Project no.:	**101095738**		
Project full title:	**6G Short range extreme communication IN Entities**		
Project Acronym:	**6G-SHINE**		
Project start date:	**01/03/2023**	**Duration**	**30 months**

D4.2 – PRELIMINARY RESULTS ON THE MANAGEMENT OF TRAFFIC, COMPUTATIONAL AND SPECTRUM RESOURCES AMONG SUBNETWORKS IN THE SAME ENTITY, AND BETWEEN SUBNETWORKS AND 6G NETWORK

Due date	30/06/2024	**Delivery date**	30/06/2024
Work package	WP4		
Responsible Author(s)	Dimitrios Alanis (APPLE)		
Contributor(s)	Keyvan Aghababaiyan (UMH), Dimitrios Alanis (APPLE), Anders Berggren (SONY), Baldomero Coll Perales (UMH), Javier Gozalvez (UMH), Christian Hofmann (APPLE), Pedro Maia de Sant Ana (BOSCH), Yasser Mestrah (IDE), Ognen Ognenoski (IDE),		
Version	V1.0		
Reviewer(s)	Fotis Foukalas (COGN), Thomas Jacobsen (NOKIA)		
Dissemination level	Public		

Co-funded by the European Union

Horizon Europe Grant Agreement No. 101095738. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or SNS JU. Neither the European Union nor the granting authority can be held responsible for them.

VERSION AND AMENDMENT HISTORY

Version	Date (MM/DD/YYYY)	Created/Amended by	Changes
0.1	03/19/2024	Dimitrios Alanis	Initial version
0.2	04/08/2024	Dimitrios Alanis	- Added Initial Introduction Skeleton with TC tables - Added consolidated reference section in the end (text to be updated)
0.3	23/04/2024	Dimitrios Alanis	- Added Introduction text and updated TC tables - Overall updates to prepare internal review
0.4	29/04/2024	Dimitrios Alanis	Added Conclusions text
0.5	03/05/2024	Dimitrios Alanis	- Added Executive Summary text and updated Introduction and Conclusions - Formatted document with English (UK)
0.6	17/05/2024	Dimitrios Alanis	- Added external reviewers' comments
0.7	10/06/2024	Dimitrios Alanis	- Added corrections external reviewers' comments
0.8	28/06/2024	Berit Hvidberg Christensen	Layout check and proofread
0.9	29/06/2024	Dimitrios Alanis	Last minor updates
1.0	30/06/2024	Dimitrios Alanis	Final version

TABLE OF CONTENTS

FIGURES

TABLES

ABBREVIATIONS

Abbreviation	Description
AS	Access Stratum
BS	Base Station
CCN	Compute Offload Controlling Node
CN	Core Network
CompN	Compute Node
CP	Control Plane
CSI	Channel State Information
DC	Dual Connectivity
DRB	Data Radio Bearer
GW	Gateway
HARQ	Hybrid Automatic Repeat Request
HC	High Capability Device
HO	Handover
IE	Information Element
LC	Low Capability Device
MAC	Medium Access Control

MgtN	Management Node
NR	New Radio
NSA	Non-Standalone
NW	Network
ON	Offloading Node
RA	Random Access
RAN	Radio Access Network
RAT	Radio Access Technology
RLC	Radio Link Control
RRC	Radio Resource Control
RRM	Radio Resource Management
PDCP	Packet Data Convergence Protocol
SDAP	Service Data Adaptation Protocol
SA	Standalone
SN	Subnetwork
SN-RP	Subnetwork Routing Protocol
SN-TP	Subnetwork Tunnelling Protocol
snCP	Subnetwork Control Plane
TMSI	Temporary Mobile Subscriber Identity
TP	Transport Block
UE	User Equipment
UP	User Plane
UWB	Ultra-Wideband
XR	Extended Reality

EXECUTIVE SUMMARY

This deliverable reports the activities carried out in Task 4.2 “Resource management within subnetwork entity and towards the 6G network” during the first 16 months of the project. Initially in Section 2, aspects of routing of data and control signaling within subnetworks in the same entity are presented corresponding to Task 4.2a. A distinction between dynamic and static subnetwork topologies is first made to tailor and fine-tune the proposed solutions accordingly. In terms of static topologies and while focusing on in-vehicle scenarios, a preliminary approach for the routing of data and control signaling is proposed in Section 2.1 leveraging traffic correlations to achieve a time-sensitive-capable integration between the hybrid subnetwork connections, consisting of both wireless and wired connections. Moving on to the dynamic topologies in Section 2.2, an enabler for seamless integration to the overlay 6G network is proposed in the form of the so-called virtual connections. Mobility procedures, i.e. when user equipment joins or exits the subnetwork, are defined so that the user equipment maintains connection with the 6G network, thus achieving network continuum. Furthermore, flexible topologies are enabled by distributing network and user equipment functionalities, achieved by the proposed novel distributed subnetwork control and user planes. Apart from these procedures in Section 2.3, the existing 3GPP framework for QoS flows is also surveyed, with respect to using relays which is relevant for subnetworks.

Moving on to Section 3, studies regarding dynamic computational resources offloading from subnetwork management nodes to 6G edge-cloud have been undertaken, corresponding to Task 4.2b. In this context, a framework-based approach is presented for traffic management mechanisms, which provides input information for offloading tasks to remote edge-nodes. Methods for coordinated planning for converging communication and computation are investigated in Section 3.1 to facilitate the distribution of computation among the network entities. Apart from computational offloading to remote edge-nodes, the prospect of local computational offloading within the subnetwork is also investigated in Section 3.2. For this reason, the roles of the subnetwork elements are defined to enable local computational offloading within the subnetwork entity as well as convergence of communication and computation. Additionally, a preliminary framework for service-aware joint planning and allocation of communication and computational resources is introduced in Section 3.3 with a special focus on in-vehicle subnetworks.

Finally, the dynamic spectrum sharing between 6G overlay network and in-X subnetwork mechanisms are studied in Section 4, corresponding to Task 4.2c. The challenges and opportunities of spectrum usage in both licensed and unlicensed bands for in-X subnetworks are presented. More specifically, an outline of the intricacies of spectrum sharing, the feasibility of various spectrum-based applications, and the potential for technological innovations is made. In this context, blockchain for decentralized spectrum management and big data processing for dynamic spectrum management are studied.

1 INTRODUCTION

This document presents preliminary studies on the management of resources in subnetworks located in close vicinity as well as of resources shared with the overlay 6G network. The term “resources” is generalized in the context of 6G networks and inherently subnetworks, as it is not limited to only spectral resources for transmission purposes, but also includes resources to enable functional and computational offloading. Naturally, coordination mechanisms to achieve the best performance possible in terms of the KPIs defined in [1], e.g., data rate and latency, are required. This deliverable provides an outline of how the methods described herein correspond to the project’s Technology Components (TCs) as set out in [13]. Additionally, we present an initial mapping of the proposed methods to the project's use cases. The definition of these use cases is described in deliverable D2.2 [1]. Furthermore, it is indicated how the proposed technologies can contribute to environmental, economic, and social sustainability, as the latter are rendered as key value indicators (KVIs) in the design of technologies in 6G-SHINE.

1.1 RESOURCE MANAGEMENT WITHIN A SUBNETWORK ENTITY AND TOWARDS THE 6G NETWORK

The densification of the nodes expected in 6G networks imposes a large overhead on both the core network (CN) as well the Radio Access Network (RAN) side. Naturally, under the current cellular paradigm, individual direct connections should be established for each UE, thus imposing significant control overheads at the network side. To accommodate the increasing traffic demand and the denser topologies, the concept of *networks of networks* [2] could be applied, which is deemed as one of the pillars of 6G. Under this principle, the nodes form smaller networks, referred to as *subnetworks*. The subnetwork nodes are served and controlled by a special UE node, referred to as *Management Node* (MgtN) [3], which provides connectivity to other subnetworks or to the 6G overlay network. Hence, it becomes evident that the nodes are organized within hierarchical networks. In this way, the subnetworks can offload control and data overheads from the overlay 6G network.

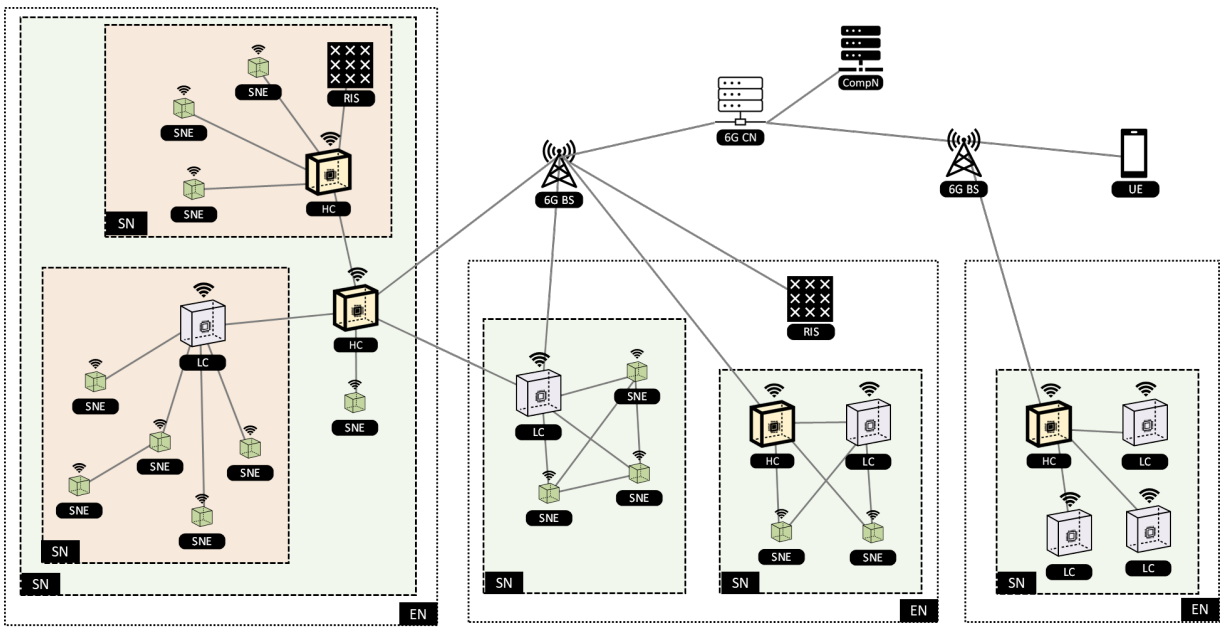

Figure 1 Reference architecture from [1].

However, a meticulous design of new protocols is required to enable the subnetworks offload control functionality and data routing from the overlay 6G network, otherwise there is an inherent risk of imposing further overheads due to the distribution of control and data within the subnetworks. In this deliverable, the control and data processes within the subnetwork, among subnetworks, as well as with the overlay 6G network will be investigated. The design target is to achieve a seamless integration of the subnetwork functions with the overlay 6G network as well as with neighbouring subnetworks, while, at the same time, to impose the minimum control and data routing overhead. To achieve this delicate balance, adaptive and efficient coordination mechanisms for distributing RAN-related *Base Station* (BS) and UE functionality among subnetwork nodes should be designed, while considering the different topologies. For instance, in-vehicle topologies can be viewed as rather static, while industrial and consumer topologies are perceived as more dynamic, hence requiring additional mobility process design. These coordination mechanisms not only distribute network functionality but also enable the distribution of UE functionality as well, thus providing functional offloading. More specifically, the latter means that some procedures, functions, or tasks can be distributed among nodes of the subnetwork, while traditionally all these functions are local and are executed by a single UE. In this case, a node can execute some tasks on behalf of another node. Naturally, this makes the deployment of UEs with lower capabilities possible since the latter can extend their capabilities by utilizing functional offloading.

The coordination of the subnetworks with their neighbouring subnetworks as well as with the overlay network raises an important issue, namely that of the subnetworks' and overlay 6G network coexistence in terms of spectral resources. Therefore, efficient dynamic spectrum sharing mechanisms should be designed. As different regulatory frameworks are adopted by each country, such constraints should be considered by the design. In addition to these constraints, the adoption of licensed and unlicensed spectrums come with their own set of challenges, which will also be investigated in this deliverable.

Apart from functional offloading, the subnetworks could enable general-purpose computational offloading, thus allowing the subnetworks to also harness another type of resources. Therefore, nodes with low computational capabilities could be assisted by the subnetwork by offloading computationally expensive applications to other nodes. These computation requests could be routed through the subnetwork and be undertaken by the neighbouring nodes within the subnetwork, or even by nodes in neighbouring subnetworks or by edge nodes in the cloud via the overlay 6G network or by the MgtN. Therefore, it becomes imperative that the subnetwork protocol design requires the tight integration of computational offloading features with communications.

1.2 Summary of the Proposed Solutions

In Chapter 2, this deliverable initially addresses aspects of routing data and control signalling within subnetworks of the same entity. These proposals are distinguished between scenarios characterized by static and dynamic subnetwork topologies, which may be present in various deployment scenarios such as in-vehicle and immersive classroom environments, respectively [1]. For static subnetwork topologies, where the time-varying element is the traffic, we extend the deterministic and predictive traffic scheduling framework introduced in [4]. This framework aims to derive scheduling decisions to support multiple traffic flows, including the simultaneous support of deterministic time-sensitive traffic and best-effort traffic. It leverages predictive modelling to anticipate the demand for communication resources from subnetwork elements. In this extension, we account for the coordinated scheduling decisions required between different subnetworks of the same entity that collaborate to support

applications involving communication between subnetwork elements (e.g., sensors and actuators) located across different subnetworks.

In the consumer category and especially in the immersive education scenario, the network topology may be deemed as inherently dynamic, as nodes tend to move and thus topologies change over time. For this specific reason the end-to-end connectivity of the nodes within subnetwork towards the overlay 6G network is investigated, which is referred to as virtual connections. Those virtual connections define how end-to-end Uplink (UL) and Downlink (DL) flows from the 6G overlay network to the UEs are managed within subnetworks as well as what UE context information needs to be stored at the 6G Radio Access Network Base Station (RAN-BS). Those components enable the UE to be reachable to and from the 6G RAN-BS indirectly via the subnetwork. Also the nodes' mobility in and out of the subnetwork is considered. The implementation of a distributed Control and User Planes is then proposed to guarantee seamless integration to the overlay 6G network. Finally, aspects of inter-subnetwork coordination are investigated.

With the introduction of Extended Reality (XR) services and Cloud Gaming applications, new demanding requirements are put on the systems that needs to deliver and facilitate such services. For many of such services, the Quality of Service (QoS) framework is being put at stretch, with large amount of data to transfer between multiple users and entities (where e.g., a user is using an entity being part of a subnetwork), as well as being time and delay sensitive, and with a need for synchronized delivery. Providing XR experiences that make the user feel immersed and present, several relevant Quality of Experience (QoE) considerations are needed, e.g., providing the feeling of being physically and spatially located in the virtual environment, when using a head mounted display (HMD).

In the context of in-X-subnetwork, use cases and requirements have been defined in the area of consumer, industrial and automotive categories, all considering applications requiring low latency, high data rate, and with high reliability, namely Ultra Reliable Low Latency Communication (URLLC) type of traffic. To ensure that relevant requirements can be fulfilled for developing In-X-subnetworks with defined use cases, there is a potential for further optimizing the QoS framework, e.g., by 3GPP, both looking into the Core Network-UE interfaces as well as investigate what QoS functionality is defined for other state-of-the-art 3GPP technologies such as V2X and Sidelink for within subnetwork communication.

Moving on to Chapter 3, several approaches for dynamic computational resource offloading from subnetworks to 6G edge-cloud providing coverage across all 6G-SHINE use cases are also discussed. These are mainly preliminary framework-based approaches and depend on aspects such as the traffic characterisation, the architecture definition, and service characteristics. Contrary to computational offloading to remote edge nodes in the cloud, the consumer category also provides the possibility to perform computational offloading to neighbouring nodes, either within the same subnetwork or in neighbouring subnetworks under the same entity. The roles of the nodes are thus defined, so that distributed and local computational offloading can be enabled.

Additionally, a novel framework, problem formulation and system characterization for the joint planning and allocation of communication and computing resources is introduced. This framework is designed to support end-to-end dependable service level provisioning across the continuum from the subnetwork (deep edge) to the 6G parent network (edge – cloud). The proposed framework leverages the trend towards software-defined platforms, such as software-defined vehicles (SDV), which offer higher levels

of flexibility for managing and allocating workloads to hardware resources. The proposed framework is aimed at seamlessly integrating the scheduling of tasks and communications within the local subnetwork continuum (using the multiple computing units available in the subnetwork entity) and the end-to-end continuum (i.e., from subnetwork to 6G parent network), ensuring timely, and even deterministic, task completion, even when tasks are offloaded across multiple compute nodes in the local/end-to-end continuum.

In Chapter 4, we delve into the dilemmas of using licensed versus unlicensed bands, emphasizing the trade-offs between coexistence issues and cost implications associated with spectrum licensing. We highlight potential optimization strategies for existing technologies, such as Wi-Fi and Bluetooth, and the development of new ones tailored for specific subnetwork applications. It also discusses the dynamic nature of in-X subnetworks in various scenarios, such as automotive environments, underscoring the need for flexible data transmission approaches to accommodate different data sizes and ensure reliable connectivity. Regulatory constraints, essential for guiding the development and deployment of these subnetworks, are reviewed with respect to their impact on the design and operation of in-X subnetworks in different regions.

Based on the above, the novel solutions presented in this deliverable may be summarized as follows:

- Task 4.2a, presented in Section 2:
    - *Adaptive and efficient coordination mechanisms within and across subnetwork for deterministic service level provisioning*: deriving a framework for scheduling decisions to support multiple traffic flows, including the simultaneous support of deterministic time-sensitive traffic and best-effort traffic. Predictive modelling is leveraged to infer the demand for communication resources from subnetwork elements.
    - Subnetwork management and flexible roles of devices in the SN: the role of the MgtN is defined along with that of the HC and LC devices. In addition, a special case of a LC device, namely the non-standalone UE is introduced.
    - *Subnetwork and overlay network interactions:* the end-to-end UL and DL flows are defined for UEs that are managed within subnetworks along with the UE context information required to be stored at RAN-BS.
- Task 4.2b, presented in Section 3:
    - *XR-traffic mechanism for (re)transmissions coordination with computational and task offloading*: a new framework for reducing delay due to retransmissions for URLLC.
    - *Joint task and communications scheduling for dependable service level provisioning*: joint planning and allocation of communication and computing resources is introduced in the form of a novel framework designed to support end-to-end dependable service level provisioning both within from the subnetwork to the 6G parent network

- *Local computation offloading*: new subnetwork elements roles introduced to enable local and distributed computational offloading.

Note that in for Task 4.2c on dynamic spectrum sharing on Section 4 there are no novel solutions presented; however, a state-of-the-art survey along with the initial problem formulation are provided.

1.3 Positioning of the Designed Solutions

The research conducted in this deliverable directly contributes to achieving objective 6 of the project, which is as follows:

- **Objective 6.** Develop new methods for integration of subnetworks in the 6G architecture and efficient orchestration of radio and computational resources among subnetworks and wider network.

6G-SHINE have identified and are working on 16 technology components (TCs) relevant for subnetworks, which are listed as follows:

TC1. In-X data traffic models
TC2. Channel models for in-X scenarios
TC3. Sub-THz system model
TC4. Ultra-short transmissions with extreme reliability
TC5. Analog/hybrid beamforming/beamfocusing
TC6. Jamming-aware native PHY design
TC7. RIS enhancements
TC8. Intra-subnetwork macro-diversity
TC9. Flexible/full duplex scheduler
TC10. Predictive scheduler
TC11. Latency-aware access in the unlicensed spectrum
TC12. Centralized radio resource management
TC13. Distributed/hybrid radio resource management
TC15. Hybrid management of traffic, spectrum and computational resources
TC16. Coordination of operations among subnetworks in the same entity

Research in this deliverable covers TC15 and TC16, and the preliminary work presented in D4.2 covers both TCs. In Table 1, a list of the technology/methods studied in this deliverable is presented, and the connection with the original TCs.

Table 1: Connection of the studied methods with the 6G-SHINE TCs.

Technology/method	6G-SHINE TCs
Adaptive and efficient coordination mechanisms within and across subnetwork for deterministic service level provisioning	TC16, TC10
Subnetwork management and flexible roles of devices in the SN	TC16
Subnetwork and overlay network interactions	TC16
Coordination between different subnetworks	TC16
XR-traffic mechanism for (re)transmissions coordination with computational and task offloading	TC15, TC16
Joint task and communications scheduling for dependable service level provisioning	TC15
Local computation offloading	TC16

Subnetwork resource management and coordination methods both within the subnetwork as well as with the overlay 6G network are the cornerstones for achieving the extreme performance requirements of XR applications as well as extremely time-sensitive in-vehicle and in-factory applications. Table 2 describes the main KPIs and KVIs targeted by the presented methods. As the 6G-SHINE project is a low technology readiness level (TRL) project, it should be noted that it is not the aim to directly measure the impact of the proposed solutions in terms of KVIs, since such an assessment can only be possible, when the proposed solutions are implemented and integrated in a complete system design, with the latter lying outside of the scope of the project. However, KVIs are at the centre of our technology design, and it can be speculated how solutions can be the main enablers for addressing relevant KVIs. A thorough description of how 6G-SHINE research addresses environmental, economic and social sustainability is included in deliverable D2.2. Given the nature of the research in novel resource management for subnetworks and the new types of applications these methods accommodate, the solutions presented in D4.2 are addressing environmental, economic as well as social sustainability for future in-X subnetwork products.

Table 2: KPIs and KVIs targeted by the presented methods.

