to MLOps data flow (Fig. 2, Fig. 3):
  - T1 – Collecting raw data
  - T2 – Transforming raw data into data sets
  - T3 – Ingesting data by the ML model
  - T4 – ML algorithms execution with processing

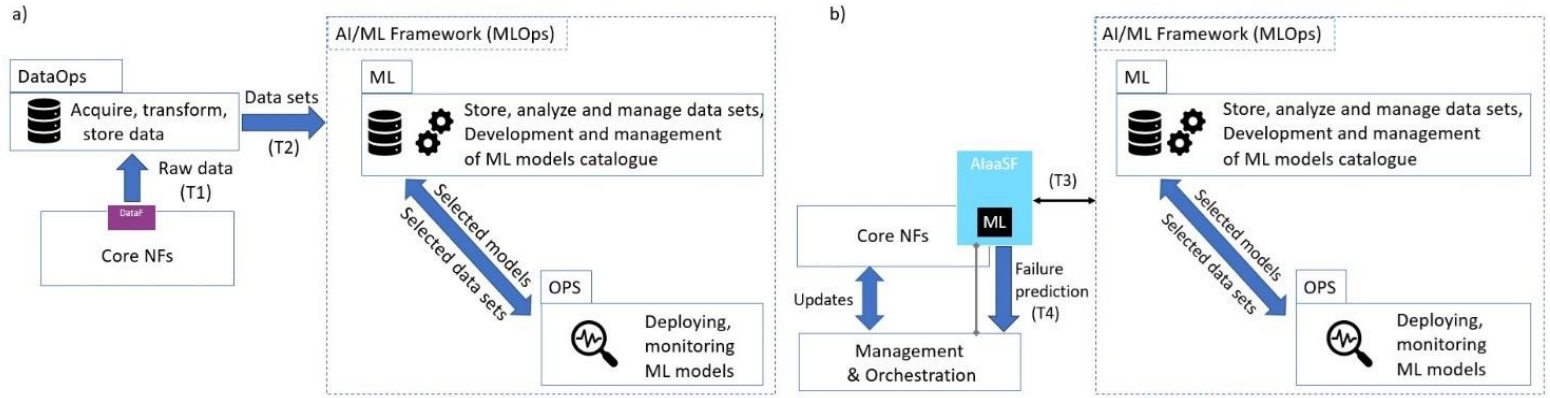


Fig. 1: Specific view of process for failure prediction in 5G/6G Core.

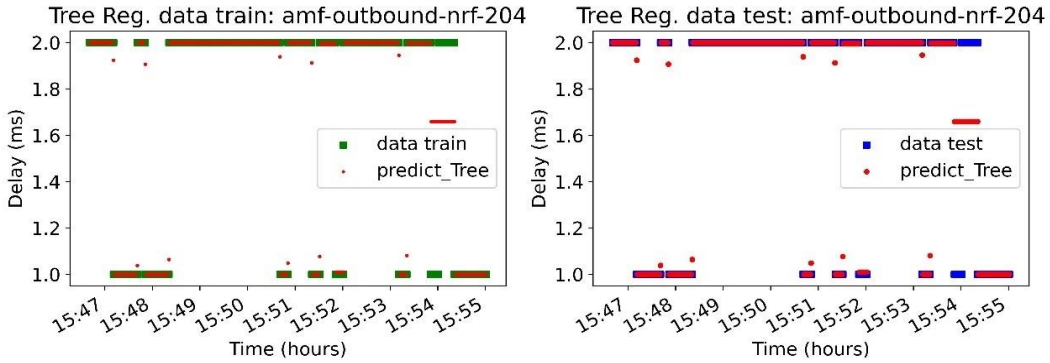


Fig. 2: TreeRegressor model prediction for training data - output for 'amf-outbound-nrf-204' attribute.

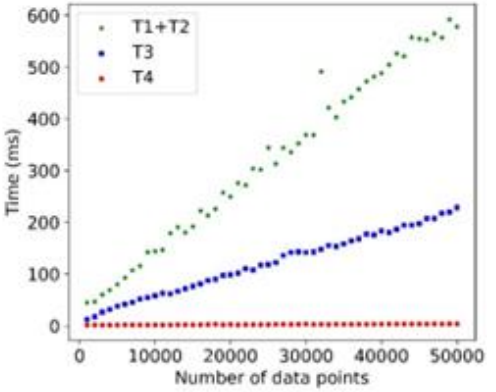


Fig. 3: Times T1+T2, T3, T4 in the function of data points.

MLOps Key take-away

The **key takeaways** can be summarized as follows:

- MLOps represents a set of tools and methods to streamline the ML lifecycle, especially in distributed settings.
- To address the challenges of distributed ML, MLOps must incorporate advanced orchestration, privacy-preserving techniques, and robust infrastructure improvements.
- The orchestration process must optimize various and possibly conflicting KPIs, including ML model efficacy, total consumed time and energy, and resource utilization.
- The introduction of privacy-preserving architectural components is essential, while carefully considering the added complexity to the network architecture.
- Establishing communication links between distributed computation nodes is imperative for ML model parameter exchange, demanding careful design as the network scales.
- Continuous monitoring and feedback loops are needed to guarantee freshness of datasets and trained models across the network.
- MLOps acts as an enabler for AIaaS by optimizing AI services throughout their lifecycle.
- MLOps and DataOps work in synergy to streamline AL/ML lifecycle.

MLOps may **improve sustainability** through several key aspects:

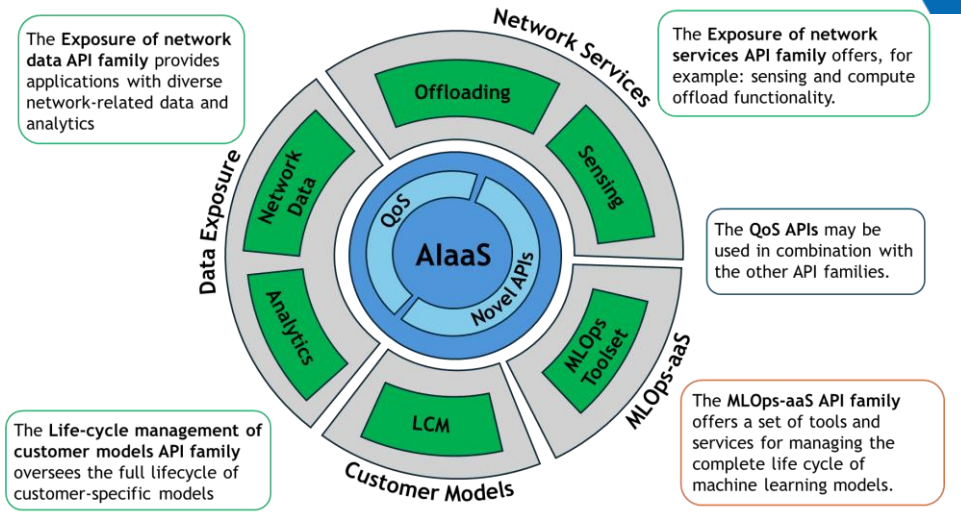
- **Optimized Network and Service Management**
  - Automation, intelligent orchestration of network resources, and proper management of distributed AI functions contributes to overall optimized network and service management.
- **Resource efficiency**
  - ML model layer offloading allows efficiently utilizing resources across the computing continuum while minimizing on-device energy consumption.
  - ML model generalization allows for serving multiple use cases via Split Learning, reducing redundancy.
  - Radio and computing resource allocation support resource efficiency in MLOps.
- **Socio-Economic Benefits**
  - Fosters innovation by automating the lifecycle management of advanced AI applications.
  - Provides trusted distributed ML model training solutions via privacy-preserving data collection and training protocols.
  - Supports distributed ML model training, reducing the decision-making and training overhead of demanding AI applications, making AI more accessible and cost-effective.

AI as a Service (AIaaS)

AIaaS description

- **AIaaS concept**
  - AI-native platform integrated within the 6G network architecture to provide AI capabilities as services in support of various applications (see Fig.2)
  - Superset of MLOps features, incorporating all MLOps APIs and additional APIs (e.g., data exposure, QoS) to enhance the network's capabilities, reinforcing the vision of the mobile network as a platform (see Fig.1)
- **Main goals and objectives**
  - On-demand and tailored functions for common AI services exposure and AIaaS lifecycle management (see Fig.2)
  - Unified AIaaS exposure APIs for AI services and functions operations and management (see Fig.1)
  - Flexible API orchestration to integrate multiple APIs and deliver efficient and comprehensive AI services
  - Cloud-native solution to support heterogeneous deployments across the continuum while enabling decentralized and cooperative AI functions
  - Support tailored AI services to meet specific customer needs, including model lifecycle management (selection, deployment, inference, monitoring) considering application-specific requirements
- **Work to be finalized in D3.5**
  - Identification of main AI services to be offered
  - Definition of AIaaS APIs to expose AI capabilities towards consumers

Fig. 1: Categories of AIaaS exposure capabilities and APIs



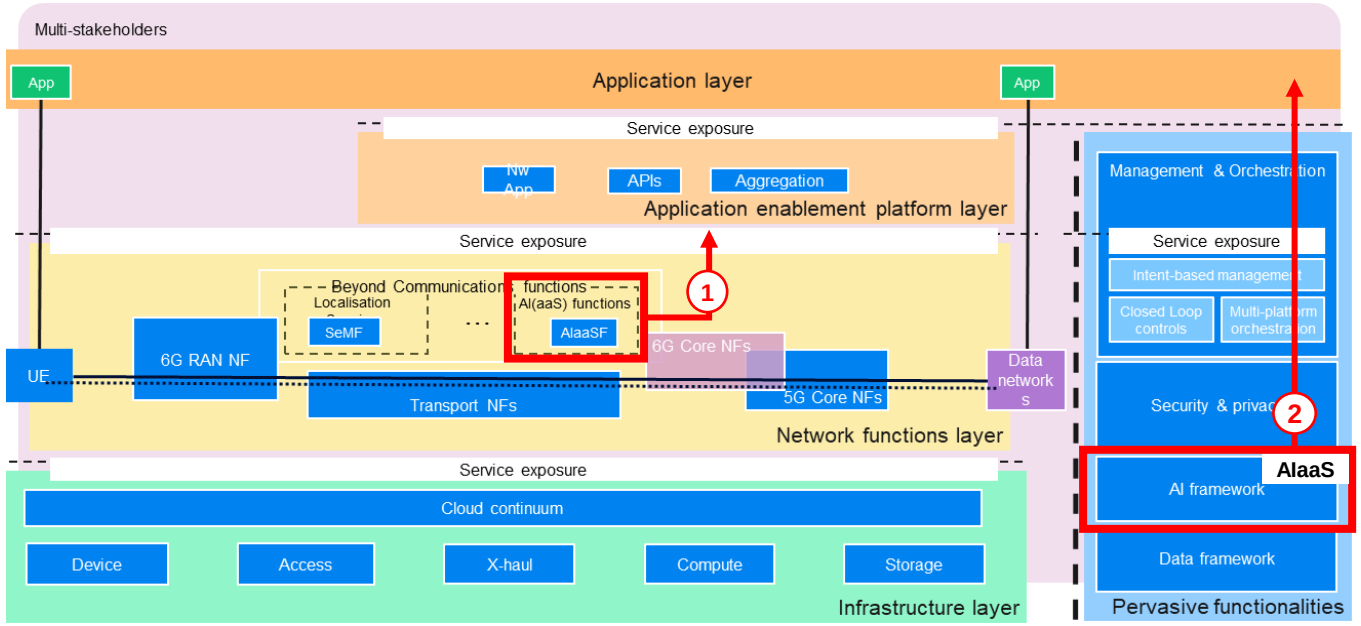


Fig. 2: Exposure of the AIaaS function(s) from any of the network domains (1) and external exposure of AIaaS to applications (2) in the Hexa-x-ii D2.4 end-to-end system blueprint.

AIaaS evaluations – prototype implementation

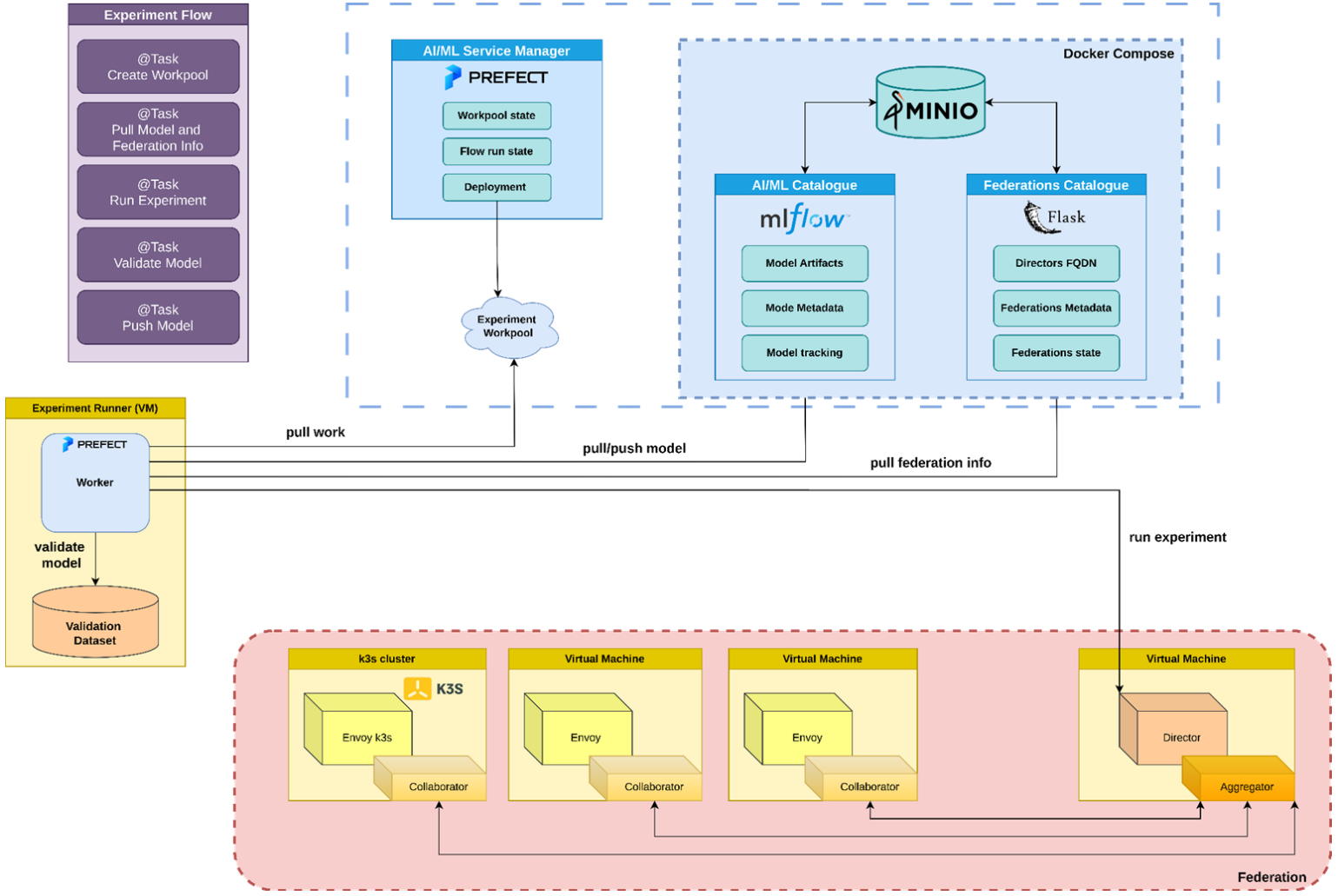


Fig.1: FLaaS software prototype components with the integration of Prefect, MLFlow and MinIO

- AIaaS software prototype capabilities:
  - Automated deployment and exposure of training and inference services
    - Work on top of heterogeneous cloud environments for distributed AI services
      - Kubernetes, docker, bare metal
  - Integration of opensource components
    - Prefect as AI/ML Service Manager
    - MLFlow as AI/ML Catalogue
    - Minio as ML model storage
- Tailored functionalities in support of AIaaS "enabled" for FL (through OpenFL)
  - FEDaaS: to establish a federation (or more than one) to launch future experiments or simply make the data the FEDaaS user owns indirectly available (more info will be provided in D3.5)
  - FLaaS: to test the performance or training an algorithm over an existing federation, indirectly using the federation data by running an experiment (see Fig.1, more info will be provided in D3.5)

AIaaS evaluations – benefits

- Easy Access to AI: applications can use AI tools without building their own infrastructure
- Unified exposure: unified interface and AI/ML model profile definition for easier exposure, storage, selection and management of different AI services
- Pre-built Model provisioning: exposes ready-to-use AI models, datasets, and algorithms via APIs, which avoids new model training and speeds up deployment
- Fosters Innovation: simplifies AI integration, encouraging new and innovative use cases.

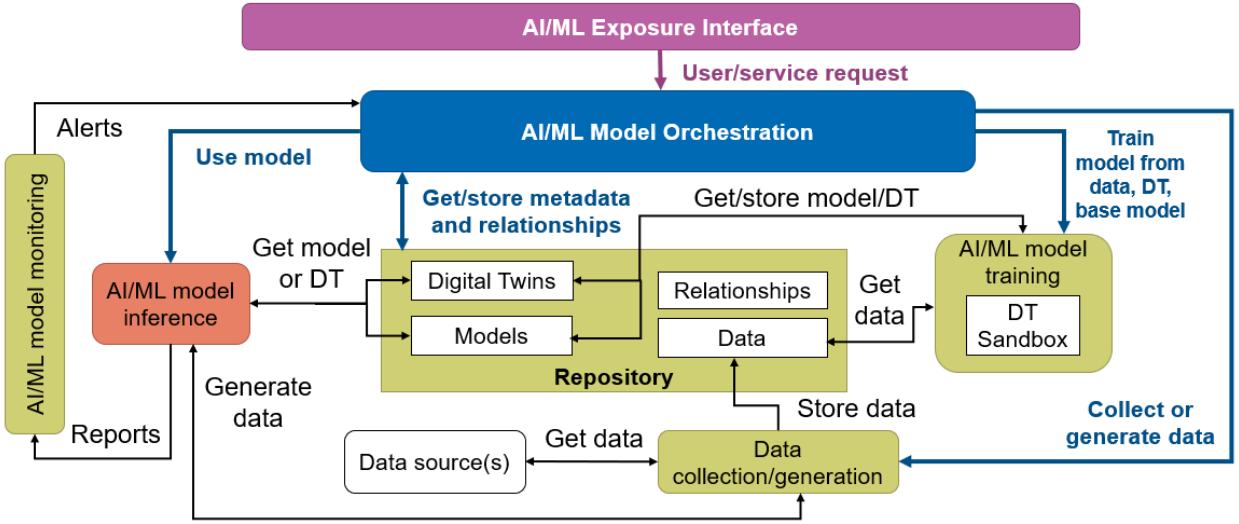


Fig. 1: AIaaS solution with orchestration capabilities for AI/ML model and Digital Twin lifecycle management and exposure to model and DT consumers.

- Cost Efficiency: reduces the cost and complexity of developing and maintaining AI systems
- Zero-touch management: more efficient and fully automated AI/ML model training, deployment and lifecycle management considering different use cases and dependencies for responsive AI service execution (see Fig.1), where model performance degradations can be proactively mitigated/prevented based on events through closed loops
- Integration in the continuum: seamless deployment and operation of AI/ML services on top of distributed cloud-native infrastructures
- Network-Specific insights: Automated model monitoring, performance degradation detection to expose insights not available from hyperscale cloud providers

AIaaS evaluations – KPI improvements

- **Impacted KPIs**
  - AI Model Accessibility
  - Performance Metrics
  - Computing resource utilization
  - Operational Efficiency
  - Automated Execution
  - User Satisfaction
- **Main KPI improvements**
  - Tailored (training and deployment of) AI/ML models reduce unnecessary processing by leveraging network data efficiently
  - Tailored AI services improve user experience
  - AI services exposure and APIs allow seamless access to AI functionalities without the environmental impact of redundant data processing
  - AI/ML model performance optimization and maintenance
  - Resilience to environment changes and AI/ML model performance degradation through proactive and reactive model update

- **Related Hexa-X-II Use Cases**
  - Interacting and Collaborating Robots (Cobots)

- **Applicable Design Principles** (introduced in [HEX224-D33])
  - Principle 1: Support and Exposure of 6G Services and Capabilities
  - Principle 2:Full Automation and Optimization
  - Principle 3: Flexibility to Different Network Scenarios
  - Principle 4: Network Scalability
  - Principle 5: Resilience and availability
  - Principle 8: Separation of Concerns of Network Functions

AIaaS protocols and APIs – AI services and exposure approach

- 3GPP CAPIF as candidate for AIaaS APIs exposure
  - Standard approach for API registration & discovery for 5GS
  - Provides native API discovery capabilities
  - Can “dynamically” expose AI/ML training, deployment and inference services & APIs
- O-RAN uses 3GPP CAPIF to register & discover available AI/ML services
- ETSI SDG OpenCAPIF open source project recently started

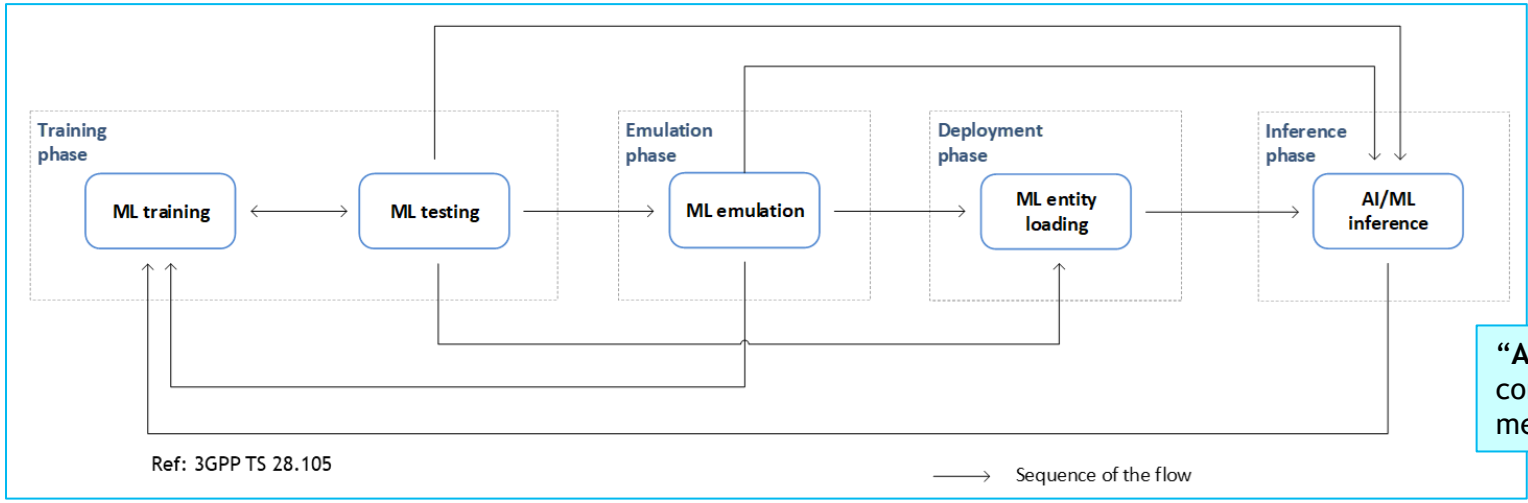


Fig.1: AI/ML entity operational workflow (ref. 3GPP TS 28.105)

- **Candidate AI services (see Fig.1)**
  - Training phase:
    - AI/ML entity training
    - AI/ML entity validation
    - AI/ML entity testing
  - Emulation phase:
    - AI/ML inference emulation
  - Deployment phase:
    - AI/ML entity load and deployment
    - AI/ML entity monitoring and notifications
  - Inference phase:
    - AI/ML inference control
    - AI/ML inference performance evaluation
    - AI/ML update

“AI/ML entity”: an AI/ML model or an entity containing an ML model and the ML model-related metadata → *manageable as a single composite entity*

AIaaS protocols and APIs - mechanisms for AI/ML model lifecycle management (inference phase)

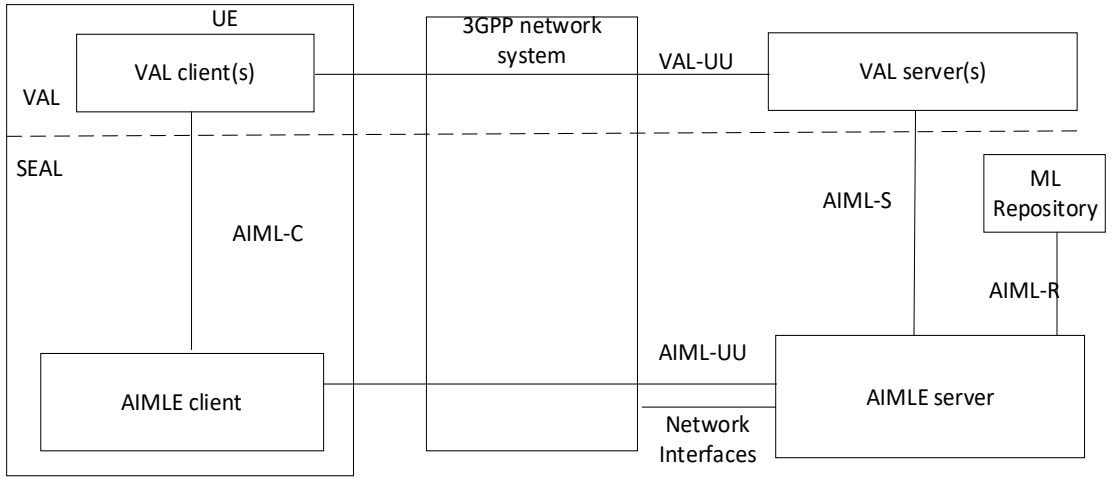


Fig.1: 3GPP architecture for AI/ML application enablement layer (ref. TR 23.700-82)

- 3GPP SA6 Working Group defines functionalities for the application enablement layer
- SA6 introduced in Release 19 the AI/ML application enablement layer AIMLAPP to define the required architecture, protocols and APIs (see Fig.1)
- Support for AI/ML model lifecycle management has been added with model performance degradation detection, update, and proactive update of dependent models (see Fig.2)

Contributions from Hexa-X-II accepted in 3GPP TR 23.700-82

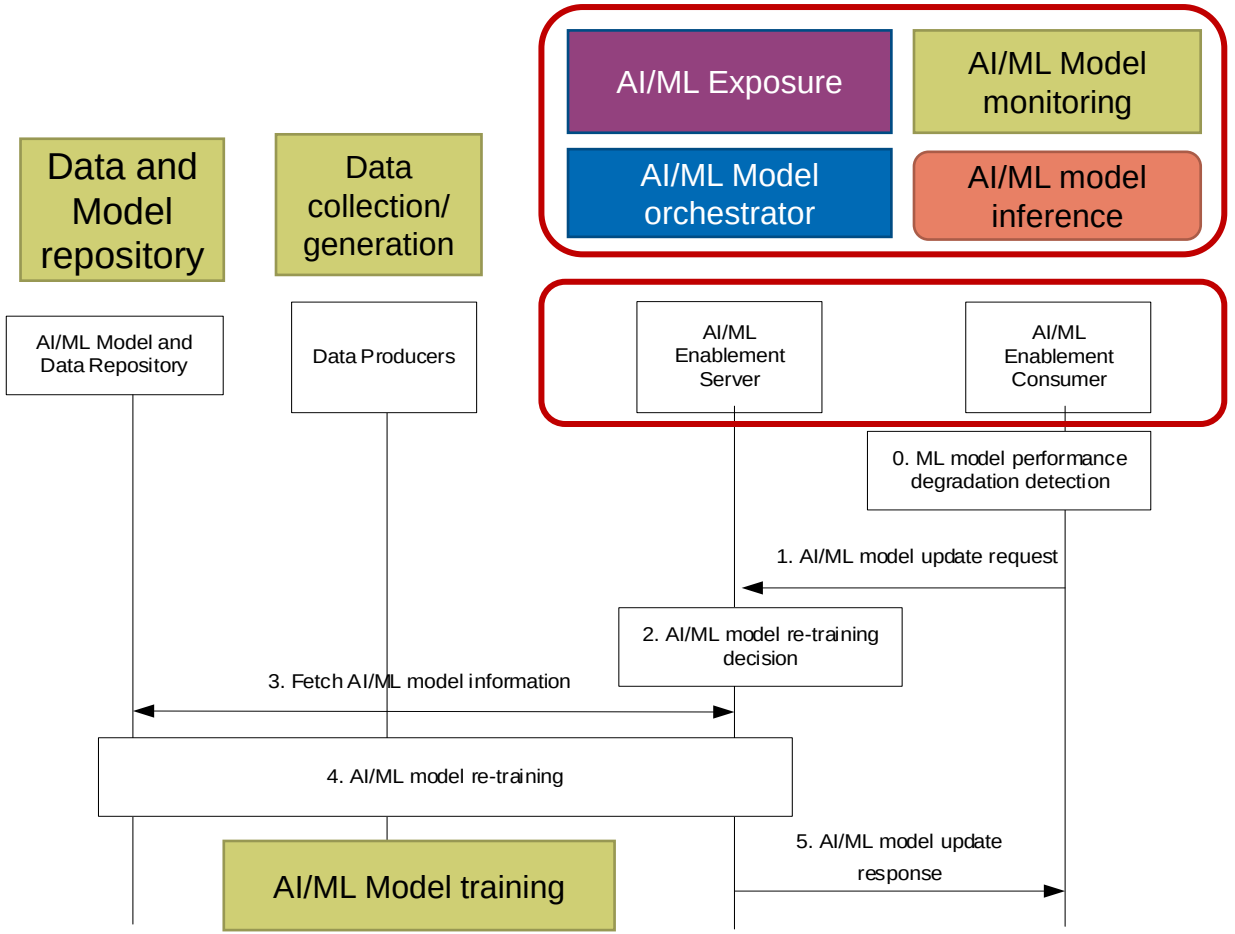


Fig.2: Support for AI/ML model lifecycle management in 3GPP (ref. TR 23.700-82)

AIaaS Key take-aways

The **key takeaways** can be summarized as follows:

- AIaaS as a one-stop shop solution to ease the deployment, operation and execution of AI/ML services, pre-built AI models, datasets, facilitating their consumption through a set of open APIs
- AIaaS allows to operate AI/ML services on top of distributed cloud-native infrastructures, integrating with the extreme-edge/edge/cloud continuum
- AIaaS helps in simplifying the transition to AI-enabled services, which is key for those verticals and application providers with limited AI skills
- AIaaS is a key enabler for AI and data monetization and new business propositions, and for the realization of new AI-driven vertical services and applications
- Contribution to 3GPP and alignment with architecture and API specifications ensures wider adoption of the developed concepts

Implications:

- Complexity of internal "aaS" logics to satisfy heterogeneous use case and AI task requirements (e.g. for deployment constraints, performance needs, etc.)
- Need to regulate ownership and interactions in case of multi-stakeholder scenarios (for AI/ML models, data for training and inference, access to services)
- Need to define new APIs, protocols, data models for AI services (and their exposure) management with lack of standard specifications

AIaaS may **improve sustainability** through several key aspects:

- **Optimized Network and Service Management**
  - Optimized training, deployment and performance of AI/ML models used for energy-efficient network and resource management
- **Resource Efficiency**
  - Automated model training and updates can be optimized to achieve energy efficiency in AI tasks
  - Provides pre-built AI tools and models, while reducing redundant infrastructure and optimizes resource use
  - Reduces resource demands, improving overall resource utilization
- **Socio-Economic Benefits**
  - Fosters innovation by offering advanced AI capabilities without requiring infrastructure investments
  - Distributes AI capabilities, making advanced functionalities more accessible and cost-effective

Relation with other enablers:

- **MLOps:** to automate AI/ML models lifecycle management
- **DataOps:** to efficiently distribute and manage data for AI services
- **Compute Offloading:** to optimize the placement of AI tasks and functions

Integrated Sensing and Communication (ISAC)

Integrated Sensing and Communications (ISAC)

- Integrated sensing and communications (ISAC) supports detection of the presence/absence, location, and some characteristics of an object as well as the direction and velocity if the object is moving.
- A radio (Tx) transmits a sensing signal and one or multiple receivers (Rx) detect the waveform after it has interacted (reflected, scattered) with the environment, see Fig. 1.
  - Line-of-sight (LoS) between object and Tx/Rx is often assumed.
  - Based on the location of Tx and Rx, radar systems are categorized into monostatic and bi-/ multi-static.
- The objective of this enabler
  - Definition of the interface for the request of sensing services and the provision of sensing outputs to a sensing client.
  - Description of configuration and control signaling for sensing operations, coordinated by the Sensing Management Function (SeMF).
- An exemplary scenario: The mmWave radios periodically exchange angle and distance estimates (self-sensing) by bouncing the signal in the environment, thus enabling accurate estimates of the target object/material surface, see Fig. 2.

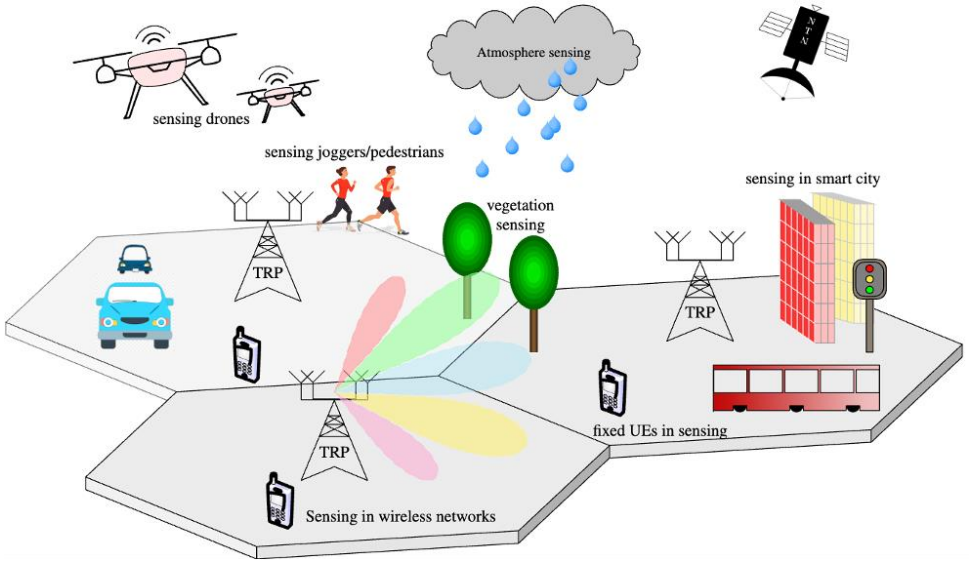


Fig. 1 ISAC topology

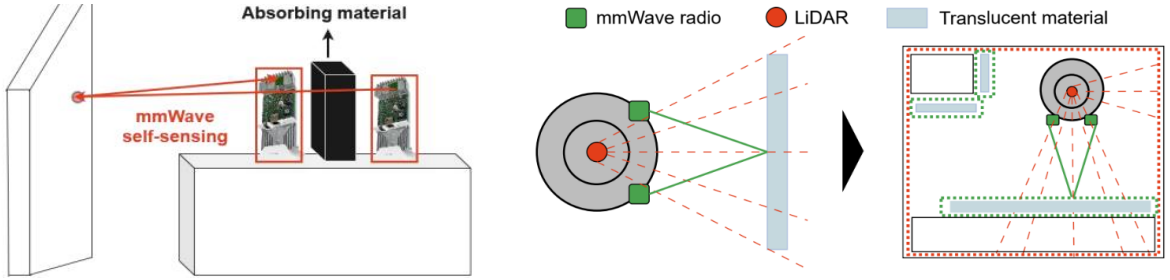


Fig. 2 Exemplary case study of mmWave sensing

Protocols and APIs for sensing - UE involvement

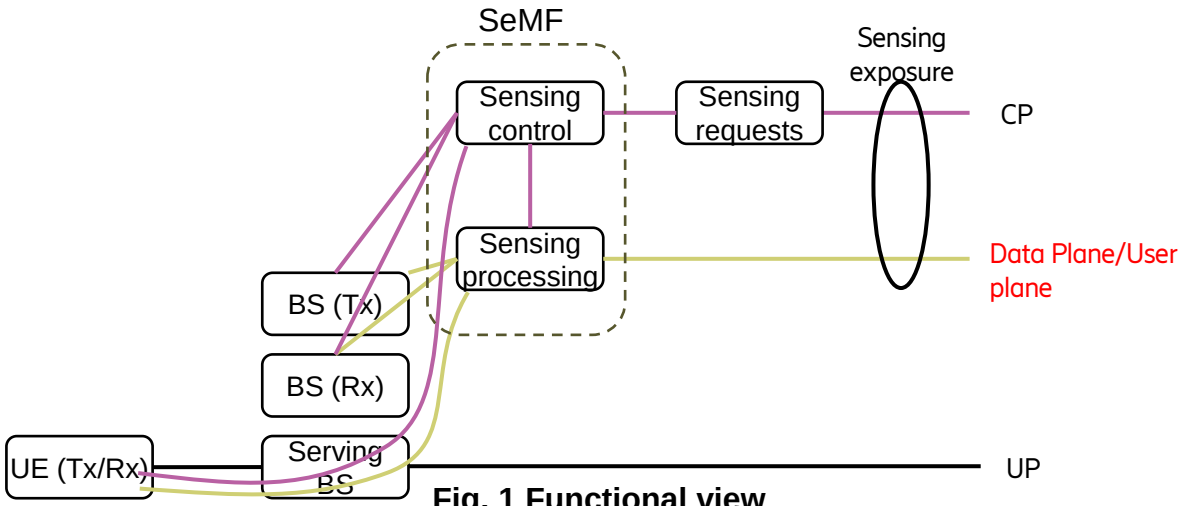


Fig. 1 Functional view

- Basic functional entities for sensing are (see Fig. 1 too):
  - Sensing control
  - Sensing processing
  - Sensing request
- When the UEs are involved, the following functional entities are added to the architecture:
  - UEs with sensing capabilities (capabilities comprise radio, battery, privacy, user consent, etc.)
  - Serving Base Stations (BSs)
    - serving the UEs involved in the sensing with connectivity
    - possibly involved in sensing measurement reporting from the UE
- Sensing data can be transmitted over user plane or possibly a new data plane
- The control task flow for a sensing scenario is depicted in Fig. 2
- The message sequence chart (MSC) is shown in Fig. 3 (see next slide).

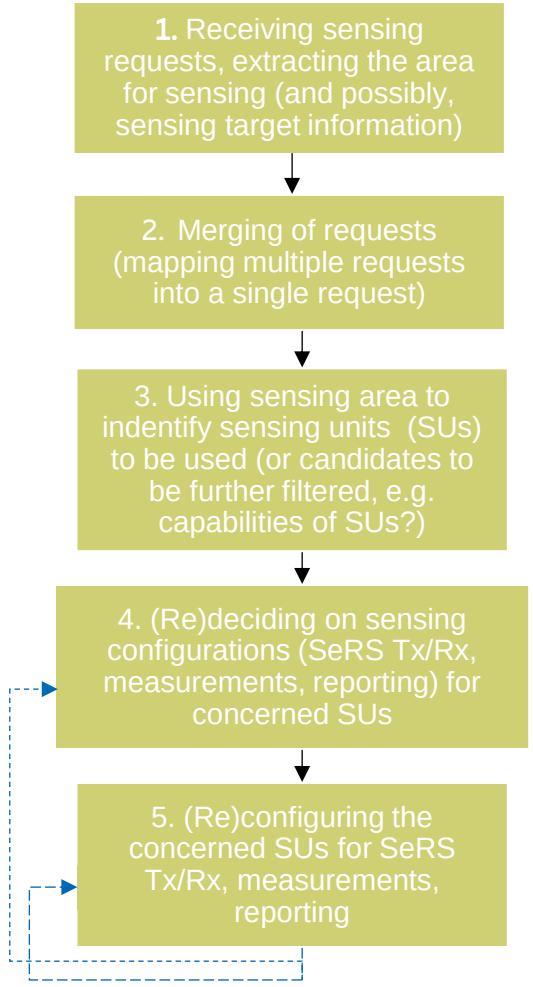


Fig. 2 Sensing control task flow

Protocols and APIs for sensing - UE involvement

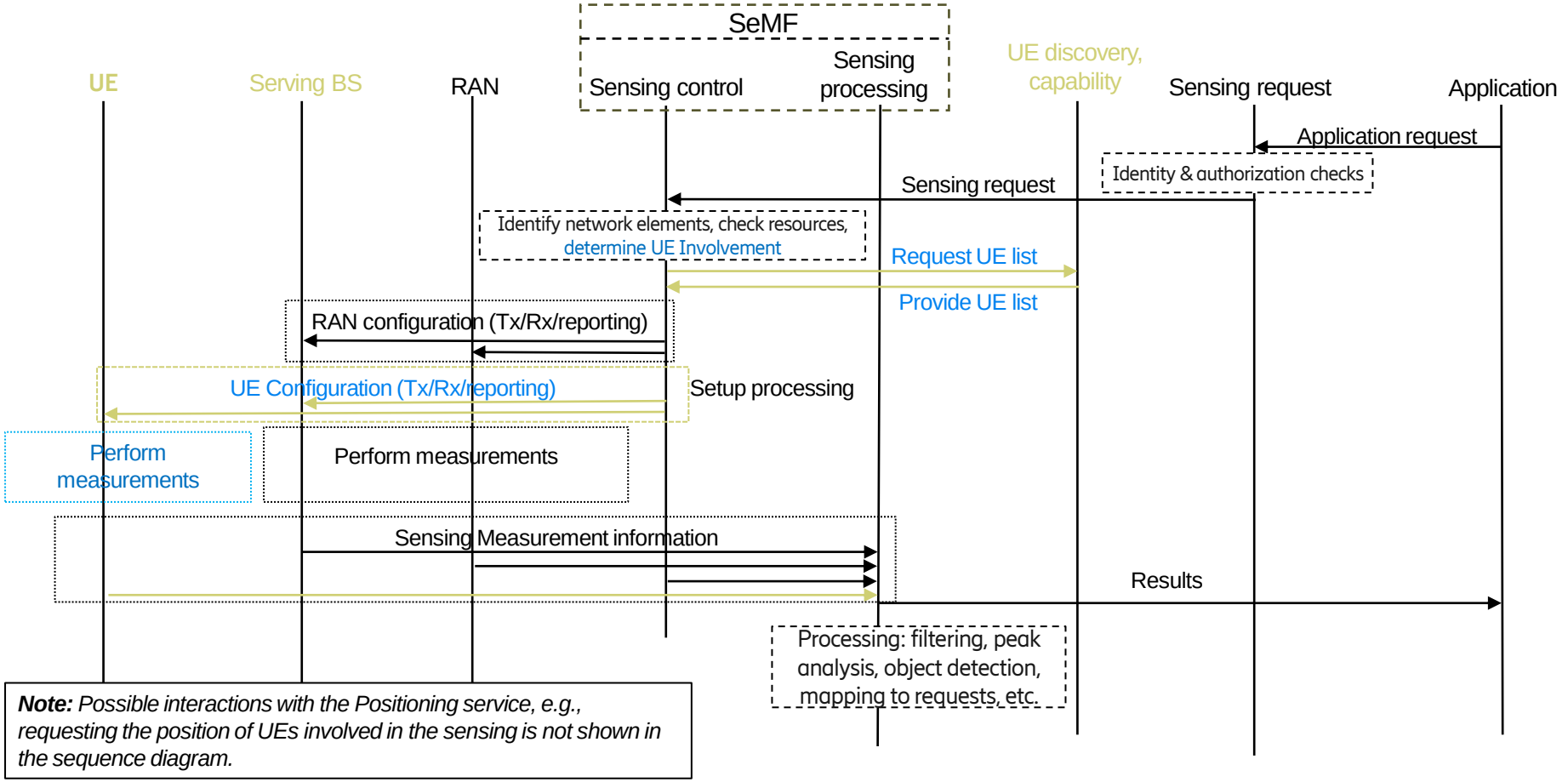
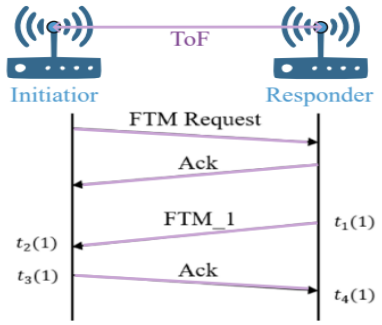


Fig. 3: Application (from the UE or on the internet) initiated sensing message sequence chart, including UE involvement: The light green functions are involved when UEs are involved in sensing.

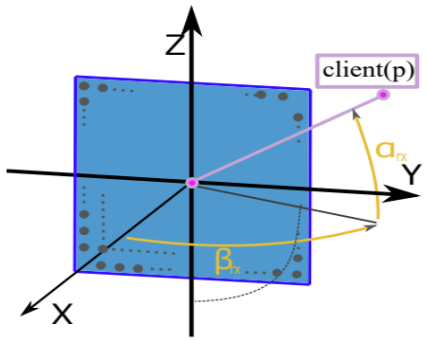
Exemplary case of mmWave Sensing

Fine Time Measurement (FTM)



(a) ToF extracted from FTM

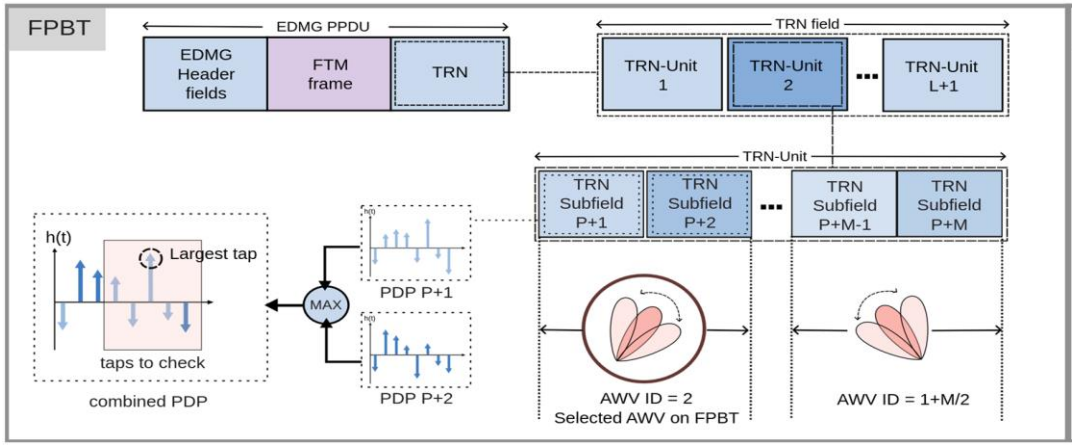
Channel State Information (CSI)

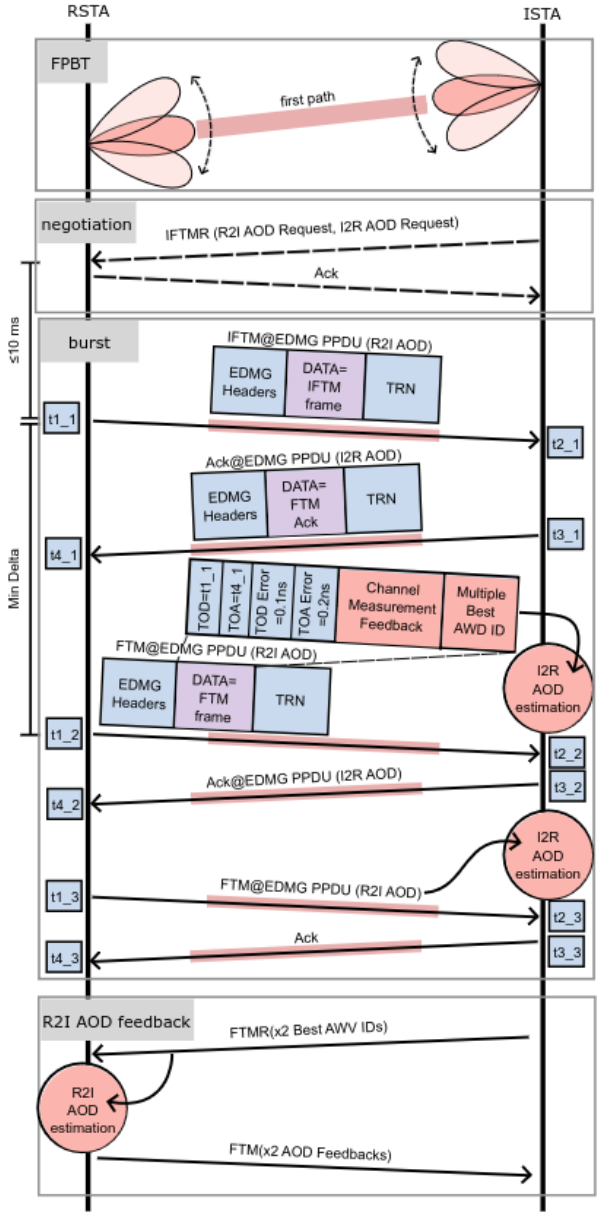


(b) AoA estimations with URA

Fine Time Measurement (FTM) over mmWave can accurately calculate the distance between two stations without need of being previously associated

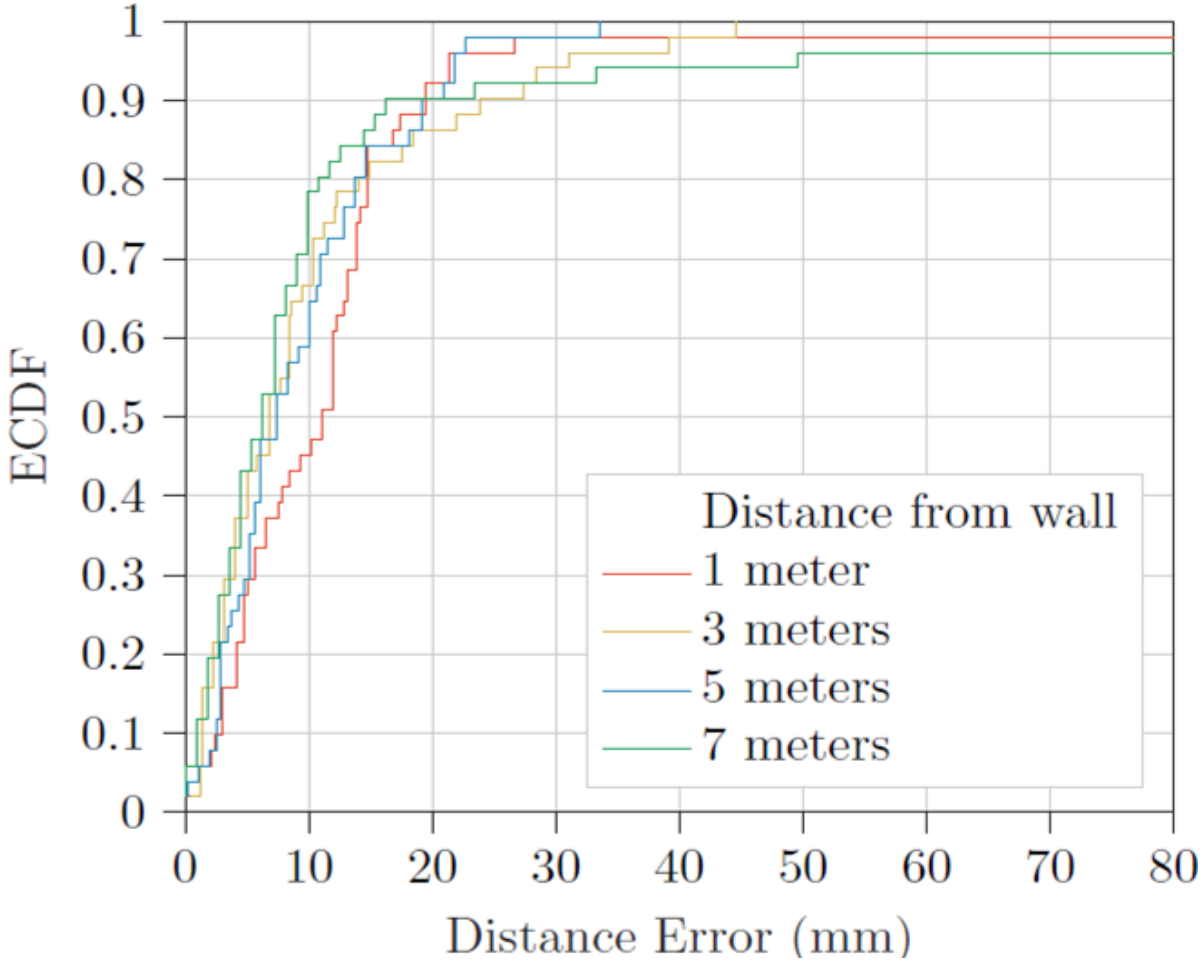
Channel State Information (CSI) is used in First Path Beamforming Training (FPBT) to derive the best path (azimuth and elevation) between the two stations. This can be used to estimate the position of environmental obstacles





Evaluation of mmWave sensing

- mmWave sensing information can be reused to perform tasks as simultaneous Localization and Mapping (SLAM)
- Achieves cm level accuracy in scenarios where other solutions do not work e.g. in situations with no light or fog where cameras cannot work
- Detects translucent materials that Lidars cannot detect



Picazo Martinez, P., Groshev, M., Blanco, A., Fiandrino, C., de la Oliva, A., & Widmer, J. (2023, November 21). waveSLAM: Empowering Accurate Indoor Mapping Using Off-the-Shelf Millimeter-wave Self-sensing. IEEE Vehicular Technology Conference, Hong Kong, China. https://doi.org/10.5281/zenodo.10171206

Benefits of ISAC

- Reutilization of the communication network to provide sensing service
  - Taking advantage of communication sites (large coverage)
  - Min/no effect on the communication
- New monetization opportunities for MNOs
- Potentially high volume of data collection and processing
  - Privacy-aware collection/processing
- Additional use of radio infrastructure until now used for communications
- Needed changes to architecture should be limited to not affect performance of communications

- Enabling sensing capability in the existing communication network would
  - Allow the exposure of network sensing services,
  - Ensure privacy checks on triggered sensing operations and avoid exposing privacy sensitive data
  - Support the fulfillment of sensing QoS requirements i.e., defined by the sensing clients
  - Enable efficient coordination of BS and UEs that will be involved in a sensing operation

Compute offloading

Compute offloading Description

- Compute offloading, or device compute offloading: A mechanism to move computation from one device to another with more suitable capabilities (see Fig. 1)
  - Offering offloading of critical tasks or functions to app developers, exposed as a network service
  - Supports offloading of customized application modules
  - Taking advantage of network resources, processes, and information.
  - Offering in-network computation
- Motivations are for example:
  - reducing computation times
  - balancing compute and energy tradeoffs, see Fig.2
  - balancing performance and cost tradeoffs
  - reducing device heat
  - facilitating synchronization and coordination
  - optimizing network utilization
  - increase scalability and availability

Compute offloading

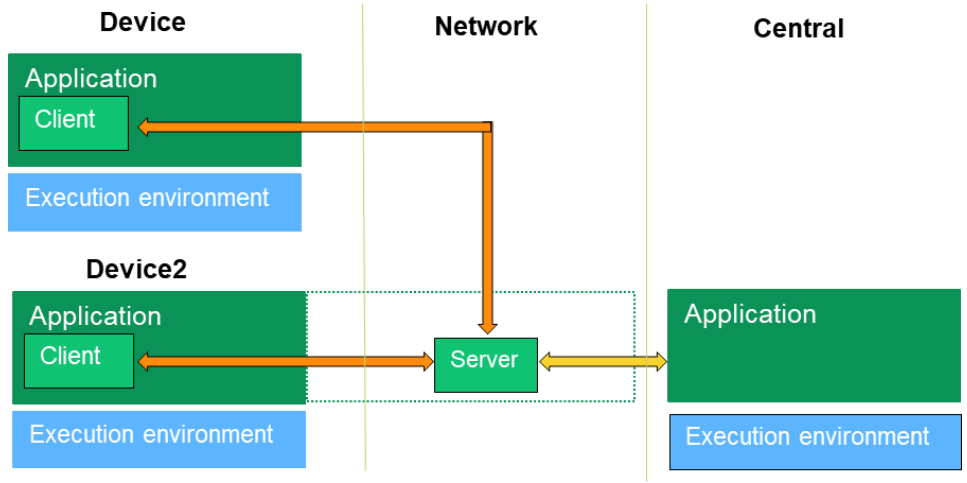


Fig. 1: Device offloading principles: move computation from a device to the network

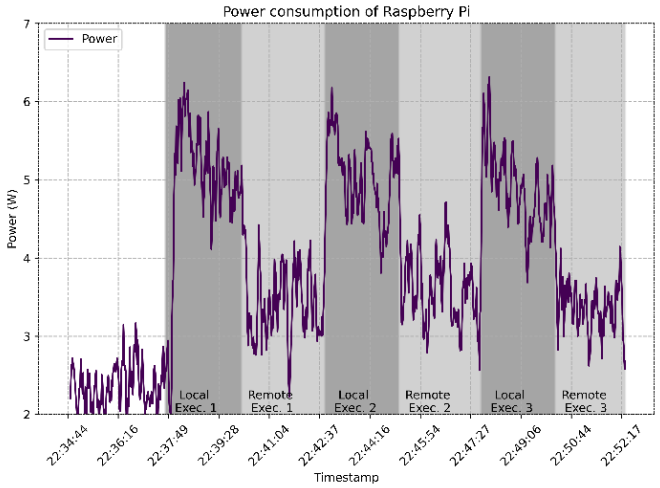
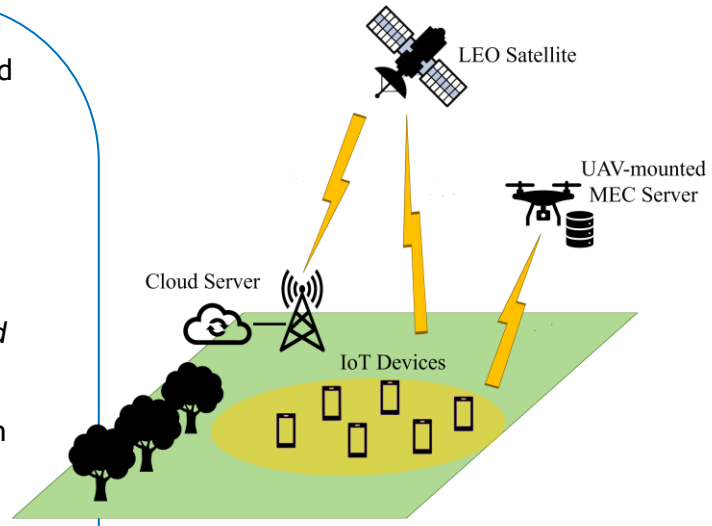


Fig. 2: Device compute offloading example: power saving of the offloading device (local vs. remote execution phases), see [HEX224-D33] for more results.

Compute offloading application in integrated TNs/UAV Evaluation

- The objective is to investigate methods on how to offload computation tasks from IoT devices to UAVs or satellites:
  - UAVs and NTN is used to achieve coverage in the rural area.
  - The UAVs are equipped with MEC servers.
- Distributed solution concepts investigated:
  - **Nash Equilibria (NE)-based solution** - each IoT device aims to *minimize its utility, i.e., expected aggregate time and energy overhead (i.e., the mean)*.
  - **Satisfaction Equilibria (SE)-based solution** - each IoT device aims to achieve a *minimum possible acceptable value for its expected aggregate time and energy overhead (i.e.*, minimum value).
- For the performance evaluation:
  - Consider 10 IoT devices with increased intensity of compute [#bits] and task computation [CPU-cycles/bit]
- Conclusions from the evaluations
  - With increasing device intensity, a smaller portion of its task is offloaded to the UAV for both types of equilibria, see Fig. 1.
    - Higher-intensity tasks incur greater overhead to IoT devices in case of UAV failure.
  - The SE achieves a more balanced utilization than NE
  - However, the NE method results in lower energy consumption than the NE for the IoT devices, see Fig. 2.



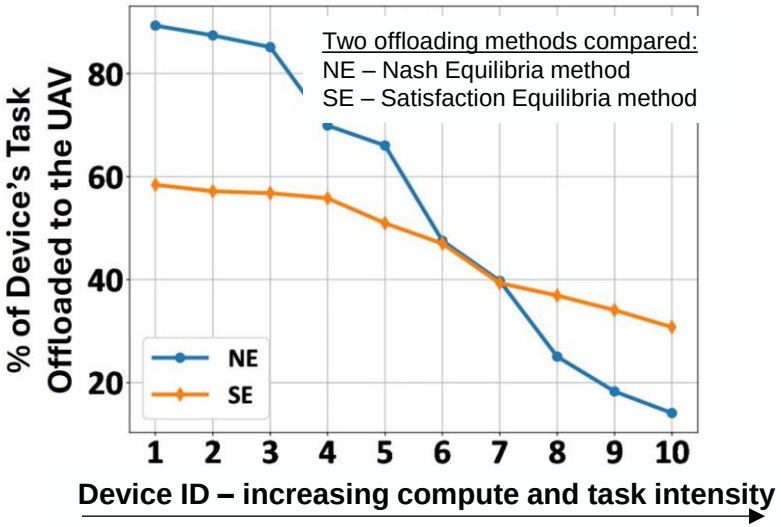


Fig. 1: Percentage of offloading to UAV by each device under different equilibria.

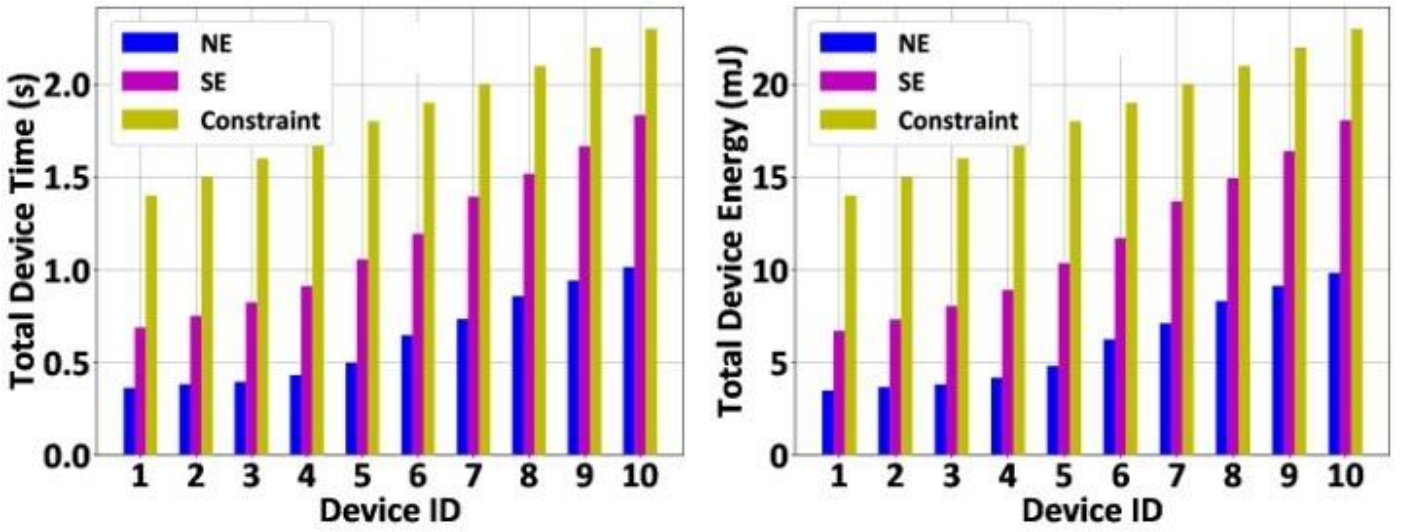


Fig. 2: Total time (left) and energy (right) overhead experienced by each device under different equilibria.

Loose integration of a compute offloading service Protocols & API

- Below is **a draft architecture of the compute offloading (Fig. 1)**
  - Using 3GPP SA6 EDGEAPP as a basis
  - All Computational offloading components act as 3GPP application functions (AF)
  - 3GPP SA6 EDGEAPP would not satisfy all requirements, but would be a good start for common issues like service discovery, user plane influence (QoS), and other network interaction requirements
  - Requires no significant changes to RAN protocols or NAS but new interactions with the compute network need to be developed (e.g., registration, see to the right)

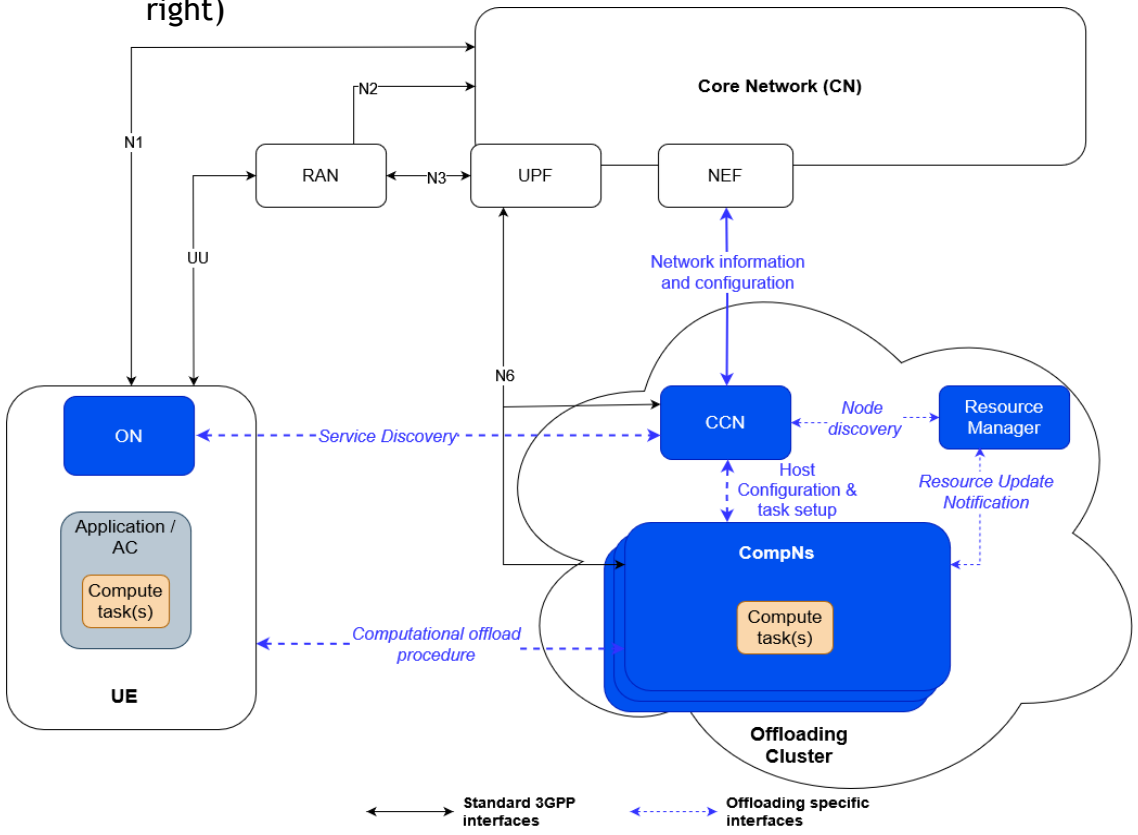


Fig. 1: Draft architecture: offloading components interacting with the network

- **Initial compute offloading registration procedure to Compute Control Node (CCN), see Fig. 2.**
  - Different nodes register either for compute offloading **request** (i.e., ON) or **offering compute support** (i.e., CompN) or both.
  - The Compute Resource Management Service (CRMS) in CCN is responsible for checking the Node registration request and whether the Node is allowed to provide or request computation support.
  - Each node can update its state with the CRMS and switch between requesting or offering compute resources.
  - Each node can deregister from the CRMS when it no longer wishes to request offload or provide compute support.

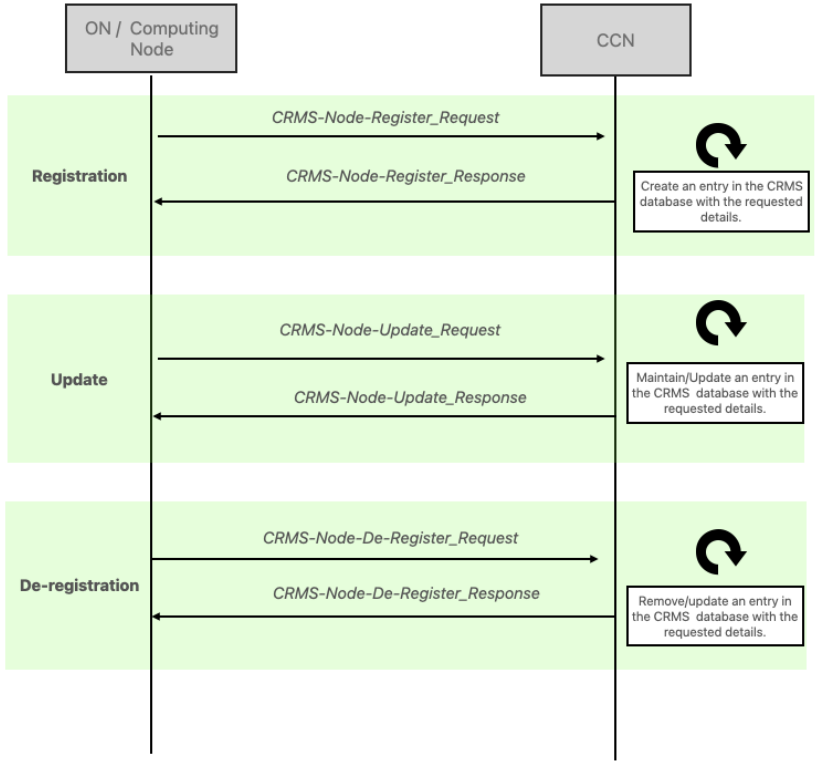
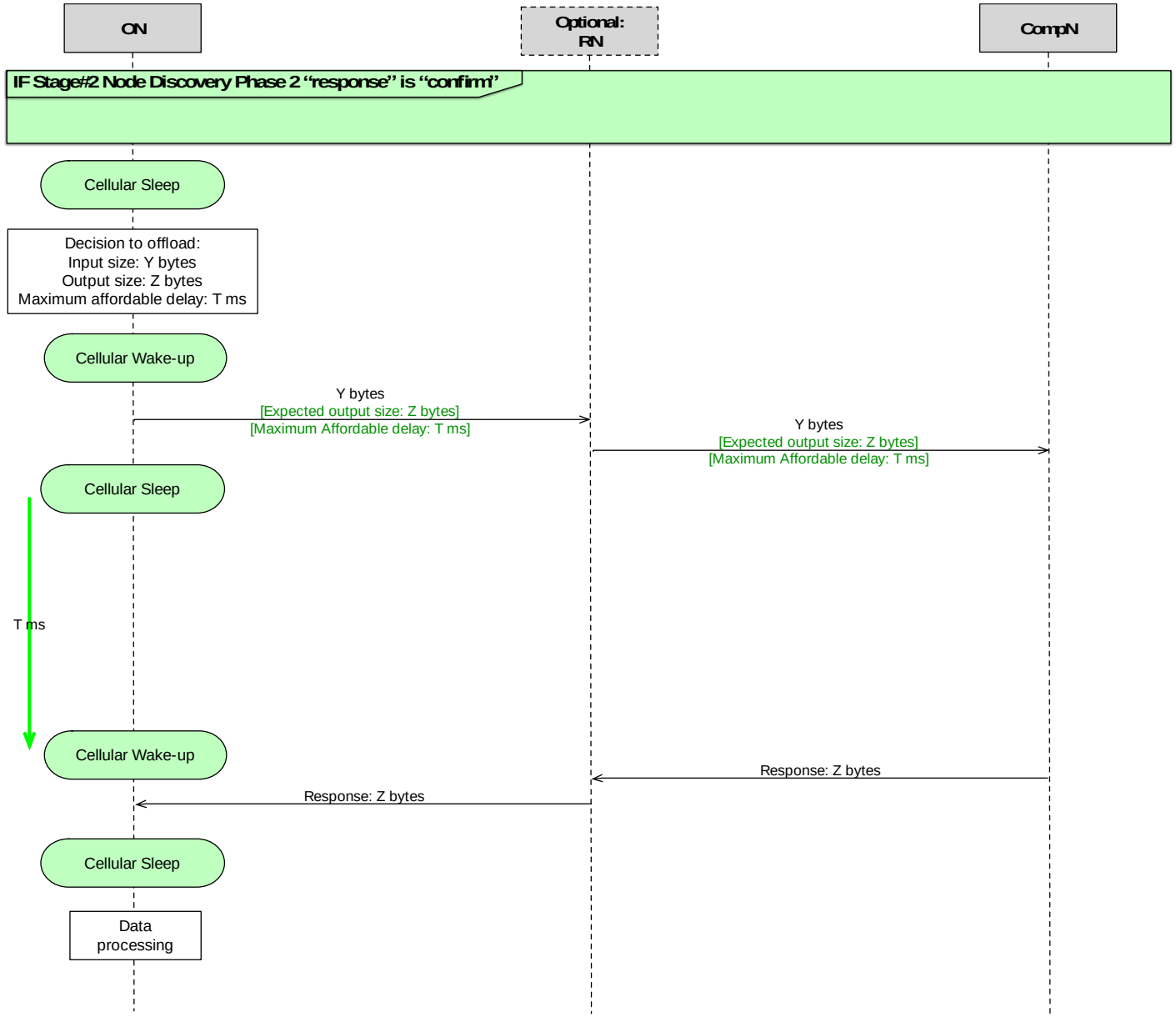


Fig. 2: The initial registration procedure in CRMS.

Compute Offload Procedure Protocols & API

- High-level messaging and Initial registration
- ON sends a compute task Y
- ON expects the compute result Z within T
- ON can enter Power Saving state after Y was sent and until Z is expected
- After receiving compute results Z, data processing starts immediately.



Compute offloading
Key take-away

- Compute offloading moves computation from one device to another with more suitable capabilities
- 3GPP SA6 EDGEAPP is one working assumption for common architecture.
  - Requires no significant changes to RAN protocols or NAS but new interactions with the compute network need to be developed like
    - service discovery,
    - user plane influence (QoS),
    - compute offload registration
    - and other network interaction requirements
- We show solutions for compute offloading registration and power saving techniques for computing devices
- Compute offloading between IoT devices and UAVs
  - Different offloading approaches evaluated, aiming to reduce energy consumption of the IoT devices

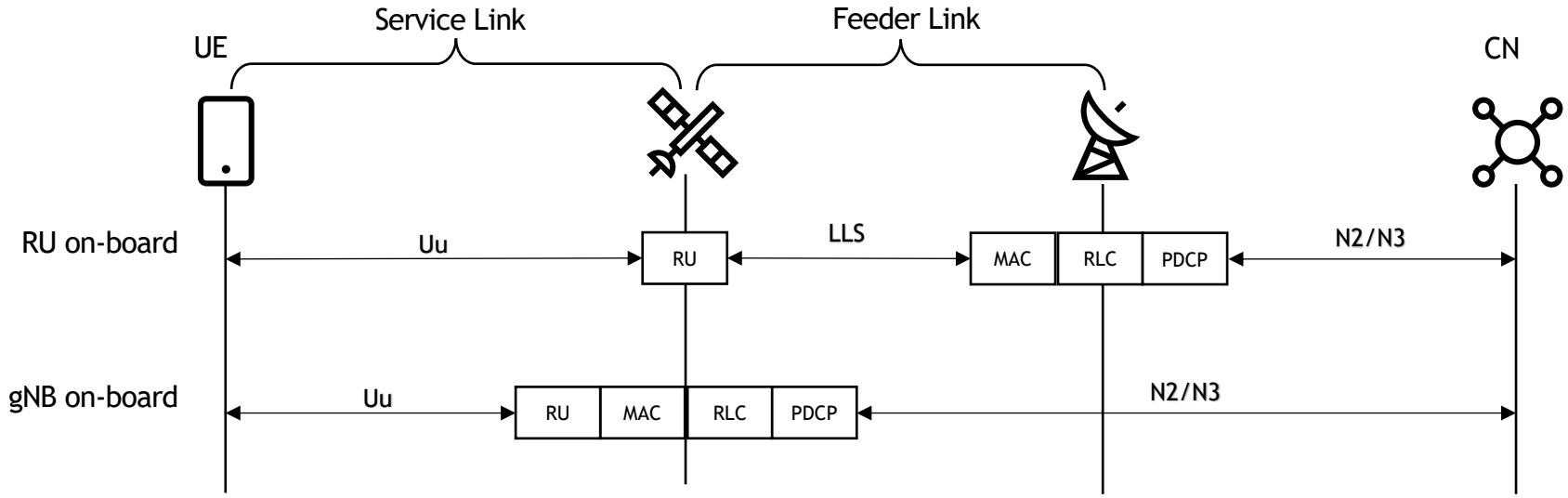
- Decreases the power consumption for a task of the offloading devices.
- Can utilize power efficient computing nodes in the network
  - Note that offloading may also increases the traffic overhead over the air interface to some extent
- Inclusion
  - Compute offloading may allow all units to have processing techniques with high computability
- Relation to other enablers
  - MLOps in Task 3.1 and the PoC B.2 shows that energy consumption reduces with the offloading of (protocol stack) layers from 22J to 5J (see slide 24)
- Relevant use case:
  - Seamless Immersive Reality
    - E.g., devices offloading immersive reality processes

Flexible topologies and new access methods

Network of Networks (NoN)

NoN Description: Non-Terrestrial Networks (NTN)

- One of the main technologies for achieving the 6G service coverage requirements
- NTN architectural options (see Fig. 1)
  - gNB on-board: The whole gNB is located on the satellite
  - Remote Radio Head (RRH) on-board: The RRH is on the satellite, the rest of the gNB functions are on the ground
- Inter-Satellite Link (ISL)
  - Focus on the usage of optical switches for ISL
  - Investigated if Radio Network Layer (RNL) / Transport Network Layer (TNL) protocol stack can be optimised.
- Terrestrial Network (TN) - NTN dual connectivity architectural options
  - Master Node: TN; Secondary Node: NTN
  - Master Node: NTN; Secondary Node: TN



Service Link: UE-Satellite link

Feeder Link: Satellite-ground station

Fig. 1. NTN architecture options

NoN Description: Trustworthy flexible networks

- Development of dynamic, adaptable network structures for 6G, formed based on trust established between the nodes
- Subnetwork architectural options (Fig. 1):
  - (a) Management Node Standalone (MgtN-SA)
    - UE is connected only to the MgtN
    - MgtN has access to the core network, via the BS
  - (b) Base Station - MgtN Dual Connectivity (BS-MgtN DC)
    - Main connection: BS; Secondary connection: MgtN
  - (c) BS-SA with MgtN Support
    - UE uses MgtN to support in different CP functionality
- AI-embedded solutions could enhance energy efficiency, reliability, and scalability of these networks

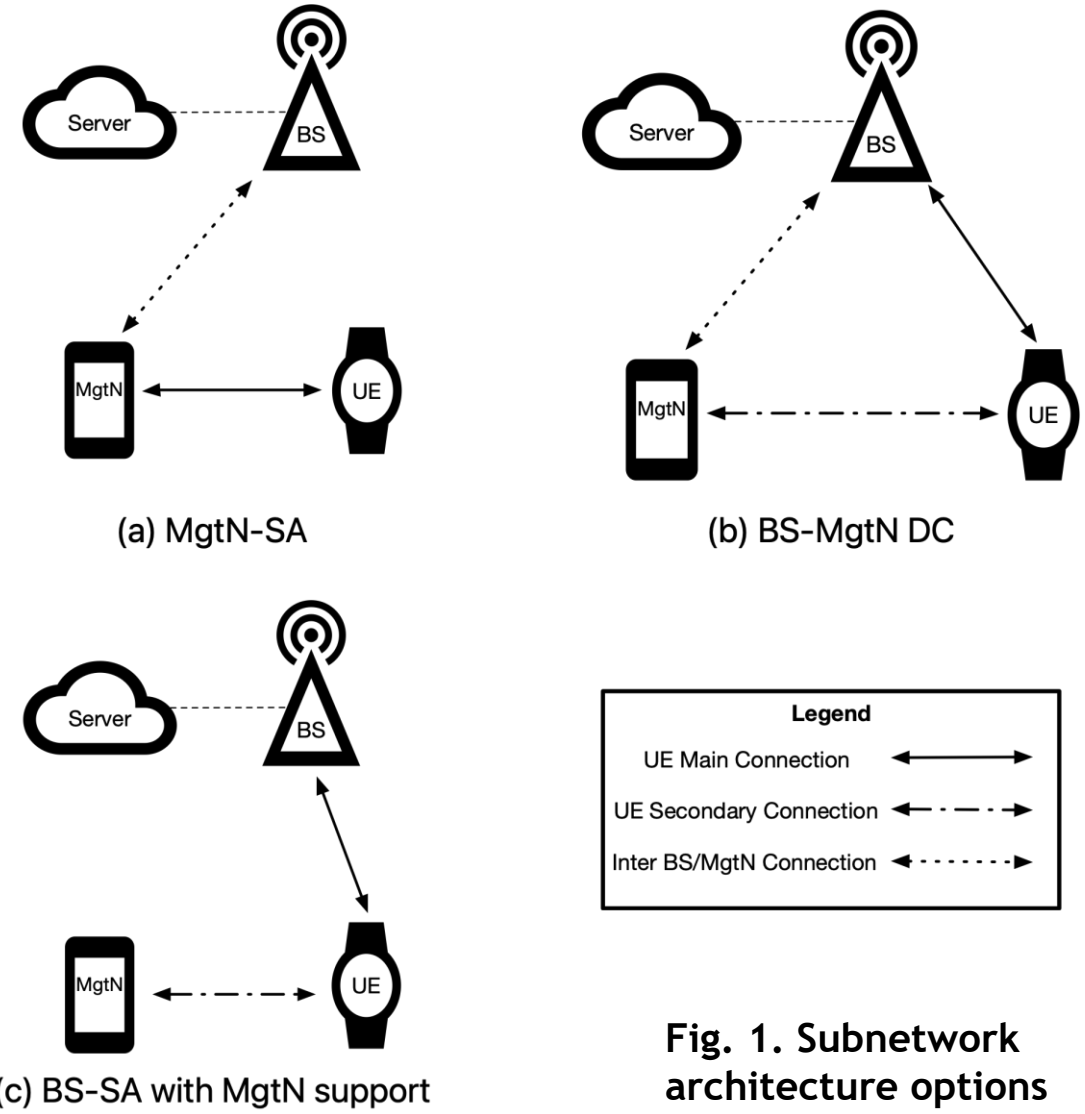


Fig. 1. Subnetwork architecture options

NoN Evaluation: Coverage Inequality Index

A new *quantitative index* that combines a 6G coverage map and a rurality map into a single scalar index that reflects coverage inequality

Features/properties of the index:

- compare fairness in regions
- compare fairness over years
- compare fairness of 6G operators
- a new regulation tool
- a 6G network planning tool

Main KPI improvements:

- *Large-area Coverage Inequality and Fairness*: New means to measure coverage inequality for fully integrated NTN/TN networks. A critical tool for 6G ubiquitous coverage

Applicable WP2 Design Principles:

- **#3 Flexibility to Different Network Scenarios:** Adaptability to various network topologies (e.g., NTN)
- **#5 Resilience and Availability:** Architecture supports high resilience and availability with TN-NTN

Cellular coverage data from the Swedish National Regulator (PTS).
Population data from Sweden's National Statistics Agency (SCB).

Fig.1.
Example coverage map: local service quality in joint NTN/TN

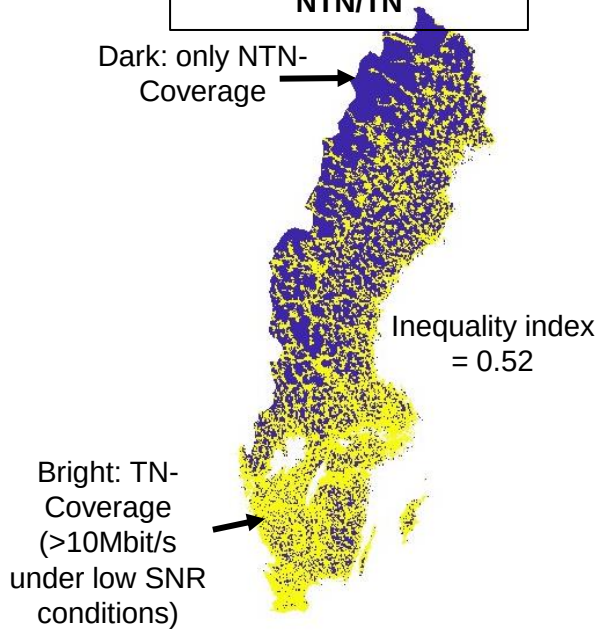
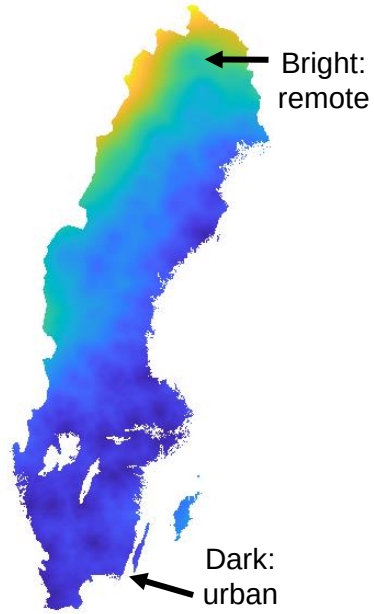


Fig. 2.
Rurality map: weighted distances to population clusters



Cellular coverage systematically favours urban regions over rural regions. Operators deploy networks where people live denser, rather than in rural/remote regions. It is desired to measure and quantify this lack of fairness across a population and across a region. There is a need to quantify the extent of this.