Technology/method	**Main target KPIs**	**Targeted KVIs**
Adaptive and efficient coordination mechanisms within and across subnetwork for deterministic service level provisioning	Deterministic low latency and high reliability. Dependable service level provisioning on end-to-end subnetwork connectivity	Improved economic and environmental sustainability thanks to the gained flexibility in service provisioning and configuration by replacing rigid cables with configurable wireless subnetworks.
Subnetwork management and flexible roles of devices in the SN	Data Rate and Improved Latency to achieve sufficient Service Quality and Continuity	Improve economic, environmental, and societal sustainability by enabling more affordable, lower complexity devices and more energy efficient devices to participate in wireless communication.
Subnetwork and overlay network interactions	Data Rate and Improved Latency to achieve sufficient Service Quality and Continuity	Improve economic and environmental sustainability by improving the architecture towards more scalability and reducing the complexity in the NW cause by the densification.
Coordination between different subnetworks	Data Rate and Improved Latency to achieve sufficient Service Quality and Continuity	Improve economic and environmental sustainability by improving the architecture towards more scalability and reducing the complexity in the NW cause by the densification.
XR-traffic mechanism for (re)transmissions coordination with computational and task offloading	Overhead, number of retransmissions, access resource utilization, latency.	Improved economic and environmental sustainability thanks to the gained flexibility in service provisioning exploiting resource availability in the local/end-to-end continuum from subnetwork to 6G parent network
Joint task and communications scheduling for dependable service level provisioning	Optimize CPU and communication resource allocation, guarantee timely task completion in the local/end-to-end continuum under diverse load scenarios	
Local computation offloading	Data Rate and Improved Latency to achieve seamless compute offloading	Improve economic, environmental, and societal sustainability by enabling more affordable, lower complexity devices and more energy efficient devices to

		participate in wireless communication.

Table 3 presents the mapping of the presented methods to the use case categories (and specific use cases) as defined in deliverable D2.2. It should be noted that the authors' intention is not to evaluate each presented method for all the mapped use cases. Instead, performance evaluations are evaluated based on the respective main use case or use case category of interest, since the extension to different use cases with similar KPIs could be viewed as straightforward.

Table 3: Mapping between presented methods and use cases as defined in D2.1[13]

Technology/method	Main use case category(ies)	Relevant use cases
Adaptive and efficient coordination mechanisms within and across subnetwork for deterministic service level provisioning	In-vehicle	Collaborative Wireless Zone ECU
Subnetwork management and flexible roles of devices in the SN	Mainly Consumer, but also Industrial could benefit	Immersive Education, Indoor Interactive Gaming
Subnetwork and overlay network interactions	Mainly Consumer, but also Industrial could benefit	Immersive Education, Indoor Interactive Gaming
Coordination between different subnetworks	Mainly Consumer, but also Industrial could benefit	Immersive Education, Indoor Interactive Gaming
XR-traffic mechanism for (re)transmissions coordination with computational and task offloading	Consumer subnetworks, but can be extrapolated for others	All categories
Joint task and communications scheduling for dependable service level provisioning	In-vehicle	Collaborative Wireless Zone ECU, Virtual ECU
Local computation offloading	Mainly Consumer, but also In-vehicle and Industrial could benefit	Immersive Education, Indoor Interactive Gaming

An outlook of the standardization potential of the proposed methods is presented in Table 4.

Table 4: Standardization potential of the presented methods

Technology/method	Standardization potential
Adaptive and efficient coordination mechanisms within and across subnetwork for deterministic service level provisioning	Potential 3GPP RAN1 and RAN2 impact for Release 20 and beyond and IEEE 802.1DG (TSN Profile for Automotive In-Vehicle Ethernet Communications), including integration of TSN and wireless for the end-to-end support of time sensitive traffic.
Subnetwork management and flexible roles of devices in the SN	Potential 3GPP SA1, SA2 and RAN2 impact.
Subnetwork and overlay network interactions	Potential 3GPP SA1, SA2 and RAN2 impact.
Coordination between different subnetworks	Potential 3GPP SA2 and RAN2 impact.
XR-traffic mechanism for (re)transmissions coordination with computational and task offloading	Targeting 3GPP RAN2 for Release 20+, in particular WI/SI on XR enhancements for NR
Joint task and communications scheduling for dependable service level provisioning	Potential 3GPP RAN2/3 and SA2 impact for Release 20 and beyond with architectural functionalities for enabling the joint scheduling of computing and communication resources in the local/end-to-end continuum from subnetworks to 6G parent network.
Local computation offloading	Potential 3GPP SA1 and RAN2 impact.

In the description of the methods, the nomenclature currently being defined in WP2 [1] is adopted, when referring to the relevant subnetwork components.
For the work carried out in D4.2, the following elements are relevant:

- Element with High Capabilities (HC). An element with high capabilities is a device/node with increased capabilities in terms of networking and computation. Such a node might act as the

central communication node in a subnetwork and also might offer compute resources to other devices in the subnetwork. Multiple such HCs can be installed in a single subnetwork. An HC device can be a user equipment as defined by 3GPP or a non-3GPP device.

- Element with Low Capabilities (LC). An element with low capabilities is similar to an HC but has limited capabilities in terms of networking and computation. This can reduce the functionalities this device provides to the subnetwork and even there might be no connection to the 6G base station. In a hierarchical or nested subnetwork, the LC might act as an aggregator. An LC device can be a user equipment as defined by 3GPP or a non-3GPP device.
- Subnetwork Element (SNE). Subnetwork elements are computationally constrained devices that have limited form factor, cost footprint, and include devices such as sensors/actuators. A SNE device can be a user equipment as defined by 3GPP or a non-3GPP device.

For a thorough description of all subnetwork elements, refer to deliverable D2.2 [1].

2 ROUTING OF DATA AND CONTROL SIGNALLING WITHIN SUBNETWORKS IN THE SAME ENTITY

This chapter presents preliminary 6G-SHINE solutions for the routing of data and control signalling within subnetworks of the same entity, which is mainly related to 6G-SHINE TC 16 (see Section 1.3). The proposed solutions address different subnetwork topologies of the 6G-SHINE's use cases [1], which can be divided into statics (e.g. in-vehicle networks) and dynamics (e.g. consumer). For static topologies, Section 2.1 presents and describes the architecture of a collaborative in-vehicle subnetwork and identifies architectural elements and roles for supporting the routing necessary to interconnect subnetworks located in different zones within the vehicle. Section 2.1 also describes a framework aimed at providing TSN-capable performance in hybrid wireless and wired in-vehicle networks, and discusses its alternative deployment options (centralized, hybrid and distributed) for providing dependable services levels provisioning in the end-to-end communication between elements (sensors, actuators) located in different subnetworks. Section 2.2 focuses on routing solutions tailored to dynamic subnetwork topologies, such as the ones faced in the deployment of the immersive education use case. Section 2.2 considers very diverse and dynamic environments where subnetwork elements can flexibly change their roles within the subnetwork (Section 2.2.1). Alternative solutions to manage the potentially changing subnetwork topology are proposed to efficiently manage the interactions within the subnetwork and between the subnetwork and the 6G parent network (Section 2.2.1.1). This includes solutions to enable devices with lower capabilities and limited (cellular) functionalities to be part of the subnetwork by exploiting functionalities offered by higher capabilities subnetwork devices (Section 2.2.1.2). Furthermore, the establishment of virtual connections between the subnetwork and the 6G parent network is defined (Section 2.2.1.3) and the mobility in and out of the subnetworks (Section 2.2.1.4) is described. In Section 2.2.2 novel distributed user plane (UP) and control plane (CP) functionalities among subnetworks are defined, implying adaptations in the interaction between 6G Network (NW) and the subnetwork (Section 2.2.2.1) and within the subnetwork (Section 2.2.2.2). Finally, Section 2.2.3 addresses scenarios requiring the coordination between different subnetworks. This chapter concludes with an analysis of required extensions for QoS support in subnetworks (Section 2.3). The investigations initially consider the interactive gaming use case, whose traffic flows within the subnetwork are analysed in Section 2.3.1. Then, in Section 2.3.2, a review of the state of the art on QoS support is presented. Finally, the challenges for supporting QoS in subnetworks and future research directions are presented in sections 2.3.3 and 2.3.4, respectively.

2.1 STATIC TOPOLOGIES

2.1.1 Adaptive and efficient coordination mechanisms within and across subnetwork for deterministic service level provisioning

6G-SHINE's in-vehicle subnetwork category [1][13] considers the collaborative wireless zone ECUs use case to support automotive systems and applications that require for their execution the collaboration between functions or offloading between sensors and actuators located at different zones of the in-vehicle network (IVN). This use case is represented by the subnetwork architecture shown in Figure 2 which exemplifies the scenario where different zones of the vehicle form individual subnetworks (SN); depicted in light brown. Each of these SNs integrates different (static) sensors and actuators represented as subnetwork elements (SNE) and a zone Electronic Control Unit (ECU) represented by the low capability element (LC). The LC in each zone aggregates the input and output interactions to and from the subnetwork and the central computing unit of the vehicle or High-Performance Computing Unit

(HPCU), represented by the high capability element (HC) in Figure 2. The HC is also responsible for coordinating and managing the interactions between SNEs of different subnetworks, ensuring end-to-end service level provisioning for automotive systems and applications executing in different subnetworks.

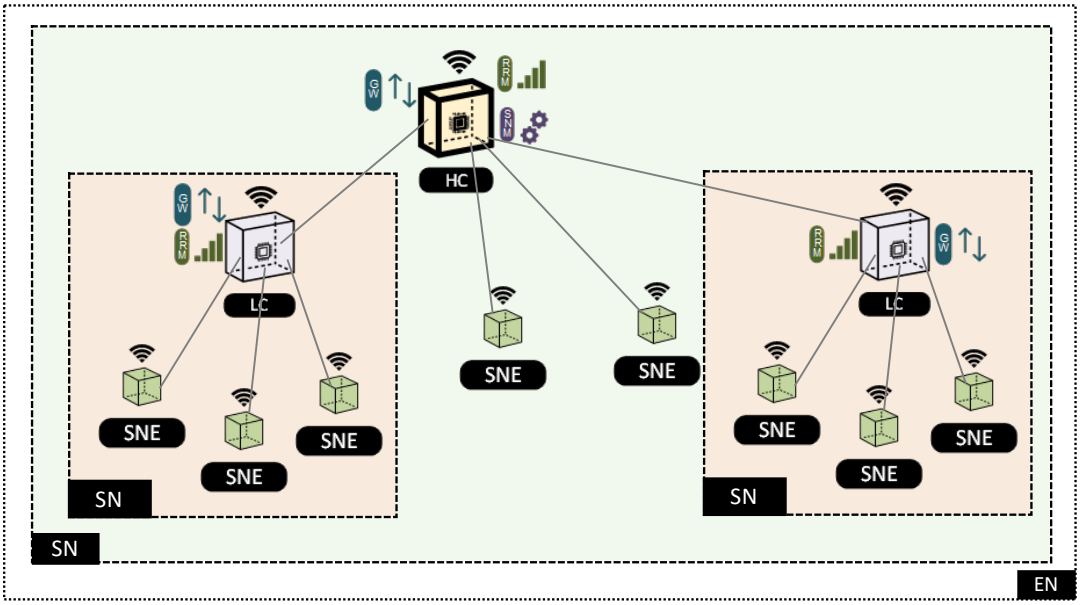

Figure 2 Exemplified subnetwork architecture of the collaborative zone ECU use case. In the figure, the SNs shadowed in brown represent IVN zones. IVN allow SNE to reach the HPCU directly without any support from the zone ECU (LC in the figure).

A framework for predictive-based and TSN-capable hybrid wireless and wired in-vehicle networks has been presented in [4]. This framework is designed to sustaining determinism over wireless links in individual subnetworks. That same framework is depicted in Figure 3. Figure 3 highlights in red the three main newly added blocks in 6G-SHINE aimed at achieving predictive-based and TSN-capable hybrid wireless and wired in-vehicle networks, compared to traditional TSN. First, a wireless scheduler in the data plane co-located with the multiplexer (MUX) and represented with a red trapezium. This newly added module is aimed at supporting the traffic generated/addressed to/from sensors/actuators with stringent and deterministic requirements. Second, the hybrid wireless and wired scenario involves replacing a subset (e.g., *X* number of links) out of the total *N* wired links with wireless connections, seamlessly integrating them into the data and control planes for effective flow management. Finally, Figure 3 shows a new block labelled 'Predictor,' which plays a crucial role in managing and scheduling incoming (and possibly outgoing) wireless transmissions. The 'Predictor' module is closely integrated with the 'Traffic Shaping & Scheduling' module available in traditional TSN networks to ensure that the integration of wireless links does not interfere with operations supporting data flows, including those for deterministic and time-sensitive applications. This implies, for example, that the scheduled wireless transmissions must arrive at the TSN's queues before the queues' gates open. 6G-SHINE envisions feeding the 'Predictor' module with context control information derived from processed traffic (e.g., closed feedback on channel state, BLER statistics, traffic-flow characteristics, and their correlation) and environmental factors (e.g., vehicle speed, road type). Additionally, AI/ML models could contribute to the predictive capabilities. In the context of 6G-SHINE, the "Predictor" module is envisioned to play a crucial role in the joint management and coordination of both wired traffic shaping and wireless link scheduling mechanisms. For instance, the "Predictor" module could coordinate the opening timing of

queues' gates for wired traffic (e.g., advancing when the queue opens if the traffic is predicted to be waiting on the queue), while simultaneously relaxing the time requirements for the wireless link.

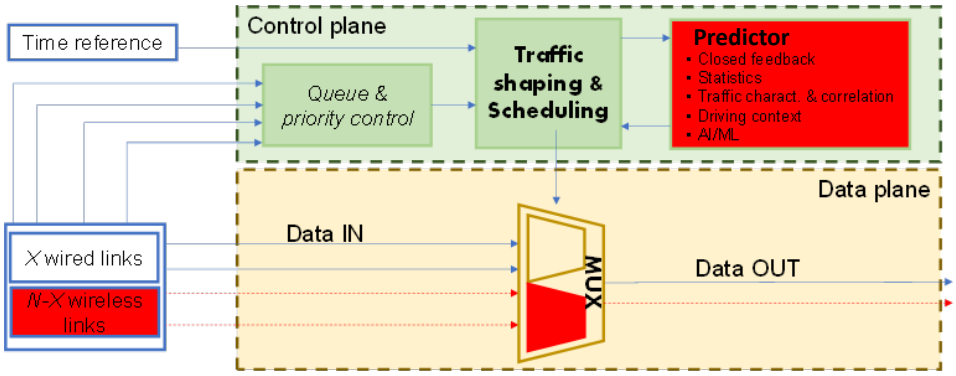

Figure 3 Framework for predictive-based and TSN-capable hybrid wireless and wired in-vehicle networks [4].

Figure 4 shows how the framework presented in Figure 3 is deployed to support applications requiring collaboration or offloading between functions, sensors and actuators located at different zones of the in-vehicle network. Coordination between Zone ECU (LC) in different collaborative subnetworks and the HPCU (HC) can result in distributed (i.e., only relying on zone ECU), centralized (i.e., only relying on HPCU) or hybrid (i.e., both) approaches. These approaches manage and schedule the transfer of data through the in-vehicle network from the zone of origin to the zone of consumption. Centralized and hybrid approaches implemented at the HPCU benefit from higher communication, management and computing capabilities - see Subnetwork Management (SNM), Radio Resource Management (RRM), and Gateway (GW) capabilities of the HC in Figure 2. However, they require for their implementation control information to be available at the HPCU, potentially increasing the load burden on the in-vehicle network. Distributed approaches implemented at the zone ECUs (LC) for interconnecting different subnetworks reduce the control load burden but may face challenges in meeting stringent Quality of Service (QoS) requirements, such as ensuring bounded end-to-end latencies [5].

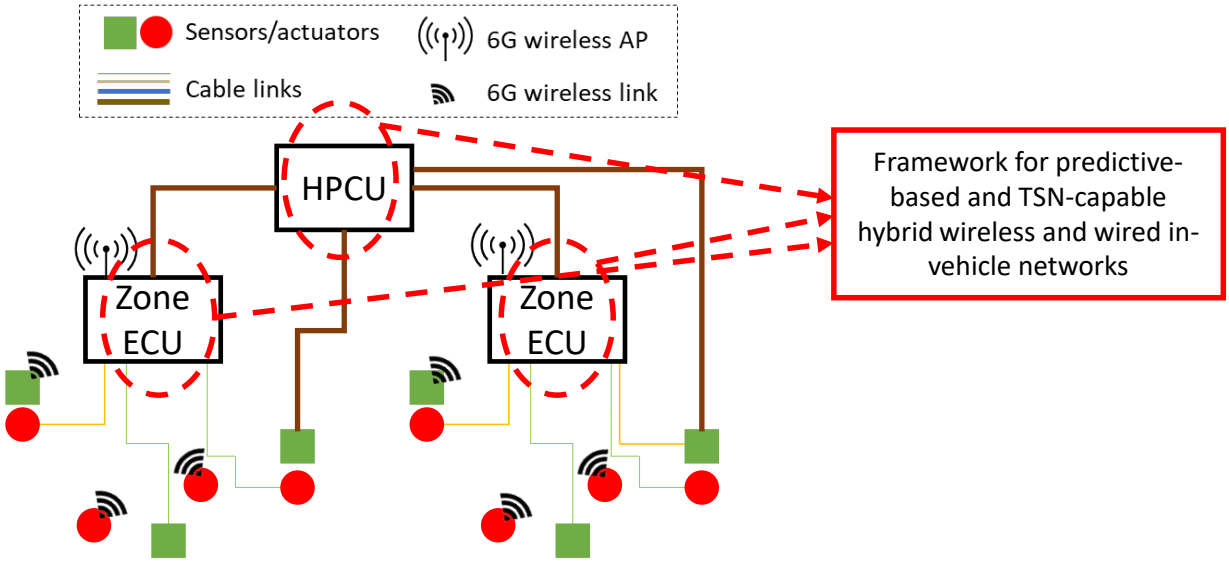

Figure 4 Deployment of the framework represented in Figure 3 for scenario requiring the interconnection of different subnetworks.

Future work will focus on assessing alternative deployment options to support E2E functions or services requiring the collaboration between SNEs located in different subnetworks. This future work will rely on the distributed, centralized or hybrid approaches between the zone ECU and HPCU indicated above. To achieve this goal, the framework and problem formulation presented in [4] are planned to be extended, taking into account solutions such as programmable and adaptive redundancy and multi-path links. These solutions aim to meet end-to-end dependable service level provisioning, with a specific focus on deterministic service level provisioning.

2.2 Dynamic Topologies

The immersive education use case involves a certain degree of mobility. Explicitly, the devices have access to limited resources of power, students may physically move during their *extended reality* (XR) classroom sessions and the *Access Points* (APs), which act as GWs to their connected devices and are considered regular *User Equipment* (UE) devices, may opt for acting as APs or stop to do so.

From the current 3GPP framework, Integrated Backhaul Access (IAB)[9] and Sidelink (SL) with SL relay [6] could potentially be used for multi-hop communications in subnetworks. IAB allows BSs to connect in a tree formation with the intermediate BS configurations being controlled by the root node, referred to as IAB donor. Despite the deployment mobile IAB leaf nodes, which are fully capable BSs, this framework is not flexible enough for the collaborate use case of immersive education, where latency and power consumption requirements are dynamic. As for SL Relay, it has primarily been deployed for coverage extension use cases. However, even in this case the SL link configuration is controlled by the BS the SL Relay is attached to, thus failing to tackle the dynamic latency requirements. A degree of flexibility in terms of RAT technologies is given by the Personal IoT framework [9]. Nevertheless, all the aforementioned technologies focus solely on the user plane functionality and routing of data, in the form of Backhaul Adaptation Protocol (BAP) [11] and SL Relay Adaptation Protocol (SRAP) [12], without allowing for control plane functionality offloading.

Based on the above, a more versatile framework with flexible node roles shall be defined.

2.2.1 Flexible Roles for the Subnetwork Nodes

However, before delving into defining the specifics of those flexible roles, let us set out the design assumptions to consider. To begin with, while still in coordination with an overlay network, the subnetwork should be an distinct entity formed by devices. Locally within the subnetwork, the devices may use the same or a different technology or spectrum compared to the overlay 6G NW, e.g. they could use licensed spectrum granted by the NW, unlicensed spectrum or license-exempt resources. Additionally, subnetworks shall be formed among UEs without or with limited NW configuration or awareness. In this way, not only is the complexity in the overlay NW constrained, especially in dense deployments, but also the dependency on the overlay NW is reduced enabling the subnetwork to continue functioning, even when going out of coverage of the overlay NW.

As far as the subnetwork nodes are concerned, two classes of devices are considered, namely the *High Capability* (HC) and the *Low Capability* (LC) devices as depicted in Figure 5. Note that SNEs as defined in [1] are considered even lower capability devices and are part of the LC device class further on. The LC devices are considered to be constrained in their communication, computation or power capabilities or are even, in some cases, incapable to establish on their own a direct connection to the overlay NW. The

HC devices have higher computational and communication capabilities, while they are capable of connecting, by themselves, to both the subnetwork and the overlay NW. Note that this distinction can be rather dynamic for a specific device. Explicitly, a specific device could potentially switch from LC to HC mode or vice versa depending on e.g. access to adequate power resources.

2.2.1.1 Subnetwork Management

Based on the aforementioned design considerations, in juxtaposition to the state of the art [6][9], the intention is to move the SNM functionality towards the subnetwork side. For this reason, a new role is defined, namely that of the *Management Node* (MgtN) [1][3]. This node is responsible for the local subnetwork control and routing, thus guaranteeing subnetwork operation in the absence of an overlay NW (e.g. a 6G BS). The MgtN acts as a gateway connecting subnetwork nodes and the overlay 6G NW, with the latter offloading some of its routing and control functionalities to the MgtN. Naturally, this role can be exclusively taken by HC devices due to the additional communication and computational resources required for managing the subnetwork. MgtN devices must also have direct access to the 6G NW in order to act as a gateway for the devices within the subnetwork. It should also be pointed out that this role is a rather dynamic one, since a device might opt for switching from/to a MgtN role or to/from an LC or HC device, depending on its internal and mobility states.