Assumptions: NTN coverage exists where no TN coverage exists.

A *new coverage inequality index* reflects the rural-urban service fairness for an NTN/TN-served large-scale region and reveals residual inequalities and remaining nuances of the '6G fully-connected world' vision. Full equality: index=0; Full inequality: index=1

NoN Evaluation: Trustworthy flexible networks

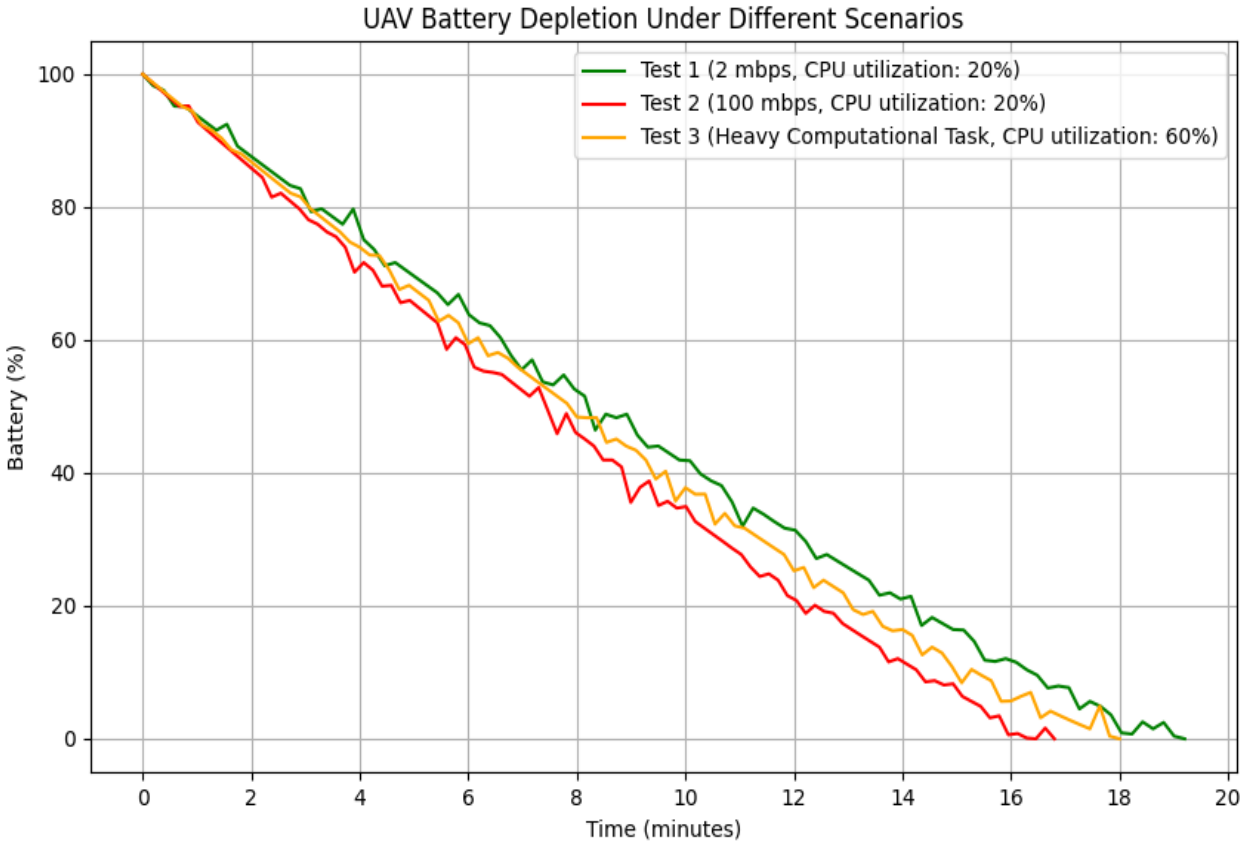
The development of dynamic, adaptable network structures for 6G enables real-time reconfiguration in remote and underserved areas. By prioritizing power consumption as a key metric, AI-driven algorithms adjust node positions and resource allocation to maintain efficiency and coverage. This ensures optimal performance and energy use, even during changing network demands. In this system, metrics of a drone that is serving one robot are monitored for various scenarios.

Main KPI improvements:

- *Energy Efficiency*: Dynamic optimization leads to reduced energy consumption.
- *UAV Autonomy*: High-data rate communication depletes battery faster than a heavy computational task performed locally
- *Latency/Reliability*: <15ms, >99.999% reliability in critical scenarios, based on PoC-based measurements in the lab
- *Adaptability and Flexibility*: Enhanced real-time network reconfiguration for underserved areas.
- *Reliability and Trustworthiness*: AI-driven trust management ensures reliable and secure communication.

Applicable WP2 Design Principles:

- **#3 Flexibility to Different Network Scenarios:** Adaptability to various network topologies (e.g., mesh networks)
- **#5 Resilience and Availability:** Architecture supports high resilience and availability with flexible topologies



Test 1: Transmission Load: 2Mbps, CPU Utilization: 22%, CPU cores: 7, Battery Autonomy: 0.32h

Test 2: Transmission Load: 100Mbps, CPU Utilization: 22%, CPU cores: 7, Battery Autonomy: 0.28h

Test 3: Transmission Load: 0Mbps, CPU Utilization: 60%, CPU cores: 7, Battery Autonomy: 0.30h

NoN Protocols and APIs: NTN

Architecture option	Dynamic function relationships	ISL / multi-hop	Capacity / performance	HW/SW impact on satellites	Standard impact
Transparent	Up to satellite operator	Satellite RF techniques (outside 3GPP)	Difficult to scale (outside 3GPP)	Relay amplifier with minimal HW impact	No impact
RRH on board	RRH-DU dynamic association to be implemented in LLS	Extension of LLS routing for multi-hop	The data over Lower Layer Split (LLS) using L1 split option "7-x" [38.801, Ch. 11] requires several times more capacity than the gNB onboard option, (depending on the actual LLS for NTN), this may be difficult to scale the LLS for ISL	Minimal	No functional split impact
gNB on board	- Dynamic gNB association to AMF - Handling of dynamic gNB configurations	Multi-hop via satellite using full RAN protocol stack	Control plane traffic and user plane traffic	High with full gNB	Impact to N2 and their L3 to support dynamic association

- Characteristics of the NTN architectural options

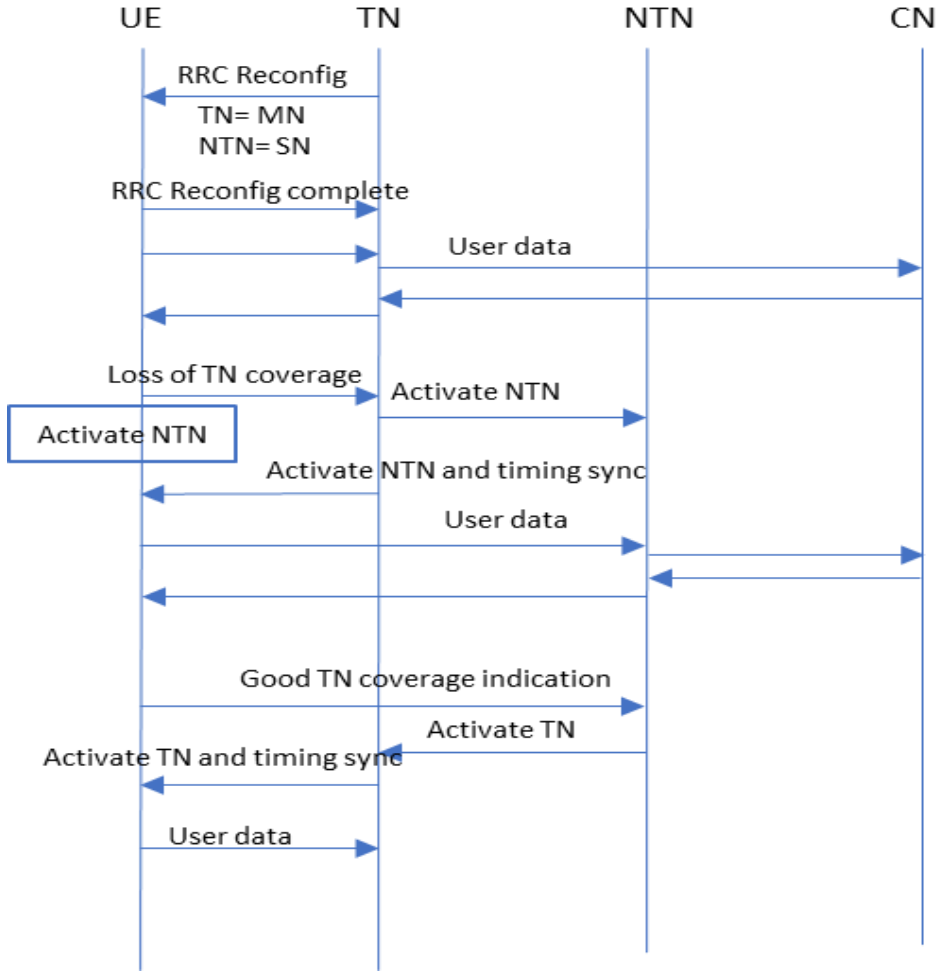
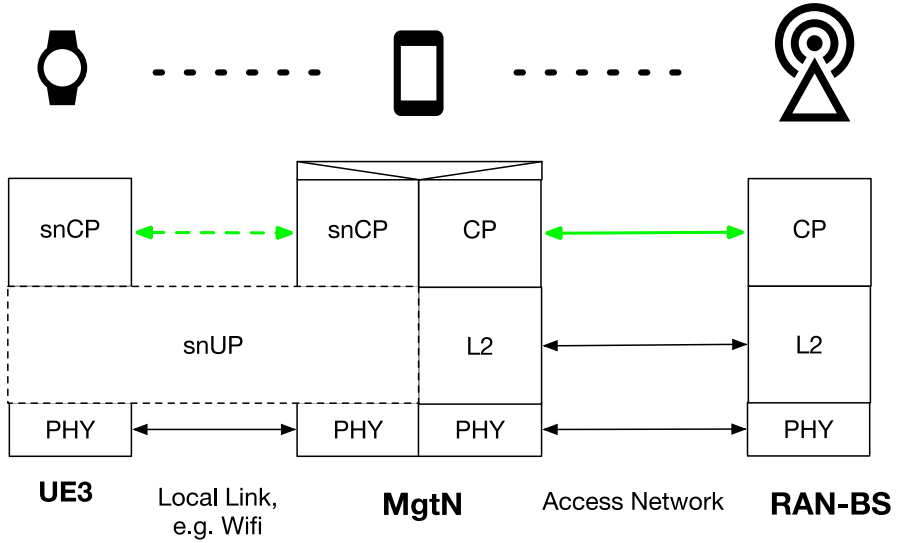


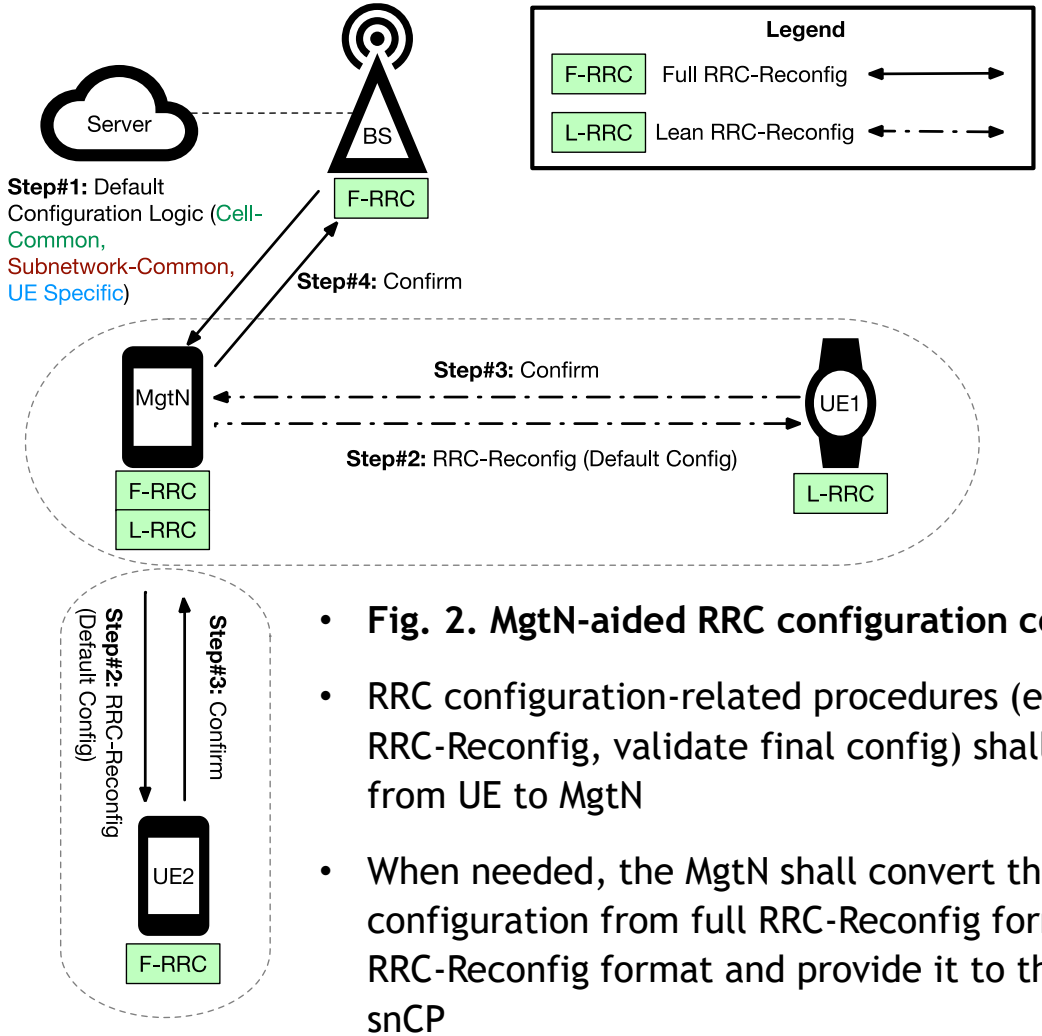
Fig. 1. TN-NTN DC with TN as MN and NTN as SN

- Activation of NTN SN due to loss of TN coverage (indicated by the UE)
- During NTN deactivation, the UE does not have to monitor NTN signals for mobility

NoN Protocols and APIs: Trustworthy flexible networks



- **Fig. 1. Subnetwork CP: UEs CP aggregated at the MgtN**
- UEs offload their control plane (CP) or part of it towards MgtN
- Subnetwork uses a new lightweight subnetwork CP (snCP) between UE and MgtN
- snCP is transparent to the NW and includes the configuration and procedures within the local subnetwork, as well as the offloaded UE CP information



- **Fig. 2. MgtN-aided RRC configuration concept:**
- RRC configuration-related procedures (e.g., decode RRC-Reconfig, validate final config) shall be delegated from UE to MgtN
- When needed, the MgtN shall convert the final configuration from full RRC-Reconfig format into lean RRC-Reconfig format and provide it to the UEs via the snCP
- Subnetwork-common configuration should be defined in 3GPP RAN2.

NoN Protocols and APIs: Trustworthy flexible networks

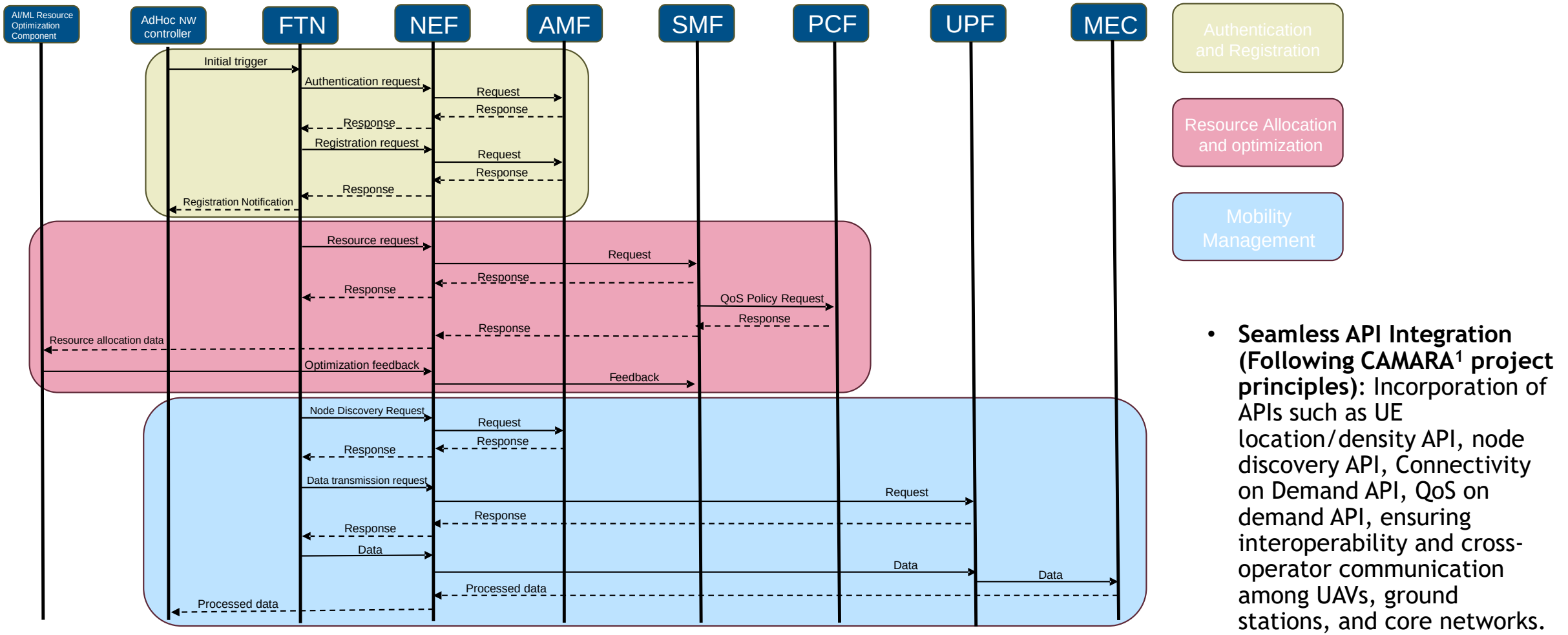


Fig. 1. Message sequence chart for authentication, registration, resource allocation and mobility management of a trustworthy, flexible and unstructured network

- **Seamless API Integration (Following CAMARA[1] project principles)**: Incorporation of APIs such as UE location/density API, node discovery API, Connectivity on Demand API, QoS on demand API, ensuring interoperability and cross-operator communication among UAVs, ground stations, and core networks.
- **Alignment with 3GPP**: TS23.501, TS23.222

[1] https://camaraproject.org/

NoN Key take-aways

- Multiple architectural options for NTN and subnetworks are needed to achieve the coverage requirements
- Novel procedures for NTN and flexible topologies to reduce UE complexity and enable seamless mobility
- Standards will be affected (e.g., RAN2)
- New index for quantifying coverage inequality, which can be used for network planning and by regulators
- Implications
  - New RAN equipment (e.g., satellites, drones) and types of UEs (e.g., MgtN) are required for the proposed systems to work
- Relation to other enablers
  - T2.2: Proposals affect the RAN protocols
  - T4.1: NTN
  - T5.4: Flexible topology architectures may be used for zero-energy devices

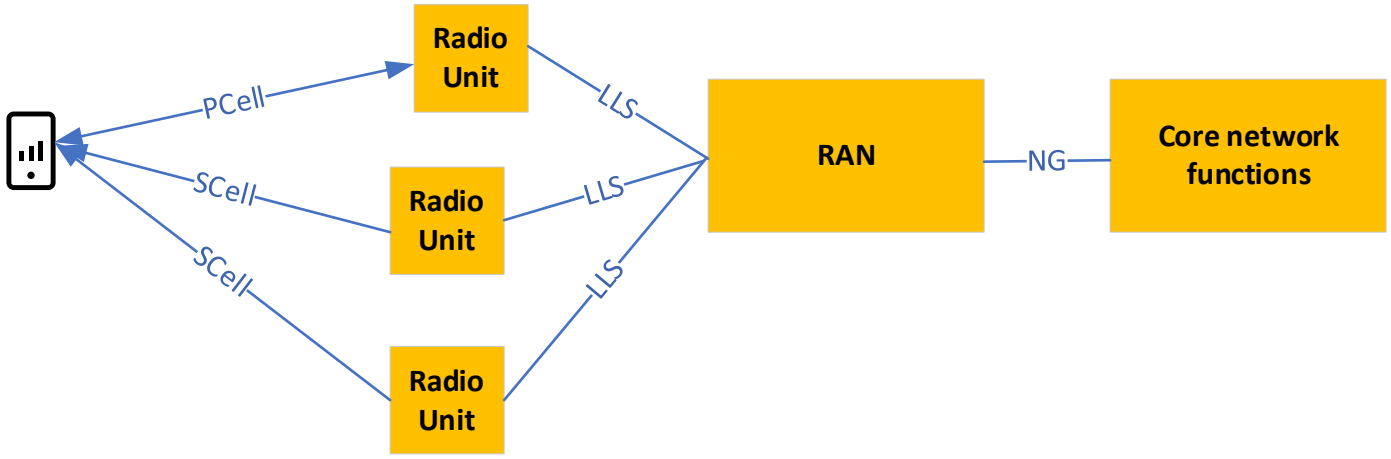
- **Related use cases and sustainability risk mitigations**
- **Ubiquitous network use case**
  - NTN and subnetworks will provide affordable global coverage to ensure digital inclusion
  - TN-NTN dual connectivity may enable finding a better balance between building a TN network and relying on NTN, to minimize the footprint
  - Coverage inequality index can quantify coverage (un)fairness and identify which remote area networks should be adapted to have essential service coverage
- **Seamless immersive reality use case**
  - Flexible trustworthy topologies will have procedures to ensure that data is exchanged between secure and authenticated nodes
  - XR devices can be part of a trustworthy subnetwork, allowing the network to offload procedures to their MgtN, resulting in lower power consumption in the system
- **Cooperating mobile robots use case**
  - Usage of flexible trustworthy networks will provide tailored and optimized solutions for specific industries

Multi-Connectivity (MC)

MC Description

- Multi-connectivity (MC) increases the reliability and enables aggregating data from more than one connection in certain scenarios (see figure)
- One goal with the 6G MC is to avoid having two similar solutions, such as Carrier Aggregation (CA) and Dual Connectivity (DC) and aim for one MC solution
  - Use CA as base for the new 6G MC solution
  - Improve CA by developing new features and procedures based on DC's features
- Another aspect is to develop an integrated WLAN and 6G solution for the indoor use case (e.g., in-home).
  - Applicable when the UE is in WiFi coverage of a WLAN Terminal (WT) and may or may not be in cellular coverage of a BS. The WT can act as a relay
  - The WT is connected to the BS wirelessly (e.g., via the Uu interface).
  - The WT-BS link may use a different Frequency Range (FR) than the UE-BS link.

Fig. 1. 6G multi-connectivity high-level architecture



The UE can connect via several non-co-located Radio Units and the Lower Layer Split (LLS) to the network, aggregating the resources from each RU

PCell: Primary Cell; Used for initial access and is the main cell in the master cell group
SCell: Secondary Cell; There may be one or more SCells configured in Connected mode. SCells can be activated / deactivated based on certain conditions (e.g., traffic)

MC Protocols and APIs: Carrier Aggregation (CA) & Dual Connectivity (DC)

- For simplicity, only one architecture option (i.e., CA or DC) should be standardized.
- The selected solution should have the benefits of both architecture options.
- Possible 6G CA improvements
  - Enhancements to use conditional handover (CHO) for CA.
  - Faster L1/L2 handover-based solutions for CA.
  - Increase the robustness by:
    - Removing PCell/SCell separation,
    - **PCell recovery via SCell (Fig. 1)**
    - Using CHO and a set of inactive connections for fallback.
- Possible 6G DC improvements
  - Improve the UE feedback in flow-control
    - Master Node reacts “proactively” to avoid “stalling” the UE buffer
  - Decouple Downlink (DL) and Uplink (UL)
    - One UL connection and two DL connections should be allowed
    - Control signaling such as ACK/NACKs only via one UL connection, need to be routed via Xn
- The improvements mainly impact 3GPP RAN2.

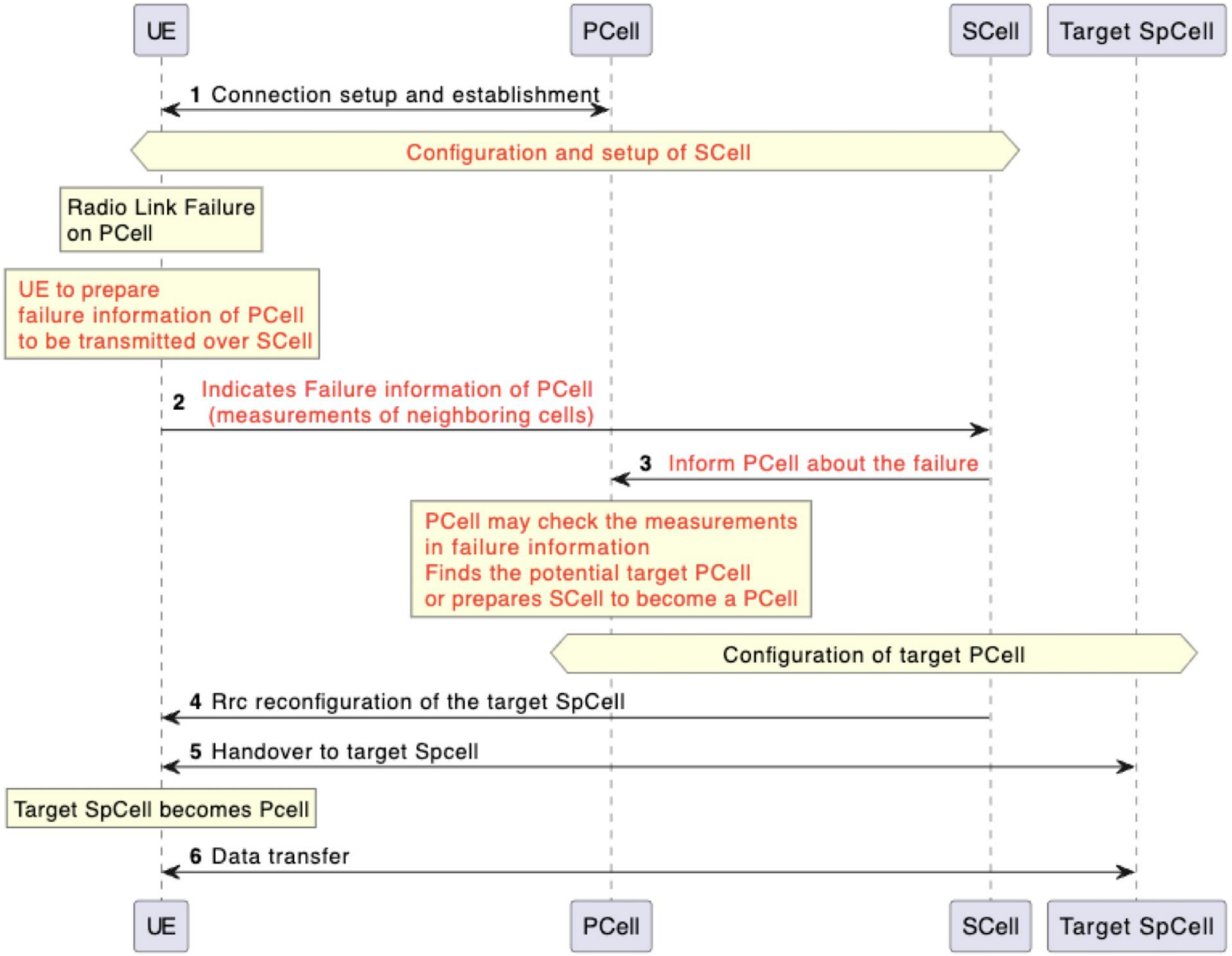


Fig. 1. PCell recovery via SCell

MC Protocols and APIs: WLAN-Cellular Aggregation (WCA)

- The **WLAN Relay Adaptation Protocol (WRAP)** is used when a packet is transmitted over the WT.
- The WRAP header includes information needed for the nodes to be able to map the packet to the correct destination and radio bearer
- A security layer may be optionally added for the WLAN Terminal (WT) - Base Station (BS) link
- These layers should be specified in 3GPP RAN2

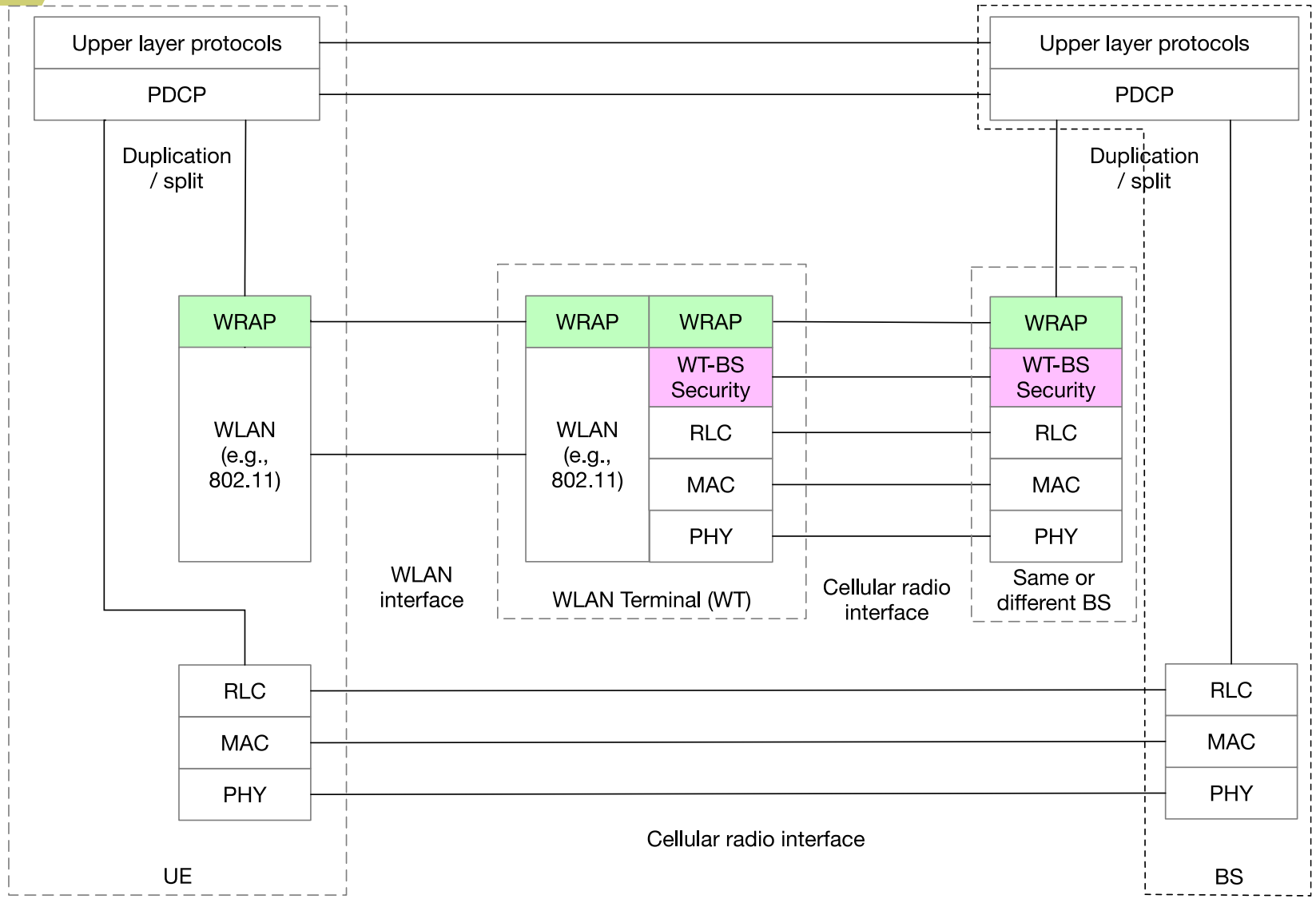
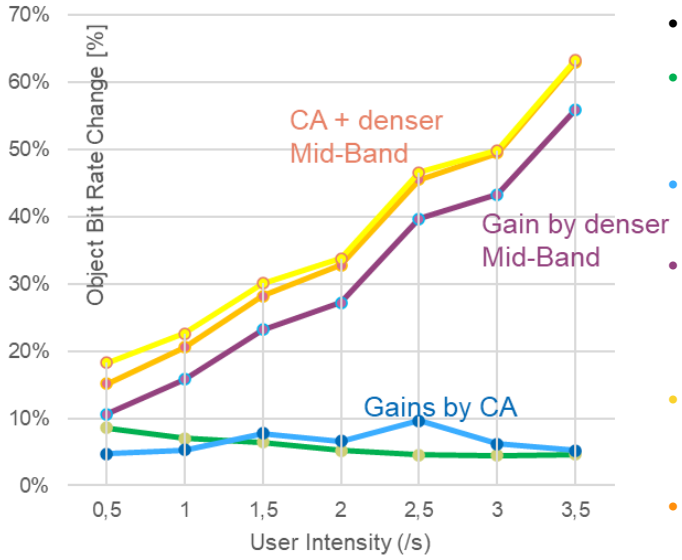


Fig. 1. UE, WT and BS protocol stacks for WCA

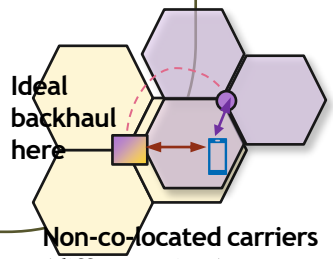
MC Evaluation



- **Deployments:**
- Non-co located, system wide scheduler (CA)
- Co-located, site scheduler (CA)
- Densification of mid-band cells (ISD from 500m to 288m)
- Midband densified, CA, site scheduler
- Midband densified, CA, system-wide scheduler

Fig. 1. System simulations for different multi-connectivity deployments options

- Ideal backhaul, PCell: 3.5 GHz TDD with 100 MHz BW
- SCell: 800 MHz FDD with 10 MHz BW
- 100 MB FTP DL, 80% Indoor Users
- CA with and without wide-scheduler have gains in similar range (5-10%) regardless of load
- Densification high gain at high load



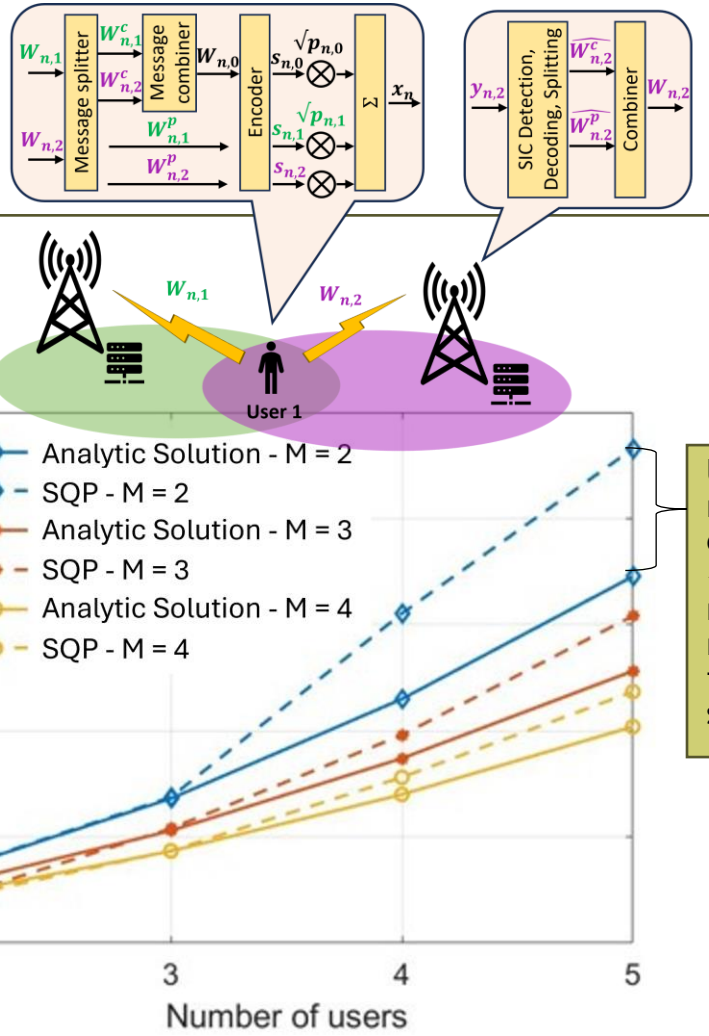


Fig. 2. System evaluation of the Rate-Splitting Multiple Access (RSMA)

- User delay: Sum of the communication delay and the remote processing delay of the offloaded task on the server.
- Downlink multi-server offloading for each user
- Compare the proposed analytical solution with the Sequential Quadratic Programming (SQP)
- 5 MHz BW, 2-5 users and M = 2-4 servers

MC Key take-away

- Only one architectural option (e.g., only CA) should be defined in the specifications
- Proposed CA improvements inspired by the DC advantages and vice versa
  - PCell recovery via SCell solution
- Coverage extension or increased reliability via WLAN-based UE Relay for Remote UEs
- Standards will be affected (e.g., RAN2)
- RSMA technique to allow low-delay multi-server offloading for each user
- **Implications**
  - Timing synchronization can be an issue for non-co-located cells using CA instead of DC
  - New device type needed: WLAN terminal with WCA support
- **Relation to other enablers**
  - T2.2: Proposals affect the RAN protocols
- **Related use cases and sustainability risk mitigations**
- **Ubiquitous network use case**
  - Increased coverage, high service availability (i.e., percentage of time the service can be delivered) and high reliability (i.e., success of transmission) by having different paths to the NW and/or by using different radio access technologies (e.g., cellular, WLAN)
- **Human Centric Networks**
  - User devices would be able to reuse their WLAN-related components
- **Seamless Immersive Reality use case**
  - Given the high data rate requirement, CA/DC could help the device and the network to complete the session faster and hence switch to low-power mode sooner

Context-aware management

Context-Aware Management Description

- **Description**
  - E2E context-aware management enables each network and compute component to dynamically adapt to the context to ensure the expected E2E QoS for the services, by leveraging on effective automation and orchestration mechanisms.
- **Main goals and objectives**
  - Guarantee the QoS of a certain network slice by creating an abstract view of transport resources and network functions to simplify network management as the network scales (see Fig. 1).
  - Implement context-aware path selection, switching, and packet processing using P4 programmability based on the context provided by an SDN Domain Controller.
  - Design optimal semantic orchestrators for robotic use cases [HEX223-D21] to maximize the number of allocated robotic functions in the system while minimizing the consumed energy at the robot ends.
  - Perform intelligent task offloading decision-making and seamlessly switch between different computing options, such as (a) delayed computing and (b) approximate computing to minimize the time and energy consumed across the network.

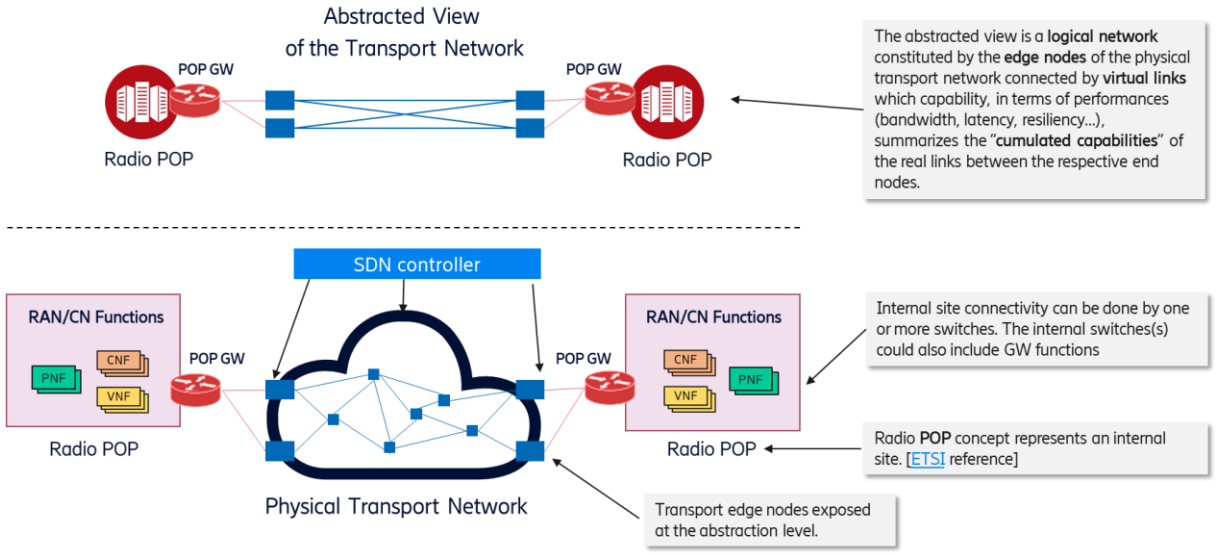


Fig. 1. High-level overview of transport network abstraction toward context-aware management.

Context-Aware Management Evaluation

Main KPI improvements:

- *Adaptability and Flexibility*: Extraction and utilization of the context allows for adaptable and flexible reconfiguration of the network tailored to the application.
- *Improved resource utilization: C*ontext-aware management of transport and compute networks enables more tasks to be handled with the same resources.
- *Energy Efficiency:* Context-aware switching and computation offloading reduce energy consumption at end-user devices, while context-aware management of end-to-end transport and compute networks enables more tasks to be handled with the same resources, improving overall energy efficiency.
- *Reliability*: Context-aware management guarantees the QoS requirements of applications by intelligently reconfiguring network slices.

Applicable WP2 Design Principles:

- **#2 Full Automation and Optimization:** Automated and optimized resource utilization by extracting and utilizing context from the network.
- **#3 Flexibility to Different Network Scenarios:** Adaptability of transport and compute network infrastructure based on the application context and capabilities of end-user devices.
- **#4 Network Scalability:** Efficient abstraction of transport resources and network functions that simplifies network management, especially as the network scales.
- **#10 Minimizing Environmental Footprint and Enabling Sustainable Networks:** Efficient resource utilization across the network and reduced energy consumption at the end-user devices owing to smart and flexible transport and compute network management.

Context-Aware Management Evaluation: Delayed vs approximate computing

- Consider users offloading tasks of an application (e.g., object recognition, video editing, natural language processing) that are diversified in (a) computing intensity, i.e., CPU cycles to be executed, (b) tolerable processing delay, and (c) acceptable loss in accuracy.
- To maximize the efficient utilization of the computing resource across the continuum, while minimizing the time and energy consumed, consider that the users have the following two options (see Fig. 1):
  - **Delayed computing:** The whole task is transmitted to the edge and then forwarded to the cloud for exact computation, experiencing a delay due to forwarding from the edge to the cloud.
  - **Approximate computing:** The data of the task are initially compressed and then, a certain percentage of the task is transmitted to the edge for approximate computation. While this approach accelerates execution, it reduces the achieved accuracy.
- Consider the following notation:
  - k_n: number of tasks generated (Poisson process) for each user n
  - x_n: number of tasks offloaded for delayed computing
  - s_n [%]: data compression percentage for approximate computing
- The problem of joint computation task offloading x_n and compression percentage s_n optimization for each user is formulated and solved in a distributed manner as a Game in Satisfaction Form, concluding a Satisfaction Equilibrium solution. Each user targets to satisfy a minimum acceptable value for its time and accuracy requirements.
- For increasing amount of data and computation task intensity in CPU-cycles/bit with the user ID in the horizontal axis, smaller amounts of tasks are offloaded to the cloud so that each user can meet its latency constraint by avoiding the high transmission time. For the same reason, the data for approximate computing are compressed more, i.e., a lower portion s_n of the task is offloaded (see Fig. 2).

Up to ~20% reduction of processing time and ~15% increase of accuracy in particular scenarios

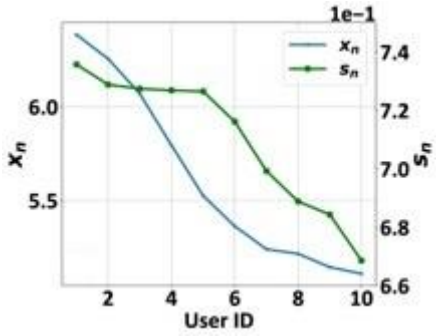


Fig. 2: Each user's number of tasks offloaded for delayed computing and compression percentage for approximate computing.

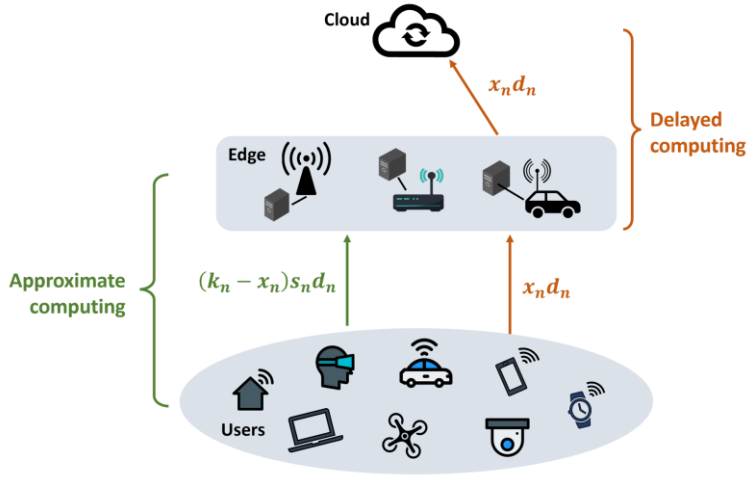


Fig. 1: Overview of delayed vs approximate computing framework.

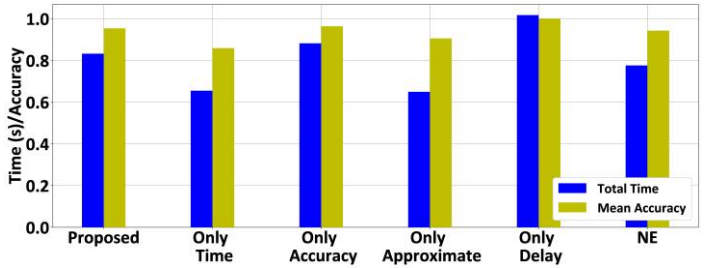


Fig. 3: Total processing time and task accuracy of the proposed framework.

- Compared to other baseline scenarios, (a) considering only time constraints, (b) only accuracy constraints, (c) only the approximate computing option, (d) only the delayed computing option, and (e) the Nash Equilibrium solution, the proposed solution yields the best time vs accuracy tradeoff (see Fig. 3).

Context-Aware Management Evaluation: E2E context-aware RAN semantic system optimization

- This work will be finalized in D3.5.
- Consider a **robotic application** π, such as security surveillance or packet delivery, characterized by certain KPIs, e.g., latency, accuracy
- A **function** $f \in F_\pi$ contributes to the execution of application π under **configuration** $c \in C_f$, using a certain set of components, e.g., autonomous navigation with LIDAR or depth cameras (see Fig. 1).
- There exist different **offloading policies** $p \in P$: (a) full offloading to the edge, (b) local execution, and (c) local preprocessing and then, offloading to the edge.
- The goal is to optimize the resource usage in the system, by minimizing the
  - consumed energy at the robot end,
  - allocated resources per robot function at the edge.
- To this end, the following parameters are optimized:
  - the number of robotic functions allocated in the system, considering a total amount of K resource types,
  - the data compression factor.

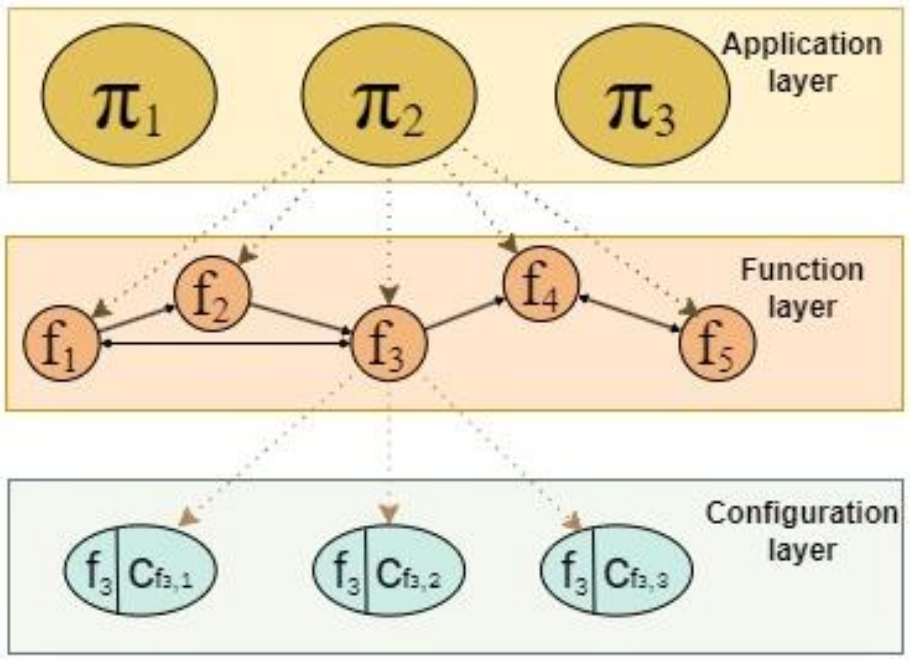


Fig. 1. High-level overview of context-aware RAN semantic system optimization for robotic use cases.

Context-Aware Management Protocols and APIs: Context-aware resource orchestration

- To enforce the context-aware transport, a **resource orchestrator** creates an **abstracted view** of the transport resources and triggers the **SDN transport controller** for resource handling to satisfy the QoS associated to a slice. It also performs E2E admission control to ensure the expected QoS for active and incoming services.
- An E2E **service orchestrator** places all network functions on the abstract view to guarantee the QoS of the considered slice (see Fig. 1).

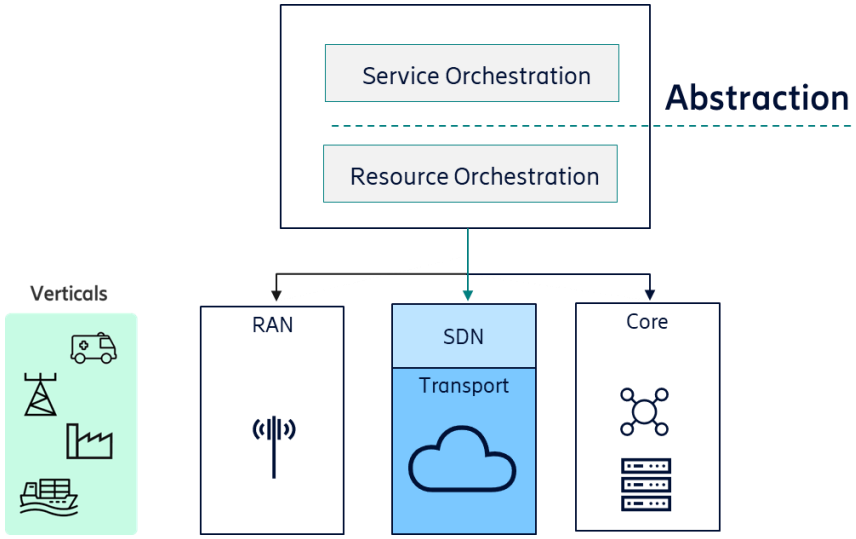


Fig. 1: High-level overview of protocol and APIs for context-aware resource orchestration.

- An example of transport network abstraction is presented in Fig. 2, where each link of a transport network, from a **node X** to a **node Y** is characterized by three parameters: a cost parameter, the link supported bandwidth (throughput), and the maximum latency.
- The definition of "cost" is out of scope of this example. Cost and latency are considered cumulative across links while the bandwidth of a sequence of links is equal to the minimum bandwidth among such links.

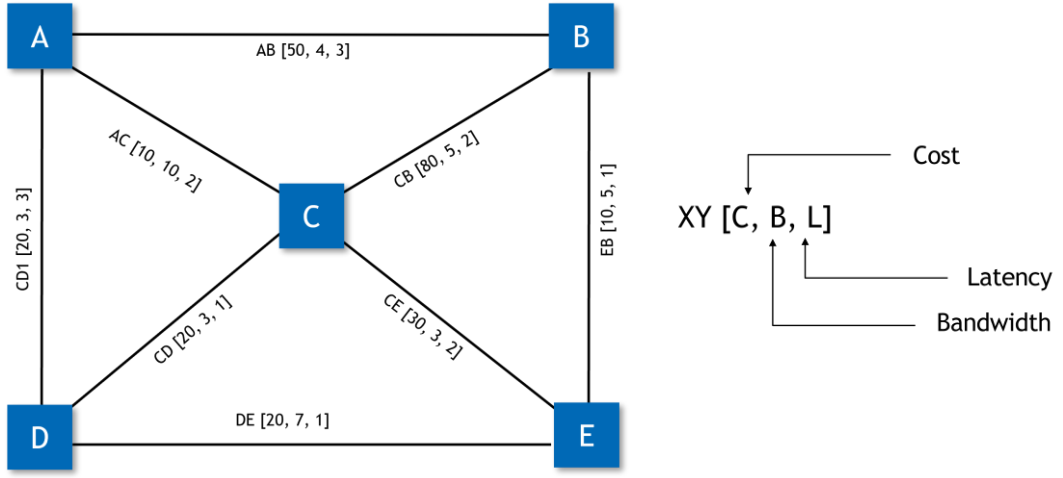


Fig. 2: Example of transport network abstraction.

Context-Aware Management Implementation: Programmable and context-aware transport

- The aim is to offer flexibility in path selection based on the context in the border nodes of the network, which further means:
  - To specify the path selection based on the context provided by an SDN Domain Controller;
  - To implement context-aware switching using P4 programmability (Fig. 1);
  - To implement context-aware packet processing pipeline (Fig. 2).

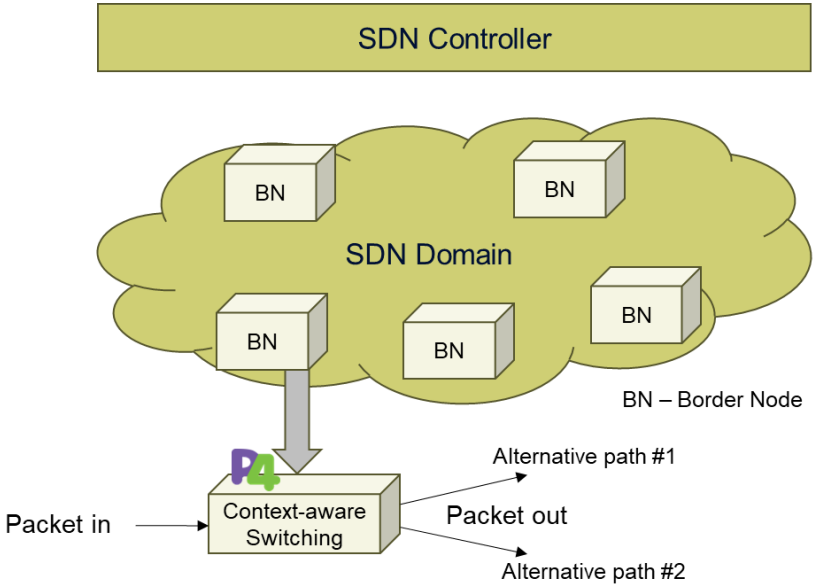


Fig. 1. High-level overview of context-aware transport implementation.

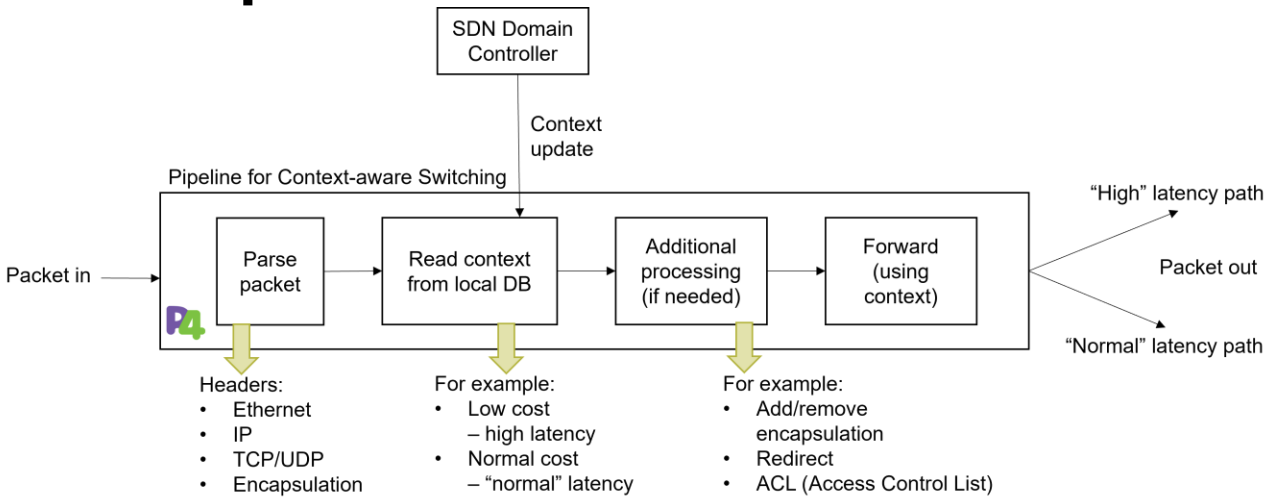


Fig. 2. Details of context-aware switching.

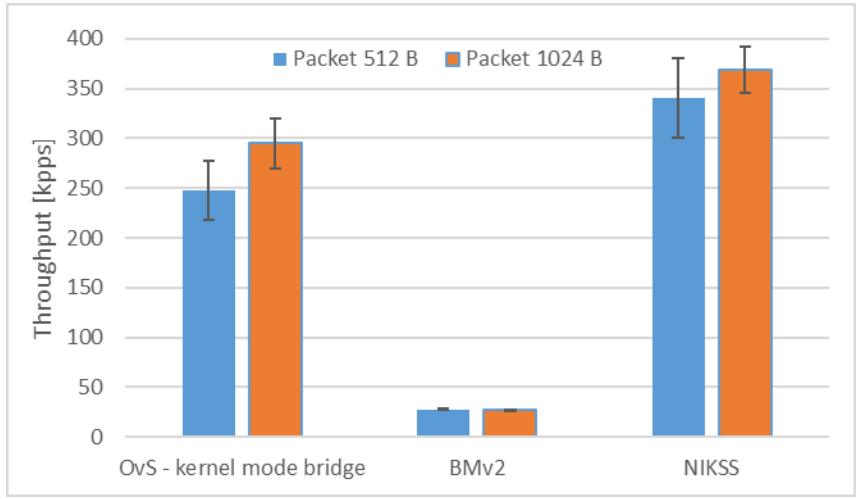


Fig. 3. Comparison of switches' performance for context aware transport

- P4 programmable switches give us more flexibility to offer context aware transport
- and, depending on the switch (e.g., NIKSS), comparable performance to Open vSwitch (OVS), Fig. 3.

Context-Aware Management Key take-away

- Optimized use of transport network infrastructure and packet switching based on the context.
- Flexible allocation of edge resources via computation offloading based on the context.
- Different computation offloading options available: (i) delayed computing and (ii) approximate computing.
    - Overall goal is to maximize the number of served users and tasks over the same resources, while guaranteeing their QoS requirements.
- **Implications**
    - New APIs are required to enable different devices to (i) indicate their state (context), (ii) exchange information (e.g., targeted KPIs, QoS) among computational nodes, and (iii) orchestrate network resources based on the available context.
    - Cross-layer interaction is required for different network components to become aware of the context (e.g., RAN, CN, transport), implying signaling and synchronization.
    - Resource allocation and orchestration mechanisms that operate even when incomplete or partial context is available are required.
- **Relation to other enablers**
    - T3.1: MLOps and AIaaS tools can be leveraged to enable distributed cross-network decision-making and optimization using AI/ML techniques.
    - T3.5: Extreme-edge connection for the context-aware RAN
- **Related use cases and sustainability risk mitigations**
- **Cooperating Mobile Robots**
    - Utilizing the context within the network will provide tailored and optimized solutions for specific industries, offering gains in terms of *social and economic sustainability*.
    - Increased production flexibility may, in turn, contribute to increased *resource and energy efficiency*.
- **Seamless Immersive Reality**
    - Context-aware management will contribute to guaranteeing the required QoS for seamless immersive reality, fostering *social, technological, and economic sustainability*.
    - Coordination between multiple devices and the 6G network can be facilitated via efficient network abstraction. Context-aware management will reduce the overhead of supporting advanced immersive reality applications, improving *resource and energy efficiency*.

Transformed architecture for 6G

System architecture for 6G CN

Architectural aspects of Migration

Migration from 5G to 6G
Overview

Motivation

- To accommodate and meet the requirements from new use cases (including, AI, JCAS, etc.) evolution of the 5G CN is inevitable. However, this evolution from 5G to 6G needs to be smooth and timely.
- Migration from 4G to 5G had multiple non standalone options. This increased the complexity and caused delays in the migration. However, now NSA coexists with the standalone 5G architecture.
- To reduce the 5G-6G interworking and deployment complexity, the goal should be having fewer architecture options specified for the deployment of 5GC as a 6G requirement.

Key Requirements

- Sustainability: social (e.g., trust, technology diffusion), economic (i.e., business and monetization) and environmental sustainability (Energy efficiency and carbon footprint)
- Simplification: decreased number of standardized APIs, network functions, signalling
- Interoperability: backward compatibility, interworking with 5G/4G, CN procedures support, combined capacity and coverage bands (i.e., through MRSS)
- Flexibility: deployment and operational flexibility

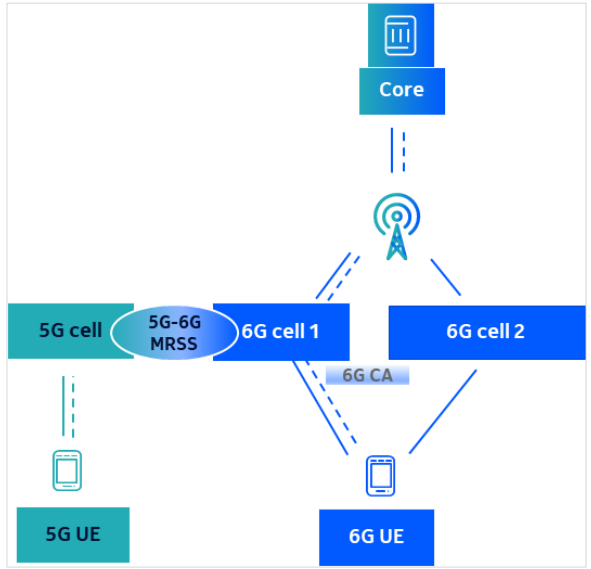


Fig. 1 Preferred architecture option for 6G, i.e., SA with Evolved 5G Core and using MRSS

Preferred architecture option (cf. Fig 1)

- 6G RAN is deployed standalone and connected to an Evolved 5GC (E-5GC). The migration is done via MRSS between 5G and 6G RATs.

Migration from 5G to 6G
5GE CN vs 6GCN

G-Agnostic Core

- To avoid delays in introducing key 6G services, 6G CN could be an evolution of the 5G CN so that the networks can gradually extend the support for new 6G services without the need to replace the CN
- This gradual evolution would also continue in the next generation networks, bringing the characteristic of the G-Agnostic core (Single core network, cf. Fig. 1)
- To ensure smooth and timely migration to 6G, 6GCN can share selected network functions (NFs) with 5GC.
- New or non-backward compatible changes can be introduced by adding new NFs or new services to existing NFs.

6G RAN

- The most viable solution is that 6G RAN is deployed as a standalone RAT
- 5G RAN was connected to 5G CN with the p2p N2 interface. 6G probably need to support a similar interface between 6G RAN and core network.

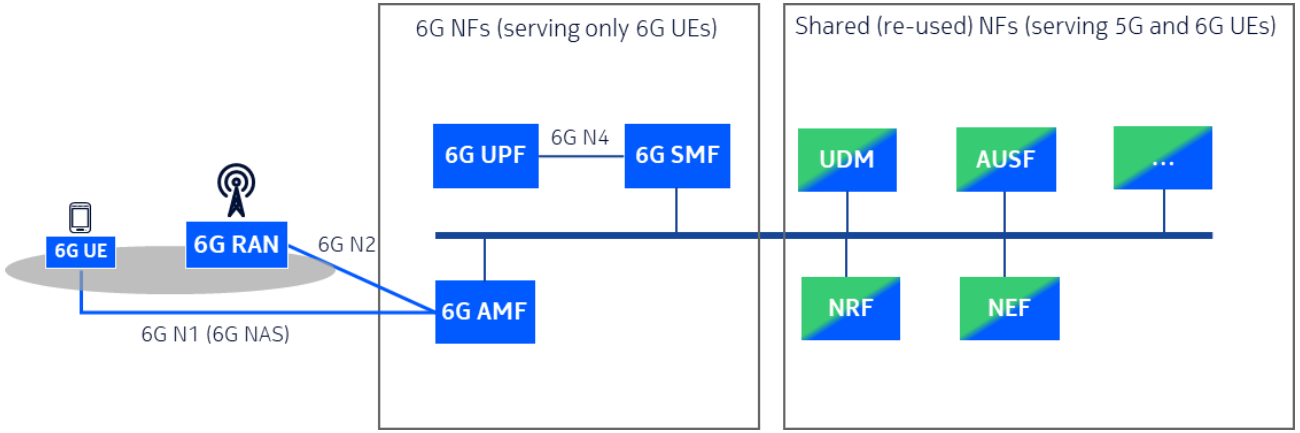


Fig. 1 6G Functional Architecture (Figure illustrates an exemplary 6G architecture. Research is ongoing)

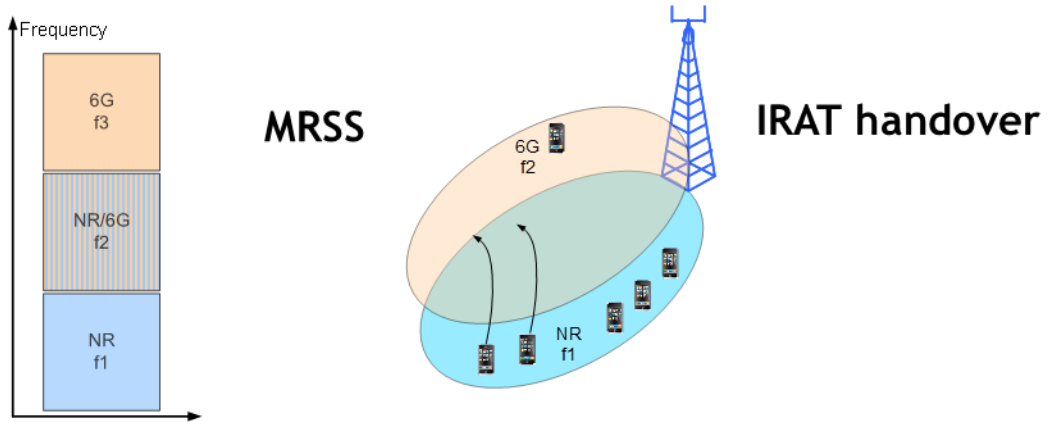
Migration from 5G to 6G
5G-6G Interworking, MRSS

6G CN & 5G CN interworking

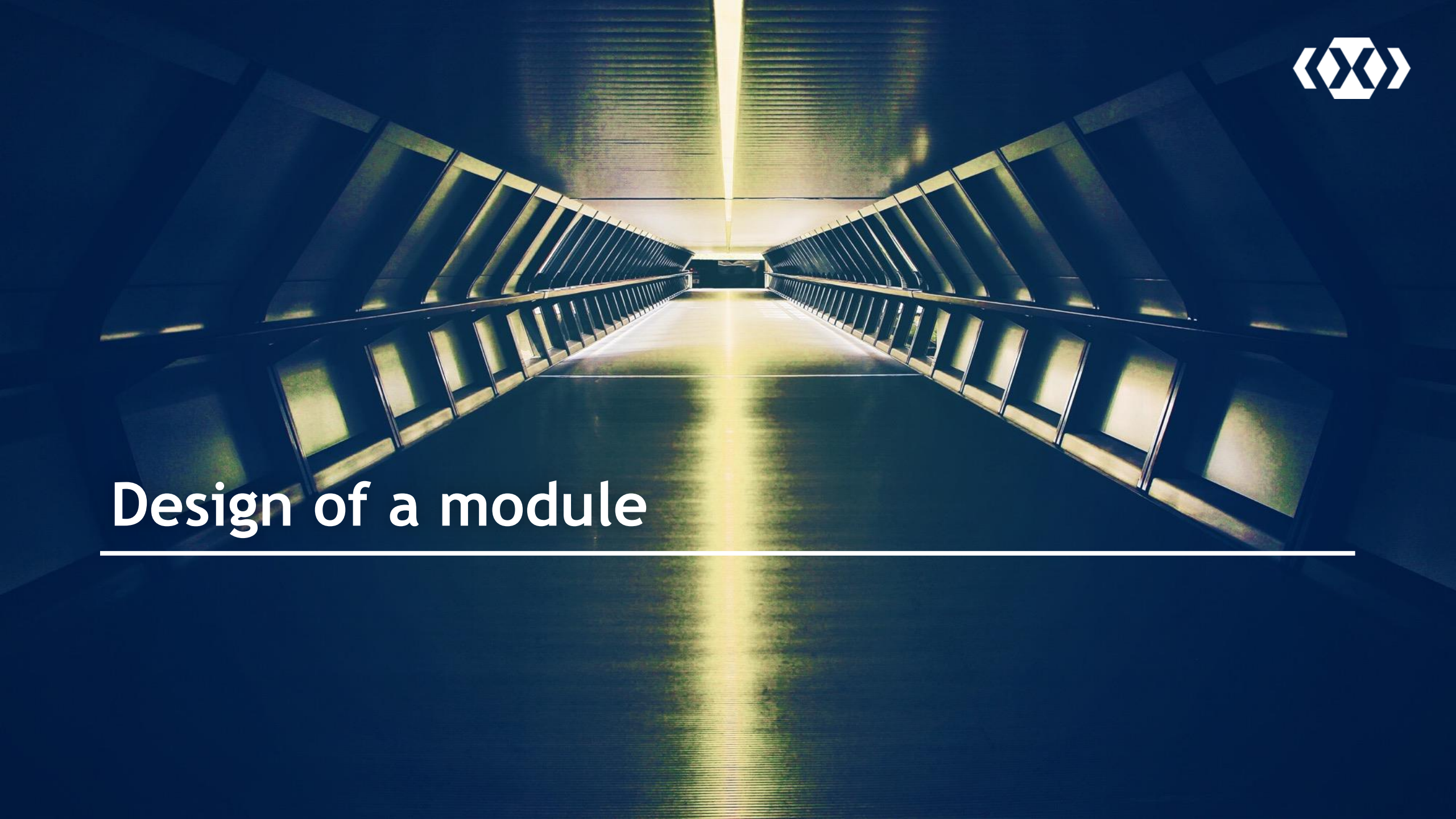
- Enable support for Single Registration(SR) based interworking (IWK) between 5G-6G.
- Single Registration based IWK must be mandated for all UE(s) that require IWK between 5G and 6G UE(s).
- Context transfer needed between AMF and 6G AMF to enable mobility between 5G and 6G network
- Combo 6G SMF and SMF, common UPF IP anchor needed for 5G and 6G Network to ensure service continuity.
- Handover support needed between 5G radio and 6G radio to enable SR based IWK.
- New 6G spectrum and/or refarming of 5G spectrum to 6G
- IRAT handover between 5G and 6G, with load balancing, can enable a smooth migration
- May lead to inefficient use of spectrum due to non-granular way of splitting (refarming) the spectrum
- Can be combined with all other solutions

5G - 6G MRSS

- 5G and 6G can “co-exist” in the same frequency band
- If MRSS is employed, the time-frequency resources in a cell are dynamically assigned to either 5G or 6G according to traffic demands
- The main disadvantage is that the control signaling for both 5G and 6G must be on simultaneously, causing a certain overhead.
- Inter-RAT (IRAT) HO between 5G and 6G still need to be supported for scenarios without MRSS support.



Modular 6G design



Design of a module

Design of a module
Overview

- Background:
  - There is a tradeoff between performance and flexibility when considering the design of modular 6G architecture
  - As we adopt virtualization and CNF-based deployments to enable more flexible networks, the complexity level rises in line with modularity. Moreover, as the networks are comprised of multiple dynamic and ephemeral elements, reliability becomes harder to guarantee
  - In 5G networks, sources of delay in procedure completion time are (1) Inter-NF networking, (2) Insufficient compute resource allocation and implementation inefficiency (3) Resource contention and I/O Bottlenecks (4) Node-related hardware problems (CPU scheduling, Memory allocation)
- Two studies are contributed:
  1. **Performance impact of disadvantageous conditions in CNF-based 5G Core:** it offers:
     - identifying bottlenecks (latency) and guaranteeing reliability for 5G/6G Core networks
     - performance evaluation of parameters such as latency (E2E), procedure completion time, CP signalling
     - main potential KPI improvements: E2E latency, reliability
  2. **Procedure-based functional decomposition:** it proposes a new design for core NFs (see Fig. 1) where the logic for executing a complete procedure, such as PDU session establishment, UE registration, UE deregistration, etc. is consolidated into one NF [GSH+22]. It results in the following:
     - reduced signaling between NFs
     - shorter time for executing a functional procedure
     - easier management of the UE context and NFs state
     - but limits flexibility
- As a fundamental evolution of the 6G CN design, this enabler does not depend on any of the other enablers although it can be a basis for all other enablers.

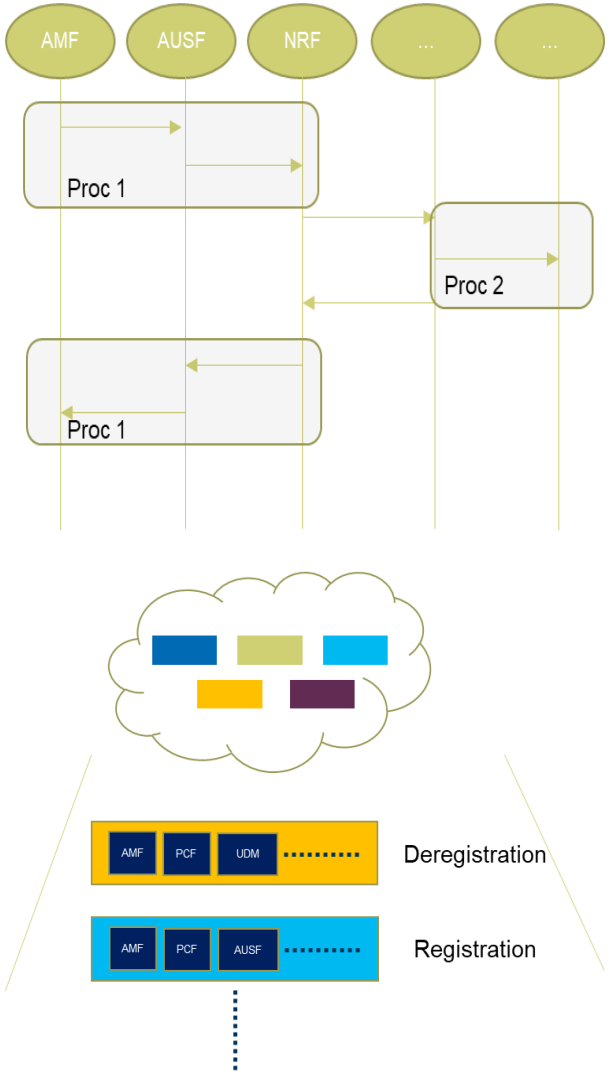


Fig. 1 Procedure based functional decomposition

Design of a module
Guiding principles

- Assumptions:
  - An NF supports one or more services.
  - Modules are groups of functions.
  - Group functions based on certain requirements such as latency.
  - Assume architecture is service-based in the core network, with service-based interfaces, having service-based communication patterns
- To the right, we list the principles that should guide the module design in 6G

Principle	Description	Motivation
• Direct communication between NFs (logically)	• Low-latency NFs can talk directly to low/high latency NF	• Prevent long signaling chains
• Info pull	• The NF interested in information ensures that it gets it	• Reduces dependencies between NFs/services. Improves reuse and separation of concern
• Smaller transactions	• If a procedure requires that more than two NFs need to communicate; try to split it into separate procedure parts	• Reduce coupling between different procedures and services
• Loose coupling and modularity	• Organize the system around the objects it handles, e.g., UEs, cells, radio resources, etc. • Avoid splitting the responsibility for a certain object or feature across different NFs	• To have better manageability • Achieve independent scaling
• Performance	• Ensure good performance (latency, robustness) for key procedures, e.g., Handover, Resume, Re-establishment, RRC reconfiguration, etc. • Trade loose coupling for performance	• 6G needs to perform better than previous Gs incl. support for critical services

Design of a module

Study #1 - Performance impact of disadvantageous conditions in CNF-based 5G Core

- Goal
  - To detect bottlenecks and to evaluate performance of a 5G Core network for **UE Registration + PDU session** establishment procedure
- Sources of delay in 5G procedure completion time
  - **Inter-NF networking**
  - Insufficient resource allocation and implementation inefficiency
  - Resource contention and I/O Bottlenecks
  - Node-related hardware problems (CPU scheduling, Memory allocation)

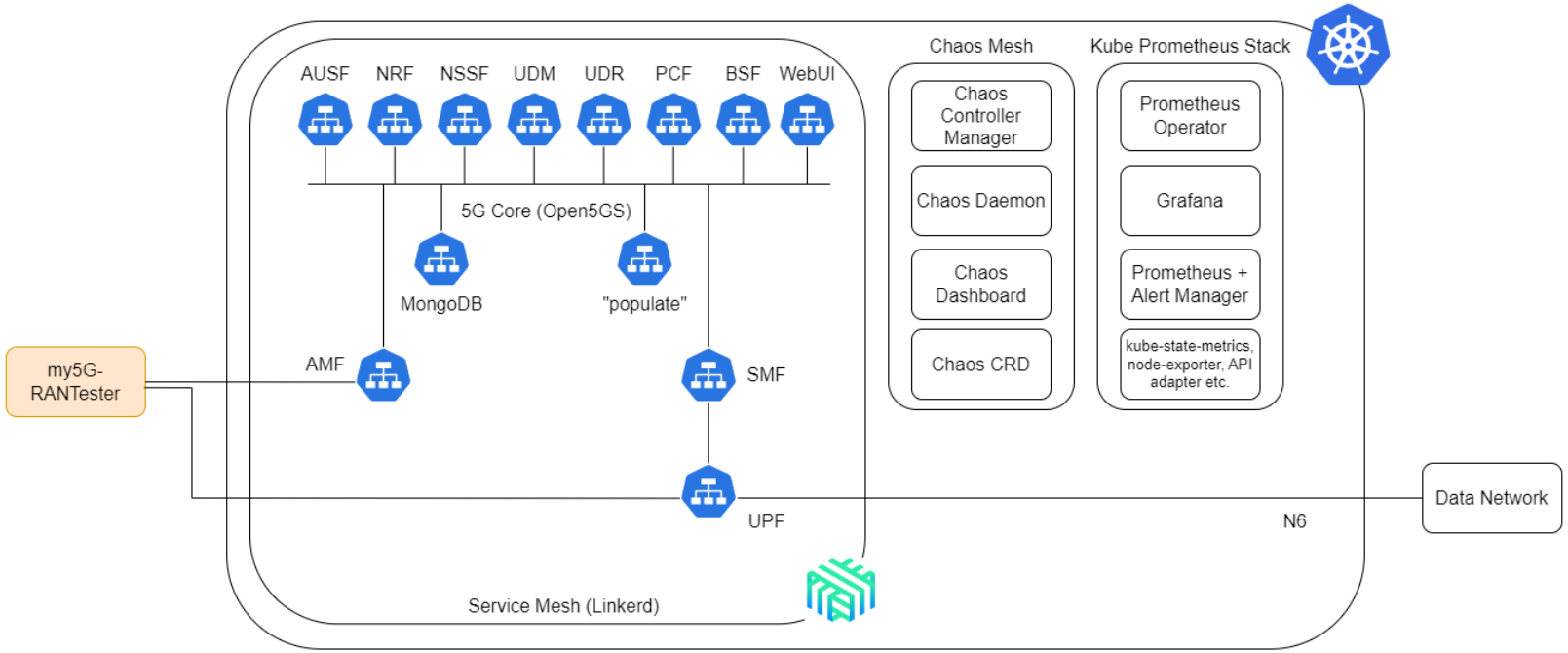


Fig 1. Experimental Setup

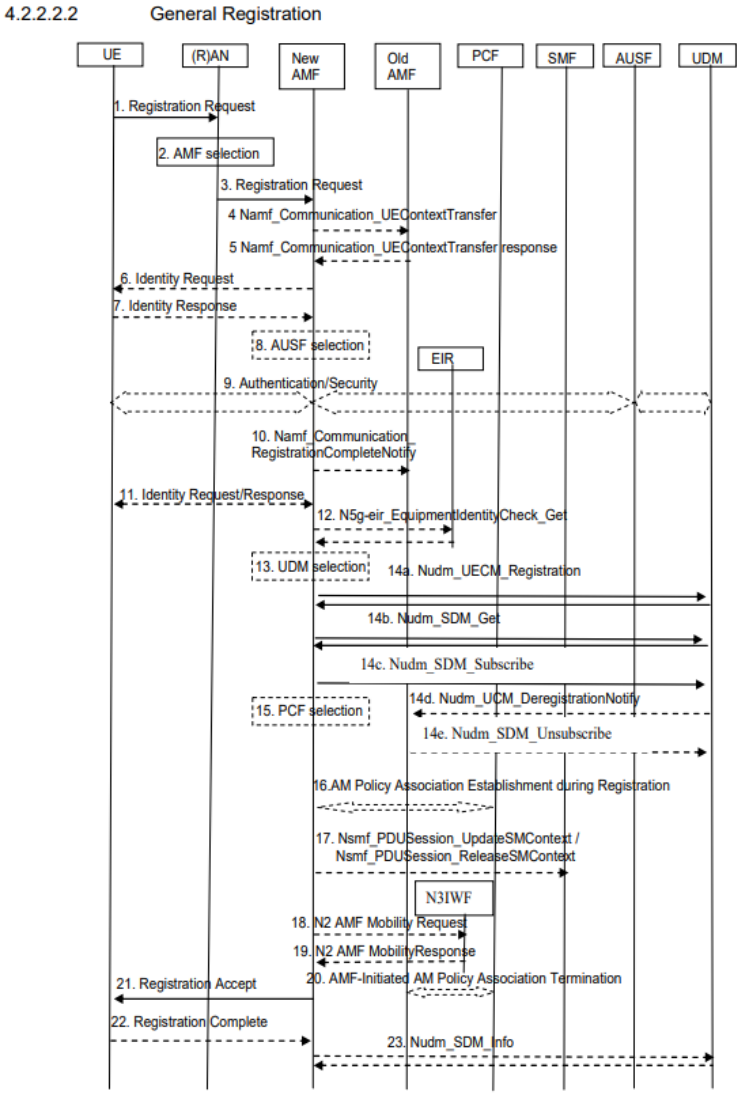
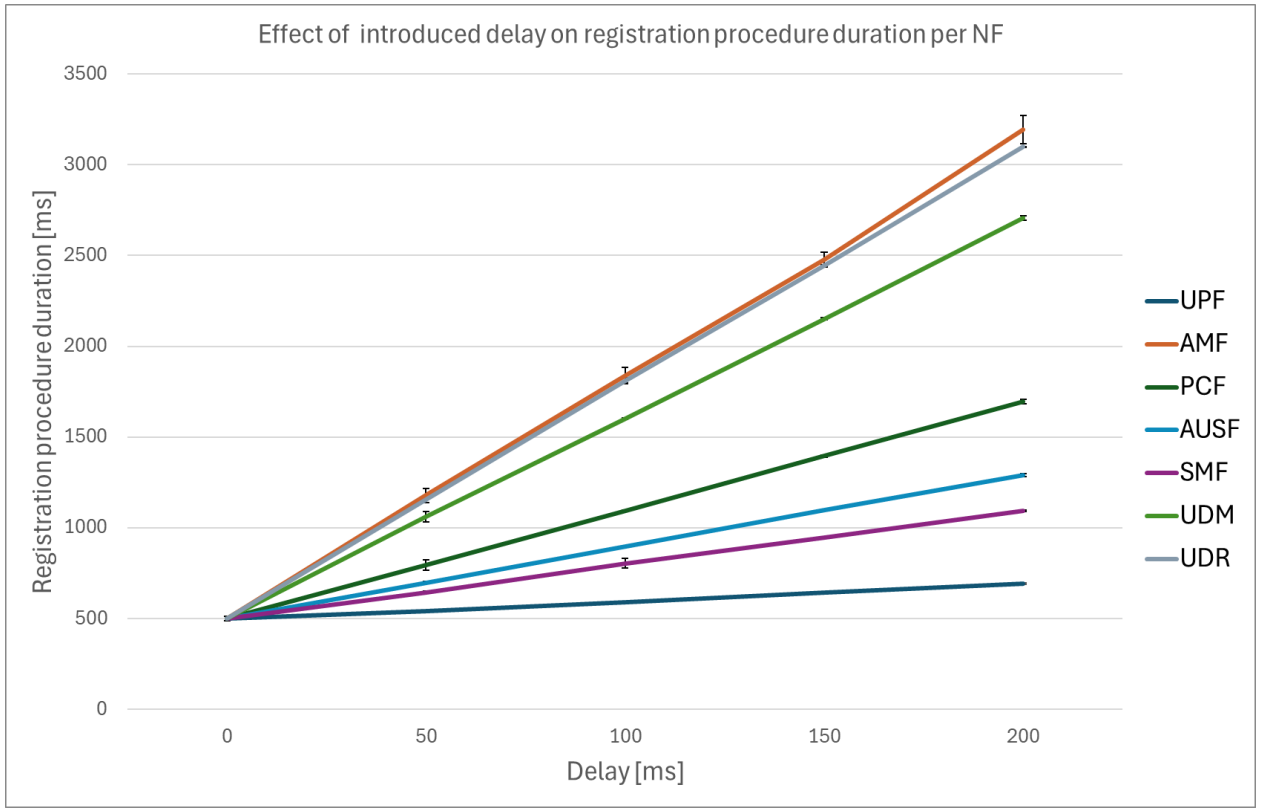
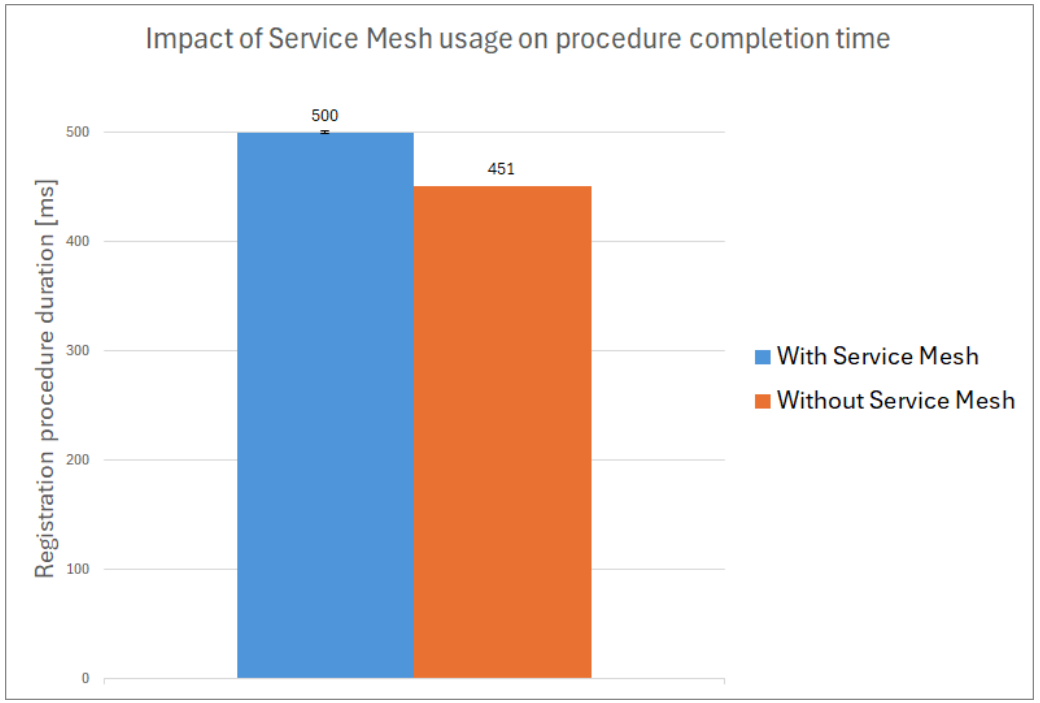