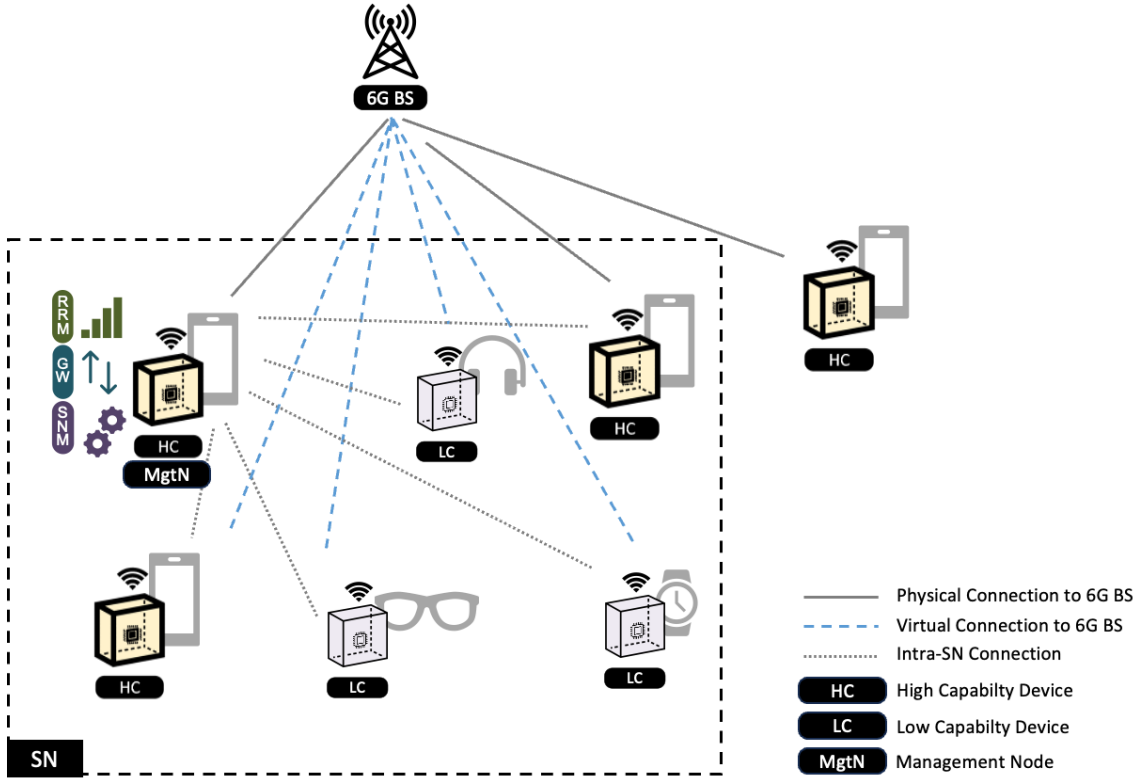

Figure 5 Exemplified single subnetwork architecture for the case of immersive education.

Figure 5 portrays an exemplified subnetwork for the case of the immersive education. In this specific example, an HC UE is acting as a MgtN, managing the SN entity illustrated by the in the dashed square. Connected to this MgtN are three LC devices and another HC device. In this specific scenario, local subnetwork control for local intra-SN links is undertaken by the MgtN. Furthermore, the devices have been registered and managed by the overlay NW and they maintain a logical connection to the 6G BS, while in the subnetwork. This is crucial for maintaining with the 6G overlay NW, when traffic exchange happens outside the SN.

Explicitly, physical connection from the subnetwork devices to the overlay 6G NW is undertaken indirectly though the MgtN, who is acting as a GW, while being responsible for maintaining these virtual connections towards the subnetwork devices. Devices within a subnetwork shall remain associated with the 6G NW, ensuring seamless mobility in and out subnetworks (see subsection 2.2.1.4). This is essential for the immersive education use case [1], where network continuum is required as students may move in the classroom. Ideally, the design consideration would involve complete transparency of the subnetwork nodes making the NW completely unaware of the subnetwork. In this case, the MgtN would impersonate the UEs on their behalf. However, this would involve additional overheads, since the MgtN would have to instantiate multiple User Plane (UP)/ Control Plane (CP) entities for all impersonated UEs. This makes the MgtN capabilities rather demanding, rendering this deployment with complete subnetwork transparency as impractical. Consequently, a delicate balance needs to be maintained with the NW involvement in order to guarantee the subnetwork's decoupling from the 6G NW, while at the same time imposing the least amount of architectural overhead possible.

This trade-off can be achieved by making the 6G BS that is attached to the overlay 6G NW aware of the presence of the MgtN and its role of managing its local devices but not internal state, i.e. the interconnections of the devices associated with the MgtN and its subnetwork. The 6G BS maintains UE context(s) while the UEs are within or outside the subnetwork. The overlay 6G NW remains in control of the resources and mobility at the global scale, e.g. for Handover (HO) between the 6G BSs, while the local mobility remains within the subnetwork. In fact, no RAN interaction is needed, when a UE of the subnetwork is in idle mode, since it can be exclusively handled between MgtN and UE, e.g. MgtN listens to paging for a UE or reads System Information (SI) on UEs behalf. In contrast to that, when entering connected mode, RAN gets involved. Note that in case of a node joining a neighbouring subnetwork the RAN will need to be involved at the last stage by being notified of the change; however, the mobility process will still be handled among subnetworks with minimal NW involvement.

Moreover, the architectural option of having the subnetwork not fully transparent towards the 6G overlay NW, meaning that the 6G BS is aware of the specific devices associated with a specific MgtN, comes with the following implications on the NW side:

- The 6G BS has to serve all UEs within the subnetwork via the MgtNs physical connection, while User Data, Access Stratum (AS) and Non-Access Stratum (NAS) control information etc. needs to be sent via the MgtN towards the UEs, depicted as *Virtual Connection* in Figure 5, above, which are rather logical links between UEs and 6G BS via the MgtN.
- 6G BS requires only a single physical connection for all UEs within the subnetwork, which is the link towards MgtN, which enables the 6G BS to reduce the per-UE maintenance effort, e.g. on channel estimation, CSI measurements, link adaptation, timing advance or mobility measurements [6][7].

2.2.1.2 Non-Standalone LC devices within the subnetwork

The described flexible subnetwork architecture consists of HCs and LCs devices being UEs that may connect directly towards the 6G NW in a standalone fashion or may join a subnetwork and connect indirectly via a MgtN. While being part of the subnetwork, LC devices can offload certain CP/UP functionality to HC devices and therefore e.g. save power or utilize HC device capabilities to improve performance KPIs, e.g. benefit from better RF capabilities of an HC device. This requires new Radio Resource Management (RRM) procedures for subnetwork management.

This standalone capability is a necessity for devices to maintain global mobility within the 6G NW, but also mandates that even the LC devices need to be UEs fully capable of establishing direct connection with the overlay 6G network. Despite the offloading opportunities within the subnetwork e.g. of CP functionality to the MgtN to optimize the power consumption, immersive education is very latency sensitive and therefore may still require direct data access of devices to the 6G BS in order to remove unnecessary hops and therefore delays in the communication.

To solve this, a new category of LC devices is proposed that have the ability to access the 6G NW for data communication while not being fully capable standalone cellular devices, but non-standalone (NSA) device. Such NSA devices can be categorized as LC devices that require only a very limited set of cellular capabilities, e.g., for transferring and receiving data from a 6G BS in order to support latency critical UCs. The remaining functionality could be provided by a supporting fully capable cellular HC/LC device in vicinity. For example, within the subnetwork, the MgtN as a HC device with standalone capabilities can complement the NSA devices by providing missing functionality. Naturally, the HC device is referred to as serving SA-UE.
This requires a new distribution of UP/CP functionality resulting in the following new definitions:

- NSA User Plane (on the NSA LC device): A reduced set of cellular UP functions only for transferring and/or receiving data from a 6G BS (e.g. only PHY and reduced set of Layer 2 functions)
- NSA Control Plane (on the NSA LC device): A reduced set of CP functionality towards the RAN only covering to RAN-related Uu configuration handling. All other RAN-related functionality could be removed e.g.:
    - SI Handling
    - RRM procedures for measurements and mobility
    - Access Stratum (AS) key management

Non-Access Stratum (NAS) services like mobility management, session management or identity management are also not necessary if it is associated with a MgtNs cellular subscription (serving SA-UE). In addition, a new procedure to request direct 6G connection via the MgtN is required.

- SA Control Plane (on HC device acting as MgtN): A new procedure to enable NSA LC devices to establish a direct 6G connection (see subsection 2.2.2)

This concept of NSA devices enables power efficient small form-factor devices that are capable of low latency communication. In Figure 6 an example scenario is shown, where UE2 and UE3 are LC devices within the subnetwork that have NSA capabilities.

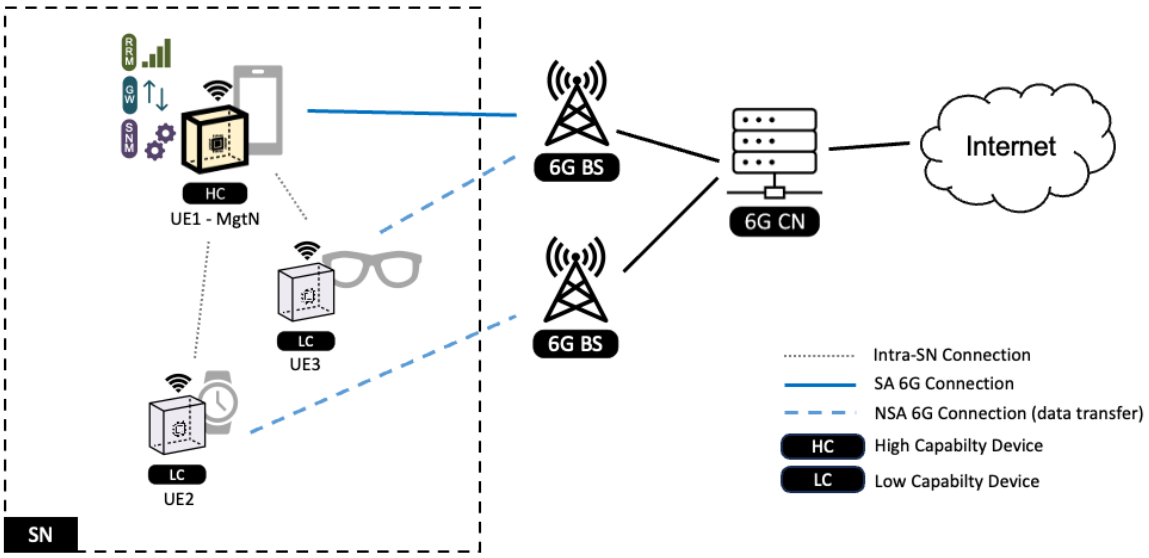

Figure 6 Subnetwork enabling Non-Standalone LC devices for direct 6G connections.

The UE1 is an HC device acting as MgtN managing the subnetwork and supports all mandatory cellular functionalities. It has a cellular subscription and provides cellular connectivity to all devices in the subnetwork acting as gateway (GW). For latency critical UCs the LC devices UE2 and UE3 are still able to perform direct data transfer with external network (e.g., Internet) via the 6G BS. This is achieved by utilizing the full capabilities of UE1 to e.g. establish a connection for a certain data bearer between the certain LC devices and the NW.

In summary, the cellular NW may consider the whole subnetwork as a single logical UE. During the lifetime of such a subnetwork (logical UE) the 6G connection situation may change from single physical 6G connection where the MgtN is the GW to having multiple physical 6G connections to/from multiple physical devices within the subnetwork.

2.2.1.3 Virtual Connections and UE Contexts

On the NW side, while a UE within the subnetwork is in connected mode, the UE context and the related procedural efforts for devices within the subnetwork can be reduced, since all link related information can now be derived from the MgtN context, as portrayed in Figure 7. Still referring to the same figure, the 6G BS maintains a new UE ID, namely the SN UE ID, valid in RAN and assigned with the aid of the subnetwork, while the UE is connected to the NW via the subnetwork. Contexts of the specific UEs within the subnetwork are linked to MgtN context via the SN UE ID at the 6G BS side.

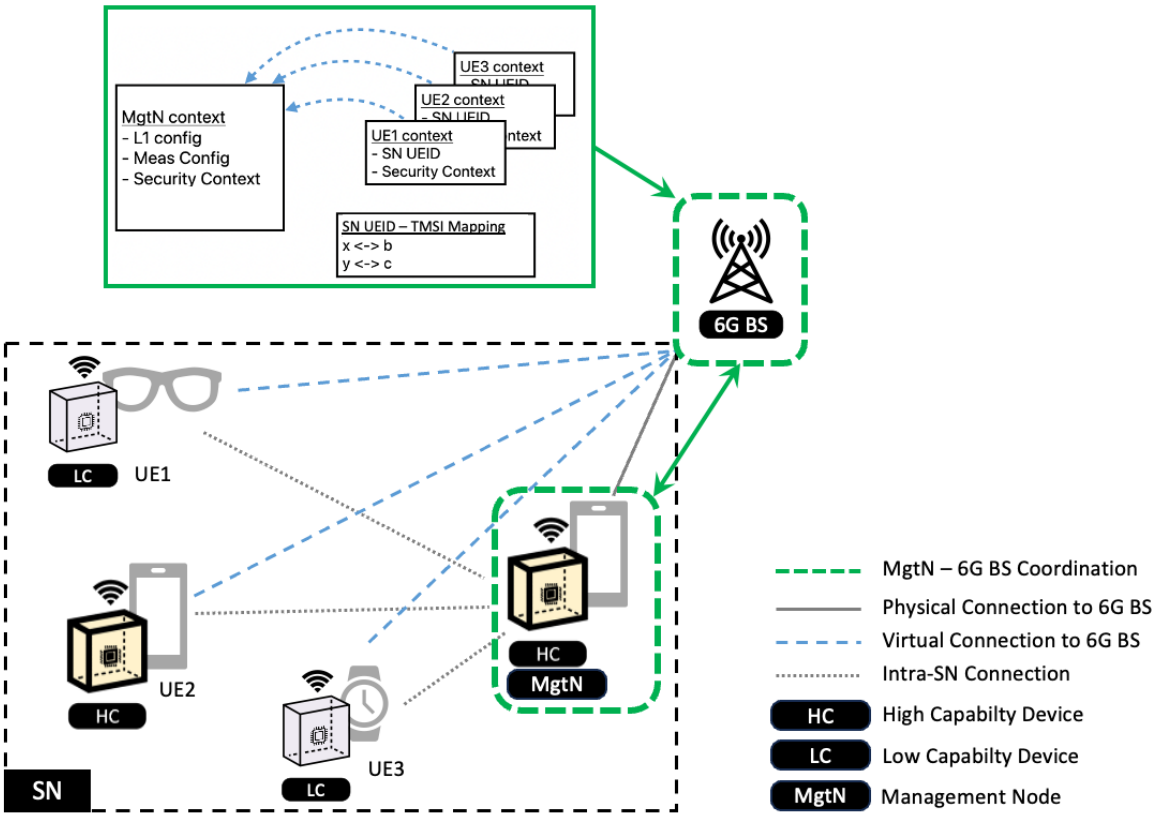

Figure 7 MgtN context stored at the 6G BS side. The green annotations refer to the entities involved in the subnetwork setup

More specifically, this identifier shall be mapped to the global NW UE ID of the particular UE in a 1-to-1 fashion, similar to the Temporary Mobile Subscriber Identity (TMSI) in 5G NR (see Figure 7) in order to allow seamless mobility in and out of the subnetwork as depicted in subsection 2.2.1.4. The SN UE ID is also used by the 6G BS to identify and address the UEs within the subnetwork as well as for routing data within the subnetwork. A *Message Sequence Chart* (MSC) on the flow of the both Uplink (UL), i.e. the direction from UE1 to the BS, and Downlink (DL), i.e. the direction from the 6G BS to UE1, is portrayed in Figure 8, where an example to RRC message transmission initiated by UE1 is shown.

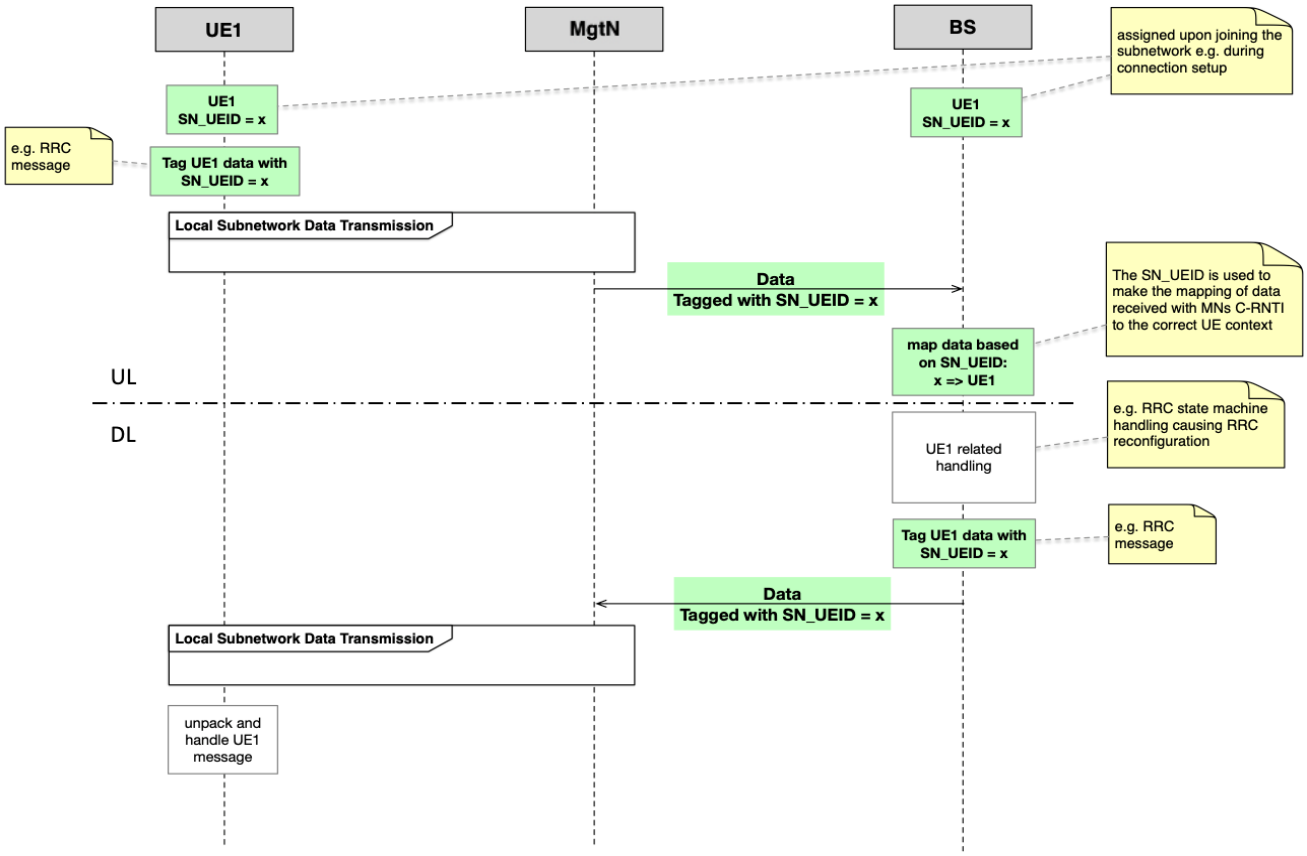

Figure 8 Message sequence chart for Uplink (UL) and Downlink (DL) flow for UE1 of the subnetwork of Figure 5.

In this specific example in Figure 8, UE1 is associated with SN UE ID and this knowledge is shared among UE1, the MgtN and the 6G BS. As far as UL is concerned, UE1 tags its data with its SN UE ID and sends it to MgtN using the local subnetwork data transmission scheme. Examples of the latter could be transmission e.g. via Wi-Fi, Sidelink, *Ultrawideband* (UWB) or similar. The MgtN forwards the received packet(s) with SN UE ID towards the 6G BS. The latter receives the packet(s) tagged with SN UE ID and based on that associates it with UE1's context, e.g. security context for deciphering etc. In the example of Figure 8, the 6G BS receives the RRC message from UE1 via the MgtN and then performs the appropriate handling. To respond to UE1, the DL path is utilized. The 6G BS creates a data packet and tags it as well with the SN UE ID corresponding to UE1 and then sends it to the MgtN. The latter reads the tag and forwards the packet to UE1 using the local subnetwork-specific data transmission scheme.

2.2.1.4 Mobility in and out of subnetworks

Another important aspect of the aforementioned virtual connections is when considering the mobility of the subnetwork devices in and out of a subnetwork. In fact, mobility is crucial for the immersive education use case [1]. The process of joining of a subnetwork and establishing the virtual link is portrayed in Figure 9, where a UE (UE4) who is directly connected with the 6G BS and not being part of a SN wants to join the SN managed by MgtN. At first, UE4 in Figure 9 requests a connection establishment via the MgtN of a given subnetwork. The MgtN in turn forwards the request to the 6G BS. Subsequently, the 6G BS associates UE4 with a SN UE ID as well as with the specific UE that sent the request (i.e., MgtN). More specifically, the 6G BS links UE4 context, SN UE ID, security context or other necessary information to the respective MgtN context, essentially creating the logical link between 6G BS to UE4 via this particular MgtN. The reverse direction is established by the 6G BS sending the assigned

SN UE ID towards the UE4 via the MgtN. Note that UE4 is considered "joined into the subnetwork" from the 6G BS perspective as soon as the UE acquires an SN UE ID from the 6G BS.