Fig 2. UE registration [23.502]

Design of a module

Study #1 - Performance impact of disadvantageous conditions in CNF-based 5G Core



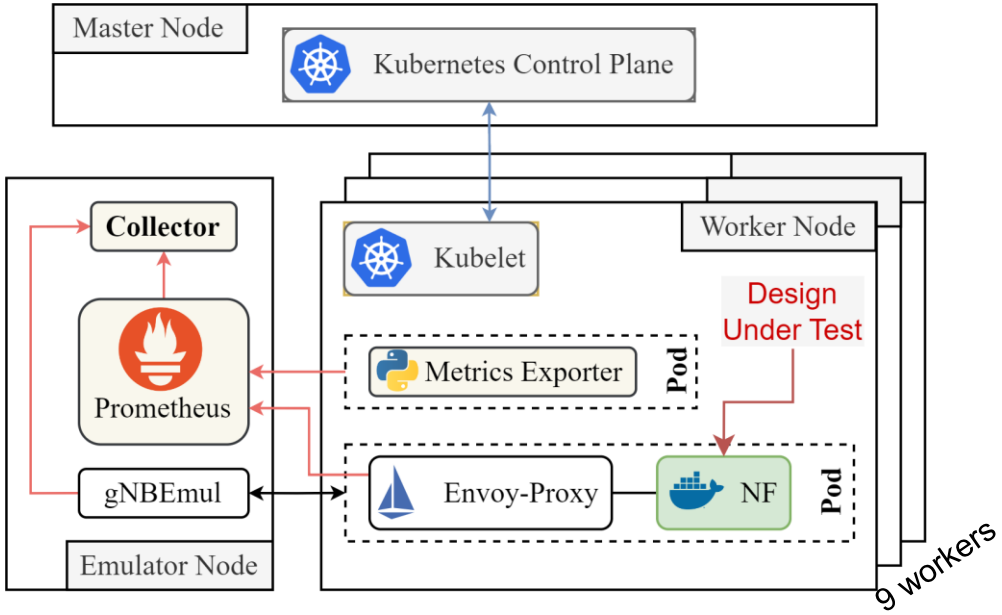


Performance evaluation conclusions

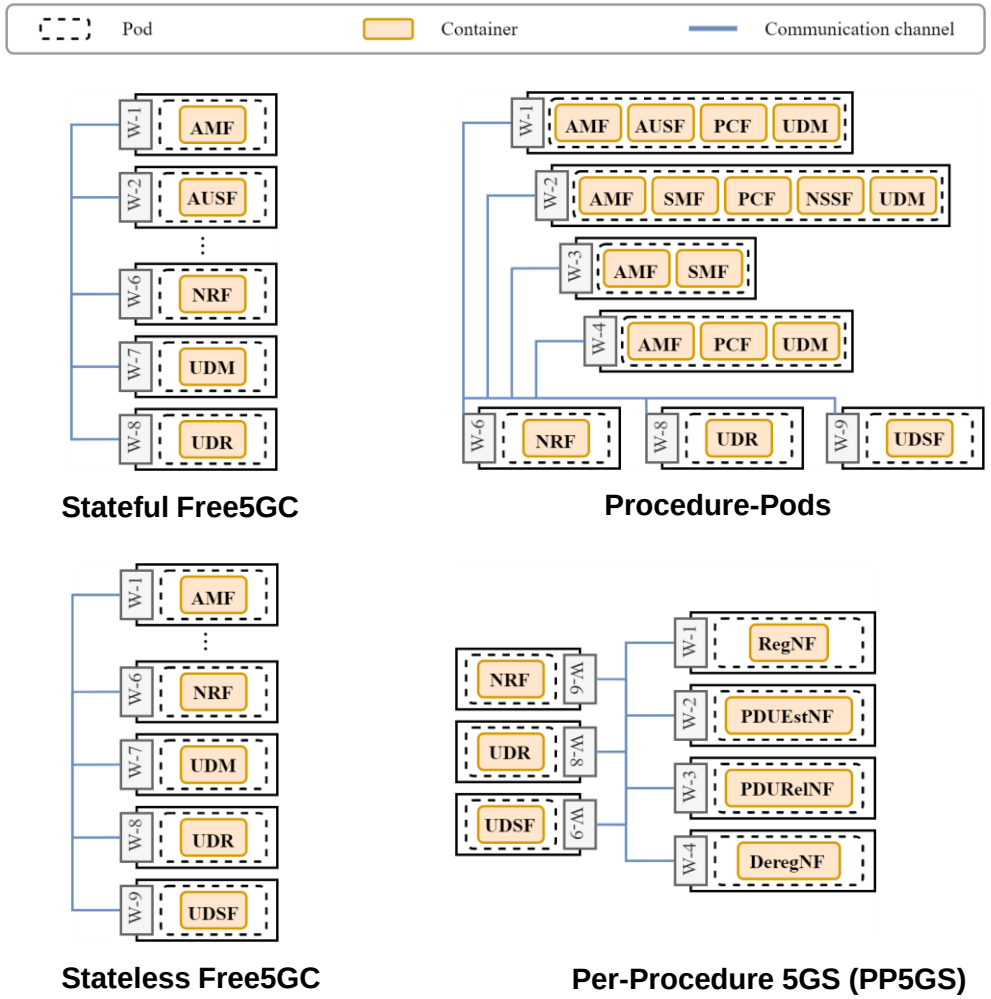
- Based on the obtained results, a significant impact of Chaos Experiments introduced towards the UDM and UDR network functions on the procedure completion time of the registration procedure can be observed
- Linkerd Service Mesh has had a considerable impact on both Procedure Completion Time as well as resource usage, which calls for an inquiry into other service mesh tools that do not use sidecar proxies, and the quality of their telemetry data
- The introduction of jitter alongside latency has not resulted in a significant increase of PCT compared to just latency being introduced alone

Design of a module
Study #2 - Procedure-based functional decomposition

- Proposal:
  - a new design where a new set of core NFs is proposed where each NF includes the logic for executing a complete procedure, such as PDU session establishment, UE registration, UE deregistration, etc. [1]
- Goal:
  - reduce procedure completion time and simplify the APIs and interactions between CP NFs compared to 5G

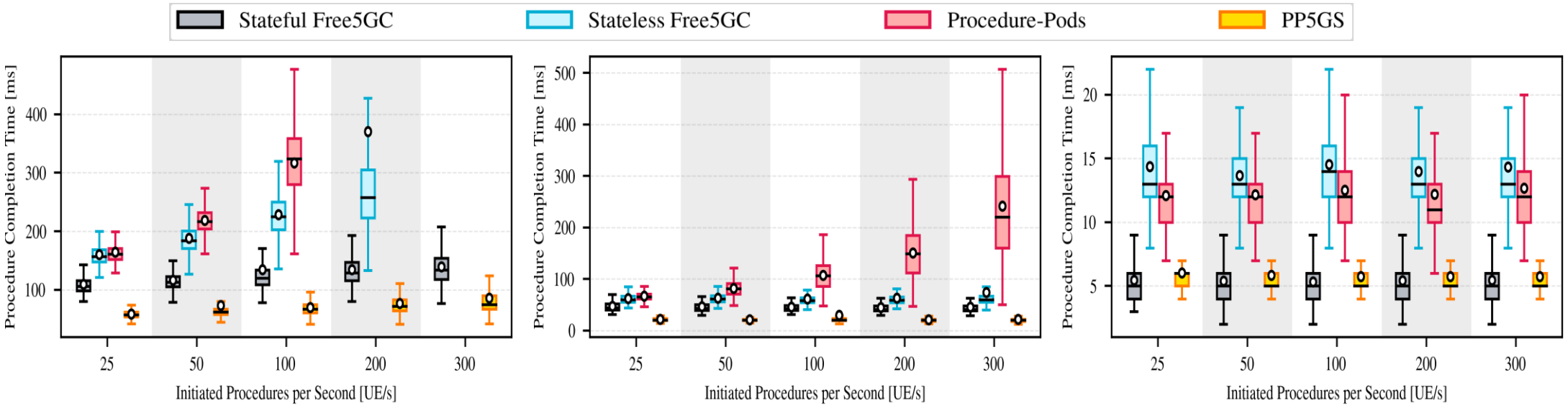


Experimental Setup



Four designs under test

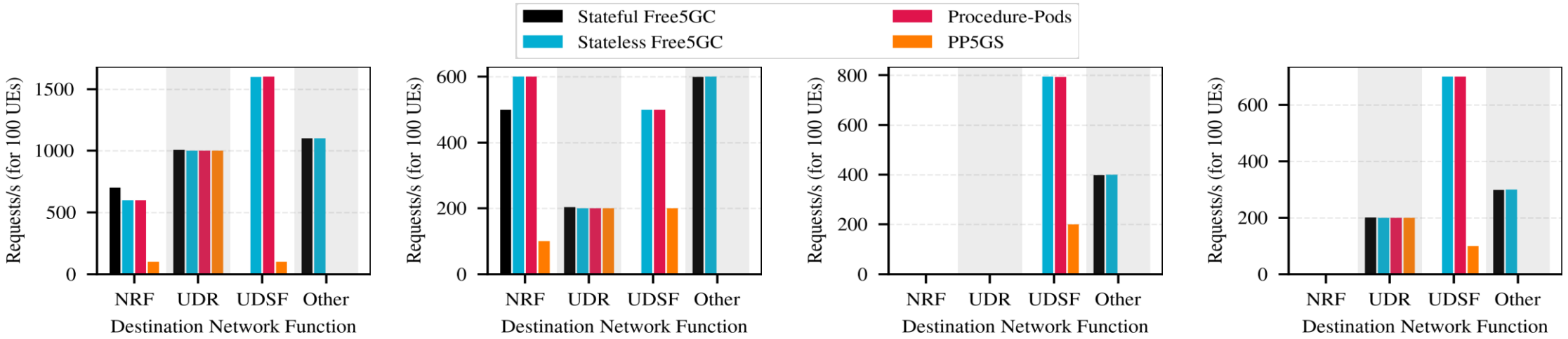
Design of a module
Procedure-based functional decomposition - Evaluation



a) Registration. *b) PDU Session Establishment.* *c) Deregistration.*

- Procedure Completion Time (PCT) for the Registration procedure and PDU Session Establishment are respectively 42% and 50% shorter in the case of PP5GS compared to Stateful Free5GC.
- PCT for the Deregistration procedure is 8% longer (~1 ms) in the case of PP5GS compared to Stateful Free5GC.
- PP5GS reduces PCT for complex procedures.

Design of a module
Procedure-based functional decomposition - Evaluation



a) Registration. *b) PDU Sx. Establishment.* *c) PDU Sx. Release.* *d) Deregistration.*

- Requests per second for the Registration procedure is 57% and 72% less in the case of PP5GS compared to Stateful Free5GC and stateless Free5GC, respectively.
- Requests per second for PDU Sx Establishment procedure is 61% and 72% less in the case of PP5GS compared to Stateful Free5GC and stateless Free5GC, respectively.
- Requests per second for the PDU Sx Release procedure is 50% and 83% less in the case of PP5GS compared to Stateful Free5GC and stateless Free5GC, respectively.
- Requests per second for the Deregistration procedure is 40% and 75% less in the case of PP5GS compared to Stateful Free5GC and stateless Free5GC, respectively.

Design of a module
Key take-away

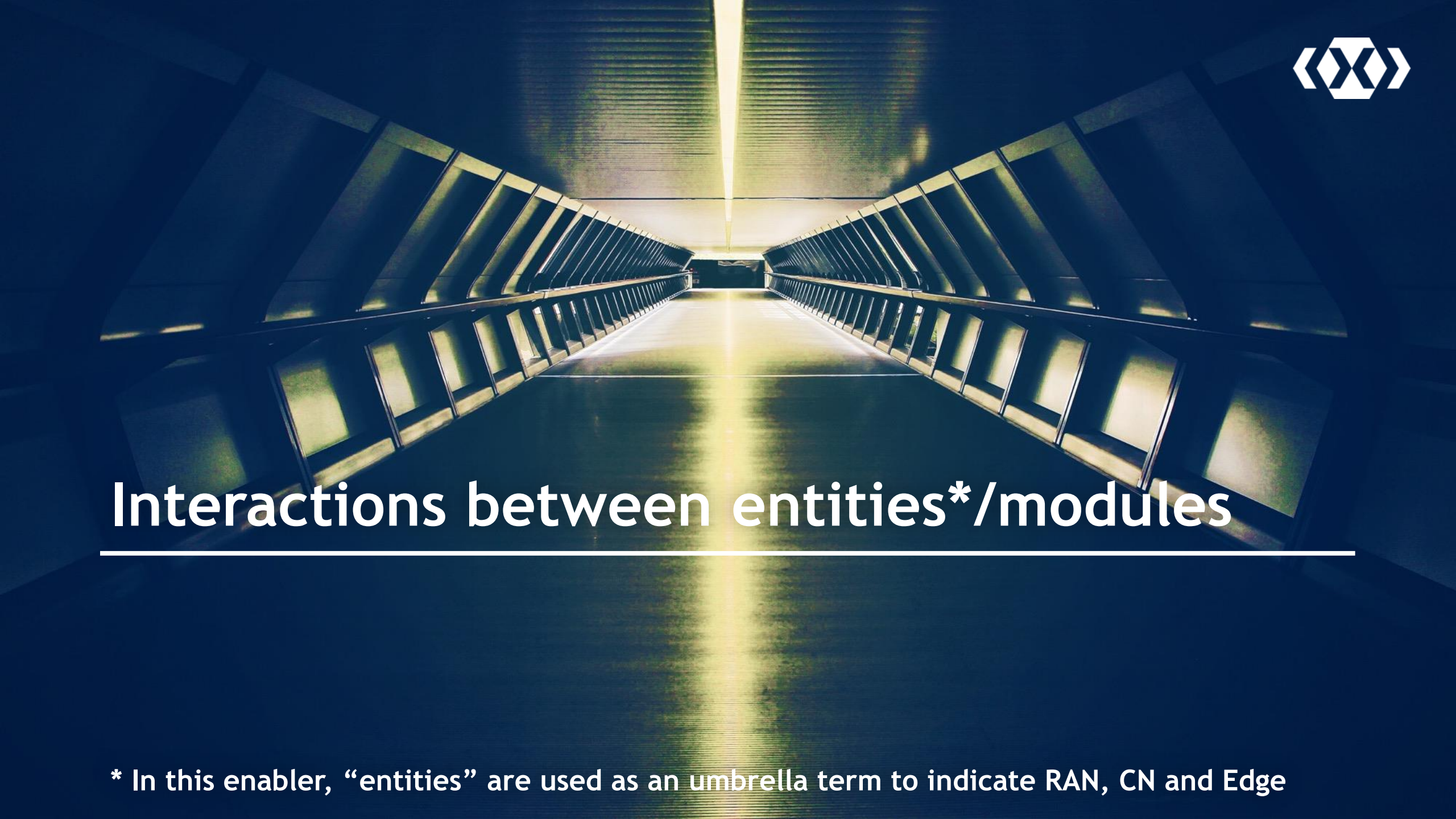
1# Performance impact of disadvantageous conditions in CNF-based 5G Core

- Observability in modular networks
  - Performance evaluation conclusions as an input for 6G CP design
- Experimental results
  - Effect of introduced delay on registration procedure duration per NF - Sensitivity to latency introduced towards AMF, UDM and UDR network functions
  - Impact of Service Mesh usage on procedure completion time

2# Procedure-based functional decomposition

- Decomposition of Core CP is designed and implemented, where the new set of NFs encompasses all the interactions between 5G Core NFs to complete a certain procedure
- Quantitative analysis of the new design revealed:
  - Gains in terms of time needed to complete the execution of a certain control plane procedure
  - Reduction in the total number of messages needed for these procedures
  - However, this design reduces the flexibility in deploying the more coarse-grained NFs

There is a tradeoff between performance and flexibility when considering the design of modular 6G architecture: More granular design results in higher flexibility in implementing and deploying modules but at the cost of reduced performance in terms of execution time and state management



Interactions between entities*/modules

* In this enabler, “entities” are used as an umbrella term to indicate RAN, CN and Edge

Interactions between entities/modules

Overview

- This enabler focuses on how different entities shall interact in beyond 5G (5GA/6G) networks
  - RAN-CN control plane interactions and interfaces (see Fig. 1)
  - Data centric service-based architectures (see Fig. 2)
- RAN-CN control plane interactions and interfaces
  - 6G services and applications, such as immersive communications, sensing, ambient IoT, etc. pose new system requirements to be met, e.g., QoS, energy saving, sustainability.
  - New devices, such as battery-less or power-connectionless AIoT devices are to be supported.
  - Simplicity and AI-nativeness are other design principles for 6G.
  - As with every G, RAN architecture (aggregation/disaggregation) is critical.
  - 6G is anticipated to operate across new frequency bands (sub-THz, 7.125 to 24.25 GHz) in addition to 4G and 5G frequency bands.
  - Reduced energy footprint across network is a key factor in the adoption of 6G networks and technologies.
- Data centric service-based architecture
  - Better fulfillment of 6G requirements, emphasizing flexible service routing for distributed resources.
  - Data Centric Networks supports dynamic stateless and loosely decoupled highly granular Network Functions (NFs).
- Implication on the 6G standards (3gpp, ETSI, O-RAN, etc.) key requirements
  - Simplification: decreased number of standardized APIs, network functions, signaling
  - Interoperability: backward compatibility, interworking with 5G/4G, CN procedures support
  - Flexibility: deployment and operational flexibility
  - Resiliency
  - Scalability
  - AI-nativeness
  - Sustainability: energy efficiency/saving enablement
  - cloud readiness,
  - load balancing,
  - community support

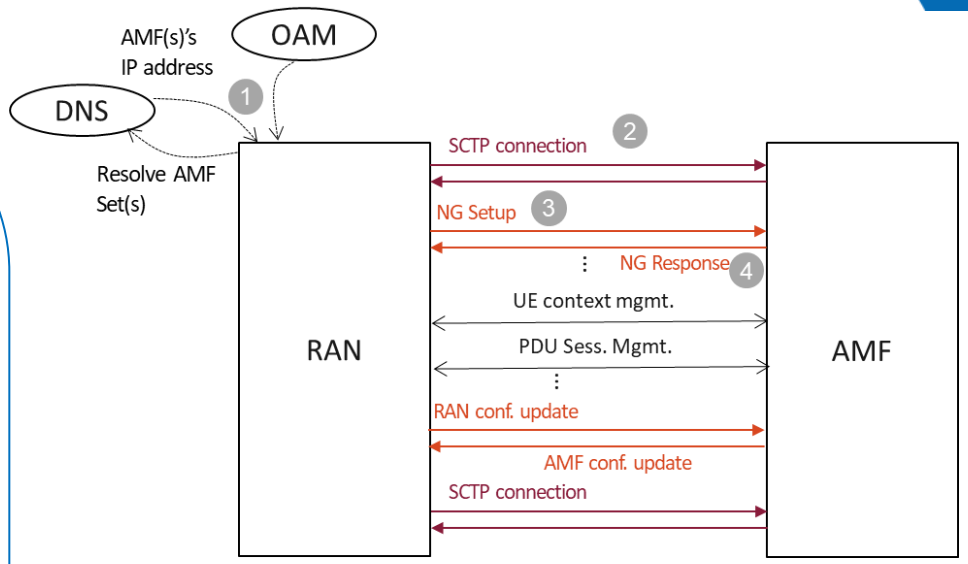


Fig. 1 RAN-CN control plane interaction

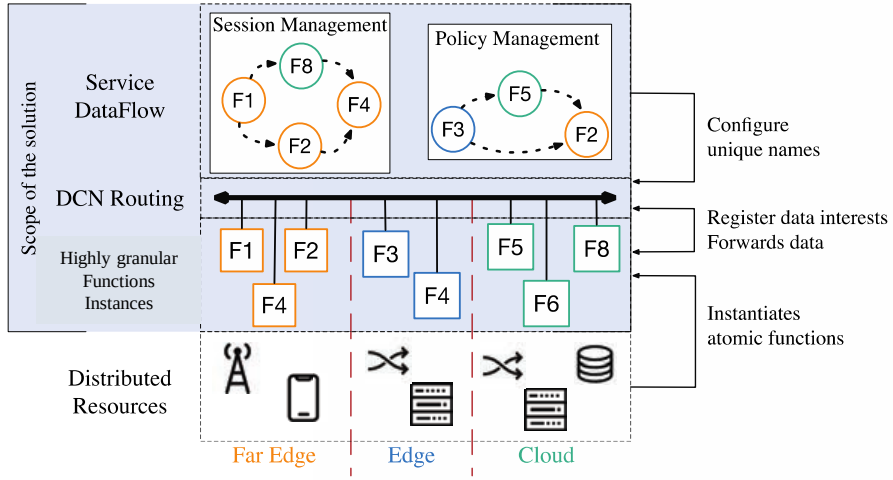


Fig. 2 Data centric SBA

Data-Centric Service-Based Architecture for Edge-Native 6G Network

Protocols and Message Sequence

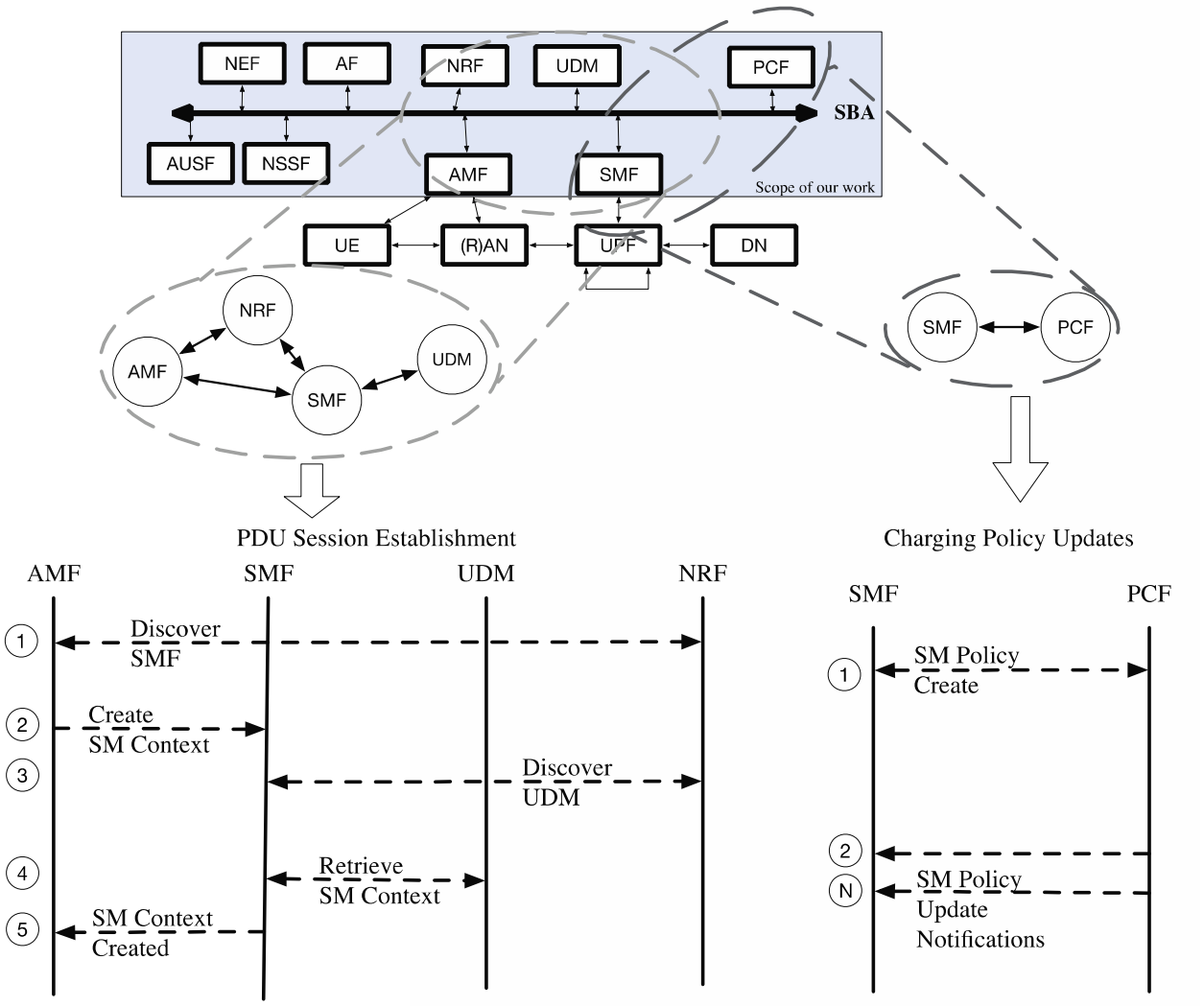


Fig. 1 5G SBA architecture, highlighting the workflows used for evaluation [BQG+24]

Data-Centric Service-Based Architecture for Edge-Native 6G Network

Evaluation of the DCSB architecture

- This research evaluates the proposed data-centric architecture by implementing a proof-of-concept prototype.
- This prototype validates the approach through selected 5G workflows, demonstrating the advantages of data-centric and dataflow mechanisms.
- Key benefits include enhanced scalability, flexibility, automation, simplification of architecture, and efficient exposure of network capabilities.
- Results in Fig. 1 shows an approach based on dataflow programming (Zenoh) shows better completion time of each procedure analysed than other approaches

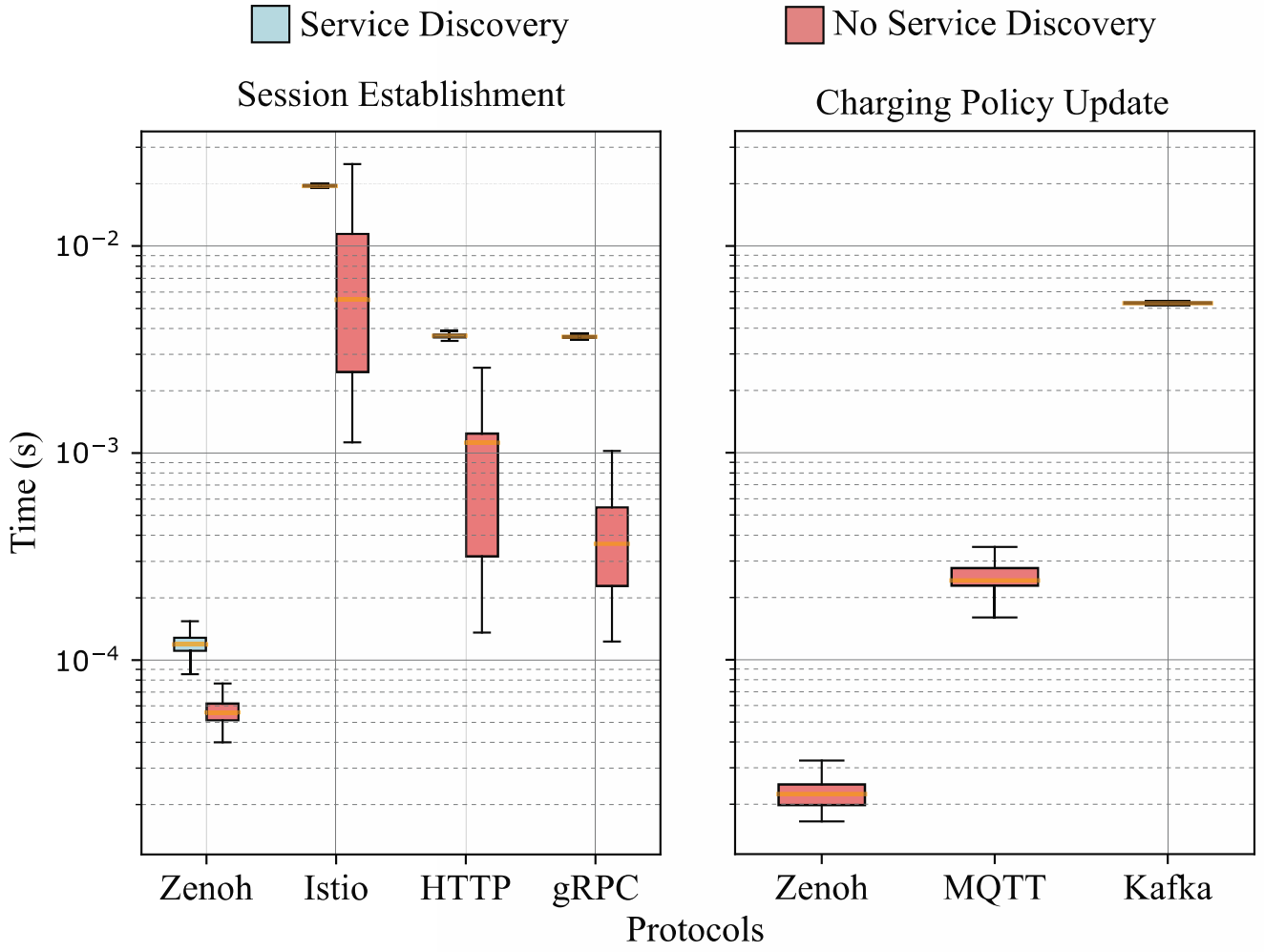
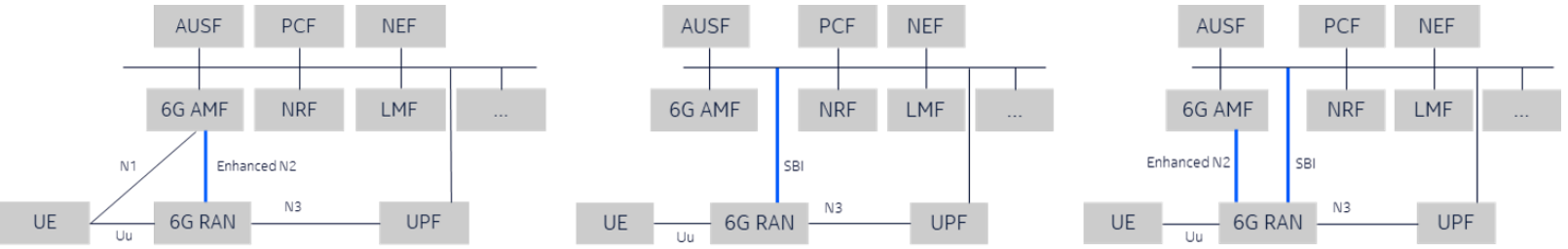


Fig. 1 Workflow completion for different protocols (lower is better) [BQG+24]

Interactions between entities/modules
Comparison of different RAN-CN interaction options



Option A: Enhancing NGAP **Option B: RAN-CN Control Plane SBI** **Option C: RAN-CN Control Plane SBI+NGAP**

Fig. 1 Considered RAN-CN CP options for 6G (Option A-C) [HEX2_D33]

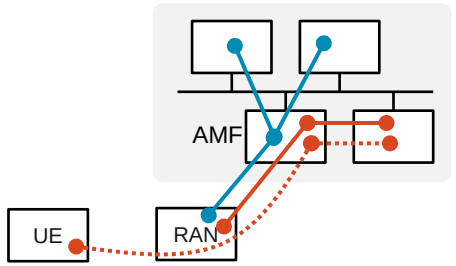
Considered Options

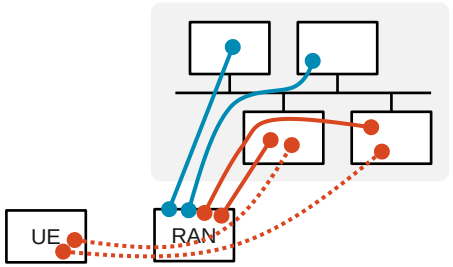
- Option A: 6G RAN communicates via NGAP with the AMF, or a 6G equivalent of it. RAN maintains ASN.1 encoded interfaces, 6G AMF maintains the gateway functionality between RAN and CN NFs.
- Option B: the RAN-CN interface is changed by introducing a Service-Based Interface (SBI) between 6G CN and RAN. The 6G RAN directly exposes and consumes services towards and from the CN.
- Option C: a hybrid approach where service-based interaction is introduced with minimal impact on the existing point-to-point interactions and without creating duplication among the two interfaces.

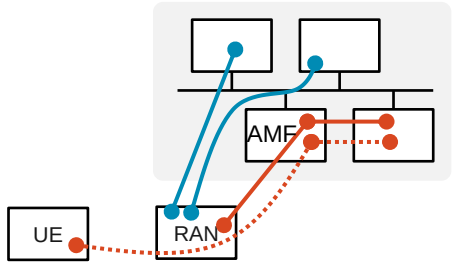
	Option A	Option B	Option C
Features	• Existing architecture and interfaces are used where 6G RAN communicates with 6G AMF via enhanced RAN-CN interface (enhanced N2) • RAN maintains P2P interfaces • 6G AMF maintains the gateway functionality	• Direct RAN-CN NF CP communication where 6G RAN and CN NFs expose services towards each other and communicate directly via RESTful APIs over SBI.	• Existing connectivity and communications support maintained over P2P interfaces, extended for 6G radio support • New services are enabled over SBI, and direct RAN-CN NF communication is enabled
Implications	• Separate domains • Similar effort in implementation and standardization compared to 5G • Possible improvement in cloud-friendliness by using enhanced Transport/Application protocols	• Cloud friendly mechanisms and protocols used • Compute/storage separation • Requires more standardization efforts • Increased testing and verification efforts • Increased efforts for handling coordination and race conditions • Performance degradation and more processing power requirement due to the overhead introduced by clear-text based JSON encoding and decoding and larger message sizes • No clear domain separation • Limited arch. benefits to direct RAN-CN NF communication because 6G AMF is the consumer to most of the current RAN services • Complicated security handling due to multiple UE anchor points in CN	• Requires more standardization efforts • Performance degradation and more processing power requirement is expected due to clear-text based JSON encoding and decoding and larger message sizes • Complex development and implementation efforts due to dual stack requirements in 6G RAN • Increased efforts for handling coordination and race conditions • Not friendly for smaller footprint RAN deployments (Pico, micro, etc.) • Additional configurations required to map services and supported service instances across the nodes
Conclusion	Option A is currently the preferred option as it has positive valuation results compared to options B and C.		

Interactions between entities/modules
Support evolution of RAN-CN communication

RAN-CN interaction reated to non-communication features (e.g., sensing, analytics)
RAN-CN interaction reated to communication features (e.g., registration, sess. mgmt.)
UE-CN interaction (NAS) related to communication features (e.g., registration, sess. mgmt.)



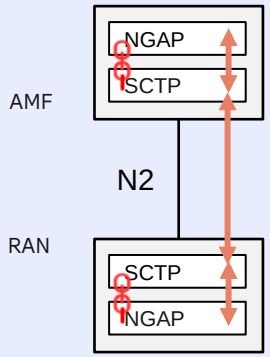
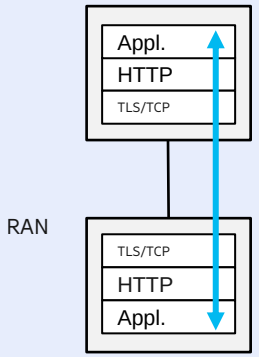
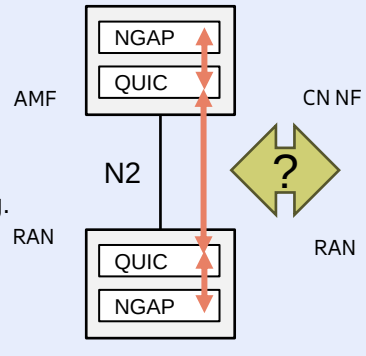




Figures show interactions, not RAN-CN point-to-point interfaces

	Via CN NF proxy	Direct	Mixed (CN NF proxy only for comms. interactions)
Extensibility	Any extension always impacts the AMF proxy as well as any other CN NF if needed (more relevant for non-comms. features).	New extensions do not need to include the proxy, only the concerning CN NF.	New non-comm features extensions do not need to include the proxy, only the concerning CN NF.
Implications	RAN/CN isolation Proxy as UE anchor point in CN Security inherently supported. The proxy may add delay to the procedure (can be solved via bundling NFs) Inherently supports 5G to 6G migration	May be possible to cut system procedure delay due to no CN NF Proxy hop No RAN/CN isolation ➔ CN hardening Higher testing and integration burden Many relations to CN NFs (establish, update at mobility, dependencies, …) Multiple NAS terminations (security risks, maintenance, dependencies, …) Migration with 5G not inherently supported.	Proxy as UE anchor point in CN No dependency between comms. and non-comm. features No load at proxy for non-comms. features Supports 5G to 6G migration for 5G services.
Conclusions	Several valuable benefits, limitations not so severe. How about new 6G features?	Several implications, major re-design with unclear gains	Of interest for independent evolution of comms./non-comms. features

Interactions between entities/modules
Evaluations on the cloud-friendliness of RAN-CN interface

	Current N2/NGAP (A.1 – base line)	Using SBI (Opt B)	Replace SCTP (e.g. QUIC) (Opt A.2)
CP application logic design is E2E	• **NGAP logic design coupled to SCTP** • Networking layer exposed to application, application design not always considering possible failures/losses, not all features are E2E • Binding between instances needed (RAN to CN)	• **Application logic designed with E2E approach, decoupled from transport** • **No binding needed, HTTP connections and cloud can do load balancing**	• NGAP logic with E2E features decoupled from transport layer. **OPTION A.2.1** • Changes to the 3GPP functionality • Improved performance and cloud native function support **OPTION A.2.2** • Utilize the enhanced native transport layer/ QUIC features • No TNLA binding/mgmt, E2E features, consider losses / failures in the whole design
Conclusion	• High level of functionality suited for N2 (see [HEX224-D33]), not cloud friendly	• Cloud friendly, but SBI needs to be extended to have the same level of functionality as current N2 ([HEX224-D33])	• Cloud friendly, current high level of NGAP functionality can be kept to large extent, no binding needed
	Baseline option 	RAN-CN SBI 	RAN-CN with e.g. QUIC Opt. A.2.1: Replace SCTP transport protocol e.g. with QUIC with minimum  Opt. A.2.2: Enhance NGAP logic with E2E features decoupled from transport layer

Interactions between entities/modules
Key take-aways

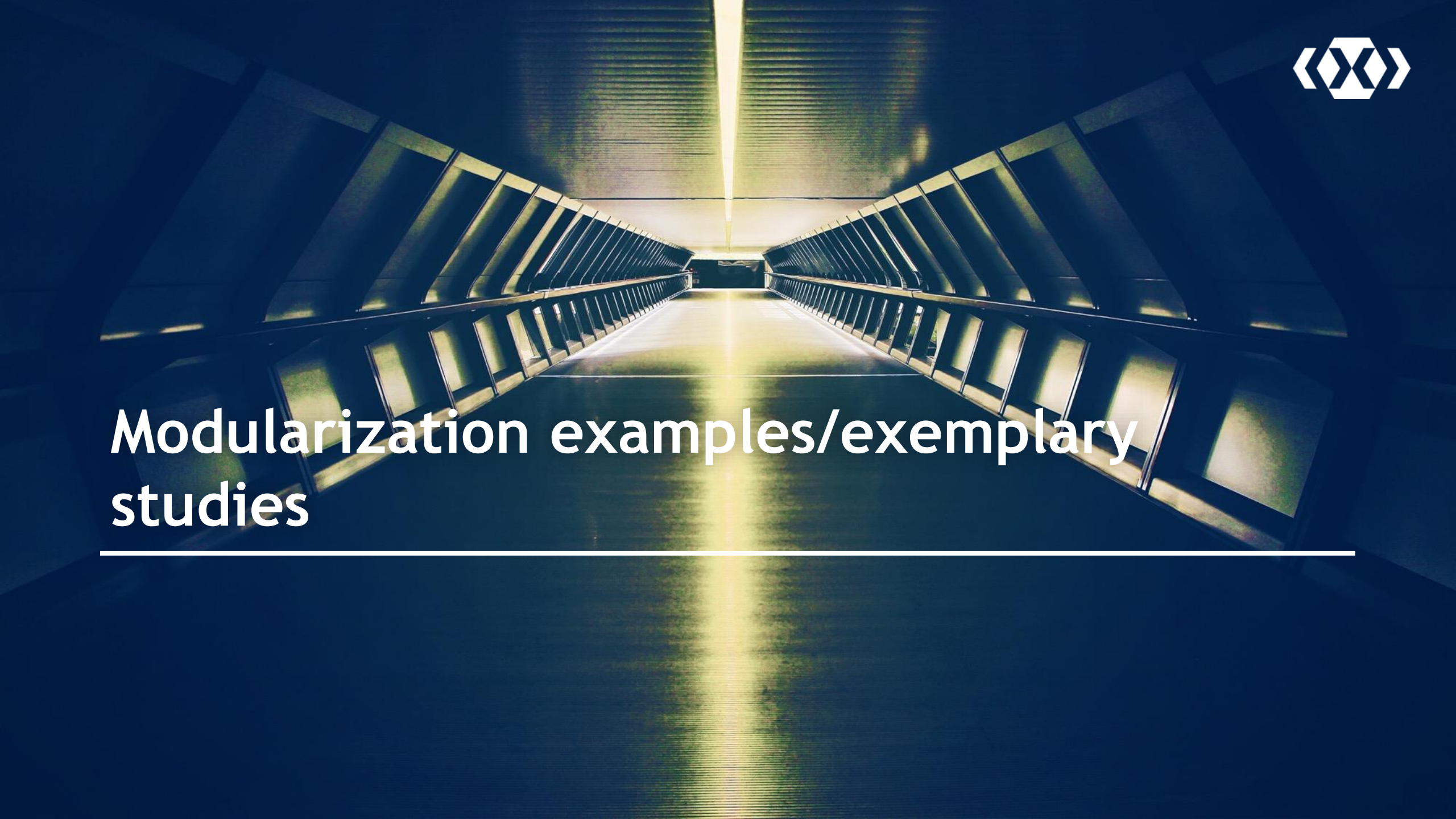
- RAN-CN control plane interactions and interfaces

- There is a need to improve the cloud friendliness of the interface between RAN and the CN (NGAP/N2 interface)
- For this to happen, we need to evolve or replace the SCTP protocol used in 5G to support better decoupling between the different layers
  - Avoid so called transport bindings, etc.
  - QUIC may be one option to use instead of SCTP
- For the 6G architecture we still foresee the main option is that the NGAP/N2 interface goes via the AMF, or a 6G equivalent of it.
  - One major reason for this is to inherently support 5G to 6G migration and security reason
- Relation to other enablers
  - T2.2: Proposals affect the RAN protocols
  - T3.2: Modularization examples (mainly the RAN architectures)
- Key performance metrics:
  - CP Signaling latency
  - Migration implications, e.g., simplicity, backward compatibility etc.

- Data-Centric Service-Based Architecture for Edge-Native 6G Network

- Transition to a fully distributed 6G system with Data-Centric Networking.
- Enhanced scalability and flexibility through dynamic stateless NFs.
- Simplified architecture with efficient resource management.
- Validation through proof-of-concept prototype using 5G workflows.
- Significant architectural benefits in terms of automation and service composability.

- Key performance metrics
  - Procedure completion times (between 10 to 100 times faster).
  - Wire overhead (between two and two-and-a- half times smaller, for PDU session establishment, and between three-and-a-half and 11 times smaller in charging policy update).
  - Lines of code (up to four times lower) for pursuing data-centric approaches as compared with traditional approaches.



Modularization examples/exemplary studies

Modularization examples
Overview

- Modularization gives the advantage of streamlining network modules and functions according to the deployment locations and the respective KPIs/KVIs (see Fig. 1). In this enabler, the focus is to demonstrate how the new modules can be designed at the different network domains (e.g., RAN or UP).
- The benefits of this enabler reflects the targeted domain and the main objectives of modularisation. This section demonstrates:
  - Enables flexible scaling and activation of functionalities on demand resulting in efficient resource utilization and energy efficiency.
  - Better user rates, compared to cellular networks.
  - Minimizing the back-haul overhead, the computational complexity and the number of front-haul connections.
- The modularization examples at this enabler impact TS 23.501, TS 23.502, TS38.401 and O-RAN.

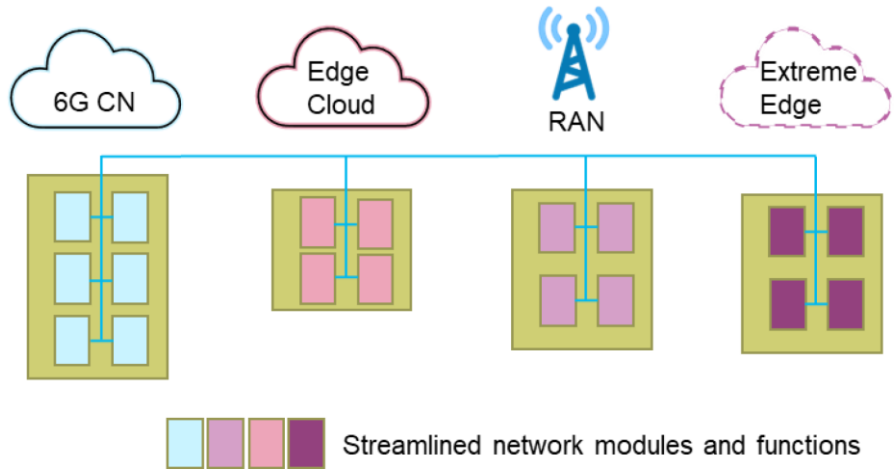


Fig. 1 Modularization examples in 6G networks, the modules can be customized according to the deployment location (RAN, CN or Edge) and KPIs/KVIs

Modularization examples
Protocol and APIs

UPF modularisation

Modularise the monolithic 5G UPF design [23.501] into several modules, enabling flexibility and scalability.

Proposed design: (See Fig. 2).

- Ingress Steering Module (ISM)
- Downlink Module (DLM)
- Uplink Module (ULM)
- On-Demand Module (ODM)

Service Function Chains (SFCs) as interfaces between modules.

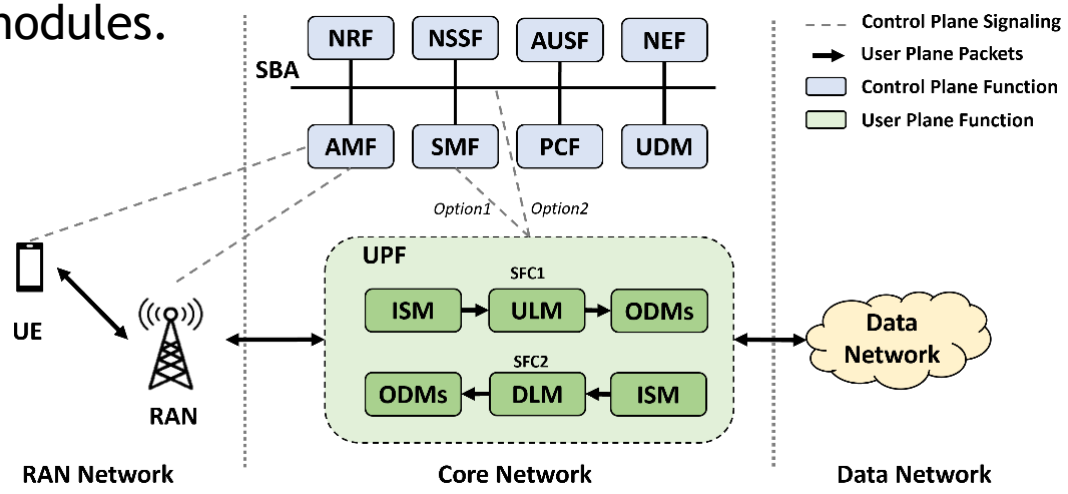


Fig. 2. UPF modules.

RAN modularisation

Disaggregated RAN to increase reconfigurability and interoperability.

Cell-free massive MIMO systems benefit from such disaggregated RAN architecture, as it provides necessary adaptability and flexibility [VRV+22].

Proposed RAN architecture for cell-free massive MIMO: [GAT24] (See Fig. 3).

- Centralized Unit (CU)
- Distributed Unit (DU)
- Radio Unit (RU)

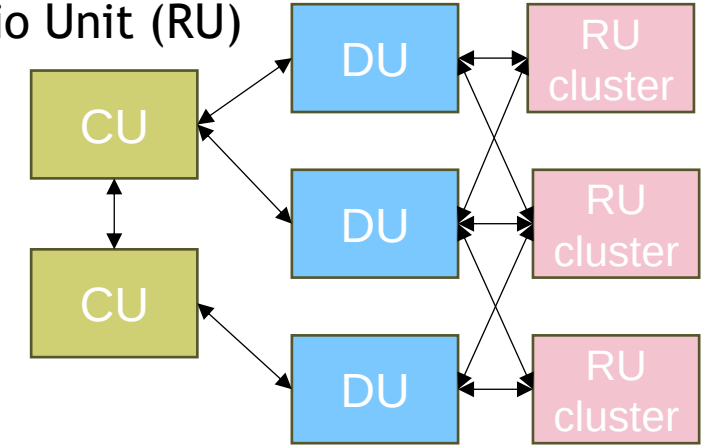


Fig 3. Disaggregated RAN architecture for cell-free massive MIMO.

Modularization examples
Evaluations

UPF modularisation

- Benefits:
  - Enhanced flexibility
  - Fine-grained scalability
  - Efficient compute resource utilization enabled by the dynamic (de)/activation of on-demand UP modules.
- Implication on the 6G standard: 3GPP TS 23.501.

RAN modularisation

- On a disaggregated RAN, aiming at increasing user ergodic rates by creating additional serving clusters of access points, requiring extra front-haul connections between Radio Units (RUs) and Distributed Units (DUs).
- All users are guaranteed a minimum expected ergodic rate by creating additional serving clusters, mutually orthogonal in frequency.
- Implication on the 6G standard: 3GPP TS38.401 and O-RAN architecture suitable for cell-free.

With the proposed architecture (in black), the user ergodic rates increase in all percentiles as compared to a cellular D-MIMO deployment [GAT24b]. This increase is especially significant for users at the cluster edges (the low percentiles). (See Fig. 4.)

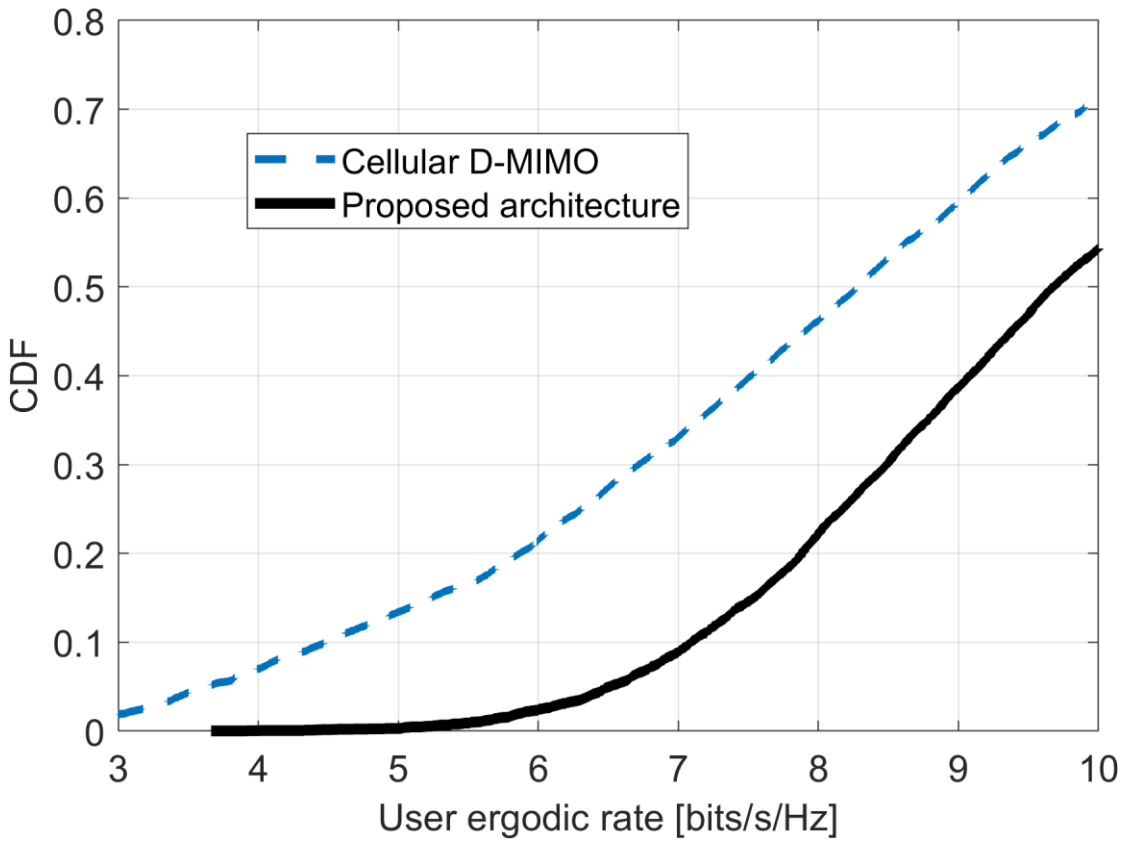


Fig 4. CDFs of the user ergodic rates of a conventional D-MIMO network, and of the proposed architecture.

Modularization examples
Key take-away

Key take aways from UPF modularisation

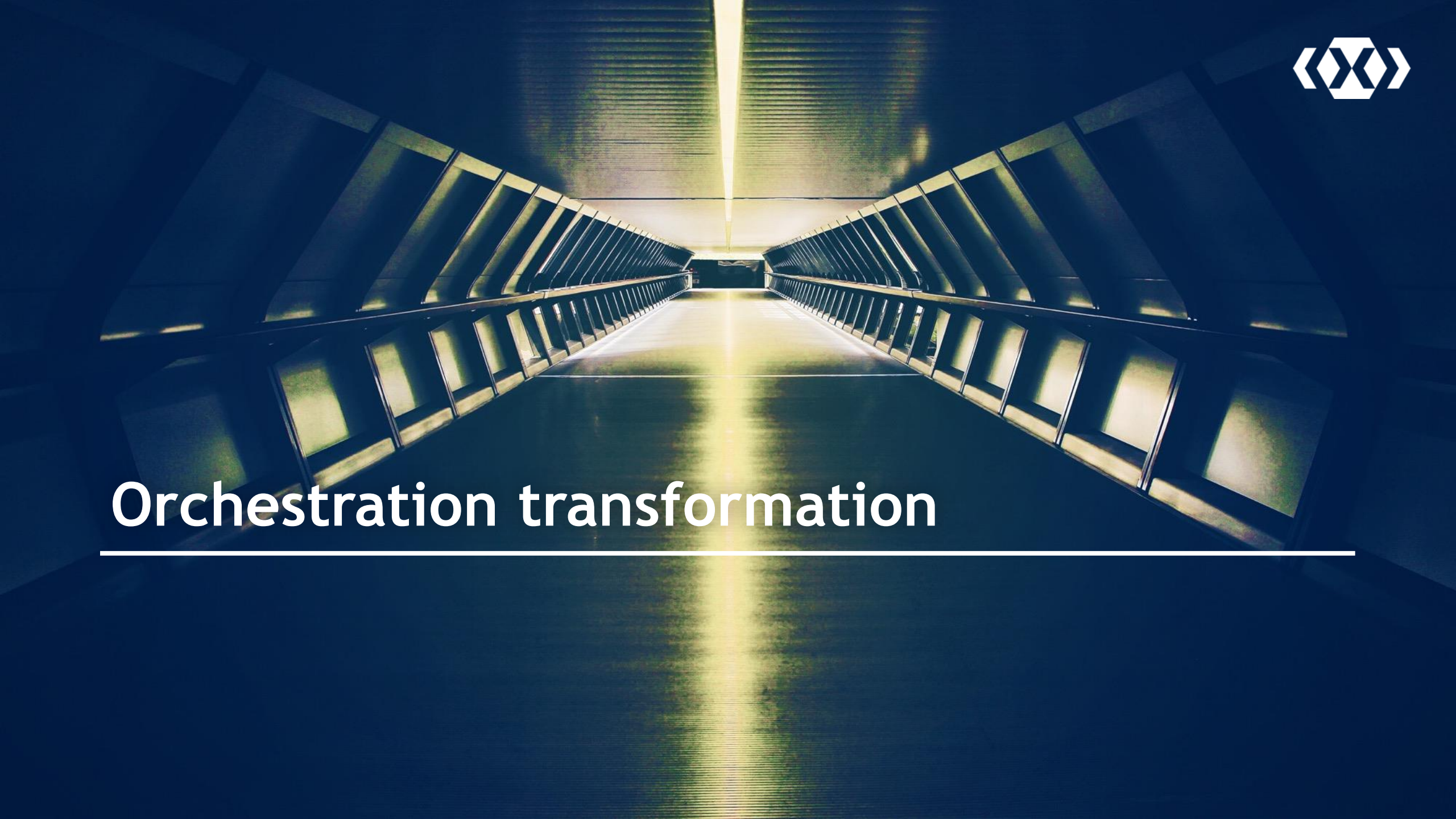
- Aiming at modularizing UPF, which is designed in 5G as one monolithic function with multiple functions and features (i.e., approximately 20 [23.501]), which hinders the scalability of different sub-functions.
- Enables flexible scaling and activation of functionalities on demand resulting in efficient resource utilization and energy efficiency – No quantitative results.

Key take aways from RAN modularisation

- Different RAN architectures have an impact on the performance for D-MIMO
  - Through Cell-free operation: Better user rates compared to cellular networks

Joint key take aways

- Better adaptability to the needs of the network, scaling the different functions as needed. (E.g. modular UPF to better handle asymmetric uplink-downlink traffic, modular RAN to allow for flexible functional splits in the RAN stack).
- Fine granular modularisation comes at the cost of an increased overhead in signaling between modules, as well as more complexity in managing more modules.
- Possible standard impact: 3GPP TS 23.501, 3GPP TS 38.401, O-RAN



Orchestration transformation

Orchestration of the cloud continuum

Orchestration of the Cloud-Continuum

With 6G, **services and network functions deployment** is envisaged to be addressed considering the whole **cloud-continuum**, including the **extreme-edge domain**, that can be **heterogeneous** in technology, massive in **scale**, **multi-stakeholder** and **volatile** in terms of availability

Three complementary solutions have been designed to address the challenges of the cloud-continuum orchestration:

- **Multi-cluster resource management**, providing
  - A unified interface for compute continuum resource management (inventory, provision, operate)
  - Enhanced placement mechanisms to distribute network functions and applications in the cloud-continuum to address resource and proximity constraints
  - Automated discovery mechanisms for virtualization platforms and extreme edge devices
- **Decentralised orchestration,** providing
  - A distributed and decentralized orchestration to handle an increased amount of network services and workloads of different shapes and sizes across multiple stakeholders and domains at scale
  - Automated extreme edge devices discovery with distributed registry of infrastructure resources to ease migration of service components across multi-stakeholder volatile extreme edge nodes
  - Integrated intelligence, with AI/ML algorithms to enable proactive distributed orchestration actions for service assurance
- **Orchestration of the extreme edge,** providing
  - ETSI MEC enhancements for next-generation hyper-distributed applications, such as edge robotics and smart agriculture
  - A lightweight constrained version of a MEC platform (cMEC) to be deployed in mobile end terminals or closest location
  - A solution to address loss of connectivity, near-zero latency requirements, and privacy concerns, while maintaining compatibility with a full-fledged ETSI MEC framework

Orchestration of the Cloud-Continuum
Multi-cluster resource management - description

A **Continuum Multi-Technology Management and Orchestration Platform** is designed to address the challenges that the coverage of the extreme-edge resources introduces on the Compute Continuum (see Fig. 1)

- Takes into account heterogeneous virtualization platforms (e.g., Kubernetes, K3s, Microk8s, OpenStack, etc.), as per the Platform Manager block in Fig. 1.
- Manages extreme edge devices (e.g., IoT devices, Sensors and Actuators, Robots and Cobots, etc.)
- Allows to deploy, migrate and distribute network functions in constrained devices, which are integrated with the Resource Manager through the NBI
- Provides discovery mechanisms of virtualization platforms and extreme edge devices introduced in the Platform Watcher

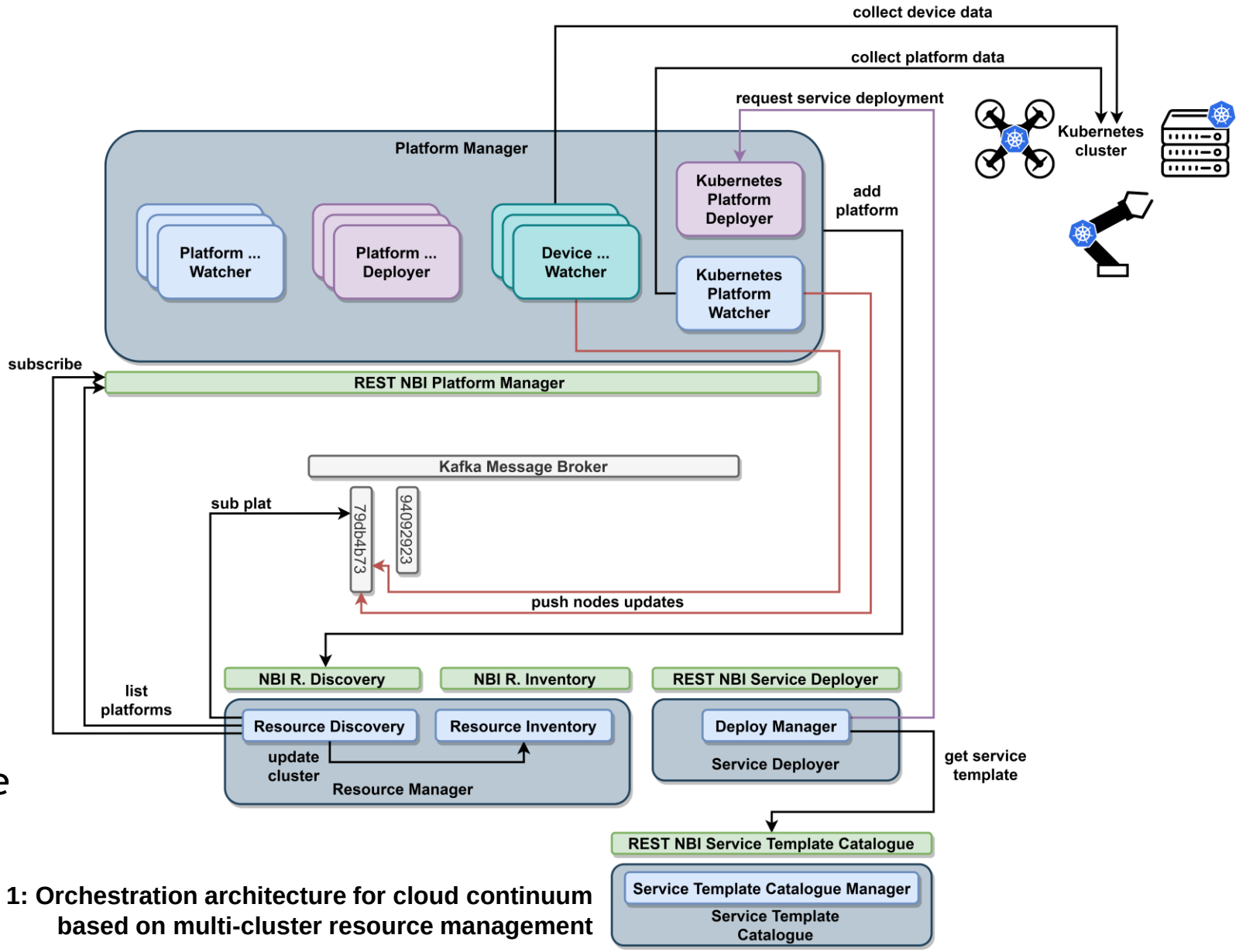


Fig. 1: Orchestration architecture for cloud continuum based on multi-cluster resource management

Orchestration of the Cloud-Continuum
Multi-cluster resource management - evaluation

- A Continuum Multi-Technology Management and Orchestration Platform implementation has been performed, Resource Orchestrator (REC-EXEC as-per D6.3), and involved in the PoC-B demonstration, with
  - Implementation of discovery and monitoring functionalities of the far-edge and edge computing resources (i.e. PoC-dedicated Kubernetes Cluster)
  - Implementation of discovery and monitoring functionalities for the device characteristics (i.e., battery level of the PoC cobots, acting as Kubernetes worker nodes)
  - Implementation of deploy and migrate a video streaming application (i.e., for a video surveillance mission) between cobots
  - Deploy and migrate the closed-loop functions (i.e., analysis, decision, execution) of the closed-loop dedicated to guarantee the continuity of the video streaming service upon detecting a low battery level of the target cobot (i.e., of the deployment)

Summary of WP6/T6.3 implementation results (D6.3)

Orchestration of the Cloud-Continuum

Multi-cluster resource management - protocols and APIs

Interfaces & APIs

- **Platform Manager** software component
  - Enables the dynamic discovery and continuous monitoring of Extreme-Edge, Edge and Cloud Continuum resources (i.e., computing and device characteristics)
  - Enables the service applications orchestration operations
  - Exposes APIs to onboard new platforms (e.g., Kubernetes clusters) to spin up the processes that carry on the discovery and monitory features
- **Resource Manager** software component
  - Creates an inventory of the computing resources and device characteristics of the platforms discovered and monitored by Platform Manager
  - Provides APIs to retrieve the information stored in the internal catalogue
- **Service Template Catalogue** software component
  - Provides CRUD APIs for managing service templates (e.g., HELM Charts, Heat Templates, Kubernetes Manifests)
  - Provides to Service Deployer the templates needed for the orchestration of service applications
- **Service Deployer software** component
  - Provides platform-agnostic APIs for service applications orchestration operations (i.e., deploy, delete, update, status)
  - Translates the platform-agnostic operations requests into platform-specific ones (managed by Platform Manager)

Protocols

- HTTP (REST APIs)
- Kafka Message Broker
  - when integrated with *Integration Fabric* (WP6, Enabler 3)

Orchestration of the Cloud-Continuum

Multi-cluster resource management - key take-aways

- The Continuum Multi-Technology Management and Orchestration Platform is provided as joint work between T3.5 (design) and T6.3 (implementation), and selected functionalities are validated in the context of PoC B
- Tailored monitoring jobs for extreme-edge devices attributes, status and behaviour to feed zero-touch closed-loop automation mechanisms
- Enhanced placement mechanisms to efficiently deploy, migrate and distribute network functions that have proximity constraints
- Automated discovery mechanisms for virtualization platforms and extreme edge devices capabilities
- Unified and abstract interface for compute continuum resource management (inventory, provision, operate)

Orchestration of the cloud continuum
Decentralised orchestration - description

Introduction:

Towards 6G, services deployment is envisaged to be addressed considering the whole network continuum, including also those network resources beyond the MNOs own domain, i.e., the so-called extreme-edge domain, which can be highly heterogeneous, massive in scale, multi-stakeholder, and highly volatile.

Description:

The upcoming Deliverable D3.5 will describe how the well-known network slicing concept would be extended to be in line with this novel Decentralised Orchestration paradigm. As a whole, and in line with the [NGMN] abstractions, the following approach will be applied:

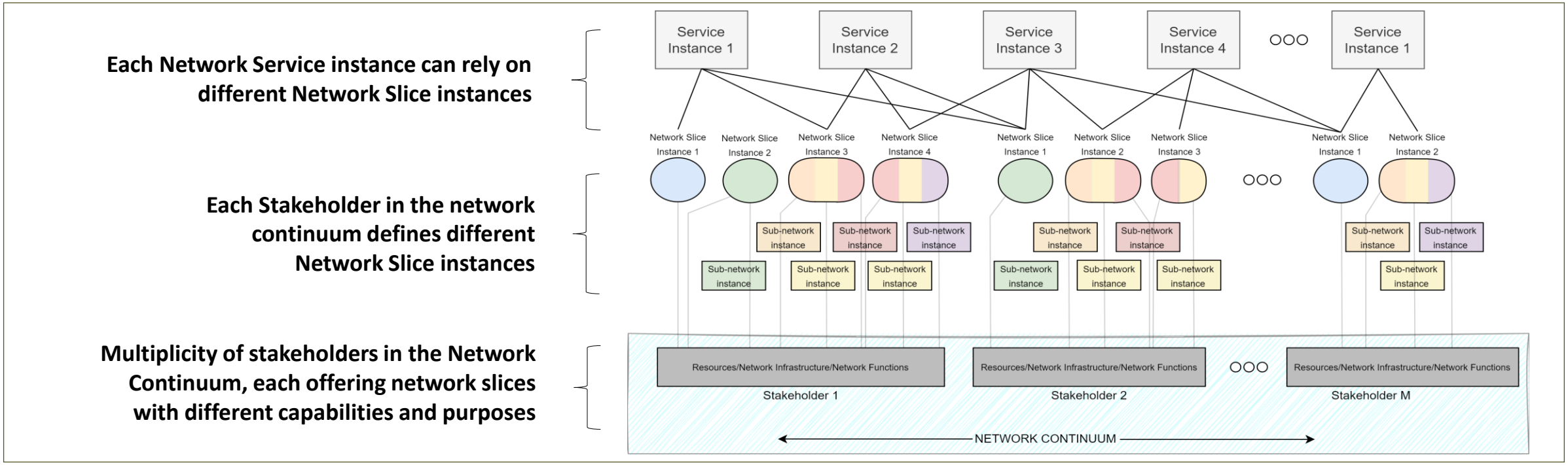


Fig. 1 Network Slicing and Decentralised Orchestration

Orchestration of the cloud continuum
Decentralised orchestration - evaluation

Concept implemented in PoC B to emulate the extreme-edge nodes' behavior and leverage AI/ML techniques to deal with the complexity of the extreme-edge domain, e.g., to predict the random conduct of the extreme-edge devices and implement proactive orchestration mechanisms to ensure service continuity.

Impacted KPIs

- **Scalability**, As a distributed system, the decentralized orchestration system would be able to manage a growing number of network services and workloads of varying types and scales, without the need for complex centralized systems that might become bottlenecks or single points of failure.
- **Latency**. The proposed approach involves deploying service components across the entire network continuum, extending beyond the MNO's own domain, and utilizing network resources from multiple data centers in different regions. This enables the relocation of service components closer to where they are requested by end users. As a result, latency can be reduced for time-sensitive applications, surpassing the capabilities of 5G relying just on the MNO own edge nodes.
- **Flexibility**, in what regards the integration of vertical parties, since they could avoid having to adapt to an external MNO-centric orchestrator, and instead, directly incorporate their own service components exposing their interfaces in a cloud-native manner.
- **Processing Capacity**, considerably expanded by integrating resources at the extreme-edge domain.

Orchestration of the cloud continuum
Decentralised orchestration - evaluation

Impacted KPIs

- **Automation**. Devices discovery processes and registry of infrastructure resources highly automated. Migration of service components through volatile extreme-edge nodes. Deployment of network services and resource placement fully automated.
- **Services Creation Time**.
- **Integrated intelligence**, with AI/ML algorithms to enable proactive M&O actions and for services assurance processes.
- **Reliability**, targeting high-volatility of resources in extreme-edge domains, which could unexpectedly vary their capabilities, move or even fully disconnect.
- **Programmability**, relying on the cloud-native principles as a whole.
- **Maintainability**, through the automation of on-boarding/off-boarding of infrastructure resources.
- **Intent expressiveness**, with network services that can be defined following intent-based approaches.
- **OPEX would be reduced for MNOs**, through delegation on a wide set of external distributed resources and M&O mechanisms.

Orchestration of the cloud continuum
Decentralised orchestration - key take-aways

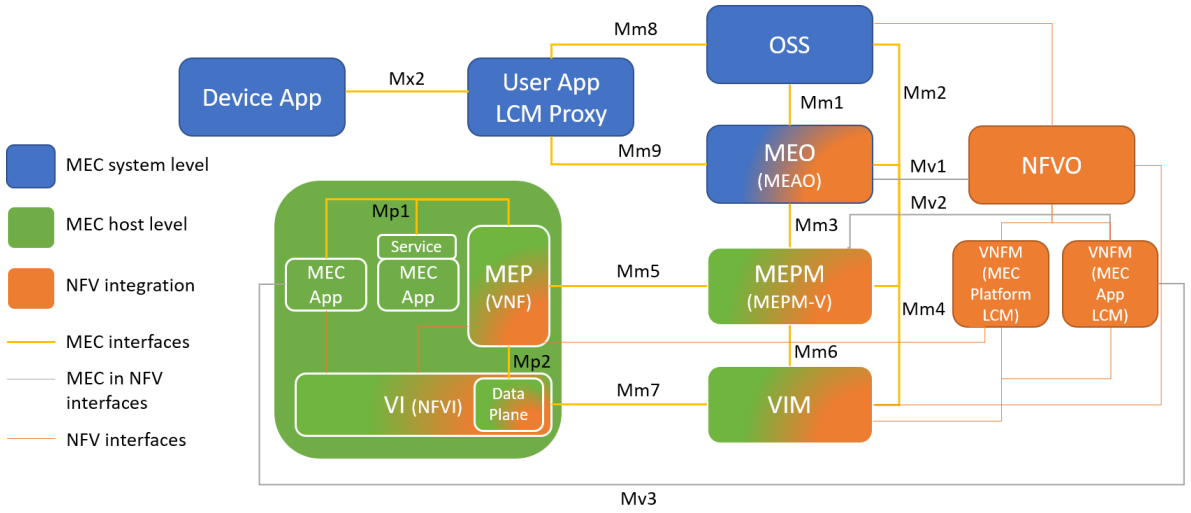
Impacted KVIs

- **Sustainability – Energy efficiency and reduced hardware costs**:
  - Extreme-edge nodes that could be already connected and consuming energy, but not hosting any workload, could be utilized to run network service components. This would eliminate the need to deploy new infrastructure nodes, thereby preventing additional energy consumption.
  - Leveraging extreme-edge devices to deploy network service components would also help reduce the hardware that MNOs or other stakeholders would need to deploy in their data centres.
  - Energy consumption for data transmission could also be reduced, as certain workloads could be executed right on edge and extreme-edge resources, eliminating the need to transmit data to central data centres.

Orchestration of the cloud continuum

Orchestration of the extreme edge - description

- The research discusses the limitations of current ETSI multi-access edge computing (MEC) for next-generation hyper-distributed applications, such as edge robotics and smart agriculture.
- Departing from MEC in NFV architecture (Figure), Introduces a lightweight constrained version of a MEC platform (cMEC) that can be deployed in UEs on their local vicinity.
- Addresses issues like loss of connectivity, near-zero latency requirements, and privacy concerns, while maintaining compatibility with the full-fledged MEC framework.



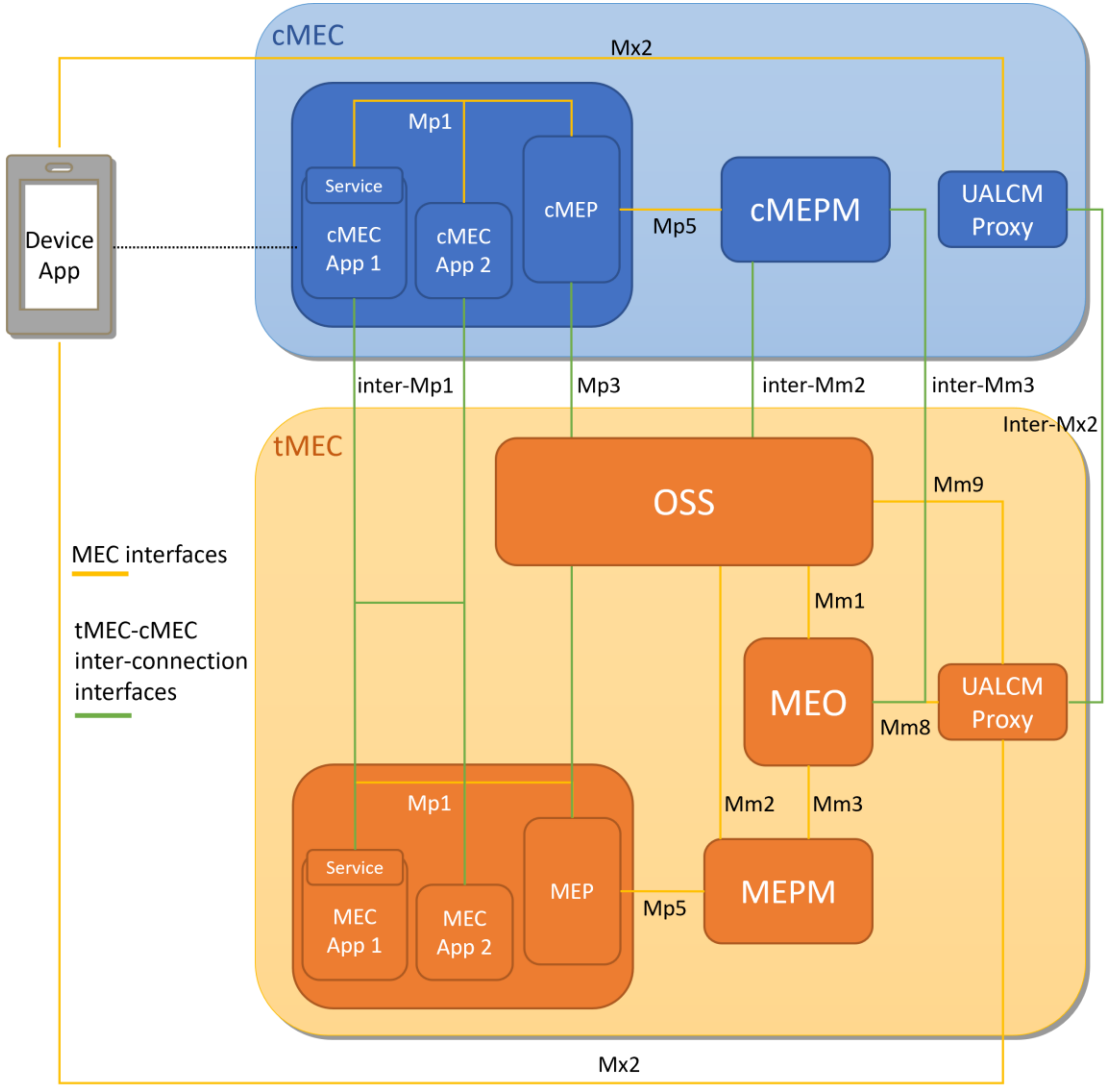
Simplified MEC/NFV reference architecture

- **LCM**: Life Cycle Management
- **OSS**: Operations Support System
- **MEO**: Mobile Edge Orchestrator
- **NFVO**: NFV Orchestrator
- **VNFM**: VNF Manager
- **VNF**: Virtualized Network Function Manager)
- **MEPM**: Mobile Edge Platform Manager
- **VIM**: Virtualized Infrastructure Manager
- **MEP**: Mobile Edge Platform
- **VI**: Virtualized Infrastructure

Orchestration of the cloud continuum

Orchestration of the extreme edge - evaluation

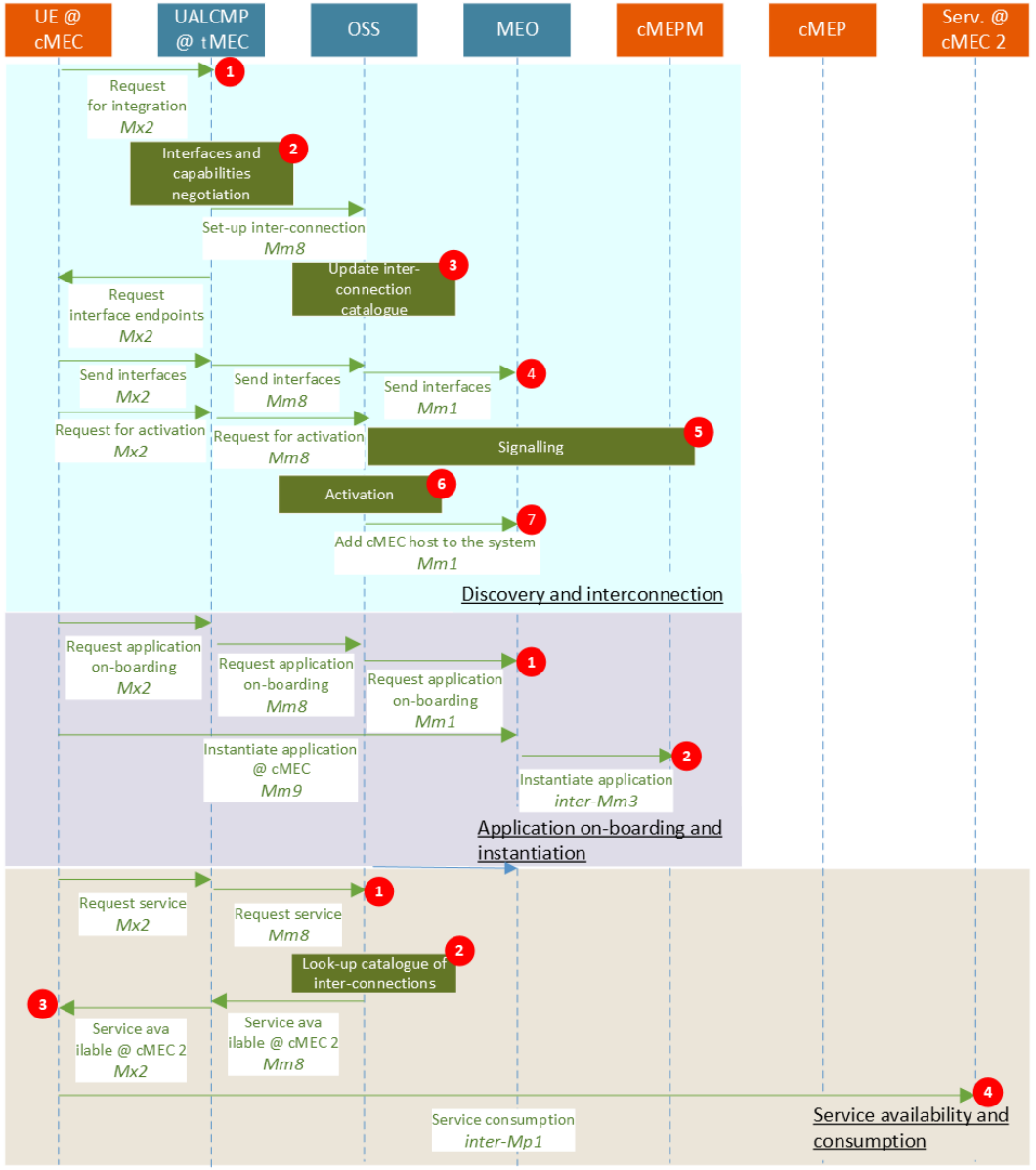
- The research evaluates the proposed cMEC architecture through design options and key use cases.
- Demonstrates how cMEC can enhance service reliability and availability in dynamic environments, reduce latency for real-time applications, and provide better security for sensitive information.
- Highlights benefits in scenarios such as remote eHealth monitoring, AR/VR applications, smart agriculture, and advanced collaborative surveillance.