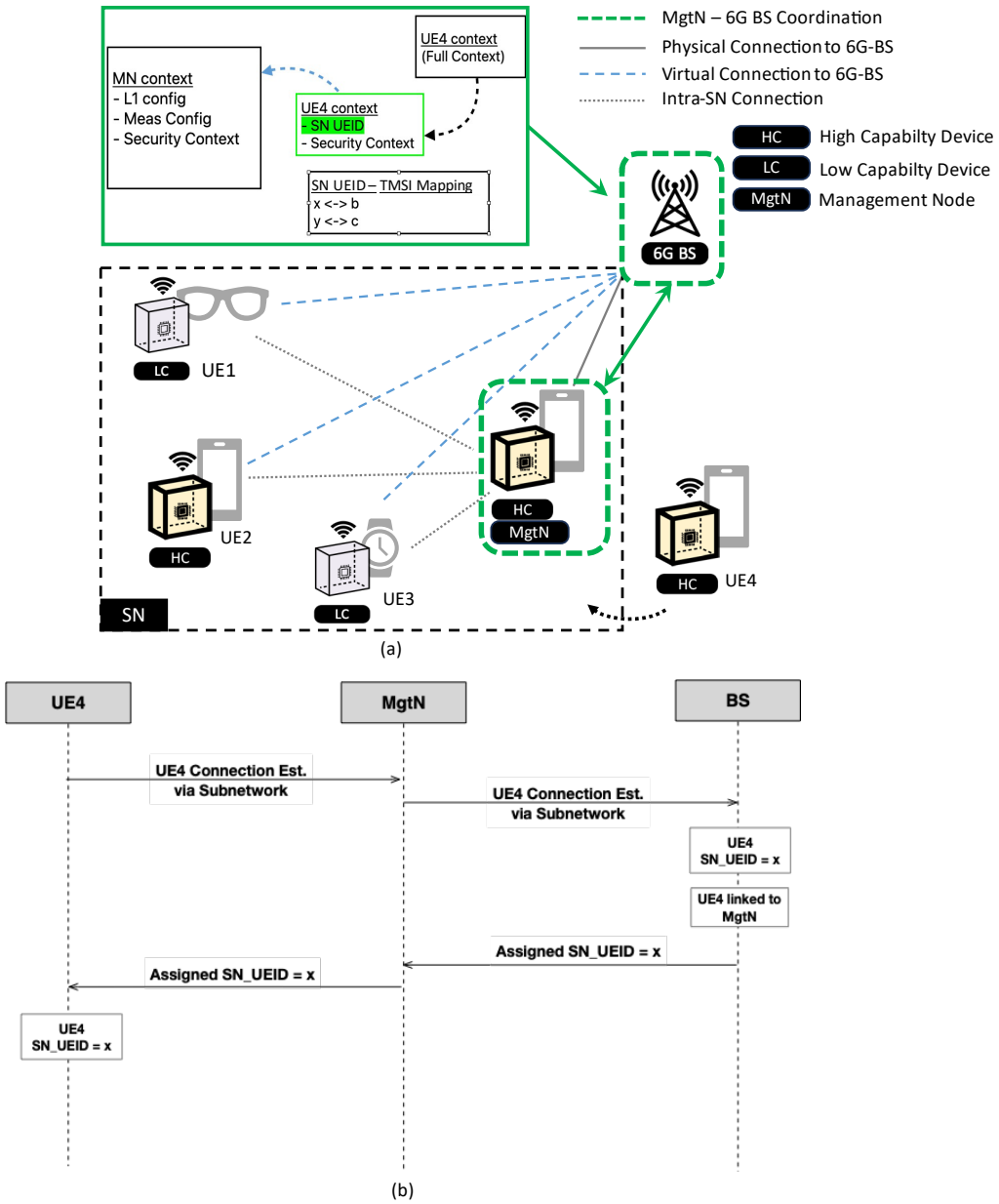

Figure 9 UE4 joining a subnetwork and establishing the virtual link by gaining a SN UE ID: (a) setup and logical configuration at the 6G BS and (b) MSC of the end-to-end configuration.

As for the case where a UE leaves the subnetwork, the configuration setup as well as the respective MSC is portrayed in Figure 10. UE4 was already associated with a SN UE ID by the 6G BS when joining the subnetwork managed by the MgtN. When the UE4 decides to leave the subnetwork and connect to 6G BS directly, it performs a Random Access (RA) procedure or follows the procedures for RA-less processes. In order to allow seamless switching of the existing connection via the subnetwork, UE4 shall send a RA Preamble (Message1) and listen to Message2 from the 6G BS similar to e.g. NR RA [7]. In Message3 the UE4 shall add its SN UE ID, to allow the 6G BS to identify the UE4 as the one that already has a context via a subnetwork and hence create a standalone context for UE4 and releases the respective SN UE ID. Finally, the RACH procedure continues, and the 6G BS may apply necessary changes via e.g. RRC (re-)configuration or reestablishment in order to complete the connection setup.

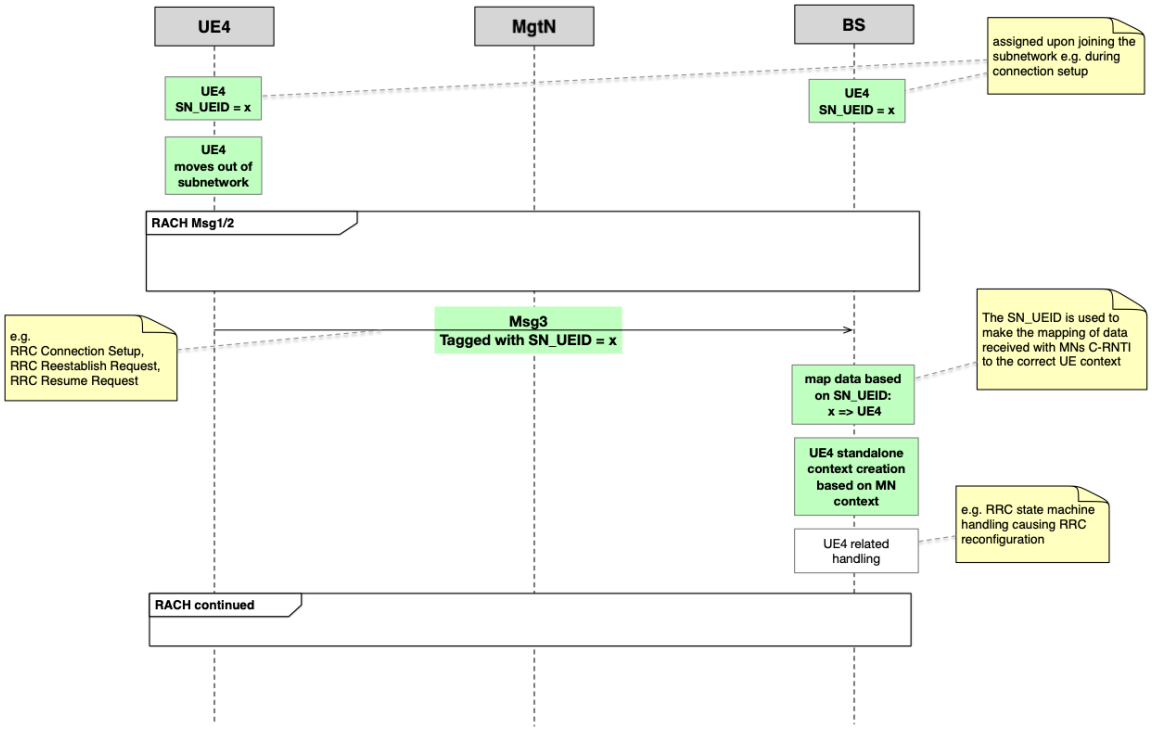

Figure 10 Message sequence chart of UE4 leaving a subnetwork to connect directly to the BS

2.2.2 Distributed User Plane and Control Plane Functionalities among the Subnetwork Nodes

The new flexible role, as it was defined in the previous subsection, requires adaptations of the CP and the UP. These adaptations can happen either at the interaction between the MgtN and the 6G BS, which are presented in Section 2.2.2.1, or at a local subnetwork level , which are resented in Section 2.2.2.2. The first would inherently impact the CP and UP of the MgtN link, which acts as a backbone link for the devices served by the subnetwork, while the latter will have implications on the way the devices interact within the subnetwork.

2.2.2.1 Base Station to Subnetwork Interaction Adaptations

The UEs in the subnetwork may have different *QoS Flows* which results in multiple Data Radio Bearer (DRB) or *Logical Channel* (LCh) entities established between a MgtN and a 6G BS, in an extreme case each QoS Flow could be mapped to a DRB in a 1-to-1 manner. In the state of the art e.g. 3GPP Sidelink and IAB [6] the aggregation happens via intermediate layers Backhaul Adaptation Protocol (BAP) and Sidelink Relay Protocol (SRAP) mainly to pass-through PDCP end-to-end. The aim is to have a more flexible scheme to enable aggregation on different levels, e.g. of different UEs or UE QoS Flows into fewer (even one) DRB/RLC channel/LCh entities. Therefore, this proposal suggests aggregating different UEs or UE QoS Flows into fewer (even one) DRB/RLC channel/LCh entities between MgtN and 6G BS for the sake of limiting the MgtN complexity in the number of DRB/LCh entities for each device in its subnetwork. In the following this protocol is referred to as *Subnetwork Tunnelling Protocol* (SN-TP).

The SN-TP provides a header containing the UE and QoS flow mapping information per packet at the link between MgtN and 6G BS. This information is portrayed in Figure 11 for both UL and DL. Explicitly, the SN-TP could be deployed on different layers depending on where to aggregate the UEs and there

respective QoS Flows within the subnetwork. Exemplary, two possible deployments are shown in the right hand-side of Figure 11. Example 1 shows the deployment of SN-TP above RLC, the whole SN traffic can be multiplexed into a single RLC channel, or a single RLC channel per UE could be established where all DRBs of a UE are aggregated. In Example 2, the SN-TP is deployed below RLC, where UEs' Logical Channels are multiplexed into a common Transport Block (TP) per UE, thus maintaining different HARQ processes for different UEs. Note that additional variants are possible, such as above PDCP or even above SDAP. Furthermore, this protocol could be an MgtN/NW capability and be agreed upon subnetwork formation and/or reconfigured during the lifetime of the subnetwork, e.g. based on the UC currently running. Depending on the layer where SN-TP gets deployed, it is required to carry different mapping information between MgtN and 6G BS within the SN-TP protocol to identify the right UE related protocol entities. More specifically for the two examples of Figure 11 the mapping information will be:

- Example 1: UE ID, DRB ID
- Example 2: UE ID, LCh ID

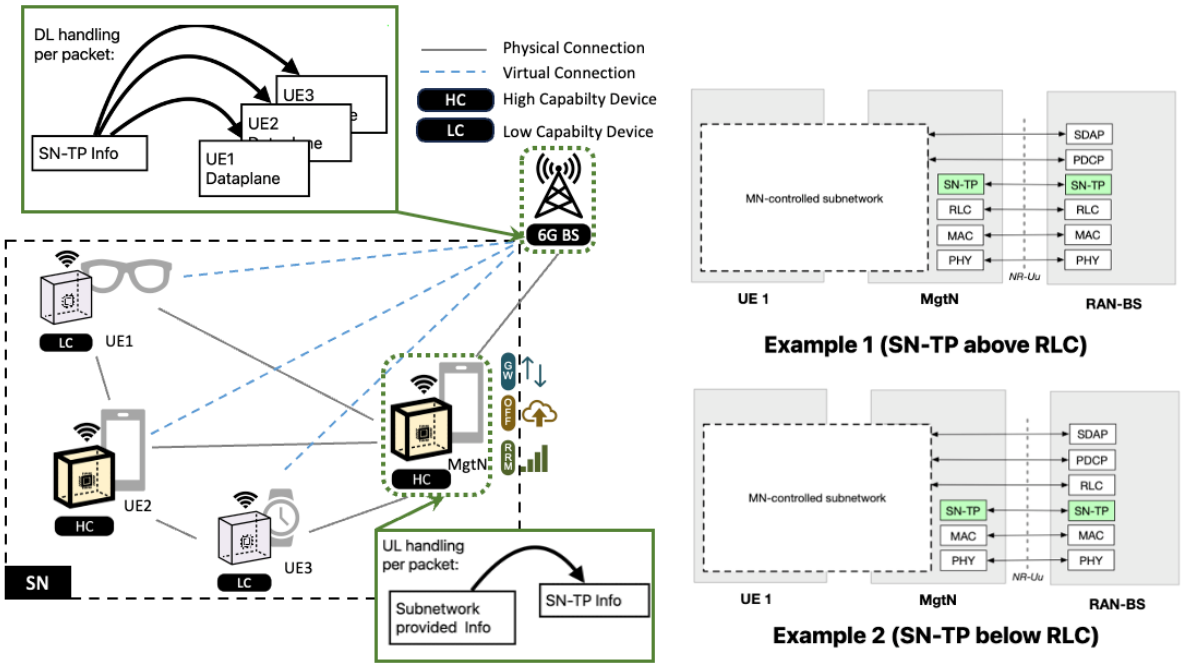

Figure 11 Subnetwork Tunnelling (SN-TP) configuration and subnetwork setup (left) and UP protocol stack (right).

2.2.2.2 Local Subnetwork interactions

In the previous subsection, a solution in the form of the SN-TP protocol is described on how to multiplex the UEs served by the subnetwork on the link established between the MgtN and 6G BS. In this section, the focus will be on the subnetwork side, i.e. the links established among the subnetwork nodes and the MgtN. In particular, the CP and UP deployment and interactions from the overlay 6G NW to the subnetwork UEs.

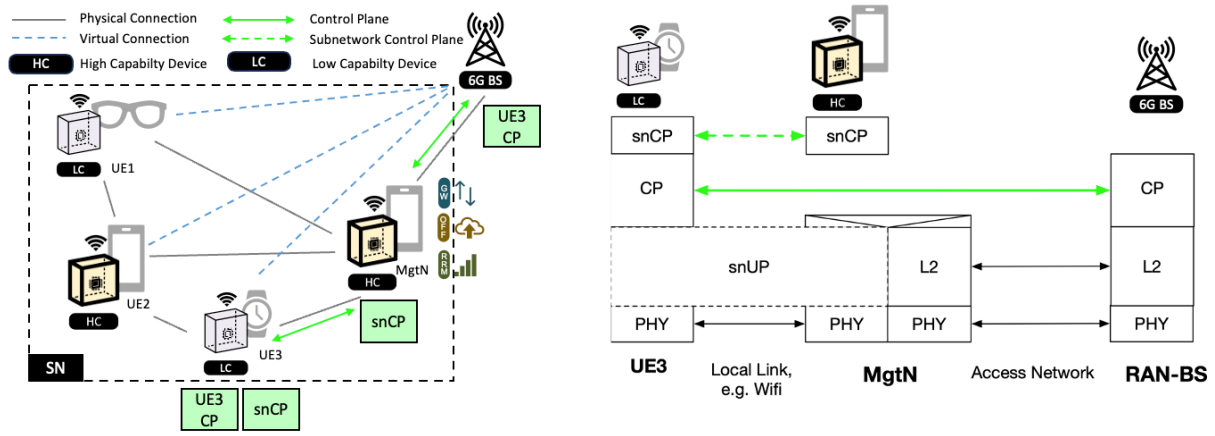

Figure 12 Option 1: the 6G BS CP terminates at the subnetwork UE.

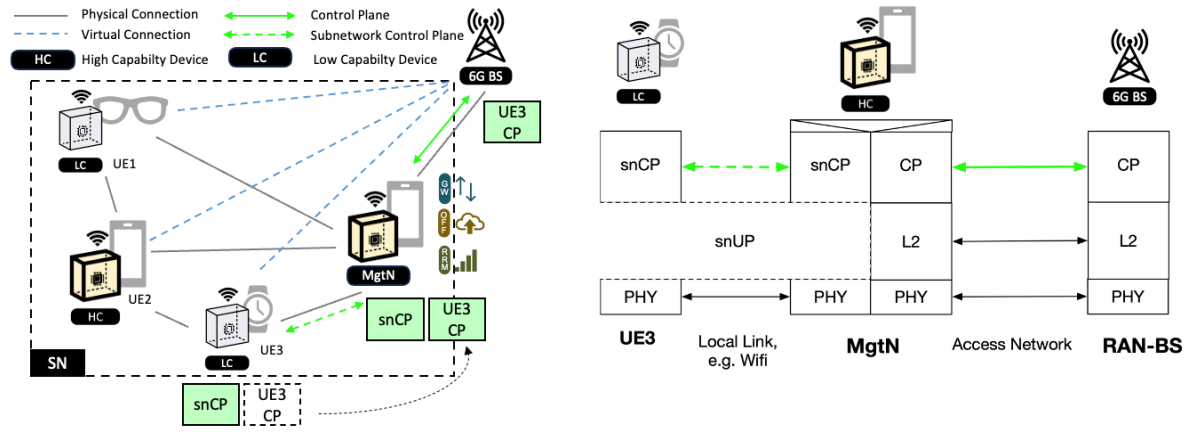

Figure 13 Option 2: the UEs CP is aggregated at the MgtN. [3]

From a CP perspective, two possible deployments are identified. The first option is referred to as *Option 1* and is portrayed in Figure 12. In *Option 1,* the CP related to the overlay network (UE3 CP) remains terminated at the UE and the transport of the related CP messages happens via the MgtN though the SN in transparent manner to the 6G BS. More specifically, the CP data is forwarded to the UEs via the MgtN and the subnetwork. The subnetwork specific control is handled in this case by a separate entity, referred to as *Subnetwork Control Plane* (snCP), which can be independent from the overlay NW in terms of e.g. subnetwork access technology and radio resources used within the subnetwork. The snCP may be a reduced version of the CP handling the overlay NW, supporting only subnetwork related functions, leaner Information Elements (IEs) and simpler state machines. A second option is shown in Figure 13, where the UE3 offloads its CP functionality for communication with the overlay NW, or parts of it, towards the MgtN, e.g. while joining the subnetwork or upon need, save power or reduce complexity of it's own operation. The MgtN aggregates the CP of all the UEs served by the subnetwork and translates it to an snCP entity. In this specific option, the snCP includes offloaded UE CP information as well.

Moving on to the UP, two possible deployments are identified as shown in Figure 14. Note that in both options of Figure 14 it is possible for CP data to be sent to UE2 both directly by the 6G BS and indirectly via UE1 which acts as MgtN. In Option 1, the subnetwork UEs use the overlay NW's UP. The UP terminates at UE2 and may be routed directly from the BS as well as via the MgtN. A possible deployment could be using either normal bearer, bearer duplication or split bearer similar to *Dual*

Connectivity (DC) in NR [8]. In terms of UP feedback, ACK responses on different levels may go to the BS either directly or via the MgtN. It should be noted that the feedback for the different UP paths, i.e. via the MgtN or directly between the BS and the UE, may be decoupled from the respective data path, e.g. ACKs can go over the direct path, while data were sent over the indirect path.

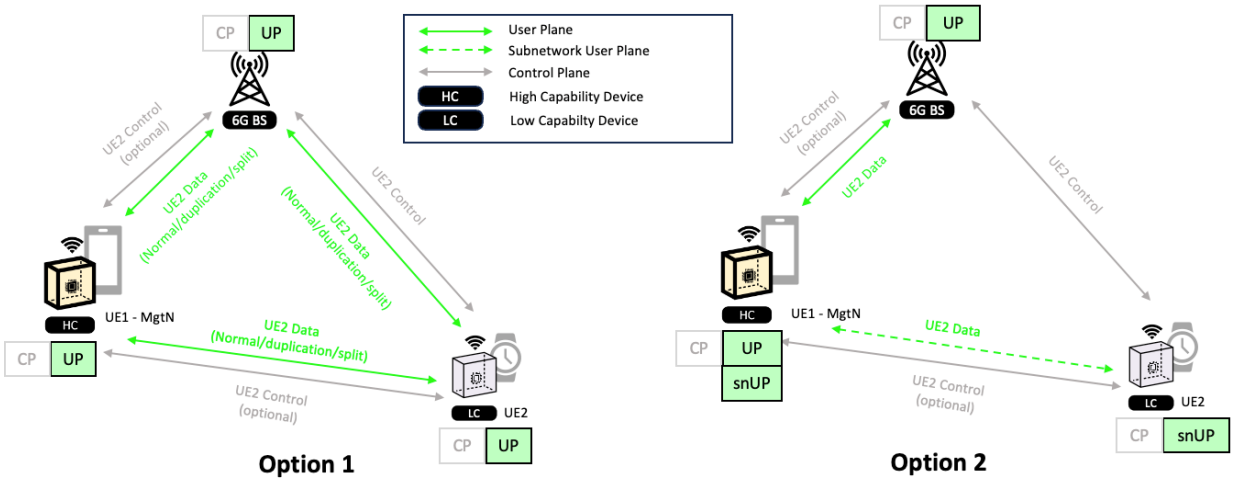

Figure 14 Subnetwork UP deployment options.

As far Option 2 is concerned in Figure 14, the subnetwork UEs are using a *Subnetwork User Plane* (snUP). The latter is a simplified UP architecture with a significant reduction to the UP procedures in comparison to the overlay NW UP architecture, e.g. reduction and consolidation of the different UP L2 layers. Therefore, data is routed from BS via the MgtN to UE2, after some post-processing and conversion to the snUP on the MgtN. In terms of feedback, User plane ACKs from UE2 go to the BS via the MgtN, e.g., RLC and PDCP entities for UE2 bearer in MgtN.