Architecture scheme of cMEC together with tMEC

Orchestration of the cloud continuum

Orchestration of the extreme edge - protocols & message sequence



This research demonstrates:

- High-level workflows for cMEC integration with tMEC, including discovery and interconnection, application on-boarding and instantiation, and service availability and consumption.
- Illustration of how cMEC devices can advertise their capabilities, request service instantiation, and consume services across the cMEC and tMEC layers, ensuring seamless operation and resource utilization.

Orchestration of the cloud continuum

Orchestration of the extreme edge - key take-aways

1. cMEC extends MEC capabilities to constrained devices (UEs hosting MEC Apps), enhancing service reliability and reducing latency.
2. cMEC addresses privacy and security concerns by allowing sensitive data to be processed locally.
3. The architecture supports dynamic environments and improves resource utilization in edge computing.
4. cMEC provides a flexible and scalable solution for next-generation applications, maintaining compatibility with existing MEC frameworks.
5. Future work includes implementing and evaluating a proof-of-concept prototype to quantify the benefits of cMEC.

Implications on sustainability

1. In home premises, implementing some form of edge with constrained capabilities, may be a suitable way to reduce edge latency in a sustainable way.
2. Some 6G applications will require even less latency than URLLC as provided by 5G.
3. It is expected such an approach will reduce the overall energy footprint
4. Related use case: gaming, XR/VR, local residential gateway

Intent based control of 6G slices

Intent based control of 6G slices

Description

- Network slicing was a key aspect in 5G, letting the network operators to customize the network resources to serve multiple use cases simultaneously
- As the research on 6G evolution considered, a key aspect is to see how different 6G slices from 5G slices. In particular, the identification of a slice, key procedures (e.g., UE registration to a slice), and the interworking between 5G slices and 6G slices.
- The coextended of 5G and 6G slices would also require orchestration of these slices (cf. Fig. 1).
- Intent-based orchestration (Fig. 2) is a way to manage resources exposing intent-based interfaces, emphasizing on uniformity in systems management and enabling the deployment, repairs, and configuration through common components and standardized workflows streamlines operations
- Orchestrator keeps comparing the live state with the user intent to make sure is always satisfied. If not, the network service is reconfigured to meet the requirements.
- In this enabler, the intent-based control of 6G slices are investigated. More specifically, this enabler presents
  - 6G network slicing
  - Intent based management
  - Envisioned impact of intent-based management of 6G slices

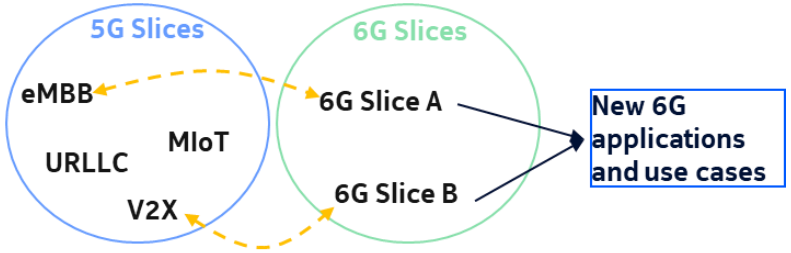


Fig 1. Interworking between 5G and 6G slices

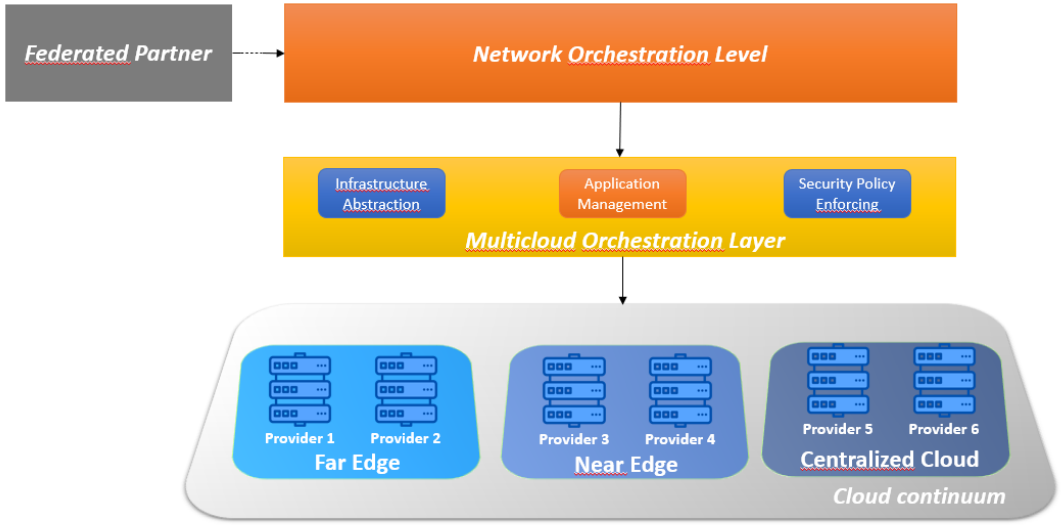


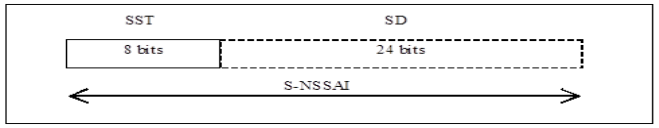
Fig 2. Intent based orchestration framework

Intent based control of 6G slices

6G network slicing problem statement

- ***Problem in 6G:***
  - An S-NSSAI identifies a Network Slice in 5G. ***We have to see how a slice is identified*** in 6G.
  - A ***UE*** in order to access a slice needs to be ***registered in the network with that slice*** - we have to see if the 5G registration is enough for 6G or additional registration enhancements are needed.
  - A network operator may have at the same time both 5G and 6G deployments. ***5G and 6G interworking*** needs to be studied.

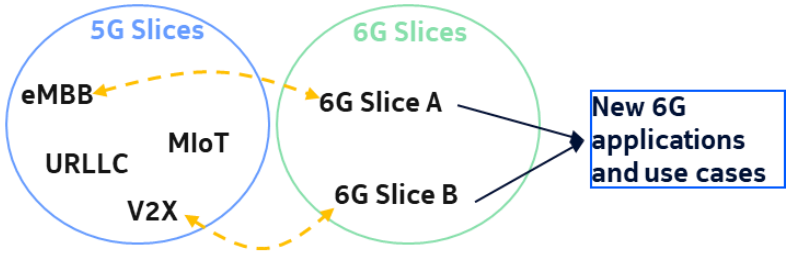
- A Single-Network Slice Selection Assistance Information (S-NSSAI) identifies a Network Slice. An S-NSSAI comprises:
  - A Slice/Service type (SST), mandatory parameter (a number) which refers to a standardized slice and its characteristics.
  - A Slice Differentiator (SD), optional parameter that identifies slices once there are more than one slice with the same SST.



Slice/Service type	SST value	Characteristics
eMBB	1	Slice suitable for the handling of 5G eMBB
URLLC	2	Slice suitable for the handling of URLLC
MIoT	3	Slice suitable for the handling of massive IoT
V2X	4	Slice suitable for the handling of V2X services
Standardized	5…127	Standardized SST range (next standardized slices under definition in 3GPP)
Operator	128..255	Operator specific range

- *KPIs:*
  - Interruption time
  - Interoperability
  - Service continuity

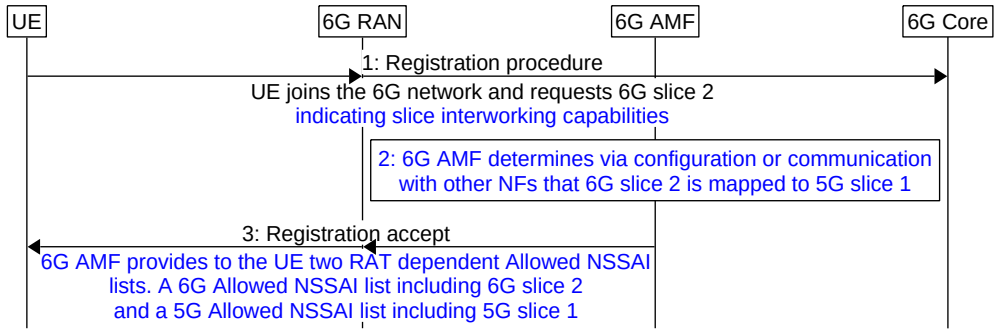
- *UCs:*
  - UE registration to 5G and 6G networks
  - Service continuity when the UE moves



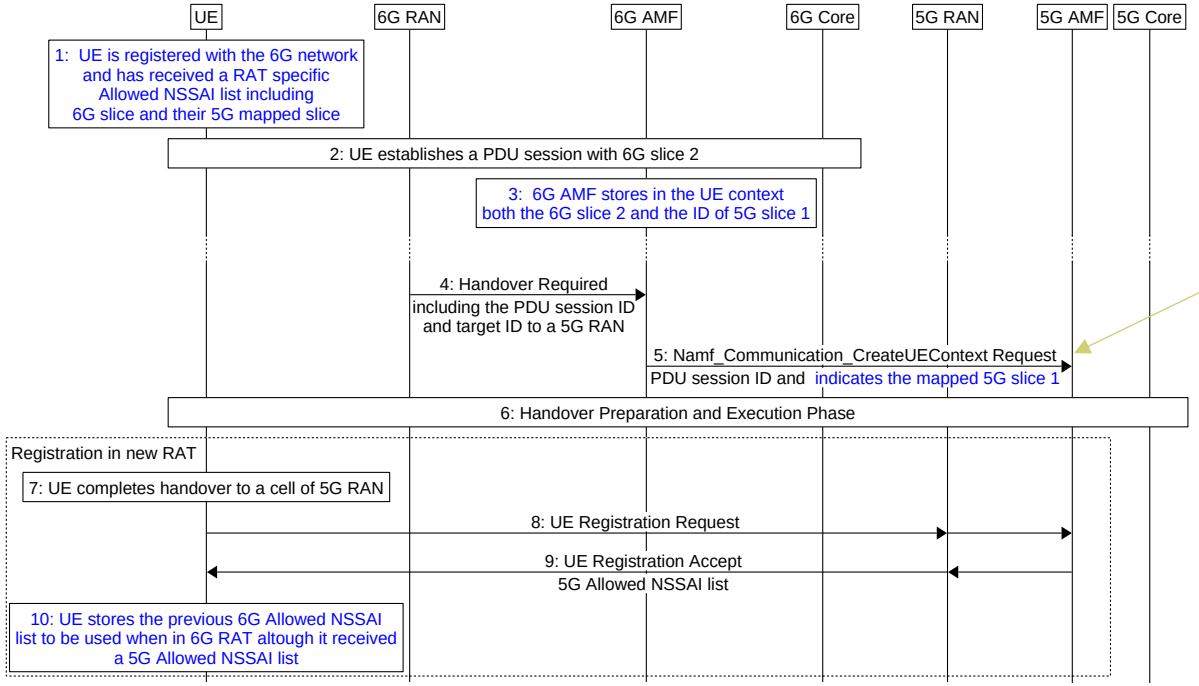
- *Protocols to be impacted:*
  - NAS
  - NGAP

Intent based control of 6G slices

6G network slicing evolution



- **Benefits:**
  - UE does not have to be re-configured with a new Allowed NSSAI list when moving to the new RAT.
  - Utilizing two separate lists per RAT avoids potential issues with the limitation of 8 S-NSSAIs for the Allowed NSSAI



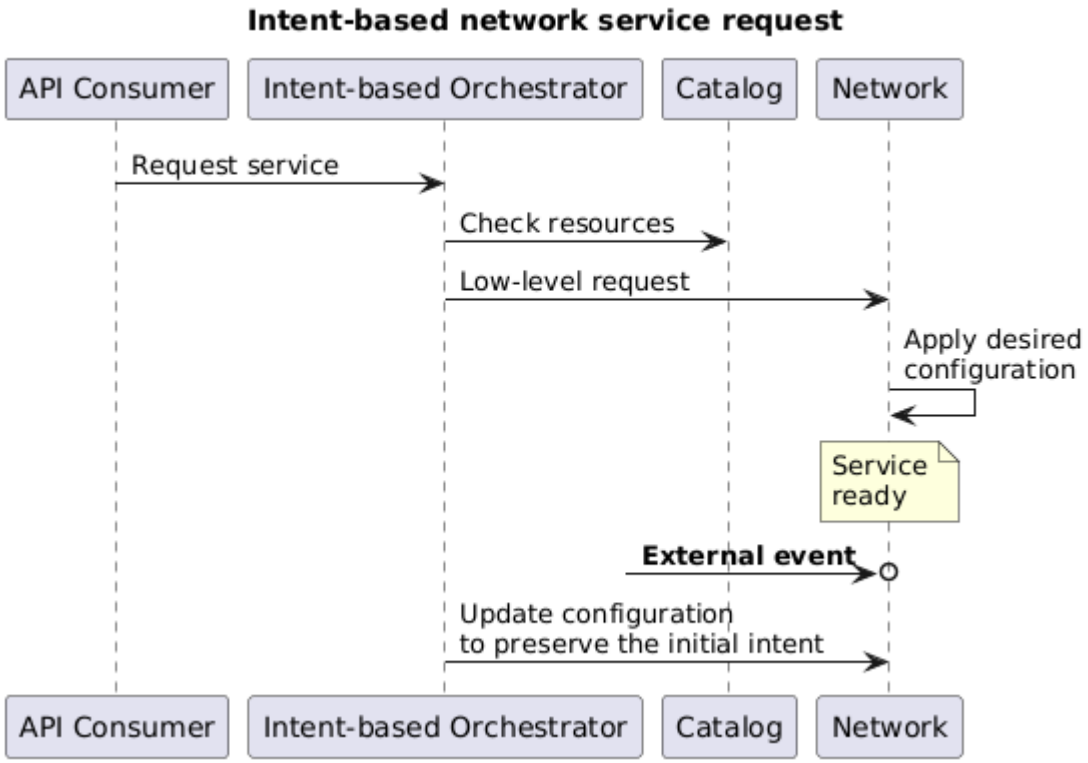
6G AMF performs Switching of 6G to 5G slice and indicates it to 5G AMF

- **Benefits:**
  - Support of service continuity in mobility between 5G and 6G
  - No core network impact

Intent based control of 6G slices

Intent based network service request

- Intent-based network service request



- An orchestrator that leverages an extendable intent-based approach can be used to provide a high-level and abstract way of expressing the desired outcomes and goals for the network and services
- The intent-based orchestrator can automatically translate these intents into network policies and configurations and enforce them across the network. The policies are dynamically applied only on those network function that are affected based on the requested KPIs.
- Orchestrator keeps comparing the live state with the user intent to make sure is always satisfied. If not, the service is reconfigured to meet the requirements.

Intent based control of 6G slices
Benefits

Supporting high level of flexibility, quick failure recovery, fast access at scale and service availability (i.e., needed by 6G use cases such as immersive experience) require not only an enhanced 6G slicing but also management and orchestration support. This enabler explored the 6G slicing concept and its possible extension with intent based orchestration (i.e., matching the WP2 principles of Support and exposure of 6G services and capabilities (#1), Full automation and optimization (#2), Flexibility to different network scenarios (#3), Resilience and availability (#5) and Network simplification in comparison to previous generations (#9).

Benefits of 6G slicing

- UE does not have to be re-configured with a new Allowed NSSAI list when moving to the new RAT.
- Utilizing two separate lists per RAT avoids potential issues with the limitation of 8 S-NSSAIs for the Allowed NSSAI
- Support of service continuity in mobility between 5G and 6G
- No core network impact

Benefits of IBO

- Increased efficiency by lower number of interfaces
- Flexibility in terms of deployment and execution
- Resiliency

Intent based control of 6G slices
Key take-away

Key take aways 6G slicing

- 5G and 6G slicing should co-exist since we cannot ensure 6G coverage everywhere.
- Similar to 5G, we are in favor of using the same the S-NSSAI structure to identify a 6G Slice, since this ensures smooth transition to 6G.
- S-NSSAI structure is sufficient.
- Current S-NSSAI structure contains 32 bits that can accommodate quite a large number of slice IDs thus no issue is foreseen for 6G slicing.
- Existing analysis does not indicate any need to extend the S-NSSAI structure; this ensures also backward compatibility with 5G slicing.

Key take aways IBO

- Gaps analysis from current technologies available in Operators networks
- Enable faster onboarding of network functions to production including provisioning of underlying cloud infrastructure with a true cloud native approach
- Reduce the costs of adoption of cloud and network infrastructure
- Manage a huge number of clusters of servers across the telco network, handling a variety of infrastructure technologies with a uniform and consistent user experience, automatically installing and configuring additional plugins

Envisioned Impact and innovation point of intent-based control of 6G slices

- Slice identification for 6G using the 5G principles ensures interworking in co-existence of 5G and 6G since it enables a smooth transition to 6G.
- UE mobility from 5G to 6G and vice versa enables service continuity in areas with limited 6G coverage.
- Intent based control brings the key innovation points of
  - Intent-based interfaces, which are REST API whose focus in input is on high-level parameters rather than technical ones
  - Native integration between orchestration and network function
  - Live state monitoring

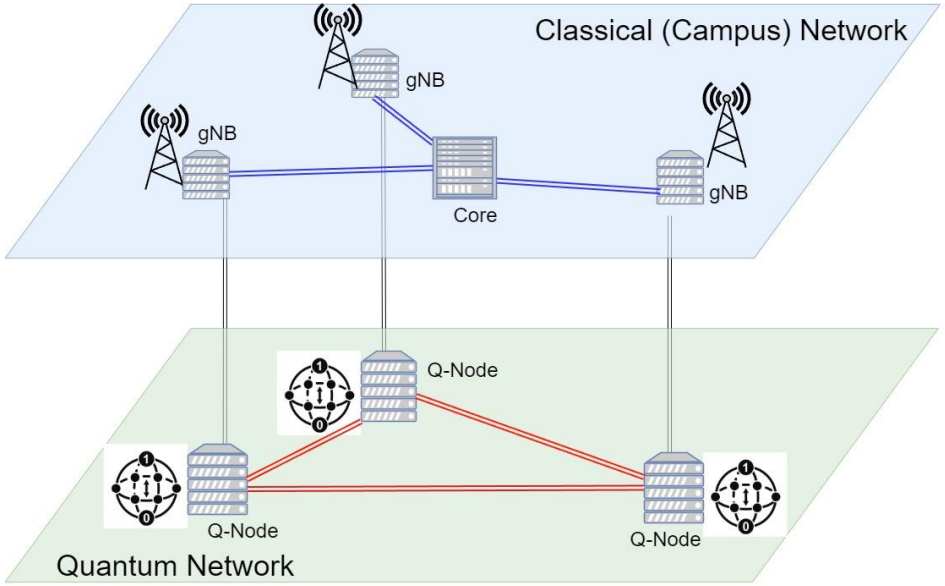
Cloud transformation in 6G quantum architecture

Hybrid Quantum-Classical Protocol Stack for an end-to-end synchronization

A hybrid quantum-classical network architecture is conceptualized. The design of a three-nodes hybrid network focuses on synchronizing quantum and classical links, optimizing resource allocation, and ensuring efficient entanglement generation and distribution. The TUD is designing and realising a full end-to-end synchronization protocol which will be integrated with 5G/6G Campus network upper layers.

The approach used here is grounded in several critical design principles that address the unique challenges of quantum-classical network integration:

- The architecture maintains synchronized quantum and classical links between all node pairs, with the system intelligently routing traffic based on the specific requirements of each process-ing task and requested resource.
- Efficient Hardware-Software Interfaces: These interfaces are responsible for real-time control of optical equipment, including precise timing control for entanglement
- On-Demand Entanglement Generation: High-fidelity entangled photon pairs are generated at the source node using quantum dot technology. This approach minimizes decoherence effects and optimizes the utilization of quantum channels.
- Memory-Free Quantum Operations: In the absence of implementable long-term quantum memories, our architecture implements a novel protocol for on-demand entanglement consumption. This approach leverages advanced scheduling algorithms that coordinate entanglement generation with immediate consumption requirements.



High-level architecture of the hybrid quantum-classical network . The Operational principles of the network are:

- entanglement as a fundamental service
- fidelity as a key metric of service quality
- hardware compatibility across different modalities

Hybrid Quantum-Classical Protocol Stack

The design of the network extends and adapts the classical TCP/IP stack to meet the demands of quantum communication.

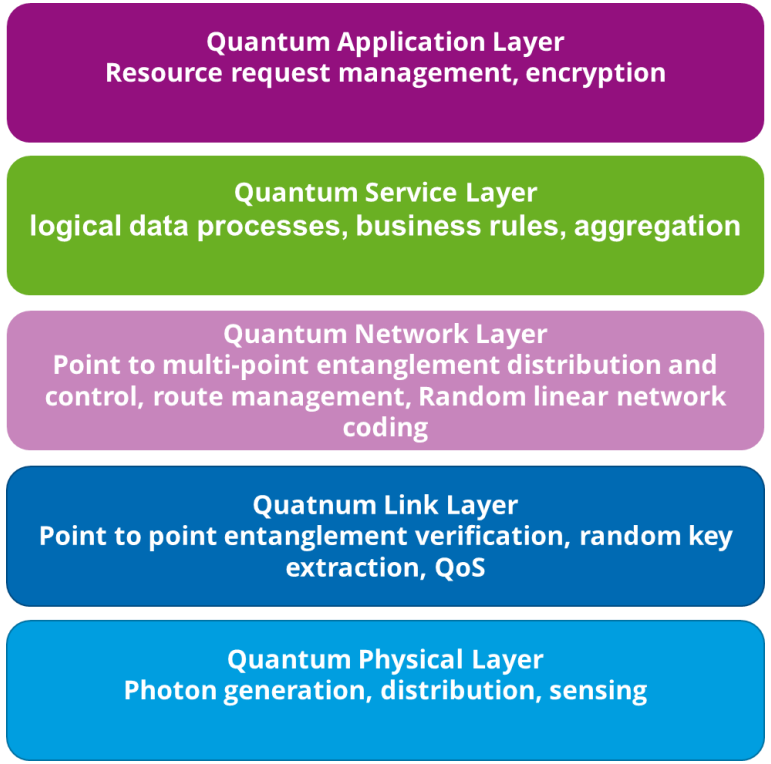
1) Quantum Physical Layer: it handles the transmission of photons between nodes to establish quantum entanglement. This layer requires precise control over the timing of photon emissions and detections to maintain synchronization.

2) Quantum Link Layer: is crucial for managing entanglement generation attempts between the nodes. This layer transforms probabilistic photon entanglement into a reliable correlation service by ensuring precise time alignment across all nodes.

3) Quantum Network Layer: it manages the coordination of entanglement generation among the three nodes. It handles the allocation of quantum resources, such as photon sources and detectors, to maintain consistent entanglement fidelity across multiple network points.

4) Quantum Service Layer: this layer is designed to manage the flow of synchronized entanglement operations and ensure that timing deviations are minimized across the network. Instead of retransmission, it relies on advanced synchronization protocols and real-time adjustments to maintain coherence and entanglement fidelity between the nodes.

5) Quantum Application Layer: in the quantum-classical stack, the application layer supports quantum applications that require synchronized photon correlations, such as secure key distribution, precise time synchronization, and high-entropy random number generation.



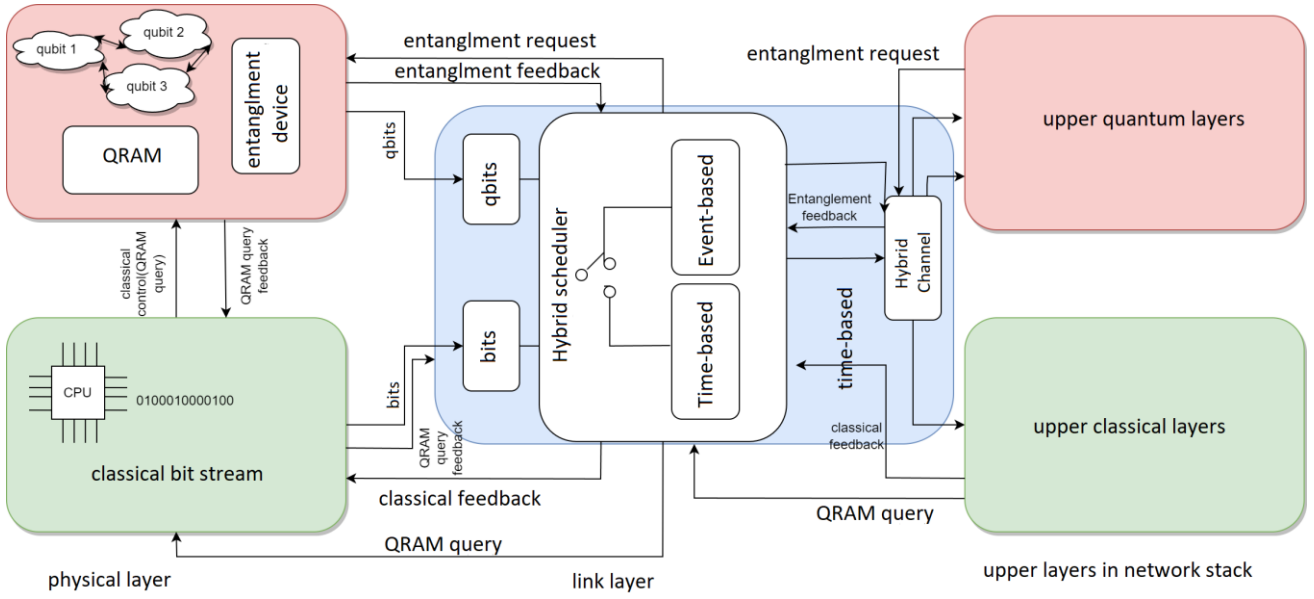
Layered Model of the quantum- classical TCP/IP stack with the key features of each layer

Hybrid Scheduler for Quantum-Classical Network

The hybrid network architecture requires specialized protocols for managing both quantum and classical data. We explore an innovative hybrid datalink layer architecture (see [HEX224-D33]) that is critical for effectively managing the diverse nature of quantum and classical data traffic and security. The main advantage of our model lies in the integration of new quantum hardware at the physical layer while leveraging existing advanced classical infrastructure.

Our approach leverages advanced scheduling algorithms that coordinate entanglement generation with immediate consumption requirements.

- A Time-based scheduler: inspired by the TSN and features the inclusion of an Entanglement slot. The scheduler employs two separated queues, one for quantum packet, associated with a Quantum Gate (QGate), and one for classical packets, associated with a Classical Gate.
- Event-based scheduler: adapts to the current state of the QRAM. When the QRAM is full, indicating readiness for quantum communication, the scheduler prioritizes quantum data transmission



Network structure and flow of information through the stack. Please refer to [HEX224-D33] for more details, see [SB+24]

(Hybrid) Scheduler for Quantum-Classical Network

We performed measurements of error in tranmission of data packets using both types of schedulers.

Measurement methodology:

- Use 16-bit data packets prepared as random bit strings
- Create data stream of 320 bits x 20 packets for: Einstein-Podolsky-Rosen (EPR), Superdense Coding (SDC) and Classical data frames, that is, a 320 * 3-bit =960 bits data streams
- Perform measurement on on 1000 data streams

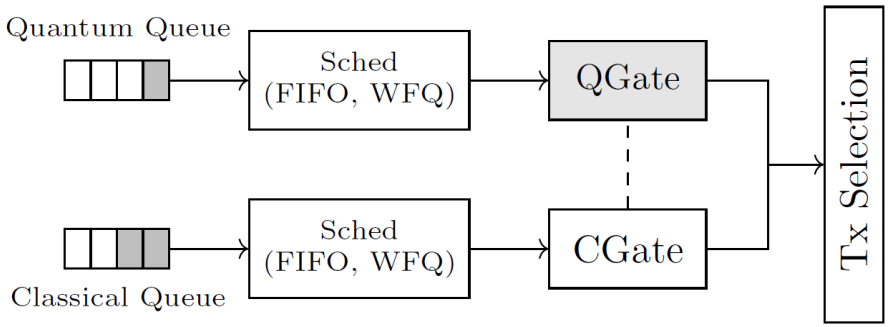
4 bits header	8 bits quantum /classical payload	4 bits ending

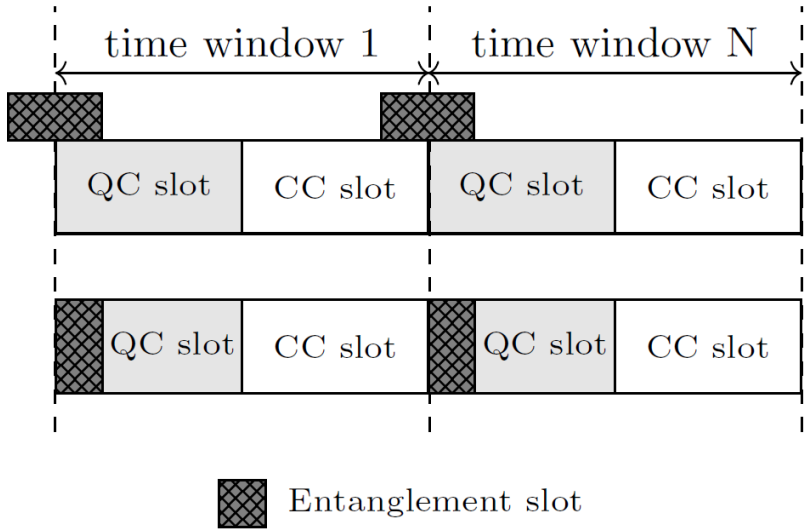
Results:

Shelduer	Accuracy (%)	# Avg. Defected packets
Time-based	96.18	4.346
Event-based	97.16	1.718

To take away:

- Re-sending of defective packets results in higher latency
- A future implementation of adding complexity like WFO and FIFO into the time based scheduling is underway.





Time-based scheduler: (above) layout / (below)window management

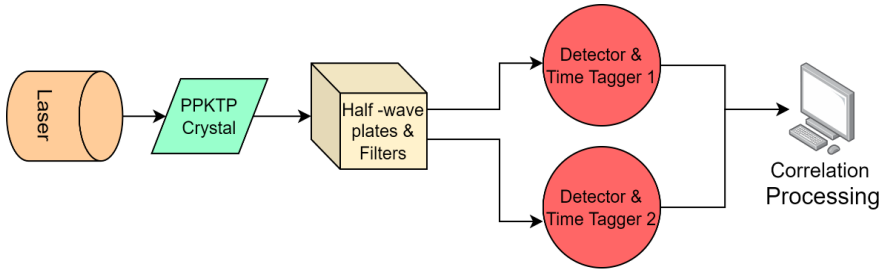
Quantum Synchronization (as a service)

Time-correlated entangled Photons (TCEP) leverages the fundamental property of entanglement to achieve ultra-precise clock synchronization across distributed systems. Two photons become entangled so that their quantum states are interdependent, regardless of the distance separating them. When a pair of entangled photons is generated, the measurement of one photon's state instantaneously determines the state of the other, enabling the simultaneous recording of events at different locations.

The fundamental principle of TCEP is that by measuring the arrival times of these photons at two separate locations (Alice and Bob) and applying the appropriate corrections for propagation delays, one can synchronize the clocks at these two locations to a high degree of accuracy.

TCEP can be applied to enhance Precision Time Protocol (PTP) IEEE 1588 standards to incorporate quantum timekeeping data, enabling the creation of quantum-enhanced PTP profiles such as G.8275.1 for full on-path support and G.8275.2 for partial on-path support.

To take away:
TCEP is efficient, but has hard pre-requisites such as high-quality photon sources, precise photon detection, advanced signal processing, and environmental control.



Time Correlated Entangled Photons (TCEP)

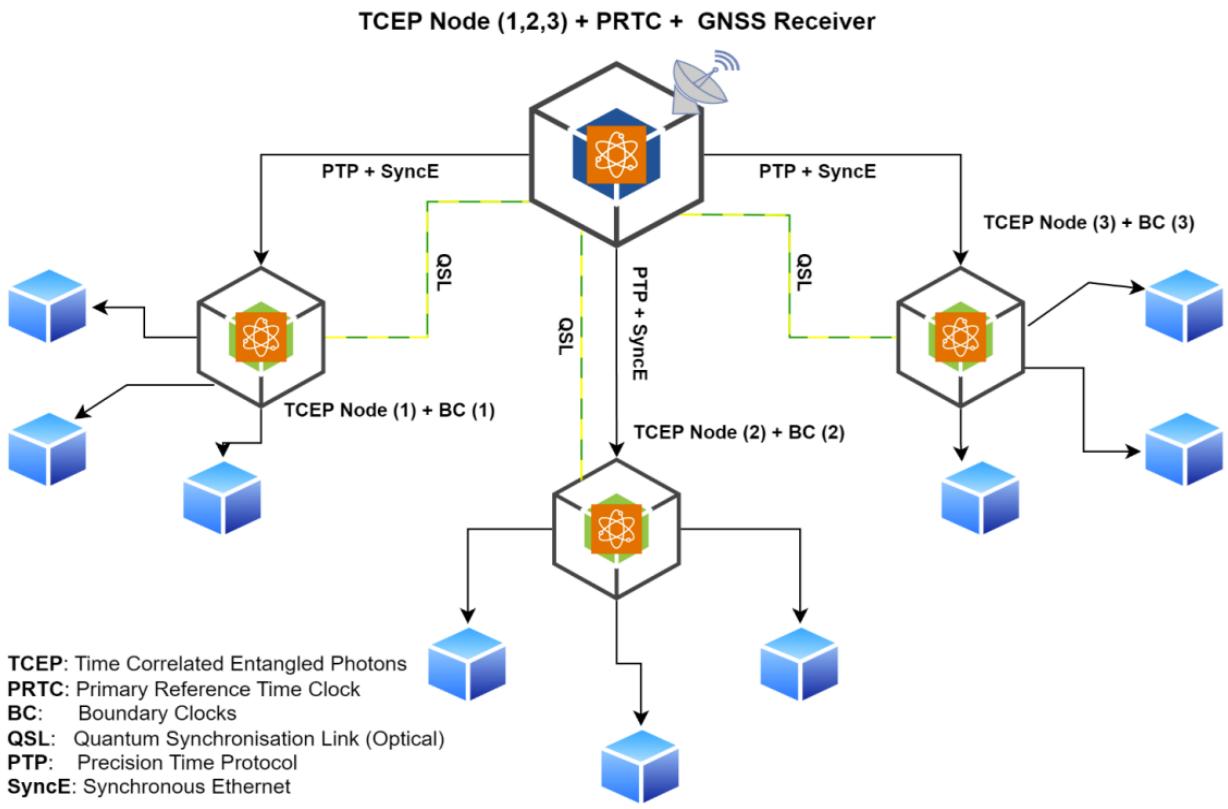
The experiment utilized a Spontaneous Parametric Down-Conversion (SPDC) crystal to generate pairs of entangled photons. The sources used in the experiments included commercial continuous wave (CW) lasers at wavelengths of 810 nm and 1550 nm, with high fidelity (97-99%). The generation rates varied, with up to 1,000,000 coincidences per second achievable, ensuring sufficient entangled photon pairs for precise timing analysis.

Quantum Synchronization (as a service)

Concept for an End-to-End Quantum Coherence, see [HEX224-D33] **for more details**

The central TCEP Node, PRTC, and GNSS receiver ensures unmatched accuracy and coverage

- Quantum-Enhanced Boundary Clocks (BCs)
- PTP&SyncE Synergy on the classical link



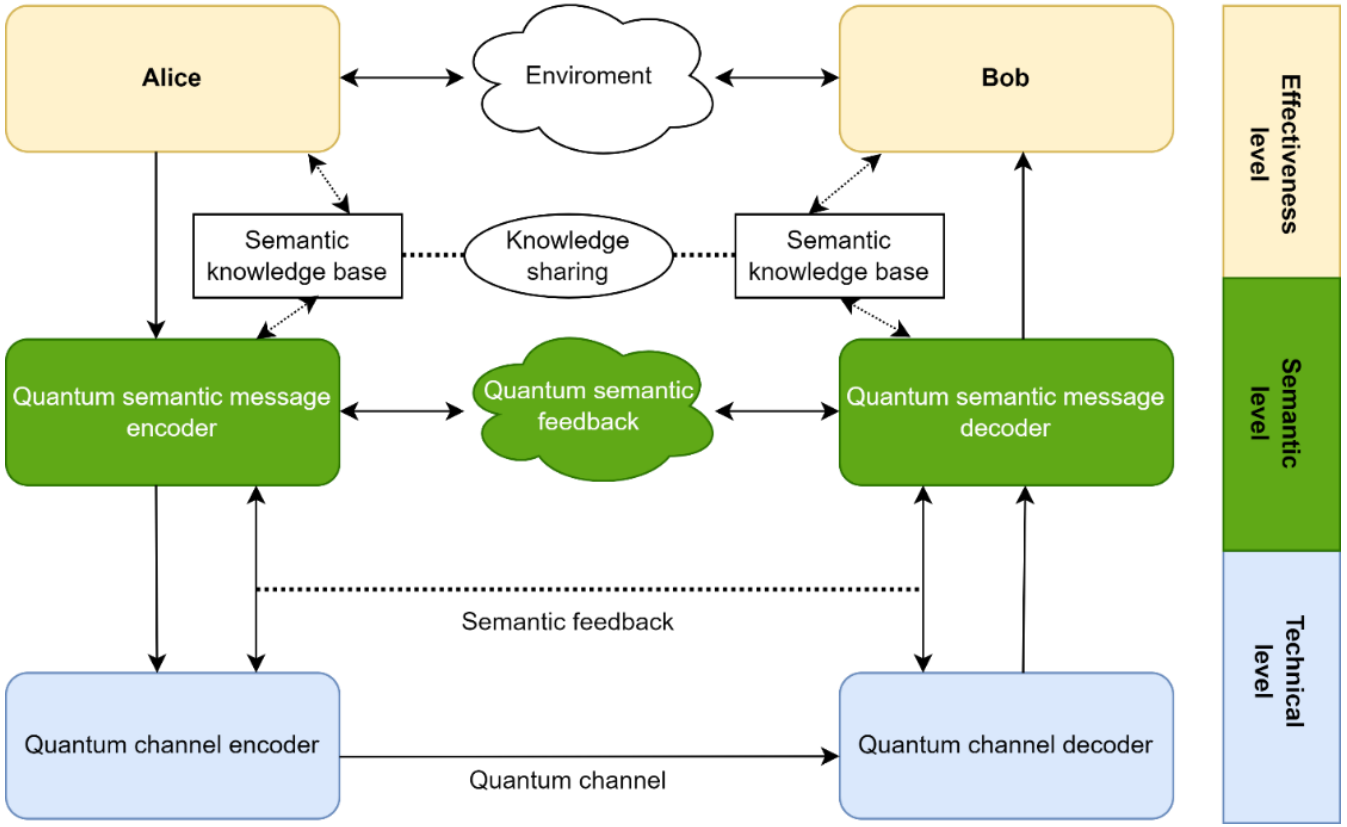
Integration of quantum modules within classical network architecture, see [HEX224-D33].

Network beyond Communication: Quantum semantic communication

Motivation - How to communicate more efficiently over an increasingly heterogeneous network?

Goals –Substantial reductions in the volume of data in the communication scenario while integrating quantum principles, in the context to 6G networks for: a) addressing the increasing data demands, b) providing scalable methods by novel techniques, and c) dealing with computationally complex data type.

Several novel functionality are introduced, aimed at establishing a quantum layer within the semantic level to address the challenges related to the complexity and cost of transmitting knowledge graphs. In the next section, we describe the brief functionalities of our approach.