The snUP is controlled by the MgtN and it is agreed within the subnetwork without any overlay NW involvement (i.e., 6G BS). For example, SDAP and PDCP, which in 5G handle QoS flows and security (among other things), can be deployed flexibly and transparently to the NW. Additionally, those functions should be deployable in a flexible manner and capable of terminating at devices within the subnetwork depending on the use case, for example:

- Depending on the used subnetwork technologies, RLC might be used to cover more than the cellular link portion, due to the fact that, e.g. the reestablishment procedure due to RLC maximum retransmissions may cover the whole path through the subnetwork.
- Some UEs might want to rely on end-to-end PDCP ciphering for privacy reasons (e.g., intermediate nodes are not trusted, or the used intra-subnetwork communication does not provide enough security). Such UEs may agree with the MgtN to have PDCP/SDAP at the UE side.
- Some UEs might have a higher trust level among each other, or the intra-subnetwork communication provides strong encryption, and those UEs want to offload the PDCP ciphering to the MgtN.
- Some UEs have tight low-latency requirements (e.g., using low-latency guaranteed bitrate bearers) and want to have SDAP terminating on the UE in order to perform proper mapping, other UEs have more relaxed requirements (e.g., only best-effort traffic with no particular prioritization or mapping to different bearers), so SDAP functionality is not required at the UE,

as the MgtN could decide on its own the respective mapping to guarantee QoS flow requirements..

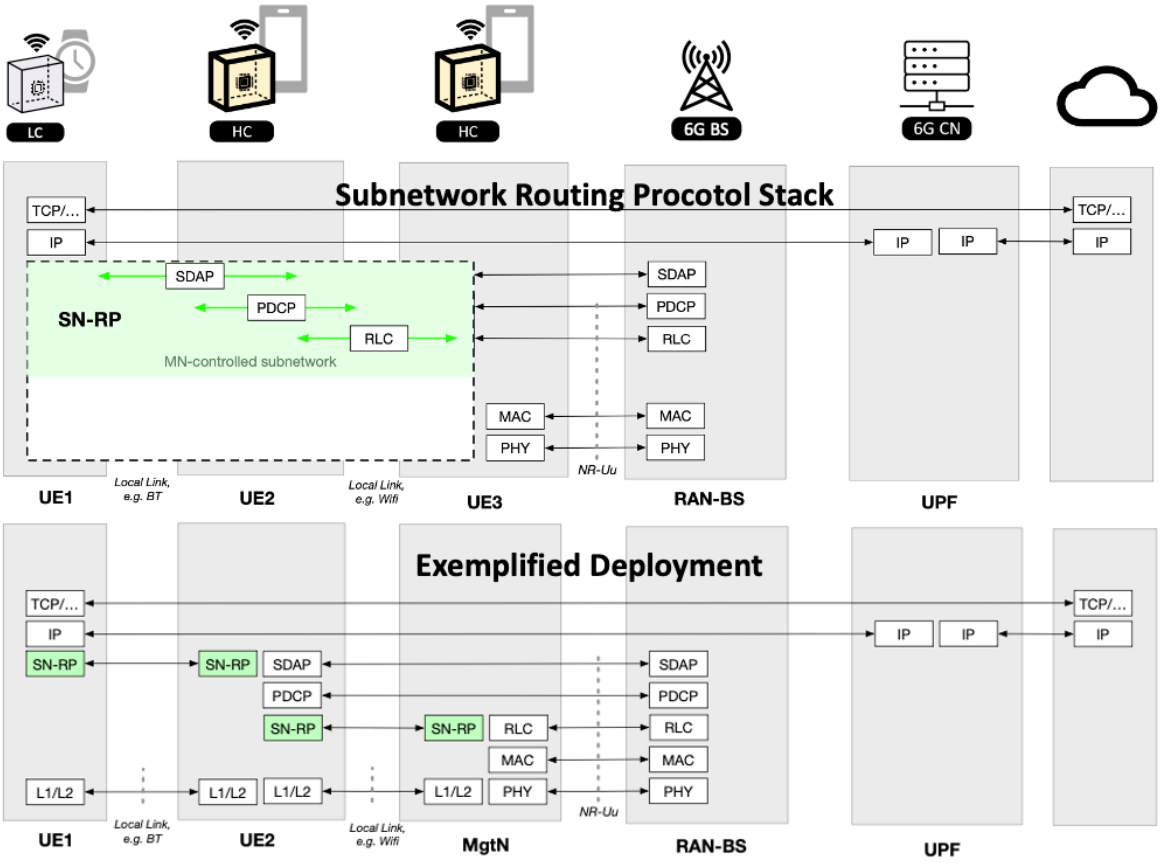

Figure 15 Subnetwork Routing Protocol (SN-RP) stack (upper figure) and an exemplified deployment (bottom figure) with a nested subnetwork and various layer deployments of the SN-RP.

For the above reasons, a novel protocol is proposed to accommodate the flexibility of the snUP, namely the *Subnetwork Routing Protocol* (SN-RP). This protocol shall be deployed above MAC and can accommodate different L2 sublayer deployments (SDAP, PDCP, RLC) across the subnetwork by carrying mapping information for the different layers between the different devices. Each UE that supports subnetwork functionality shall support the SN-RP and its UE capabilities - which can change over time to achieve e.g., power saving - define what layers it supports on the device (e.g. some devices may only support SDAP and PDCP on the device and therefore relies on RLC being terminated on another node). The SN-RP delivers routing through the subnetwork by containing routing related information per packet. It also carries mapping information in-band per packet, which is needed to perform the next mapping step(s). The protocol stack architecture of the SN-RP is shown in the upper subfigure of Figure 15, while an exemplified deployment with various upper layer termination points of the SN-RP per hop is portrayed in the bottom subfigure of the same figure.

2.2.3 Coordination between APs of different subnetworks

So far, the focus on dynamic topologies has been within the subnetwork and on its interaction with the overlay 6G BS. However, several subnetworks are expected to be located in a close vicinity to each other. For instance, a single subnetwork could be comprised by each student's mobile phone or laptop acting as a MgtN and the student's wearables. Given a typical classroom size of tens of students, where every student may form an individual subnetwork among his devices, tens of these small subnetworks will be formed [13]. This dense subnetwork deployment places a significant overhead on the 6G BS and by

extension to the *Core Network* (CN), since local inter-subnetwork data and control flows should be routed through the 6G BS. For the sake of reducing this overhead, the local inter-subnetwork data and control flows could be routed directly from one MgtN to another MgtN by using e.g. *Device-to-Device* (D2D) communication modes or even by establishing an overlay subnetwork of consisting of neighbouring MgtNs. Note that the latter is referred to as subnetwork *Entity* (EN) [1]. More specifically for the immersive education UC, the classroom could be regarded as the EN, containing several neighbouring subnetworks.

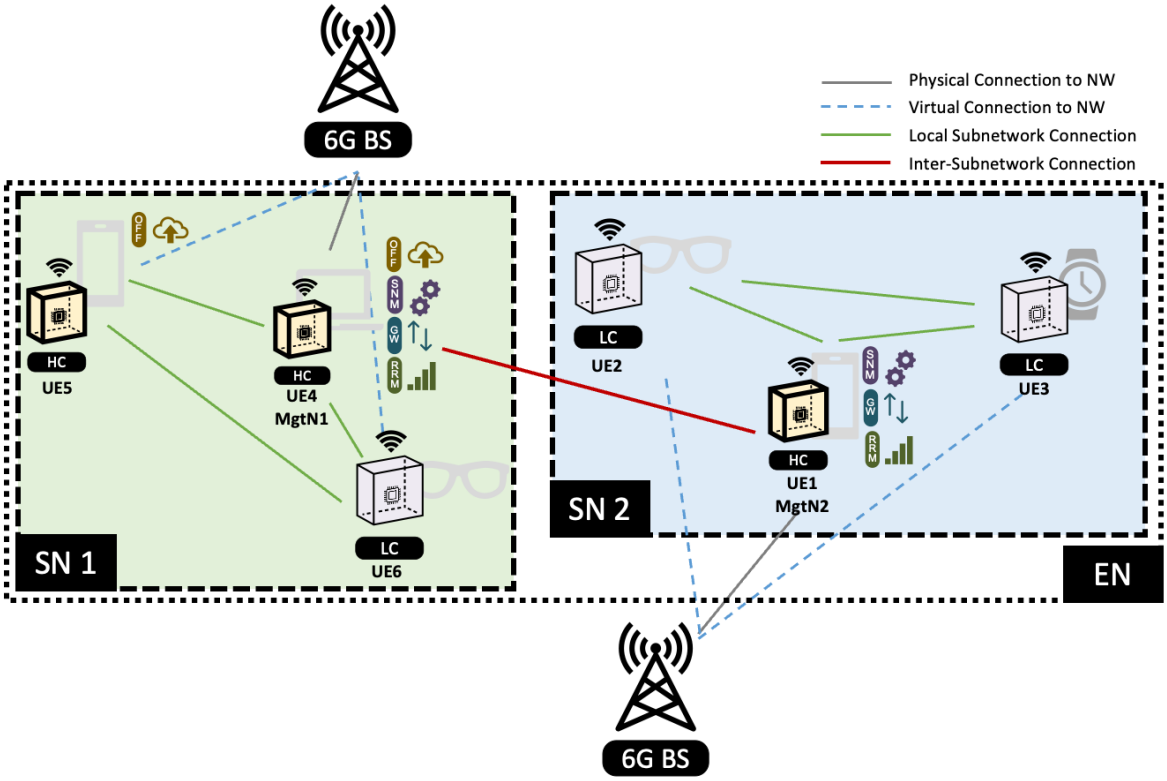

Figure 16 Entity consisting of two subnetworks, each connected to different 6G BSs; an inter-subnetwork link is also established between the two MgtNs.

A simple example for the EN is shown in Figure 16, where the EN consists of two subnetworks, namely SN1 and SN2, managed by MgtN1 and MgtN2, respectively. Note that there is no need for the subnetworks to be associated with the same 6G BS, as portrayed in Figure 16. Apart from the local intra-subnetwork connections among the subnetwork UEs and the MgtN – BS links, a new type of connection is introduced within the EN, namely that of the *inter-subnetwork connection*. This connection is used for the coordination between SN1 and SN2. This interface is similar to the X2 and Xn interfaces between eNBs and gNBs, respectively. However, in the context of the inter-subnetwork communication, this interface should involve simplified procedures and IEs, similar to the snCP and snUP design. Some notable examples of coordination using this interface are:

- Sharing of *Radio Resource Control* (RRC) measurements for the sake of improving the mobility decisions of both HC and LC UEs. More specifically, this sharing increases the side information available at the individual UEs resulting in more accurate mobility decisions and thus improving the subnetworks' performance.
- Sharing of QoS information experienced by UEs in neighbouring subnetworks. The neighbouring UEs are associated with different 6G BSs. This assists the UEs in making a more accurate *handover* (HO) decision to another 6G BS and leading to reduced HO completion time.

- Resources sharing between subnetworks for both computational and functional offloading, thus improving the capabilities offered to the UEs by the subnetworks.

In the upcoming deliverables we will delve into the problems statements and provide implementations for addressing them.

2.3 QoS FRAMEWORK FOR IN-X SUBNETWORKS

With the introduction of Extended Reality (XR) services and Cloud Gaming (CG) applications, new demanding requirements are put on the systems that need to deliver and facilitate such services.

This means that supporting such services and applications with high demanding requirements, would require a framework that can ensure the service can be provided with required level of "Quality" that is expected from the user(s). This can be in terms of timely delivering of data, with emphasise in Ultra Reliable and Low Latency communication. For many of such services, the (QoS) framework is being put at stretch, with large amount of data to transfer between multiple users and entities as well as being time and delay sensitive, and with a need for synchronized delivery. This will put high requirements on the system to ensure not only latency and data rate, but also to deliver data in a synchronized manner over different link-types, e.g., to and from 6G parent networks, and possibly in some cases via relay nodes.

For providing XR experiences that make the user feel *immersed* and *present*, several relevant Quality of Experience (QoE) considerations are needed, such as providing the experience of being physically and spatially located in the virtual environment, when using a head mounted display (HMD).

Some general and high-level requirements indicated that XR applications require highly accurate, low-latency tracking of the device movement at about 1kHz sampling frequency. The XR Viewer Pose information [14], associated to time typically results in packets of size in the range of 30-100 bytes, i.e., the XR Viewer Pose needs to have some assigned a time stamp.

Some further high-level aspects and requirements related to XR presence:

- Tracking
  - 6 degrees of freedom tracking – ability to track user's head in rotational and translational movements.
  - 360 degrees tracking – track user's head independent of the direction the user is facing.
  - Sub-centimetre accuracy – tracking accuracy of less than a centimetre.
  - Quarter-degree-accurate rotation tracking.
- Latency
  - Less than 20 ms motion-to-photon latency – less than 20 milliseconds of overall latency
  - Minimize the time of pose-to-render-to-photon. Rendering content as quickly as possible. Less than 50ms for render to photon in order to avoid wrongly rendered content.
- Persistence
  - 90 Hz and beyond display refresh rate to eliminate visible flicker.

To support the sense of presence and immersive experience, the age of the content and user interaction delay is vital. Here different types of delays need to be considered such as interaction delay (time for initial processing), how long time it takes to deliver the content to receiver, e.g., game server, and

Round-Trip interaction delay, which is the total delay before the activity is presented for the end user, e.g., wearing the HMD.

In the context of in-X-subnetwork, outlined in [1], use cases and requirements have been defined in the area of consumer, industrial and automotive categories, all considering applications requiring low latency, high data rate, and with high reliability, i.e., having URLLC characteristics-type of traffic.

2.3.1 Selected Use-case: Indoor Interactive Gaming

To develop the QoS framework for the subnetwork operation, the investigation will initially consider the indoor interactive gaming as the use-case. This use-case has stringent requirements in terms of data rates, latency, and synchronization which could be benefited by the operation of subnetwork. In Figure 17, the traffic flows within the subnetworks are illustrated. The red and green lines are representing the link for downlink and uplink direction, respectively. Furthermore, these links are between HC as the access-point/relay node with network functions (e.g., gateway, computation offloading, RRM) to the HC and/or LC devices. The link between the access-point/relay node, HC device, to the 6G parent network is represented in dash-purple line. Lastly, the link to/from SNE is represented in thin-blue line. All of the links may have different link characteristics.

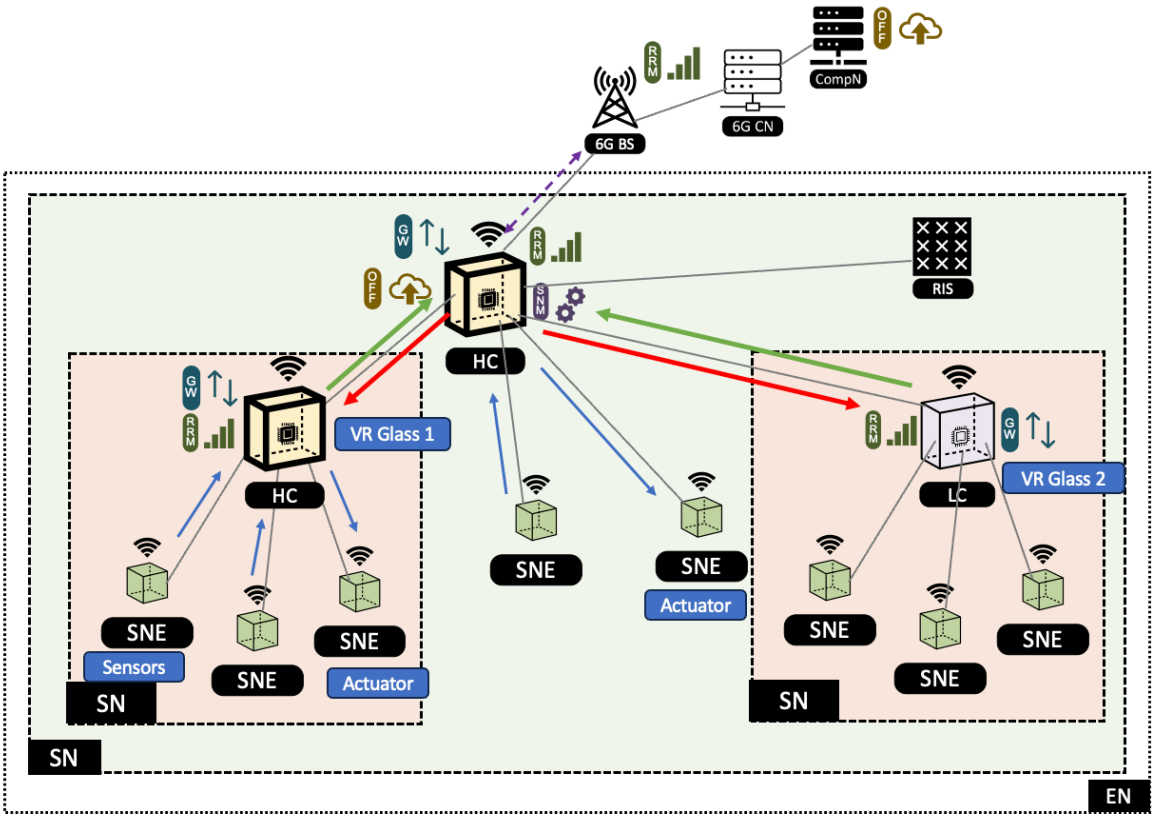

Figure 17 Subnetworks architecture and the traffic flows of indoor interactive gaming.

There are various types of traffic such as video, data, and pose. HC can represent the VR glass where the device can perform video rendering by itself. The HC can also act as a gateway to the sensors, controllers, and actuators attached to the users. There can be other users in the room with similar HC or LC device. In case of LC, we consider the device performs split rendering at the computation node. All the devices and subnetwork elements are connected to an HC which carries gateway function, RRM, and computational offload. We consider all those network elements are relatively close, or such as

within the target 6G-SHINE use cases (i.e., 10 m). The primary challenge of this use case is to support high data rate in a timely manner and with synchronized data delivery.

2.3.2 Current / Existing Solutions on QoS

In interactive gaming it is important that the video has a very low latency in order for the user to be able to react quickly to the video content. The video is in this use case received and shown on the display with high resolution in the end user device, which may be connected via the subnetwork, see Figure 16. To support this use case where the devices are connected, e.g., via sidelink relay, the end-to-end QoS between the application function in the network and the device is very important. The QoS of the feedback channel, sending the pose to the application, is equally important.

It is therefore important to guarantee the End-to-End QoS in both downlink and uplink are essential to support this kind of use cases. The existing QoS solution that are being under investigating is based on the 3GPP QoS Framework.

In 3GPP, concepts related to Quality of Service have been specified. The current QoS concept up to Rel-17, is based on that the UPF in the Core Network sends the data with different QoS flows. Each flow representing specific QoS requirements. All data in an IP flow will get the same Quality of Service requirement and thereby it will be directed to the same QoS flow. A QoS flow contains data where the QoS is described by a parameter 5QI (5G QoS Identifier) which identifies the QoS characteristics of that QoS flow [15]. RAN can, based on the 5Qis, prioritize the data in the different QoS flows when scheduling transmissions over the air interface in order to fulfil the required QoS of each QoS flow.

In Rel-18 the QoS concept for XR traffic is adapted with PDU Set handling where the PDU Set QoS parameters, PDU Set Delay Budget (PSDB), PDU Set Error Rate (PSER) and the PDU Set Integrated Handling Information (PSIHI) are added to the description for a QoS flow.

For a subnetwork using UE to NW (U2N) sidelink relay, where the communication between the base station and the device, here called a remote UE is relayed over a UE-to Network Relay, the QoS handling is somewhat changed depending on the type of relay which is used.
There are two kind or relays defined in 3GPP, Layer 2 relay and Layer 3 relay. In a Layer 2 relay as shown in Figure 18 [6], the data is handled just above the RLC in the protocol stack. Here the SRAP (Sidelink Relay Adaptation Protocol) is added in the Relay and the SDAP and PDCP layers are terminated in the Remote UE. Thereby the QoS flows in Layer 2 relays are used between the gNB through the relay to the remote UE.

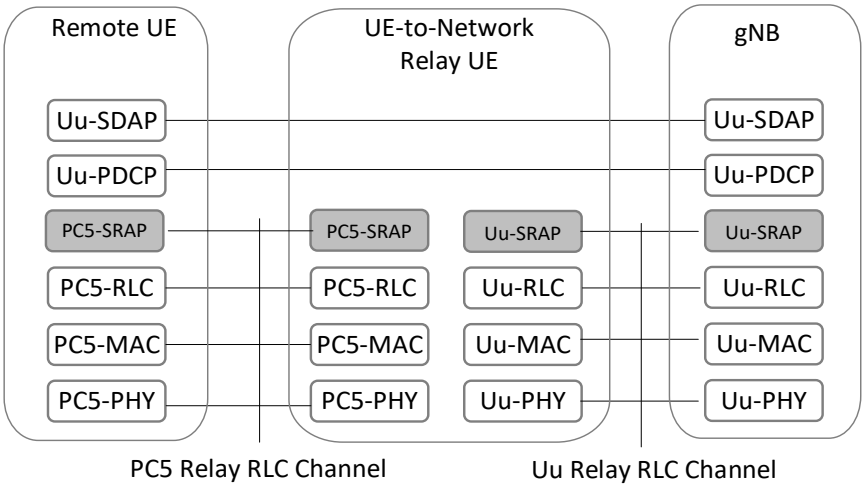

Figure 18 User plane protocol for the L2 UE to Network Relay

For a Layer 3 relay as shown in Figure 19 below, the data is handled in the SDAP layer in the relay UE where the QoS flow is terminated. A new QoS flow defined for the sidelink (PC5) communication with a new identifier PQI, (PC5 QoS Identifier) used in a similar way as the 5QI is used for the Uu interface. Thereby the end-to-end QoS is handled by two separate QoS flows with different identifiers, 5QI and PQI.

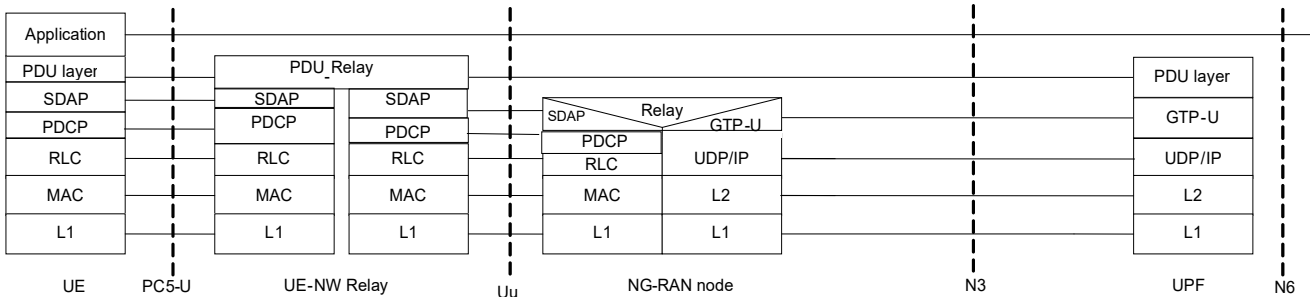

Figure 19 User plane protocol for the L3 UE to Network Relay

In Figure 20 the QoS flows are indicated for the Layer 2 and Layer 3 Relays. For Layer 2 relays, the QoS Flow defined by the 5QI is used over both the Uu interface and the PC5 interface. For Layer 3 relays, there are two QoS flows, one for the data sent over the Uu interface and then another one for the PC5 interface.

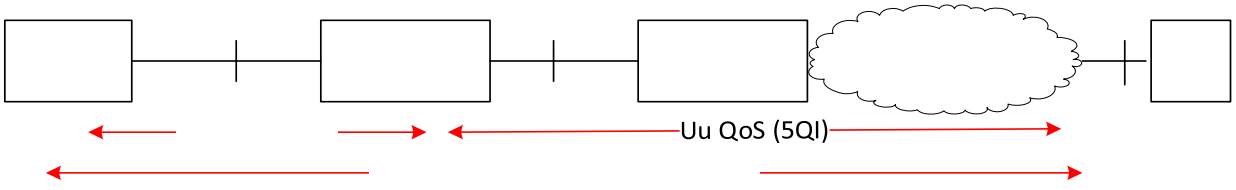

Figure 20 The QoS flows for L2 and L3 UE to Network Relays

In the case of direct sidelink communication between two UEs without any relay, the QoS framework is similar as between a UE and the network with QoS flows between the SDAP layers on the two UEs, as shown in Figure 21. Here the QoS flow is based on the PQI.

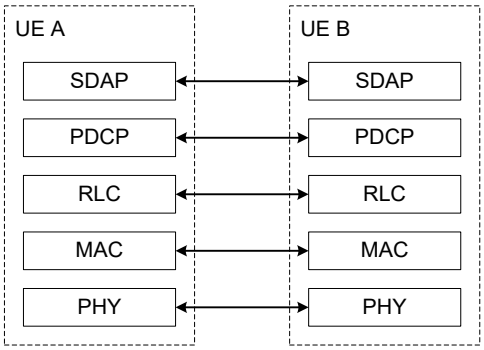

Figure 21 The User plane protocol between two UEs using sidelink communication.

for a UE to UE (U2U) Relay, when there is a relay node between the two UEs which are communicating, the QoS flow is defined end-to-end between the two remote UEs over both PC5 channels, similar to the Layer 2 U2N relay. Here the PQI is used to identify the QoS over both links.

In all cases with relays, the End-to-End QoS for U2N relays depends on the combination of the Uu interface and the PC5 interface.
For U2U relays it depends on the combination of two PC5 interfaces.

2.3.3 Challenges for QoS aspects

Supporting the mentioned use cases above, but not limited to, puts a rather high challenge on solutions fulfilling such demanding requirements in various environments.

To ensure that relevant requirements can be fulfilled for developing in-X-subnetworks with defined use cases, there is a potential for further optimizing the QoS framework defined, e.g., by 3GPP, both looking into the communication between the UE and the Core Network as well as investigating the QoS functionality for, e.g., V2X and sidelink operation within subnetwork communication. This should trigger the following questions:

- Is the currently defined framework good enough?
- If it is not good enough, can it be enhanced?, or
- is there a need for a new concept?

For example, how to define and support the QoS aspects for in-X subnetworks operating both under the control of the overlay 6G parent network and as a stand-alone subnetwork with direct communication between subnetworks.

Different types of challenges would be relevant for different use cases, scenarios, and environments. E.g.:

- Stand-alone vs operating under 6G network control
- Using licensed or unlicensed spectrum
- Operating in a public outdoor environment or indoor small cell
- Involving few or a large amount of users
- Synchronized delivery of data from multiple input sources to multiple users or receiving parties.

For XR gaming (consumer use cases) the main challenges are [1]:

- To provide various sensors outputs from multiple nodes in a synchronized manner
- To provide the scenes and information for different users in a synchronized manner
- To provide ultra-low latency and high reliability communications

- To provide high data rate communication (i.e., providing XR scene to the users)

For industrial use cases, e.g., robot control, extreme requirements on latency are foreseen to be required, for example on operating quick movements of a robot arm.

In the area of automotive with a legacy of cabled connection, a potential wireless vehicle subnetwork needs to ensure that control functionality depending on sensors and actuators are able to operate undisturbed, fulfilling all safety requirements from the car industry.

2.3.4 Investigation areas for QoS for In-X subnetworks

The existing QoS framework in 3GPP has been defined and including XR and V2X/sidelink scenarios. The QoS framework defined by 3GPP can be used as the baseline or starting point for identifying potential enhancements.

The challenge may be whether the existing framework is also suitable and/or capable of handling XR traffic that is using some sort of relay node, both between subnetworks and the overlay 6G network, as well as between devices within a subnetwork and between subnetworks.

Future work on QoS for in-X subnetworks, should include investigation of if the QoS framework can handle situations in different scenarios and environments. This should also work in a mix of different environments, supporting the different use cases and use case groups as defined in terms of consumer, industrial and in-vehicle subnetwork categories,

If it does not work in the defined QoS framework, there may be a need for new revolutionary ways of defining requirement toolset for enabling and securing the most extreme requirements from the mentioned use cases.

Investigation will be conducted addressing the selected challenges above and proposals will be made for potential enhancements and/or new solutions for QoS or policy control mechanisms in the selected use cases, particularly interactive gaming use case.

3 DYNAMIC COMPUTATIONAL RESOURCES OFFLOADING WITHIN SUBNETWORKS, AMONG SUBNETWORKS AND TO 6G EDGE-CLOUD

This chapter discusses several approaches related to the methods, procedures and architectural aspects relevant for computational resource offloading from subnetworks to the 6G edge nodes or cloud, i.e., towards the 6G 'umbrella' network. Initially, a framework-based approach for computation resources and task offloading is proposed, tailored for the 6G SHINE use cases, particularly the consumer subnetworks where XR traffic is exchanged among entities. The approach adopts as baseline several 3GPP-related concepts such as a split architecture, and within the framework XR traffic management mechanisms suitable for the 6G-SHINE architecture are discussed. Next, an approach that enables the convergence of communication and computation, particularly focusing on consumer subnetwork use cases (i.e., immersive classroom), is presented. The approach adopts the introduction of new and enhancing existing CP and UP procedures to facilitate the distribution of computation among the network entities. Lastly, a framework for leveraging coordinated planning between communication and computing technologies across subnetworks and towards the 6G 'umbrella' network is presented. The framework enables computational resource offloading for functions and services, including stringing requirements such as deterministic service levels, making it suitable for vehicular and industrial use cases.

3.1 A FRAMEWORK FOR COMPUTATION RESOURCES AND TASK OFFLOADING

This section discusses a framework for computation resources and task offloading comprised of two components. The first component refers to the RAN and encapsulates all SNEs, HC, LC devices, and 6G BS, and their interconnections, whereas the second component refers to the CN and captures all the interconnections between the 6G BS and the edge servers, and other CN functions. The proposed framework aligns with the split architecture approach proposed and discussed in [46]. Within our proposed framework, we have developed the following:

- Regarding the first component, a retransmission mechanism has been developed for a consumer subnetwork serving XR traffic to users is proposed. The aim is to generalise the mechanism for the other subnetwork cases and different traffic types. The output of the proposed mechanisms is a trigger type information that is passed over to the second component.
- Approach for the second component. In the CN, we build an emulation of CN deployment using an in-house-developed platform and use the trigger type information from the first component in a decision-making process for offloading computation resources or tasks to multiple edges. Similarly, this is done for consumer subnetworks and XR traffic, however, the approach is generalised and can be used for the other subnetwork cases and different traffic types.

The rest of this section is organized as follows. Initially, the relevant state-of-the-art of resource offloading is presented, including works on UE and edge nodes agreeing on the distribution of tasks under changing wireless conditions and resources requirements modelled via utility functions; followed by works looking at architecture for game networking where tasks from user devices are offloading to the edge; and lastly works related to the 5G extended reality evolution that capture the standardization perspective. The following two subsections discuss the proposed traffic management mechanisms for handling the (re)transmissions and the emulation with AdvantEdge [47], which is a

mobile edge emulation platform. Both combined are demonstrating the preliminary work within the proposed framework.

3.1.1 State of the Art

The relevant state-of-the-art can be grouped into three categories: 1) works related to UEs and edge nodes agreeing on distribution tasks, 2) works on architecture for game networking, and 3) works related to standardization on the topic of extended reality. In the theme under 1) and 2), the main aspect is the offloading or distribution of tasks between entities either because of variable conditions of the wireless link or due to different resource requirements, whereas the theme under 3) should provide insights related to standardization trends, in particular for XR studies conducted in 3GPP. The observations summarized in the state-of-the-art are used as basis for developing the proposed framework.

Figure 22. shows an end-to-end path between a UE and a cloud using a three-tiered model of computing [48], where in the case of the subnetworks the model can be adopted by mapping the UE representation to different subnetwork entities (e.g., SNEs, HC, or LC), and by mapping the cloud with the 6G 'umbrella' network.

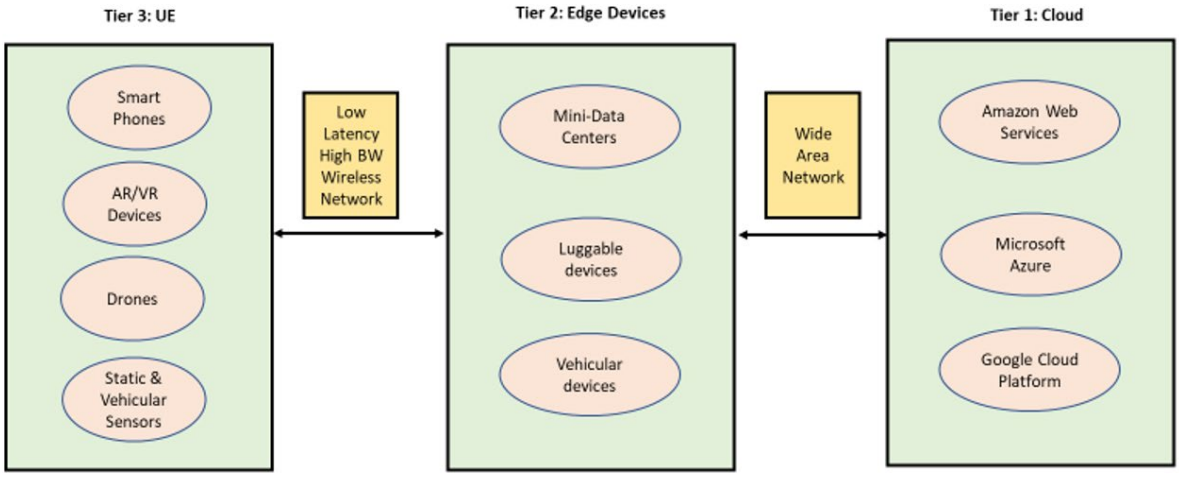

Figure 22 The tiered model of computing, as discussed in [48]

Each tier has different design considerations, such as compute elasticity, storage permanence and hardware consolidation (for Tier 1), network proximity (for Tier 2), and mobility and sensing (Tier 3), as highlighted in [48]. These design considerations can be extrapolated and applied for the case of subnetworks, for example, depending on the mobility or the changes in the wireless link of a SNEs (within a subnetwork or across subnetworks) and HCs in Tier 3. Further, different resources can be provisioned in Tier 2 and Tier 1, in order to maintain the established session, or a computationally intensive XR application running on a device can be decomposed such that it can be run across two tiers in the presented model. Further, the decomposed tasks may require multiple resources to be completed. For example, such as XR tasks may require resources such as CPU, network bandwidth, battery, cache, and memory. The work in [49] provides further details on the topic, where using utility functions that map multiple quality dimensions of decomposed tasks and corresponding resources are discussed. Additionally, an approach to exchange system parameters between UE and edge nodes is presented.

The work in [50] shows a view on the evolution of game networking technology designed to connect large numbers of users in real-time online gaming environments and discusses the technical requirements to facilitate such an environment. This maps well with multiple subnetwork use cases particularly from the consumer and industrial category where large number of SNEs are connected and require real-time updates. In addition, in such cases multiple XR devices in the consumer subnetwork need to exchange real-time updates such as in the immersive education use case. The reference [50] explores the basic principles of game networking, and highlights several design considerations such as synchronisation, entity interpolation, input prediction, and delay compensation.

The work in [46] proposes a split architecture, depicted in Figure 23, and captures relevant aspects of XR over 5G. In the split architecture, time budget examples for motion-to-render-to-photon (below 70ms) and for 5G RTT and device and server processing (below 20ms) are given. Another reference [51] states that 'more than 20ms is too much for VR and especially AR, but research indicates that 15ms might be the threshold or even 7ms' and calls for further hardware improvements to bring down the latency.

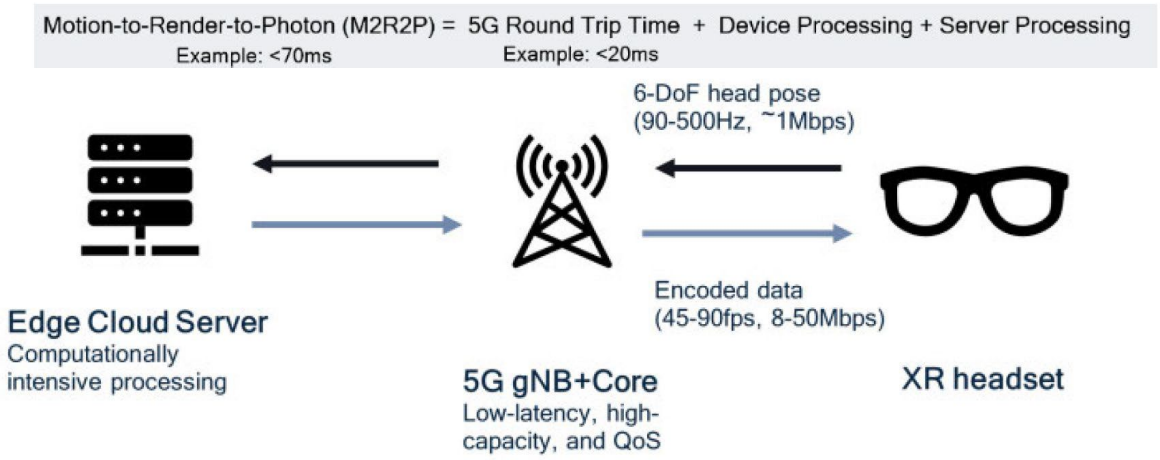

Figure 23 Split architecture for XR, as discussed in [46]

The UL and DL requirements discussed in conjunction with the split XR architecture include benchmark values for the following XR parameters: periodicity, packet size, delay bound, data rate, for VR, AR, cloud gaming and audio+data use cases. In addition, a simulation study indicating how implementation of delay-aware scheduling for indoor scenario (100 MHz bandwidth, 30 Mbps VR traffic per users) provides improvement over Proportional Fair (PF)scheduler [46]. Lastly, several XR enhancements relevant to 5G standardisation are discussed, such as: PDU sets, burst Indications, latency-aware scheduling, enhanced CDRX, enhanced configured grant, 5G directed staggering of XR traffic, 5G Awareness of XR applications and lower layer mobility. The split architecture approach as well as the XR enhancements considered relevant for 5G standardization are adopted as starting points for developing the framework relevant for the 6G-SHINE use cases.

In a nutshell, the state-of-the-art shows that for XR traffic there is a need to deploy and coordinate solutions/mechanisms both in the RAN and in the core network in order to provide sufficient end-to-end performance.

3.1.2 Traffic Management Mechanisms

XR is a broad term that includes all real and virtual combined environments and human-machine interactions, referring to several immersive experiences such as VR, AR, and MR [53]. In XR services and applications, the traffic may consist of PDUs belonging to a PDU set or data burst. A PDU from a PDU set may be associated with different segments or components of a video frame or a video slice. A data burst, on the other hand, may consist of one or more PDU sets. Further, XR traffic may be described with multiple parameters such as, e.g., periodicity, jitter, packet size, packet loss, delay boundary or PDU set related indications such as priority. The characterisation of the XR traffic should consider multiple inter-dependent PDU sets. In the following, a more detailed description of the concept of PDU sets is given, as well as the proposed mechanism for retransmissions based on the XR traffic characteristics is also demonstrated.

For XR traffic, the concept of PDU sets is considered, where in simple terms, a PDU set is a collection of PDUs/packets that carry a single media frame (e.g., video frame) generated at the application layer. The 5G Advanced system may leverage multiple PDU set information such as PDU set sequence number and PDU set delay budget information used by the RRM to ensure the QoS requirements are met. A simple representation of a PDU set is given in Figure 24, extracted from [53], where the index k identifies a user, the index i identifies the PDU set ID, and the index j identifies the PDU within the PDU set.

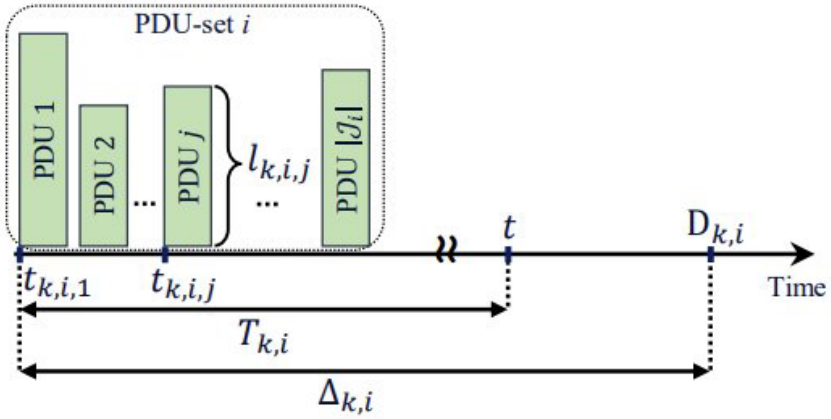

Figure 24 A representation of PDU set, as discussed in [53]

3.1.2.1 Proposed mechanism for XR related (re)transmissions

The proposed mechanism for XR related (re)transmissions are defined within the first component, i.e., in the radio access, and the novelty is that the mechanism considers application-layer subnetwork specific information to perform MAC layer retransmissions. In the diagrams, for simplicity, the message exchange between a UE and 6G BS is shown, however this can be a message exchange between different combinations of subnetwork entities such as SNEs, HC, and LC devices. The proposed mechanism is based on a cross-layer approach between the application layer metrics that describe the PDU set and the MAC layer PDU transmissions. It consists of handling the transmissions of the MAC PDUs based on PDU set information, and adjusting the access and computational resources in the core network based on the physical layer and PDU set related measurements (for example, usage of edge nodes with higher computational capability or resource elements for over-the-air transmission) by the UE and coordination with the 6G BS or 6G umbrella network. Lastly, the UE reports information about the deployed retransmission logic (i.e., type of algorithm) and relevant PDU set information. Figure 25 and Figure 26 below capture the proposed mechanism.