The schematic representation of three layer communication problem with a quantum layer, See [NN+24]

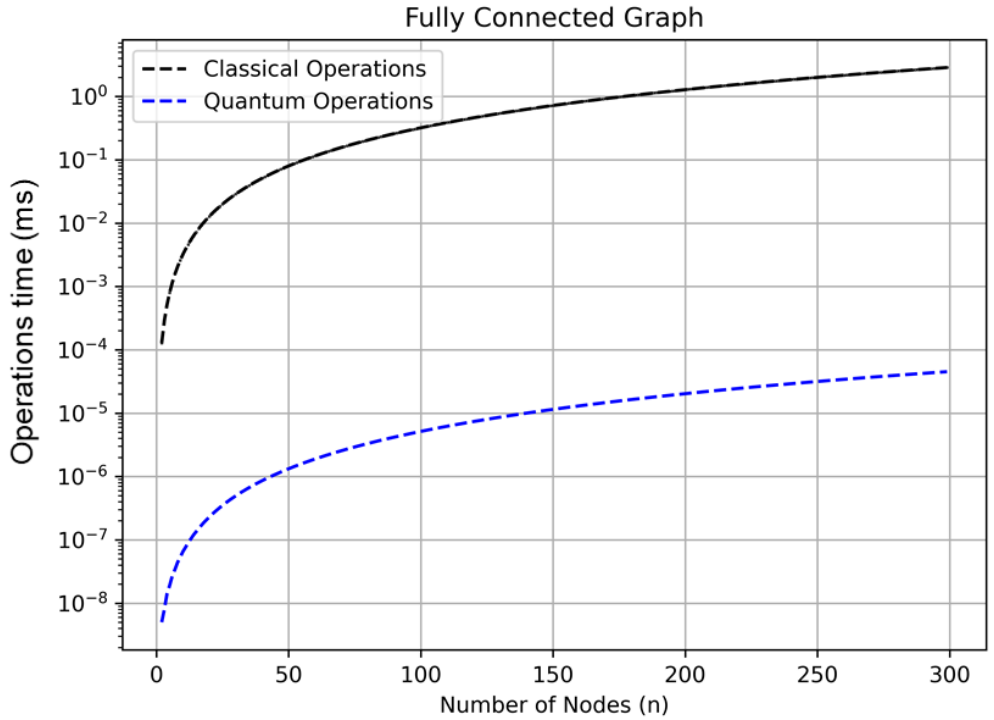
Network beyond Communication: Quantum semantic communication

In evaluating the effectiveness of our protocol, we under-score the trade-offs encountered between the rising complexity of semantic graphs and reliability.

To show the advantage of using quantum technologies, we evaluate the theoretical computational complexity of our protocol, in contrast to conventional wireless communication scenario.

Inspired by the concept of quantum supremacy, wherein a quantum computer performs tasks beyond the capabilities of classical supercomputers, we define the complex computational gain as the ratio of the time taken by a quantum algorithm to scale graphs compared to the estimated time required by the best classical computers. Our scalability analysis demonstrates a ‘quartic polynomial’ gain over conventional methods for computing complex graphs.

Protocols involved:
Semantic communications
Entangled-assisted quantum communication



The tradeoffs between increasing nodes in a graph vs computational complexity, see [NN+24]

Novel cloud functions

Multi-domain/Multi-cloud federation

Multi-domain/multi-cloud Federation - Description & Evaluation

- Multi-domain/Multi-cloud federation is the capability to aggregate cloud services provided by multiple domains and providers into a single, coherent cloud. The federation concept involves:
  - the capability of spanning across the administrative domains of different legal entities (e.g., network operators, nations)
  - the multi-cloud capability, i.e., the aggregation of underlying cloud resources built on different cloud technologies, including private and public cloud.
- Evaluation
  - Main potential KPI improvements
    - Latency, network load, reduce complexity for cross-domain deployment, improve QoE
  - Which WP1 use cases does the enabler solve?
    - Cloud Continuum
  - Which WP2 principles does the enabler match?
    - (#1) Support and exposure of 6G services and capabilities
    - (#2) Network Scalability
    - (#7) Internal interfaces are cloud optimized

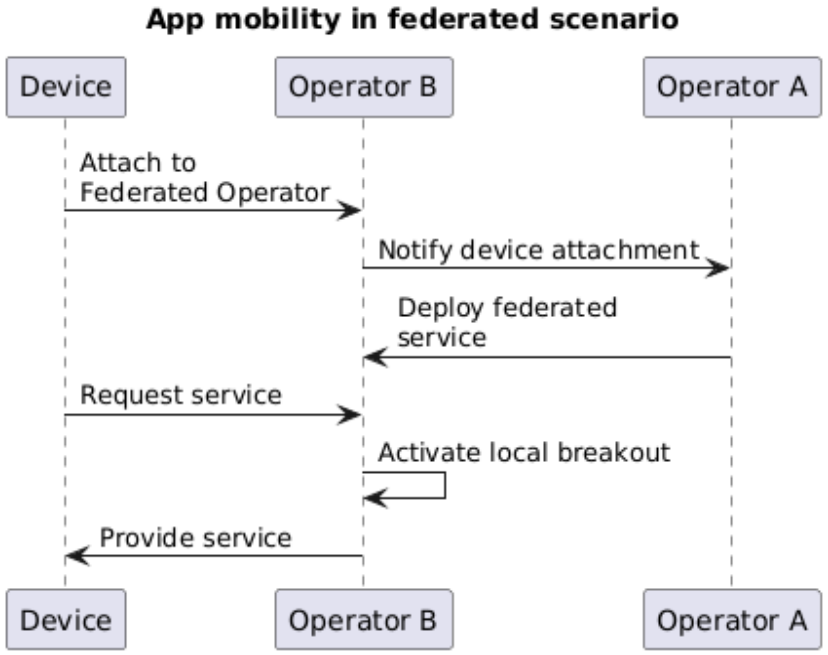
PoC

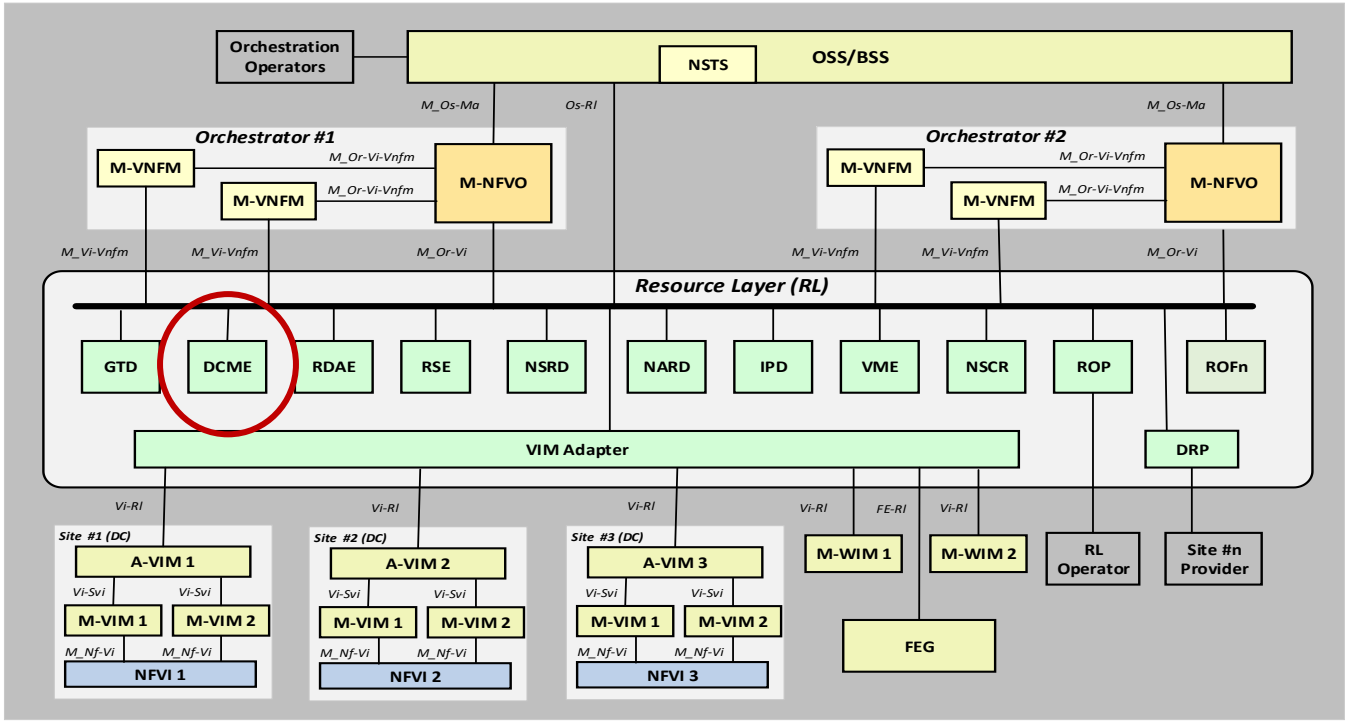
In the scope of this enabler, a PoC has been developed. This Multi-cloud PoC uses a central control plane for several cloud control planes, leveraging Cluster API, ArgoCD, GitOps

Our architecture assumes a fresh Kubernetes installation on every node in the cluster allowing for both on-premises and on-edge services to be managed by the framework

Fault tolerant storage, Secret management, Distributed and harmonized storage, as well as on-prem/on-edge complexities are still being studied

Multi-domain/multi-cloud Federation - Implication on the 6G standard





- Location updates
  - Federation module is integrated natively in core networks and devices can request services cross-operator

- Data Centre Monitoring Engine (DCME) evaluates in real-time each data centre status, i.e., energy consumption, performance, and reliability, updating its DCF (Data Centre Features)
- Data sources for DCME
  - CAdvisor (as part of Kubelet on each worker node)
  - Kepler (Kubernetes-based Efficient Power Level Exporter)

Multi-domain/multi-cloud Federation - Key take-away

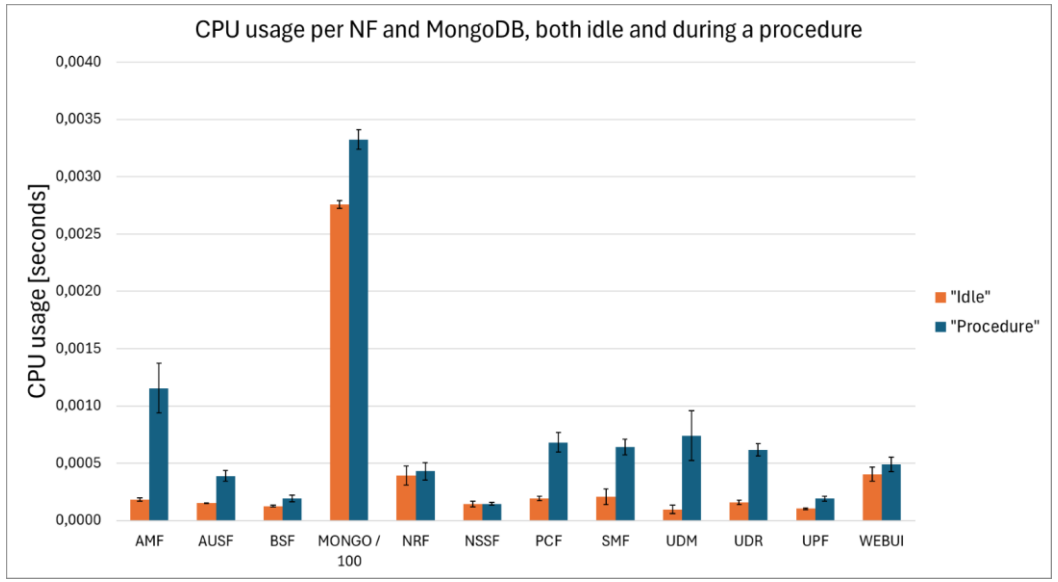
Benefits and key take-away

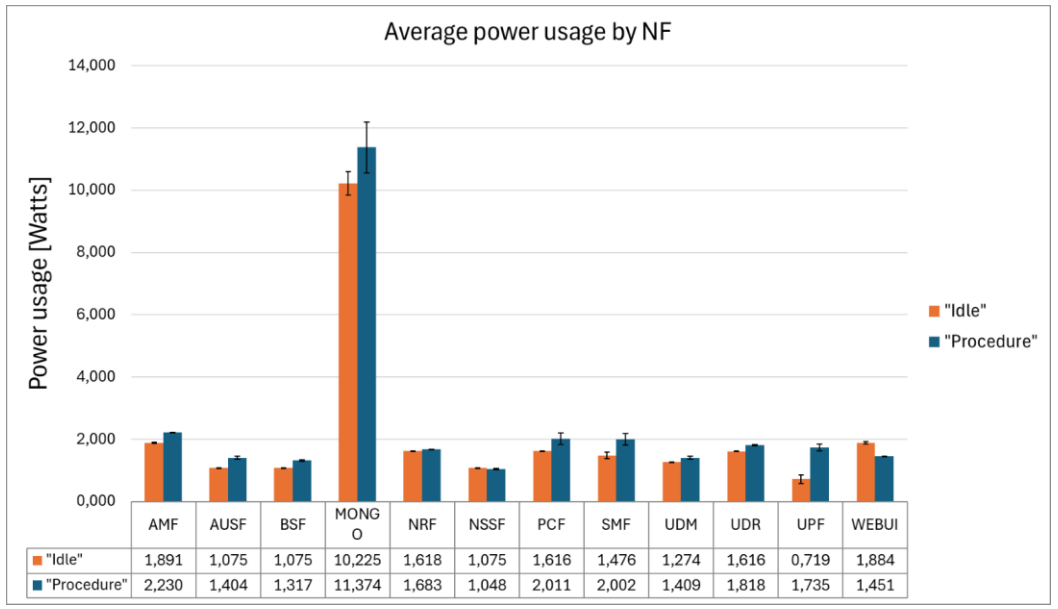
A Multi-domain framework allows for the management of multiple cloud service providers, providing the following benefits:

- Seamless user experience in all EU territory, applications work as in the home country even for edge applications (e.g, Immersive Reality App). Identify gaps from 5G architecture (e.g., authentication network functions should interact with federation modules to trust identities in federated scenarios).
- Access to more cloud resources, raising the flexibility of services that can be made available, the devices that can be reached and allow for cross-domain deployment
- Higher resource availability and flexibility also lowers the total network load and latency times in service deployment, as well the ability to leverage computing power and storage that is superior to current solutions
- A central management unit can deploy services using the best choice of resources available at the time

Experimental results:

- Resource (CPU, Memory) and Energy usage per network function per deployment for single infrastructure provider - providing data to the Data Centre Monitoring Engine (DCME)
- Obtained data show disproportionate resource usage of the MongoDB database compared to core network functions. The gap diminishes when looking at power usage
- The same principles and measurements can be carried out for other Core NFs implementations, allowing for efficiency comparison between deployments and providers





	AMF	AUSF	BSF	MONGO	NRF	NSSF	PCF	SMF	UDM	UDR	UPF	WEBUI
"Idle"	1,891	1,075	1,075	10,225	1,618	1,075	1,616	1,476	1,274	1,616	0,719	1,884
"Procedure"	2,230	1,404	1,317	11,374	1,683	1,048	2,011	2,002	1,409	1,818	1,735	1,451

Novel Services in the transformed architecture

Application-layer BCS optimisation

Application-layer BCS optimisation Description

Application-layer BCS Optimization Overview: In 6G networks, Application Functions (AFs) play a vital role in ensuring efficient resource usage. These AFs interact with **Exposed Network Functions (ENFs)** through standardized APIs to perform tasks with minimal latency and optimal resources (see Fig. 1.). Examples include Navigation AF and Object Detection AF, which rely on real-time data from the network, such as the Sensing Function and Device Location.

Joint Network and Compute Optimization (INC): The research presents an approach for integrating network and compute resources to meet the requirements of BCS applications **across the cloud continuum (see Fig. 2.)**, ensuring they meet stringent network and compute requirements. The strategic placement of Network Functions, like User Plane alike functions, at different cloud levels (edge, far edge, cloud) enhances the ability to handle traffic, particularly for critical BCS applications such as Navigation AF and Object Detection AF.

New Signaling for Constrained Devices: This research also proposes a new protocol for constrained devices, such as those used for sensing in 6G networks. These devices have limited capabilities and energy resources, and current protocols are unsuitable. A new temporary ID mechanism is suggested to improve registration procedures while maintaining security and data integrity, significantly reducing signaling between the device and the network (see Fig. 3.). The placement of functions across the cloud continuum further ensures the efficient handling of traffic for constrained devices, aligning with the INC optimization strategy.

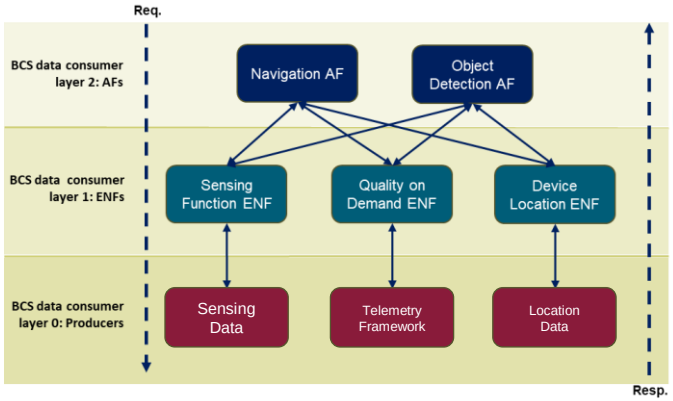


Fig. 1: Application-Layer and Network Function Interactions in 6G BCS Optimization.

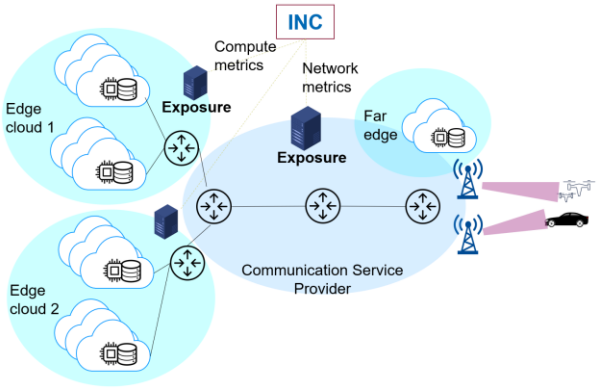


Fig. 2: Integration of Network and Compute (INC) server collects network and compute metric to decide optimized placement of application.

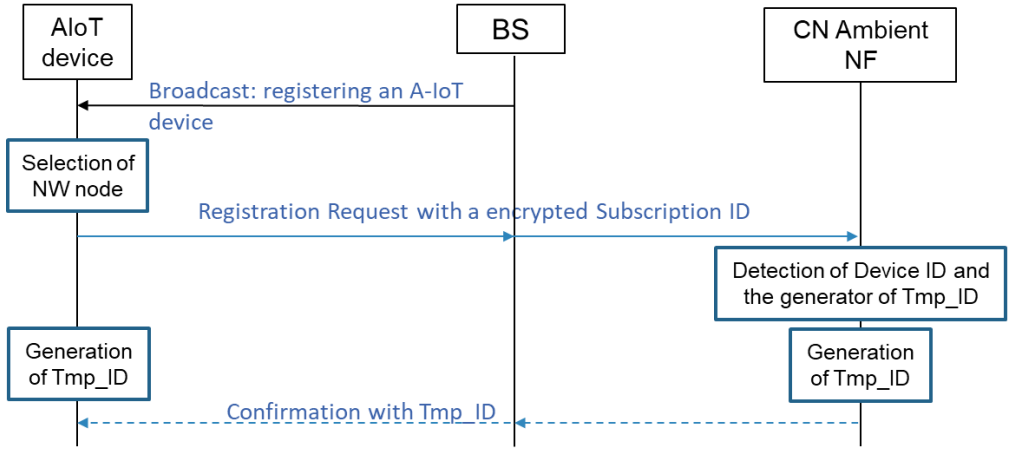


Fig. 3: Message Sequence Chart for Temporary ID-Based Registration in Constrained A-IoT Devices

Application-layer BCS optimisation Protocols and APIs (1)

Key application and network functions overview:

- **Key AFs:**
  - **Navigation AF**: Manages autonomous navigation within various environments.
  - **Object detection AF**: Detects and monitors objects in real-time.
- **Key ENFs:**
  - **Sensing function ENF**: Provides real-time environmental data.
  - **Device location ENF**: Tracks the location of devices and assets.
  - **Quality on demand ENF**: Adjusts network resources dynamically to meet specific AF requirements.

Process overview (see Fig. 1):

1. Orchestrator requests relevant ENFs though an ENF gateway upon being triggered by intents.
2. The network provides the requested information through the available ENFs.
3. The Orchestrator utilizes this information to identify the most suitable nodes to place the AFs based on their operational needs.
4. The AFs maintain interaction with the relevant ENFs mandatory for their operation.

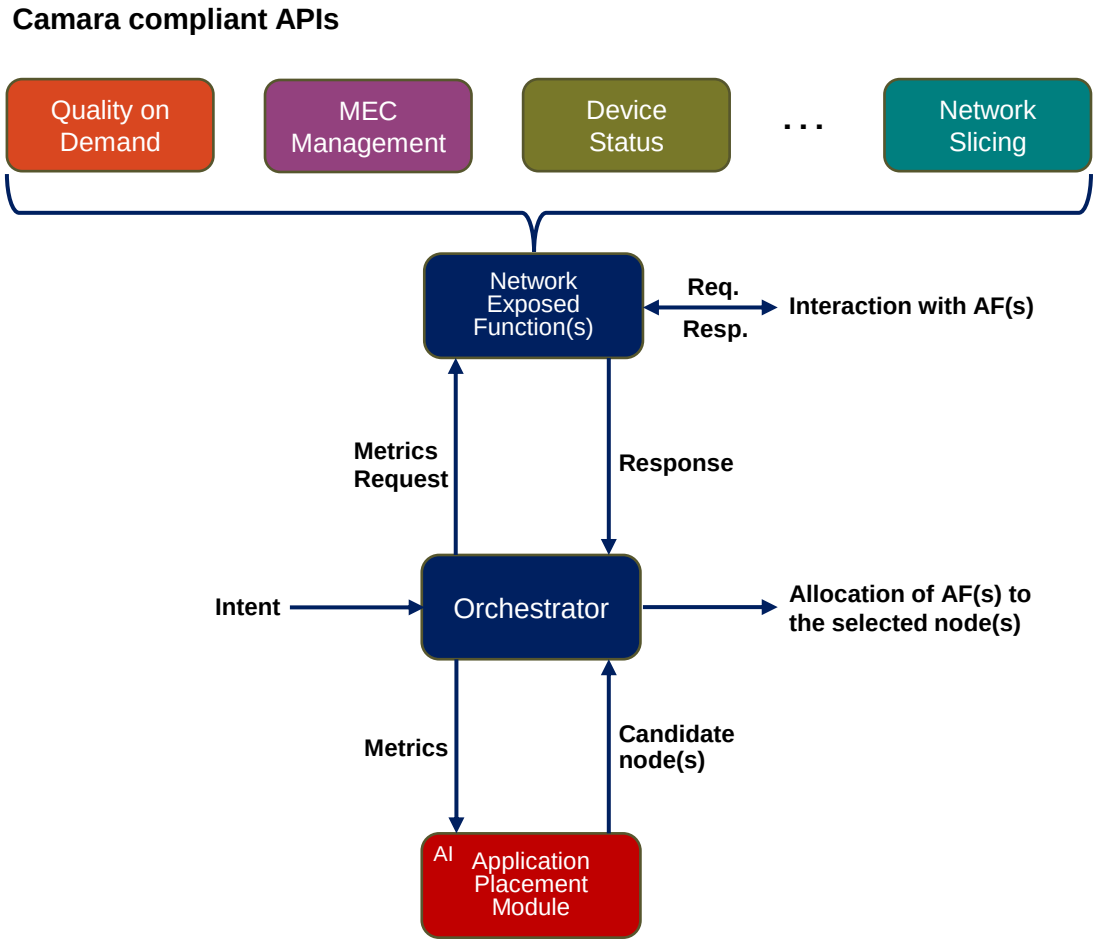


Fig. 1. Orchestration of 6G Network Capabilities Using CAMARA-Compliant APIs

Application-layer BCS optimisation Protocols and APIs (2)

INC Optimization Process:

- **Metric Acquisition:** Network and compute metrics are continuously exposed to the Integrated Network and Compute (INC) server (see Fig. 1).
- **Optimization of BCS App Placement:** The INC server processes the placement request (with network and compute requirements) and makes a placement decision based on the collected metrics. This is followed by the decision enforcement process.

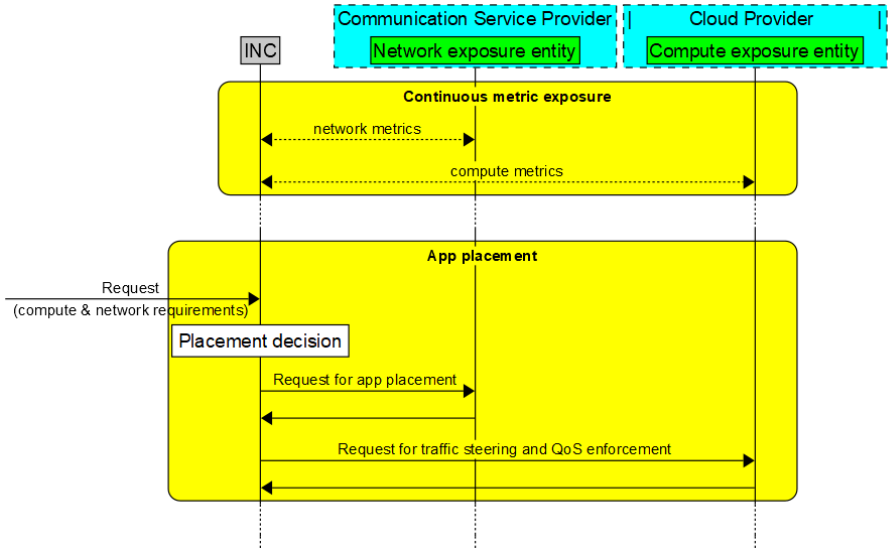


Fig. 1. A generic call flow for coordinated network and compute optimization

AIoT Device Communication Protocol:

- A new **Tmp_ID** (Temporary ID) is generated locally by the AIoT device, enhancing security by using encrypted communication between the device and the network (see Fig. 3).
- An **AIoT Function** handles both data and control of the AIoT device, while the device context (such as security capabilities) is fetched from a device database (see Fig. 2). This ensures efficient and secure communication between the AIoT devices and network functions.

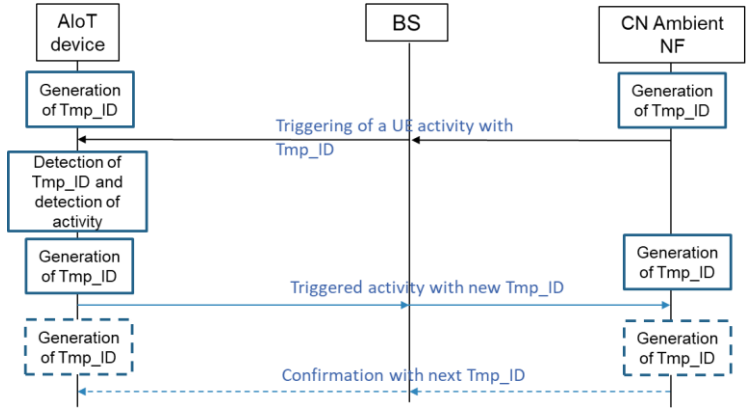


Fig. 3. Message Sequence Chart for Tmp_ID-Based UE Activity in AIoT Devices

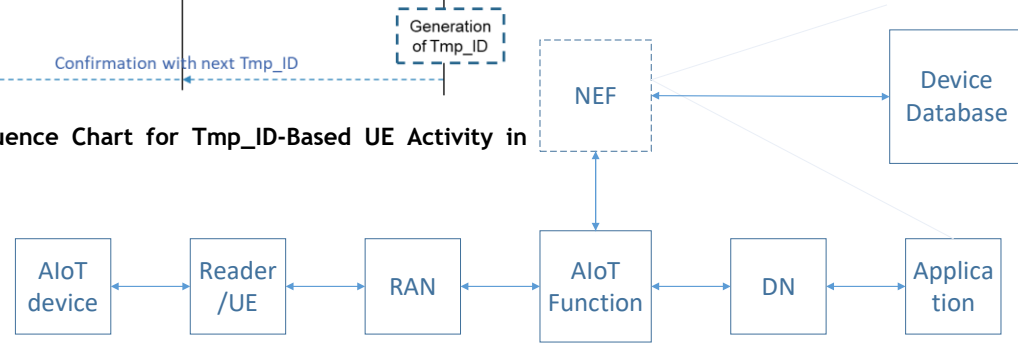


Fig. 2. AIoT Device Communication Flow with NEF and AIoT Function

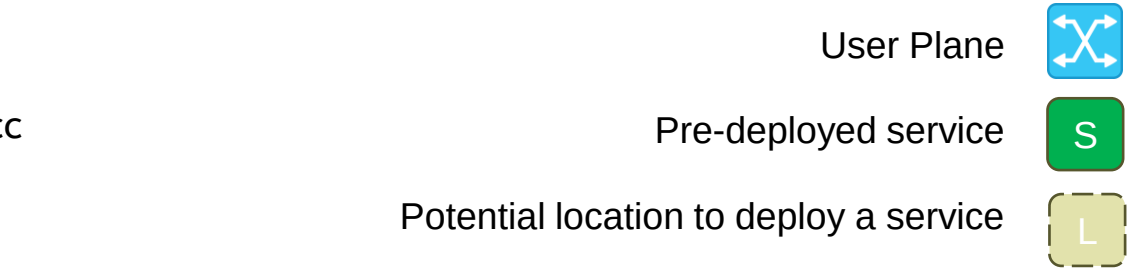
Application-layer BCS optimisation - Network module placement across cloud continuum workflow

Two scenarios are considered:

- Scenario 1: selection of a service among pre-deployed instance across cc (see Fig. 1).
- Scenario 2: on-demand deployment of a service across cc (see Fig. 2).

The related workflow is as follows:

- Offline network & compute metric acquisition.
- Request for the selection/deployment of service.
- Optimal selection of service instance/ deployment location.
- Coordinated configuration of UPs across cc.



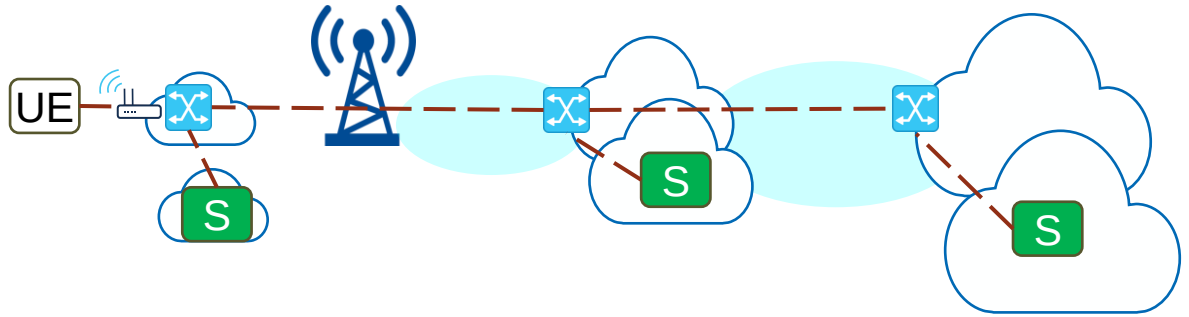


Fig. 1. Scenario 1: selection of a service among pre-deployed instances across cc

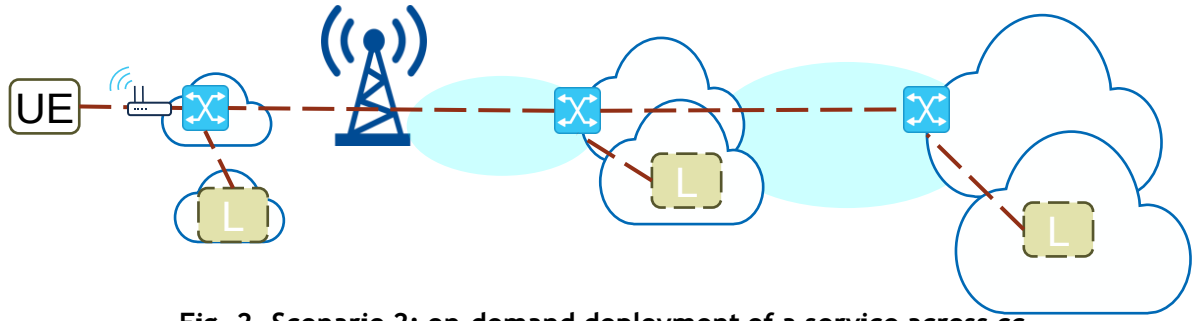


Fig. 2. Scenario 2: on-demand deployment of a service across cc

Application-layer BCS optimisation - Application Placement Optimization Evaluation/Benefits

The Genetic Algorithm utilizes the information provided by the ENFs to make informed decisions by determining the optimal node(s) to place the AFs. Benefits, include, among others:

- **Latency minimization**: The algorithm selects the most appropriate nodes to minimize end-to-end latency between AFs and network resources.
- **Energy efficiency**: By choosing energy-efficient nodes, the GA contributes to overall network sustainability.
- **Data volume exposure minimization**: The GA strategically places AFs to minimize data exposure across the network, enhancing privacy and security.
- **Resource usage optimization**: Ensures optimal use of existing network resources without requesting new allocations, maintaining efficient operations.

Results from Simulation Studies (see Fig.1)

- **Latency:** The optimized placement strategy showed an average latency reduction of more than 10% compared to non-optimized placements, vital for maintaining QoS in latency-sensitive applications.
- **Energy consumption:** The approach resulted in decrease up to 20% in network-wide energy consumption, highlighting the benefits of resource-aware application placement in supporting green 6G initiatives.
- **Data exposure:** The model achieved a 5-10% reduction in data exposure, which is critical for privacy-sensitive applications and minimizes security risks.
- **Resource utilization:** The optimization model improved the utilization of computational and storage resources by 10-15%, ensuring that resources across the network are effectively leveraged, preventing bottlenecks and enhancing overall system performance.

Benefits of the New Protocol for Constrained Devices:

- **Signaling Reduction**: The protocol reduces signaling between constrained devices and the network, lowering energy consumption and needed resources.
- **Energy Efficiency**: The reduction in signaling extends both to device registration and ongoing communication, reducing energy and spectrum resource usage.
- **Impact on Active State:** This results in lower energy consumption and decreased time in an active state for constrained devices.

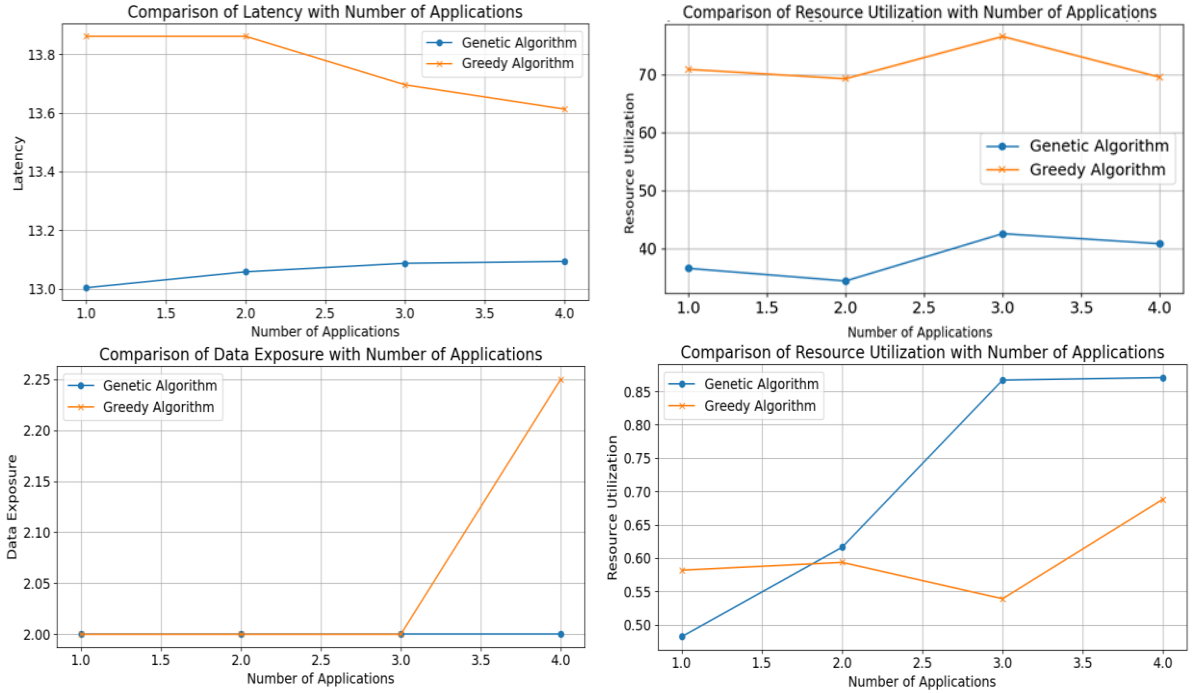


Fig. 1. Performance comparison of genetic and greedy algorithms across latency, resource utilization, and data exposure for different application loads various (see prev. slide)

Application-layer BCS optimisation
Key take-aways

- **Genetic Algorithm Performance**
  - The Genetic Algorithm (GA) significantly improves network efficiency by reducing latency and energy consumption compared to baseline approaches.
  - Through intelligent resource allocation, GA outperforms random placement strategies, optimizing resource utilization for more sustainable and cost-effective network operations.
- **INC Approach**
  - The Integrated Network and Compute (INC) approach optimizes both network and compute resources for BCS applications, considering multiple providers.
  - This joint optimization is essential for meeting the stringent requirements of BCS applications, and it relies on the collection and exposure of relevant metrics.
  - The strategic placement of User Plane (UP) functions across the cloud continuum (cc) further enhances the coordination of traffic and service deployments, ensuring optimal resource use and meeting target QoS.
- **Secure Communication for Constrained Devices**
  - Constrained devices, such as those used for sensing, require secure communication with the network. Data integrity and trust between the device and the network are critical.
  - The proposed protocol reduces energy consumption and signaling overhead by using a shared secret to create temporary IDs for secure device identification.
  - This reduces the need for spectrum resources, even with a large number of constrained devices communicating with the network.

Impact on sustainability

- **Energy Efficiency:** By minimizing unnecessary data transmission and optimizing node selection, the model reduces energy overhead, contributing to the overall sustainability of 6G networks.
- **Environmental Impact:** The reduction in energy consumption directly translates to a decrease in the carbon footprint of network operations, supporting green technology initiatives.
- The coordination of UP functions across the cloud continuum also reduces overall resource consumption by optimizing the placement and handling of user traffic at different cloud levels, contributing further to energy efficiency.

References

References

[GSH+22] E. Goshi, R. Stahl, H. Harkous, M. He, R. Pries and W. Kellerer, “PP5GS—An Efficient Procedure-Based and Stateless Architecture for Next-Generation Core Networks,” in IEEE Transactions on Network and Service Management, vol. 20, no. 3, pp. 3318-3333, Sept. 2023, doi: 10.1109/TNSM.2022.3230206.

[HEX223-D21] Hexa-X-I Deliverable D2.1 summary slides: Draft foundation for 6G system design, 2023-06-29

[HEX224-D23] Hexa-X-II Deliverable D2.3, “Interim overall 6G system design”, June 2024

[JOK+23] Jain A., Outtagarts A., Kerboeuf S. Bui T. D. “Harmonized metadata of Hexa-X-II enablers of Hexa-X-II D2.3”, public deliverable, https://zenodo.org/records/12570057

[HEX224-D33] Hexa-X-II Deliverable D3.3, “Initial analysis of architectural enablers and framework”, April, 2024.

[HEXA-I-D5.2] Hexa-X-I Deliverable D5.2, “Analysis of 6G architectural enablers applicability and initial technological solutions”, October, 2022.

[38.801] 3GPP TR 38.801, “Study on new radio access technology: Radio access architecture and interfaces (Release 14)”, March 2017.

[NGR+24] Time Synchronization in Communication Networks: A Comparative Study of Quantum Technologies," 2024 IEEE Wireless Communications and Networking Conference (WCNC), 2024, pp. 1-6, doi: 10.1109/WCNC57260.2024.10570688.

[GAT24] P. Garau Burguera, H. Al-Tous, and O. Tirkkonen, “Distributed User-Centric Cell-Free Massive MIMO with Architectural Constraints”, in Proc. EUCNC & 6G Summit, pp. 541-546 2024.

[GAT24b] P. Garau Burguera, H. Al-Tous, and O. Tirkkonen, “Remote Radio Head Multiclustering based Cell-Free Massive MIMO Systems”, in Proc. Vehicular Technology Conference (VTC2024-Fall), 2024, in press.

[VRV+22] J. S. Vardakas *et al.*, “Machine learning-based cell-free support in the O-RAN architecture: An innovative converged optical-wireless solution toward 6G networks,” IEEE Wireless Commun., vol. 29, no. 5, pp. 20–26, 2022.

[NGMN] NGMN Alliance, NGMN 5G P1 Requirements & Architecture Work Stream, End-to-End Architecture, Description of Network Slicing Concept, Online available at: https://ngmn.org/wp-content/uploads/160113_NGMN_Network_Slicing_v1_0.pdf. Accessed: 2024-10-24.

[CDP24] P. Charatsaris, M. Diamanti and S. Papavassiliou, "Joint User Association and Resource Allocation for Hierarchical Federated Learning Based on Games in Satisfaction Form," in IEEE Open Journal of the Communications Society, vol. 5, pp. 457-471, 2024, doi: 10.1109/OJCOMS.2023.3347354

[ETSI-MEC-040] ETSI GS MEC 040 V3.2.1, “Multi-access Edge Computing (MEC); Federation enablement APIs", March 2024

[SB+24] Sonai Biswas eta al., "Enhancing IoT Security with Quantum Key Distribution: A Novel Approach" IEEE Future Networks World Forum pp. 1-6, 2024

[NN+24] Nikhitha Nunavath et al, "Pragmatic Semantic Communication through Quantum Channel, ***TechRxiv August 24*** "DOI: 10.36227/techrxiv.172253927.71561050/v1

[MGB+23] Picazo Martinez, P., Groshev, M., Blanco, A., Fiandrino, C., de la Oliva, A., & Widmer, J. (2023, November 21). waveSLAM: Empowering Accurate Indoor Mapping Using Off-the-Shelf Millimeter-wave Self-sensing. IEEE Vehicular Technology Conference, Hong Kong, China. https://doi.org/10.5281/zenodo.10171206

[BQG+24] G. Baldoni, J. Quevedo, C. Guimarães, A. de la Oliva and A. Corsaro, "Data-Centric Service-Based Architecture for Edge-Native 6G Network," in IEEE Communications Magazine, vol. 62, no. 4, pp. 32-38, April 2024, doi: 10.1109/MCOM.001.2300178.

[23.502] 3GPP TS 23.502 “Procedures for the 5G System (5GS)”, V18.4.0, December 2023.

HEXA-X-II.EU //

Co-funded by the European Union

6GSNS

Hexa-X-II project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101095759.