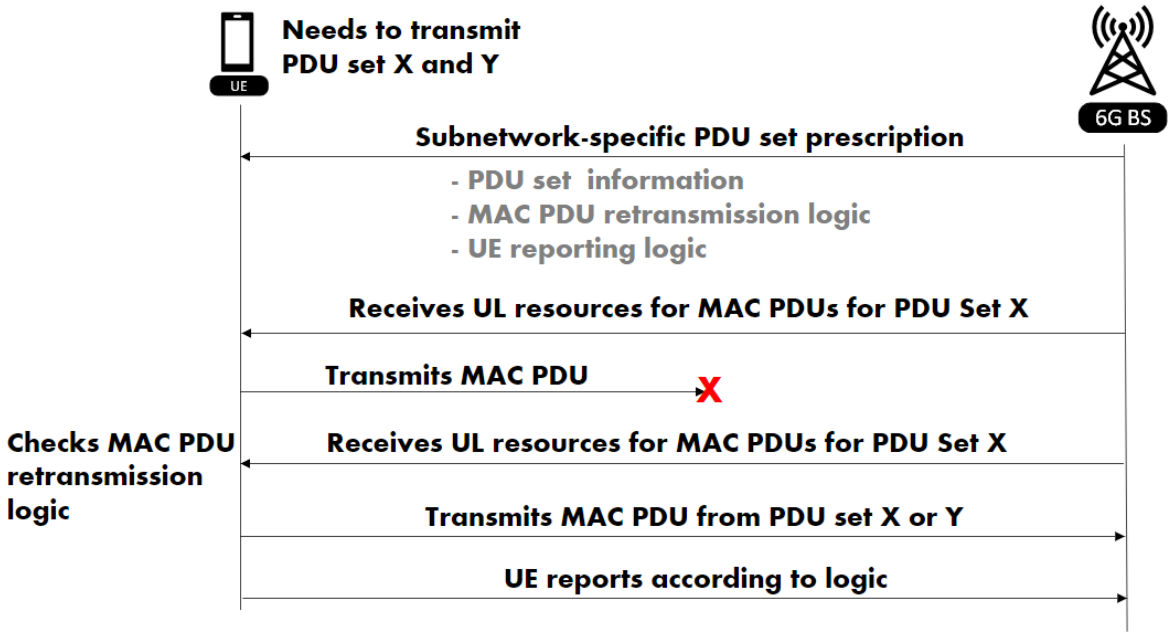

Figure 25 MAC PDU transmissions management

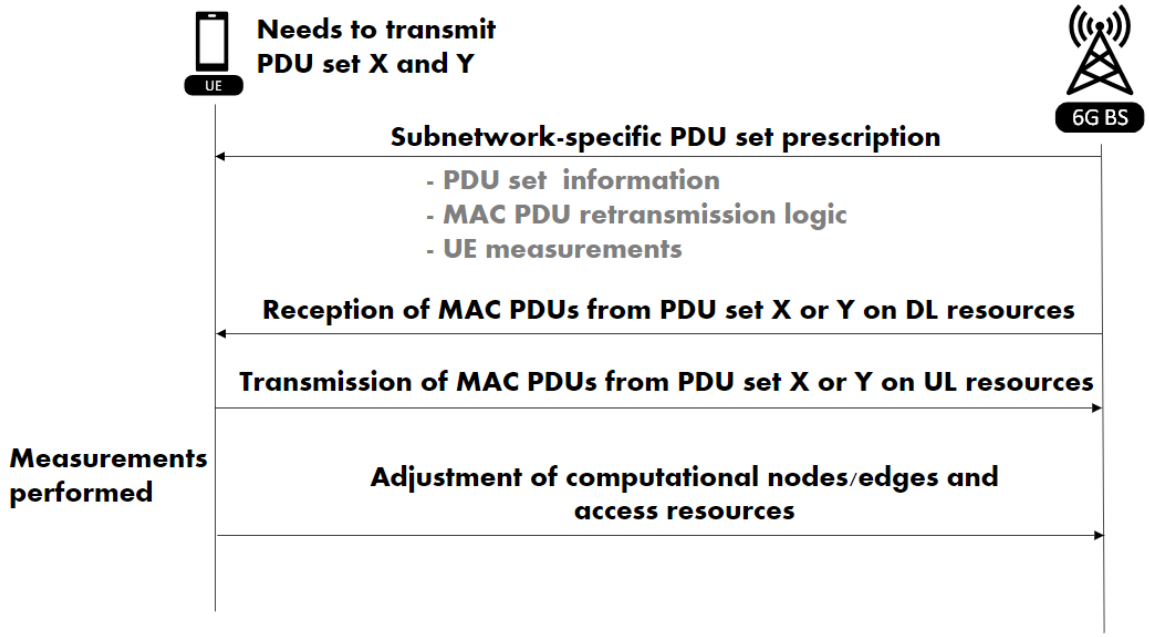

Figure 26 Access and computational nodes adjustment

3.1.3 Computational offloading using AdvantEdge

The computational offloading as part of the second component of the proposed framework is achieved using the platform AdvantEdge [47]. AdventEdge is a mobile edge emulation platform that runs on Docker and Kubernetes and provides an emulation environment that enables experimentation with edge computing technologies, applications, and services. Further, AdvantEdge facilitates exploring edge deployment models and provides the ability to investigate their impact on applications and services in short and agile iterations. In the following paragraphs, only the aspects relevant to the proposed framework from the broad set of AdvantEdge capabilities are discussed, as well as preliminary setup scenarios are demonstrated.

3.1.3.1 Description of AdvantEdge

AdvantEdge is a controller software composed of micro-services that interact together to allow deployment and testing of edge scenarios. The micro-services are packed in Docker containers that operate in Kubernetes and grouped in core platform, core sandbox, dependency, and scenario groups. A scenario is defined using a multi-layer model composed of: scenario, logical domain, logical zone, network location, physical location, and process which can be used to deploy one or multiple subnetworks controlled by entities in the 6G 'umbrella' network. Further, there is support for wireless connectivity where changes in the wireless link can be implemented, as well as changes in network characteristics expressed via latency, jitter, throughput and packet loss. The following compute characteristics are supported: CPU limits and memory limits. Lastly, there is a frontend composed of options for platform description, scenario creating, exporting/importing scenario, scenario deployment, events generation, dashboard observation and customization, and settings. A snapshot of a scenario is given in Figure 27, depicting one 6G BS (in the figure labelled as poa), a HC device (in the figure labelled as term) that runs multiple services such as video and data transfer, an Edge node (in the figure labelled as edge). We will implement the (re)transmission mechanism between the HC device and the 6G BS as discussed in the section 3.1.2.1, and based on this mechanism we will trigger offloading procedure between Edge nodes for the XR traffic exchanged between the Edge and the HC via the 6G BS.

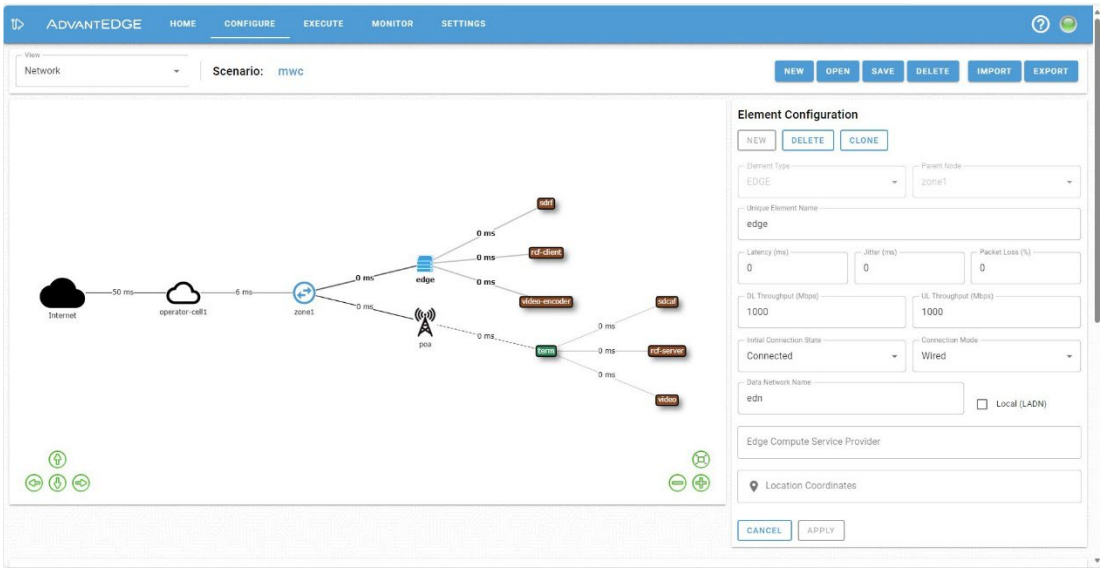

Figure 27 AdvantEdge scenario [47]

The following features within AdvantEdge are relevant for the proposed framework:

- Network mobility event. A physical node changes the location within the network, thus emulating mobility pattern/movement.
- Network characteristics event. Emulates changes in the networking conditions such as latency increase/decrease, reduction of bandwidth, error rate increase/decrease, etc.
- Compute event: emulates changing of compute environment where new edge services or terminal applications can be started at scenario runtime.
- PDU connectivity event. Facilitates creation and deletion of PDU sessions to the data network.

During the course of the project, several scenarios will be deployed and investigated. These will be driven by input information generated by the proposed mechanisms in Subsection 3.1.2, and will demonstrate the computation offloading approaches as discussed in Subsection 3.1.3. Part of the investigation will be used in the PoC work in WP5. Results will be reported in the follow-up deliverable.

3.2 Converged Communication and Computation Subnetworks

The immersive classroom use case, as defined in [1], is the basis for our study on the convergence between communication and computation within subnetworks. In this use case the use of multiple high and low capability devices would entail looking into methods for offloading some of the demanding computations among the different devices participating in the subnetwork impacting both the legacy UP and CP architectures. Contrary to the contributions of Section 3.1, where computing nodes are located in the cloud, in the immersive education use case there are devices with high computational capabilities such as laptops or HC UEs, thus enabling deployment of computing nodes within the subnetworks for local compute offloading. Moreover, the trade-offs and KPIs for decision making policies for the distribution of computations need to be analysed in this context to make sure that the decision between on-device and offloaded computation is well studied and the decision is taken allowing for an optimised overall system especially in terms of latency, power consumption, offloaded data accuracy and privacy.

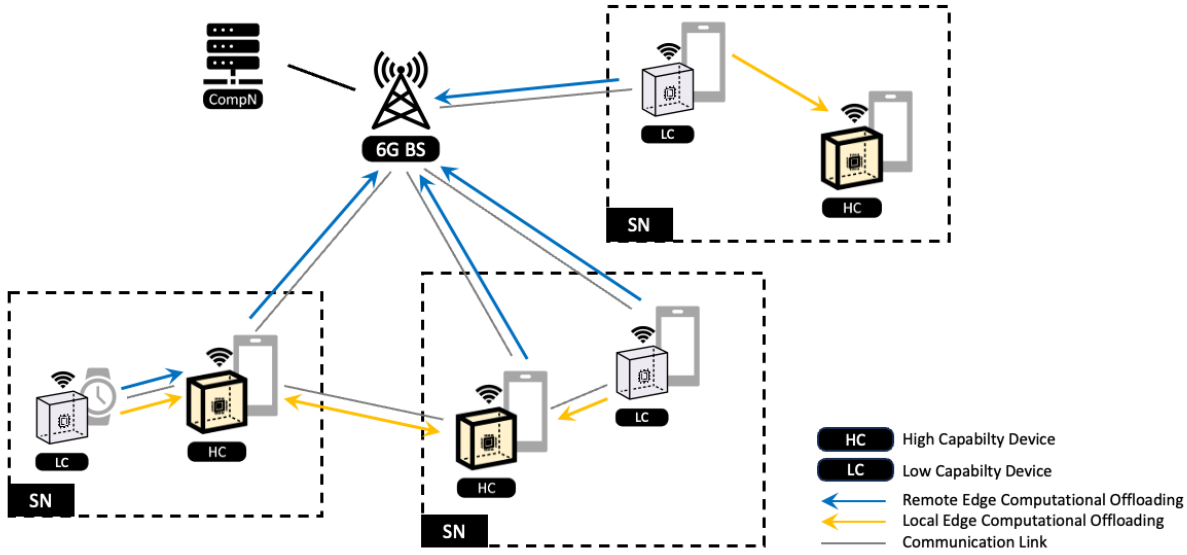

Figure 28 Overview of multiple subnetworks connected to cellular network with support for local and remote computational offload

Figure 28 gives an overview of how multiple subnetworks, comprising devices of different capabilities (i.e., low, and high capability devices), can be connected to the cellular network and enabling computational offloading either locally or to a remote edge. For this to be achieved, new CP and UP procedures need to be introduced to allow for such computational offloading to be achieved ensuring that the required latency, power consumption, offloaded data accuracy and privacy requirement are met.

Starting with the new CP and UP procedures, further investigations would be needed into: procedure for the discovery of the compute node(s) within subnetworks and corresponding signalling protocols; procedures for the transfer of the computational load; and computation-aware procedural enhancements (e.g., mobility, reestablishment, etc.). To enable these new and enhanced procedures, a set of general functional architectural components need to be first introduced [3]. Figure 29 provides an overview of the functional entities consisting of an Offloading Node (ON), a Computing Node (CompN), a Compute Offload Controlling Node (CCN) and a Routing Node (RN) with connections indicating the transfer of computation control and data between the different nodes. The ON, is a functional entity connected to a wireless network, having a compute task to be offloaded to one or

more CNs (e.g., LC or HC device from Section 2.2.1). The CompN, is a functional entity connected to the wireless network with certain processing capabilities to perform an offloaded compute task and produce compute results (e.g., HC device from Section 2.2.1, core network function, remote/edge server, etc.). The CCN, is another integral functional entity connected to the wireless network that collects all compute capabilities from all available CompNs and makes compute offload decision based on their capabilities and current load among other parameters. Finally, the RN is the last of the introduced functional entities connected to the wireless network at which the compute task(s) and result(s) gets routed between the different ONs and CompNs. Note that all those entities are just functional entities and their realization in the wireless network can vary depending on the use case's nature and the underlying requirements. In the use cases of interest, it is expected that the introduced control entity, namely the CCN, would be flexibly deployed in the network or subnetworks setup, where it can be deployed at the subnetwork edge nodes. Therefore, such CCN can be deployed at MgtNs as an extension to their SNM functionality or even to other HC devices within a subnetwork. This yields that there might be several devices within a single subnetwork and among a network of subnetworks (i.e., subnetworks connected in a direct subnetwork to subnetwork connection or connection routed via a central network-controlled entity such as a 6G BS or a core network) that has a functioning CCN entity. Such deployment, with multiple CCN entities within a subnetwork or a network of subnetworks, would entail the need for having schemes in place for negotiations between the different deployed CCNs. This is going to be further studied and analysed in the coming deliverable.

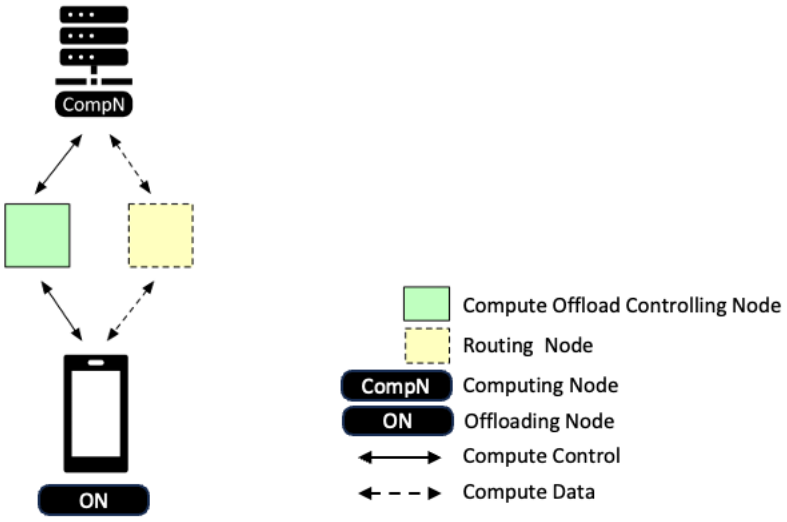

Figure 29 General functional architecture of distributed compute [3]

After having the different procedures that enable the distribution of compute tasks and results between the different nodes in the wireless network, policies need to be defined on how to make the decision on when to do on-device and off-device computation as well as the KPIs that need to be met for the different use cases to ensure that the offloading of computations doesn't entail a negative impact on such use cases. For this a characterization of the compute workload in terms of traffic class, computation complexity, communication resources, memory, precision, and the required and realizable end-to-end quality of compute service needs to be done. Additionally, compute offloading KPIs such as energy consumption, latency, communication, and computation costs need to be further investigated.

As a recap of the contribution in the context of convergence of communication and computation, an initial analysis has been provided on how to enable use cases such as the immersive classroom, where

a set of functional entities have been introduced. These entities are foreseen as essential for the introduction of new and enhanced CP and UP procedures to allow for the distribution of computation among the different entities in the wireless network, more specifically, taking the newly proposed subnetwork architecture with dynamic roles and new entities like MgtN(s) that are introduced in Section 2.2. Additionally, some details have been included on the possible deployment options for the introduced functional entities within a subnetwork and the consequent need for further study on having schemes in place for negotiations between the different control entities within a subnetwork and a network of subnetworks. Finally, a first set of KPIs and trade-offs for decision making policies for the distribution of computations have been identified. In the upcoming deliverables, a more thorough analysis into the procedures and into those policies will be undertaken.

3.3 Joint Task and Communications Scheduling for Dependable Service Level Provisioning

With the growing levels of automation in critical vertical services such as smart production and connected and automated driving, there exists an increasing number of applications with compute-intensive nature that cannot efficiently be supported within the subnetworks. This has raised the need for offloading certain tasks to the computing infrastructure (edge computing, MEC and cloud computing) available at the network side. Computational or task offloading to the computing infrastructure at the network side offers a more efficient implementation solution in terms of energy consumption and cost than the required dimensioning of the subnetworks for accomplishing compute-intensive functionalities [16]. However, deciding whether a task should be offloaded, or not, and when and where to offload a task, is not trivial.

The subnetworks ecosystem, like the in-vehicle networks and Electrical/Electronic (E/E) architecture, are introducing functional safety design principles with the softwarisation (i.e., software defined vehicles) and utilization of multi-purpose and redundant control and computational elements and systems. This flexible ecosystem provides high freedom for the local distribution and allocation of tasks' and functions' processing to computing hardware resources. There exists also a clear trend for the integration of subnetworks and (edge-)cloud infrastructures as part of the 6G 'network of networks' vision. The cloudification of subnetworks tasks relies on a deep edge–edge–cloud computing continuum that becomes a strongly interwoven technical ecosystem and can be flexibly configured to meet different service requirements [17]. Task offloading for dependable service level provisioning requires a seamless integration and planning of the communication and computing technologies. This vision is aligned with fundamental research objectives driving other SNS JU EU funded projects such as the DETERMINISTIC6G [16].

In this context, a preliminary framework designed to leverage coordinated planning between communication and computing technologies across both the subnetwork and 6G parent network is introduced. The goal is to facilitate efficient dynamic computational resource offloading for functions and services within subnetworks, including those with stringent requirements such as deterministic service levels.

3.3.1 Deterministic Service Provisioning in Subnetworks

In the domain of subnetworks, deterministic service provisioning is crucial for ensuring the reliable and predictable fulfillment of individual communication and computational tasks, as well as their aggregated outputs. For example, in vehicular and robotics subnetworks, where real-time data exchange is critical, deterministic provisioning guarantees consistent and timely delivery of information. This entails providing services with a high degree of certainty regarding factors such as timing, performance, and

resource allocation, particularly within the context of 6G networks. Deterministic services require well-defined timeframes and a predetermined level of confidence, as well as consistent and predictable performance. Such commitment is essential in the evolving landscape of subnetworks, including in-vehicle networks and Electrical/Electronic (E/E) architecture.

The need for clear service delivery schedules supports deterministic communication by ensuring precise timing and confidence levels in service provision. Moreover, the dynamic nature of subnetwork environments requires an adaptive approach toward maintaining service reliability and effectively managing workloads. Further resource utilization can also be optimized by utilizing offloading mechanisms, intelligently distributing computational tasks among computational elements, and enhancing overall system efficiency.

By dynamically monitoring system resources and real-time application workloads, and intelligently distributing computational tasks, subnetworks can uphold deterministic service levels even under fluctuating conditions. This adaptive strategy ensures that network infrastructures can respond to the evolving needs of diverse applications and workloads, thereby meeting strict reliability requirements and optimizing resource utilization.

3.3.2 State of the Art

Literature on task offloading is extensive. In this section, focus is placed on publications that aim to incorporate communication aspects into computational resource allocation for task offloading.
Some studies in this area, such as [18]-[22], aim to minimize latency for various computational tasks. To achieve this goal, the solutions proposed in [18]-[22] determine how much data should be computed locally versus offloaded for remote processing, taking into consideration factors such as communication bandwidth and processing power. In their analysis, they divide the latency associated with computational offloading into two distinct terms. The first term, known as the "latency of processing", is calculated by dividing the computational demand of a task by the processing rate of the computational element. This term represents the time required for computation. The second term, named "latency of communication", denotes the time taken to transmit data over the communication link. It is determined by dividing the length of the task in bytes by the capacity of the transmission link. This term is directly related to the availability of communication resources. The transmission time for returning the processed information is neglected due to its smaller volume compared to the transmitted data. The objective is to minimize the total of these two components. In [23], the upper bounds of latency are taken into account for different tasks, aiming to ensure compliance with these constraints. This entails ensuring that both the latency of computational and communication components remain less than the specified maximum latency thresholds.
The works presented in [24]-[26] consider a communication network implementing Time Division Multiple Access (TDMA), and they reserve a specific time slot/interval of the TDMA frame for the offloading tasks. These proposals require the latency of both computational and communication components to remain less than the duration of the allocated time slot. The works presented in [25] and [26] consider a communication system based on Non-Orthogonal Multiple Access (NOMA) for task offloading and transmission to the computational elements. Communication resources are then adjusted to this technology, i.e., power allocation across channels to control the data transfer rate.
The works presented in [27]-[28] explore different scenarios for task offloading, involving local computational elements, edge units, and cloud infrastructure. These studies consider the latency of processing within the local computational elements, with no consideration given to local communication delay. They also account for the communication and computing latency for edge units and cloud

infrastructure. The goal is to minimize the sum of all delays incurred in local, edge, and cloud computing, aiming to achieve optimal latency performance across the entire network architecture. Deciding what tasks are processed locally and what task are offloaded to the edge and cloud infrastructure is challenging. AI/ML techniques can be of help in making these complex decisions [29].

Some publications, such as [38], endeavour to minimize the maximum value of the sum of computing and communication delay, aiming to reduce latency in worst-case scenarios. This approach seeks to optimize both computational and communication aspects to achieve improved overall system performance, particularly in scenarios with stringent latency requirements.

Some studies, such as [30]-[32], extend the tasks offloading decision-making to also consider the energy consumption, in addition to the computational and communication resources. For example, [30] concentrates on optimizing resource allocation to minimize overall energy consumption, considering both energies expended for communication (transmitting data) and computational energy usage (processing tasks on devices). In [31], the goal is to ensure that the total energy consumption for a task remains below a specified maximum limit. The work presented in [32] seeks simultaneously optimizing both delay and energy consumption, combining energy efficiency with latency reduction. This approach involves defining a novel cost function that combines both energy and latency considerations.

The studies outlined in [33]-[36] introduce in the task offloading process the capacity of the communication channel as a constraint to ensure that the communication rate for task offloading and transmission to the computational elements remains within the capacity of the link. The data transfer rate on each link is determined by considering the amount of interference and the channel bandwidth. Similarly, [37] considers in the task offloading decisions the transmission channel status. The work in [39] defines an objective function to optimize the task offloading which consists of delay and reliability. Reliability is defined as the inverse of link rate.

The review of the state-of-the-art shows that there are additional considerations regarding the communication part for the task offloading that have been overlooked in previous works. Firstly, beyond energy consumption, latency, or transmission time, metrics such as bandwidth utilization, communication distance, and network congestion could play crucial roles in task offloading. For instance, while offloading tasks to edge computing elements may reduce processing time compared to using the local computational elements, it may also lead to increased communication resource usage and network congestion. Existing literature also tends to focus on minimizing latency. However, this approach might result in unnecessarily high utilization of communication and computational resources and system load when the services can tolerate higher delays. Relaxing the service's latency to near the maximum latency requirement eliminates the option for performing countermeasures, such as retransmissions or task redistribution. Moreover, many critical (deterministic) services benefit from maintaining low jitter.

Going beyond current state-of-the-art, a preliminary framework for joint task and communication scheduling for dependable service level provisioning is proposed, along with a system modelling that accounts for relevant characteristics and requirements of the services, communication network, and computational elements. This system modelling is subsequently leveraged to formulate a set of objective functions and constraints that account for the flexibility provided by the local subnetwork and end-to-end subnetwork-6G parent network continuum to efficiently manage and allocate workloads (e.g. tasks) of functions and services within subnetworks, including those with stringent requirements such as deterministic service levels.

3.3.3 Framework for Service-Aware Joint Allocation of Communication and Computing Resources

In this section, the proposed preliminary framework to support dynamic computational resources offloading using a service-aware joint planning and allocation of communication and computing resources is introduced. First, focus is placed on the subnetwork level that leverages the flexibility offered by current trends towards software-defined platforms (e.g. software defined vehicles) for managing and allocating workloads across the local subnetwork continuum. This framework is then expanded to account for the potential of the seamless integration and collaboration between subnetworks and the 6G parent network forming a deep edge–edge computing–cloud computing continuum.

The proposed framework draws inspiration from software-defined networking, which distinguishes three separate layers: the application layer, the control plane layer, and the infrastructure layer. Software-defined networking enables programmability and network customization through software-defined controllers. It also facilitates network virtualization, allowing multiple virtual networks to operate on the same physical infrastructure. It also enables centralized control over network communication and computational resources, aiming to enhance network management efficiency and empower real-time adjustments of network communication and computational resources in response to changing service demands.

3.3.3.1 Intra Subnetwork Framework Architecture

The framework proposed for the service-aware joint task and communication scheduling in subnetworks is represented in Figure 30. It comprises the application layer, control plane, and infrastructure layer that are described below.

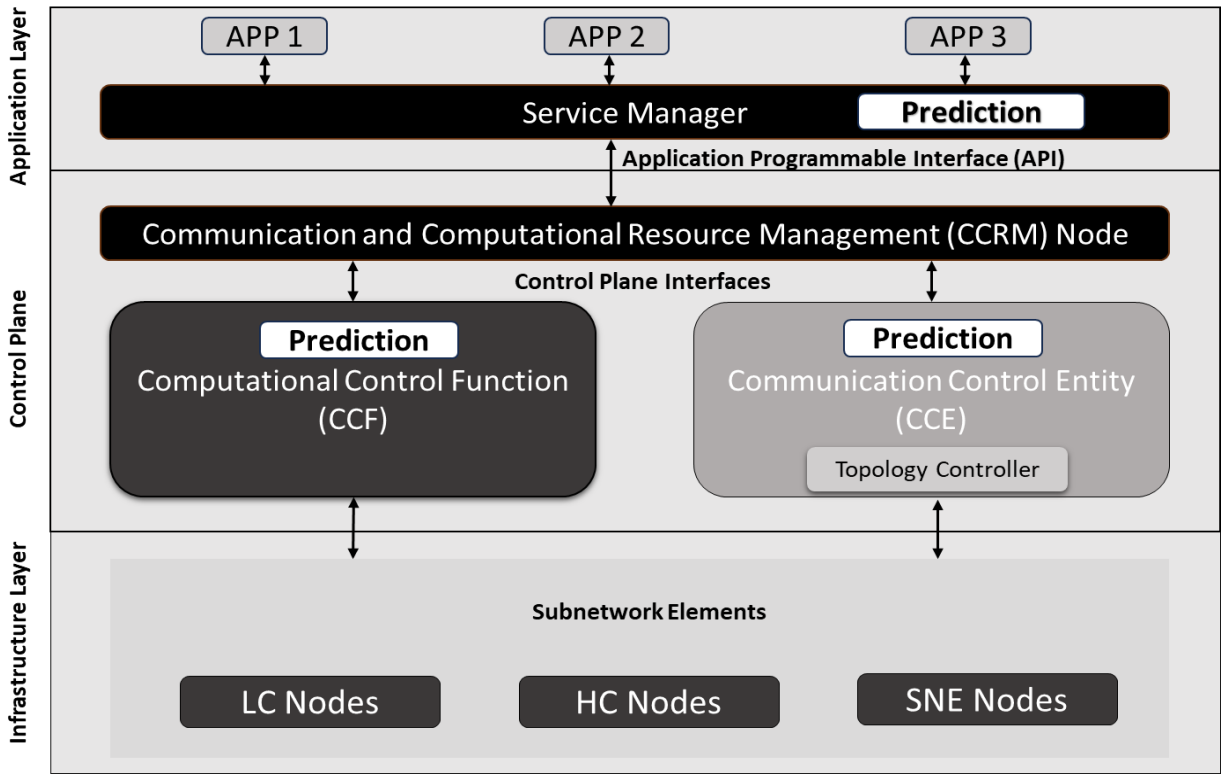

Figure 30 Intra Subnetwork Framework Architecture

The application layer serves as the point where subnetwork elements (e.g. sensors) submit their requests to access diverse services to the service manager node which could be one LC or HC node. This is represented in Figure 30 with the interaction from the APPx to the Service Manager. These requests may encompass service's QoS requirements, such as the service's maximum delay, maximum jitter, service priority or service lifetime.
The proposed framework considers that the service manager includes a prediction unit. This prediction unit enables the service manager to support Quality of Service (QoS) for both existing and upcoming services over a designated time window. The prediction unit could exploit historical data and utilize advanced AI/ML techniques to forecast available services. By predicting, the service manager can contribute to the proactive adjustment of both computational and communication resource allocation (performed in the control plane layer) to meet anticipated demands, ensuring high-quality service delivery. Anticipating service demands over a time horizon is of high importance for optimizing network performance and guaranteeing dependable service level provisioning even during periods of peak service demands. The prediction unit's forecasts can be adjusted for a designated time window. The size of this designated window could vary depending on different conditions and is closely tied to the level of confidence of the prediction unit.

The control plane layer encompasses three functional components as it is depicted in Figure 30. The Computational Control Function (CCF) node is responsible for gathering information from the computational elements within the subnetwork. The gathered information could include, but is not restricted to, parameters such as the computational elements' maximum processing capacity, processing power, power consumption, operational cost, queue length, and reserved computing resources for various tasks in this computational element.
The CCF also utilizes the capabilities of a prediction unit to forecast as depicted in Figure 30, e.g., availability of computing capacity and queue length for the computational elements within a designated time window. As was indicated for the prediction unit in the application layer, the size of this designated window could vary and is closely tied to the level of confidence of the prediction unit. Examples of inputs that this prediction unit could take include information related to services, computational elements, and network topology.
The Communications Control Entity (CCE) is tasked with gathering communication network data. This encompasses various information such as the availability of communication resources and achievable data rate. The proposed framework also considers the availability of a prediction unit that tightly cooperate with the CCE by providing forecasts about the communication networks over a designated time window as depicted in Figure 30. An important design aspect is that the prediction horizon or window size of the prediction units in CCF and CCE need to be equivalent. This prediction unit requires as input, information about services, the communication networks status and network topology.
The CCE also includes a topology controller node which collects information about existing connections between subnetwork elements. The topology controller's task is to discover the network's topology and monitor available routes between different elements (including computational elements).
Finally, a key component in the control plane layer is the Subnetwork Communication Computational Resource Management (Sub-CCRM) management node which is tasked with the simultaneous allocation of communication and computing resources to satisfy the services requirements collected from the application layer. To this aim, the Sub-CCRM management node interfaces with several other functional components of the control plane layer and application layer as it is depicted in Figure 30. The Sub-CCRM node utilizes services' characteristics and QoS requirements exposed by the application layer, as well as computing and communication resource availability (including forecasts) made available from the CCF

and CCE. In general, the information made available to the Sub-CCRM node could be exploited to dynamically plan the task offloading decisions while accounting for service requirements, communication and computing resources availability and network topology changes.
The infrastructure layer includes the physical elements of the subnetwork such as HC, LC, SNE and the(ir) computing elements. The infrastructure layer interfaces with the CCF and CCE for the communication and computational resources monitoring.

3.3.3.2 End-to-End Subnetwork-6G Parent Network Framework Architecture

In this section, the framework is extended to include collaboration among subnetworks and the 6G parent network (Figure 30). In this collaborative framework, two different scenarios arise: in the first, tasks are offloaded from subnetworks to the computational resources of the 6G parent network; in the second, tasks are offloaded from the 6G parent network to subnetworks. However, focus would primarily be on the first scenario, considering that subnetworks are entrusted with critical tasks and necessitate local computational resources for execution.

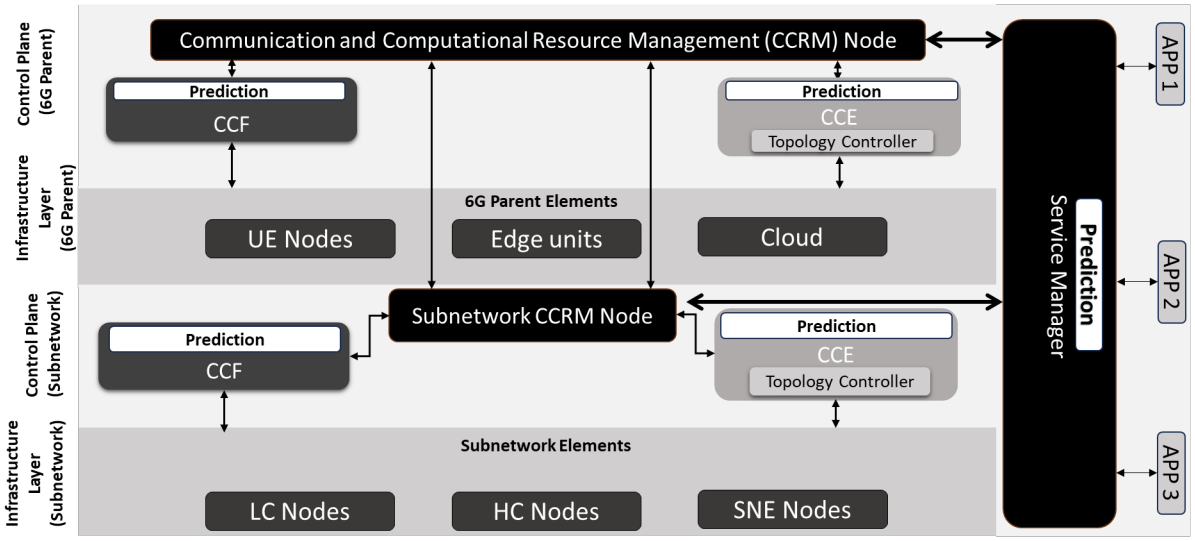

Figure 31 End-to-End Subnetwork-6G Parent Network Architecture

Similar to the framework of subnetworks, the application layer serves as the layer where subnetwork, and 6G parent network elements submit their requests to access diverse services to the service manager node. The service manager classifies services into different groups to transfer to Sub-CCRM management nodes or 6G-CCRM management node. The extended framework considers that the service manager includes a prediction unit. This prediction unit enables the service manager to support QoS for both existing and upcoming subnetworks and 6G parent network services over a designated time window.

Figure 31 depicts the control planes available at the subnetwork and 6G parent network. The CCF node in the control plane of the subnetworks is responsible for gathering information from the computational elements within the subnetwork, while the CCF node in the control plane of the 6G parent network is responsible for gathering information from the computational elements in the 6G parent network, including edge computational elements and cloud. Like the CCF node in the control plane of the subnetworks, the CCF in the control plane of the 6G parent network utilizes the capability of a prediction

unit to forecast, e.g. the availability of computing capacity within a designated time window for the computational elements in the 6G parent network, respectively.
Likewise, the CCE node in the control plane of the 6G parent network is responsible for gathering information about communication resources in the 6G parent network. A prediction unit tightly cooperates with the CCE node by providing forecasts about the communication resources in the 6G parent network over a designated time window.
The CCE node in the control plane of the 6G parent network also includes a topology controller node which is responsible for collecting information about existing connections between 6G parent network elements. The information collected by the topology controller is then made available to the 6G-CCRM node.
Finally, the Sub-CCRM and 6G-CCRM management nodes are tasked to allocate communication and computing resources within subnetworks and the 6G parent network to satisfy the service requirements collected from the application layer. The CCRM node of each layer is connected to the corresponding CCF, and CCE. The subnetwork and 6G parent network are interconnected via their respective CCRM management nodes as it is depicted in Figure 31. The CCRM nodes can exploit this interface for exchanging curated information about the subnetwork and 6G parent network topology and availability of computing and communication resources.

The infrastructure layer of the 6G parent network includes the physical elements of 6G parent network. The infrastructure layer interfaces with the corresponding control plane layer's CCE and its topology controller to obtain topology information. Additionally, infrastructure layer is connected to the corresponding CCF and CCE elements for monitoring communication and computational resources.

3.3.4 Example Deployment Architecture for the Joint Task and Communications Scheduling

Figure 32 illustrates an example of the end-to-end subnetwork – 6G parent network deployment architecture for the joint task and communications scheduling. The architecture represented in Figure 32 follows the 6G-SHINE nomenclature defined in [1]. Each subnetwork comprises of HC, LC, and SNE nodes. The integration with the 6G parent network is facilitated by the gateway functionality implemented by a HC or LC subnetwork element.

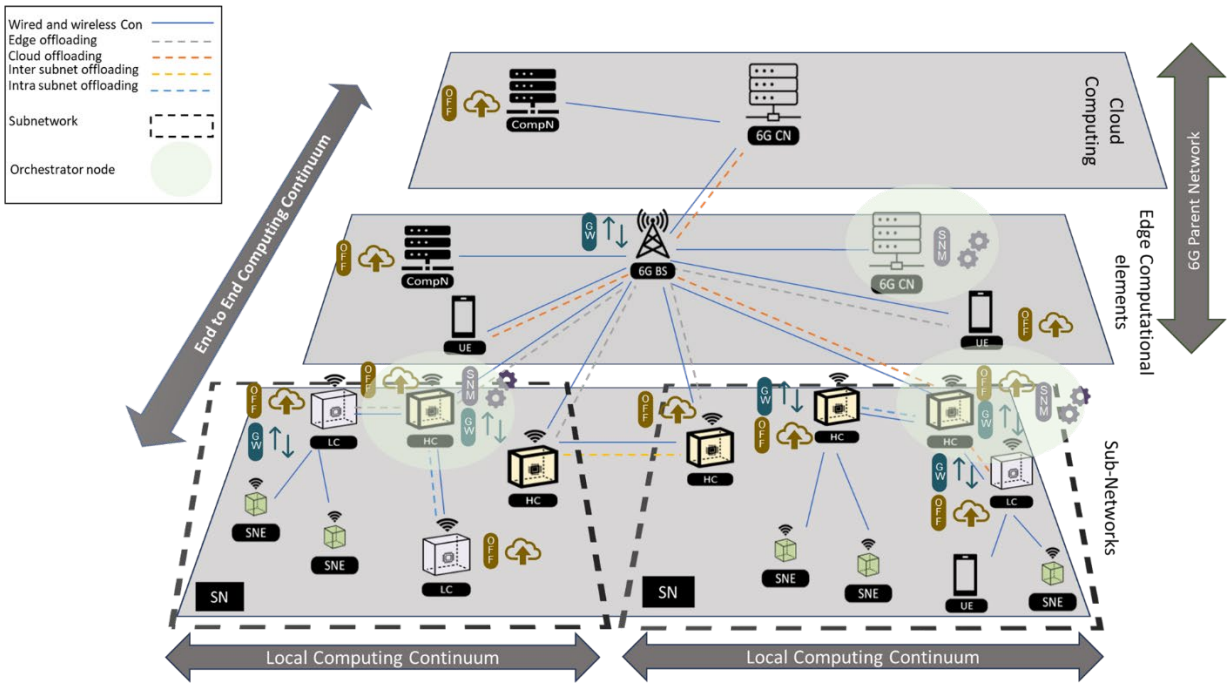

Figure 32 Schematic of deployment architecture in the end-to-end subnetwork-6G parent network scenario

Figure 32 depicts various task offloading scenarios, including intra-subnetwork offloading to HC elements (blue connections), offloading to edge units (grey connections), and offloading to cloud (red connections). Additionally, inter-subnetwork offloading (yellow connections) could be enabled with the assistance of the 6G- CCRM node.

A possible mapping of the roles and functionalities of the framework presented in section 3.3.3.2 to the network elements in a subnetwork and 6G parent network is indicated. Within the subnetwork, both LC and HC nodes can serve as CCE and CCF. For the CCF, various elements within subnetworks, such as HC, and LC nodes, could contribute computational resources. The computational resources of the 6G parent network, such as edge computational elements and cloud, are also exposed and made available to the CCF in the subnetwork. This requires a logical interface between the CCF in the subnetwork and the CCF in the 6G parent network. The Sub-CCRM role may also be adopted by the HC within the subnetwork. Thus, there may be scenarios in which a single subnetwork element such as the HC could implement the CCE, CCF and Sub-CCRM roles.

In the 6G parent network, the CCF, CCE, and 6G-CCRM roles could be considered as part of the core network functions. Then, other 6G core network functions can interface and provide information of interest to them. For example, for the CCF, different elements within the 6G parent network, such as UE, edge units, and cloud, could expose their computational resources to make them available.

Collaboration between different CCRM nodes within subnetworks and 6G parent network enables efficient resource allocation and task management across the network. This collaboration facilitates the pooling of computational resources across subnetworks, edge units, and cloud infrastructure to support diverse computing tasks, leading to the concept of an end-to-end computing continuum.

3.3.5 SYSTEM MODEL

In this section, the system model for facilitating service-aware joint planning and allocation of communication and computational resources is described. Concentration is placed on the definition of

characteristics of services and tasks, and the relationship between the characteristics of each task and its corresponding service. Next, the overall computational elements characteristics for the computational aspect and the overall link characteristics to support communication are presented.

3.3.5.1 Service Characteristics

The system model considers scenarios in which a set of services $\{s_1, s_2, \ldots, s_M\}$ may be received by the service manager. Each service can be defined by a unique set of characteristics as follows:

$$S = \{s_1, s_2, \ldots, s_M\} \quad (1)$$

$$s_m = \{c_m, c_m(t, t+\Delta t), v_m, v_m(t, t+\Delta t), \acute{v}_m(t, t+\Delta t), p_m, t_m, T_m^{max}, \Delta_m^{max}, \gamma_m^{min}, E_m^{max}\} \quad (2)$$

c_m: Computational demand. This reflects the number of computation units required to complete each service.
$c_m(t, t+\Delta t)$: Computational demand within the designated time window. This reflects the prediction of number of computation units required to complete each service within the designated time window.
v_m: Service volume. This indicates the amount of data to be transmitted for each service, measured in bytes.
$v_m(t, t+\Delta t)$: Service volume within the designated time window. Signifying the prediction of volume of data associated with each service within the designated time window, typically measured in bytes.
$\acute{v}_m(t, t+\Delta t)$: Service volume after processing within the designated time window. Signifying the prediction of volume of processed data associated with each service within the designated time window, typically measured in bytes.

p_m: Priority of service. This enumerates the relative importance or urgency of each service within the system's hierarchy, guiding resource allocation and task scheduling. Larger numbers indicate higher priority.
t_m: Lifetime of service. This characteristic defines the expected duration for the service's existence.
T_m^{max}: Maximum latency. This denotes the upper limit of acceptable delay for each service to meet performance expectations.
Δ_m^{max}: Maximum jitter. This describes the allowed variation in latency for each service, ensuring consistent and reliable performance across diverse network conditions.
γ_m^{min}: Minimum reliability. This establishes the minimal reliability threshold that each service must meet, ensuring consistent service availability and safeguarding against disruptions.
E_m^{max}: Maximum energy consumption. This provides the maximum acceptable energy consumption for each service.

3.3.5.2 Task Characteristics

It is considered that services can be partitioned into a set of distinct tasks $\{f_1, f_2, \ldots, f_{I_m}\}$, each characterized by its own characteristics. The tasks' characteristics include:

$$s_m = [f_1, f_2, \ldots, f_{I_m}] \quad (3)$$

$$f_{im} = \{c_{im}, c_{im}(t, t+\Delta t), v_{im}, v_{im}(t, t+\Delta t), \acute{v}_{im}(t, t+\Delta t), p_{im}, t_{im}, T_{im}^{max}, \Delta_{im}^{max}, \gamma_{im}^{min}, E_{im}^{max}\} \quad (4)$$

c_{im}: Computational demand. Reflecting the computational resources required to execute each task.
$c_{im}(t, t+\Delta t)$: Computational demand within the designated time window. This reflects the prediction of number of computation units required to complete each task within the designated time window.
v_{im}: Task volume. Signifying the volume of data associated with each task, typically measured in bytes.
$v_{im}(t, t+\Delta t)$: Task volume within the designated time window. Signifying the prediction of volume of data associated with each task within the designated time window, typically measured in bytes.
$\acute{v}_{im}(t, t+\Delta t)$: Task volume after processing within the designated time window. Signifying the prediction of volume of processed data associated with each task within the designated time window, typically measured in bytes.
p_{im}: Priority of task. Indicating the relative importance or urgency of each task within its corresponding service. Larger numbers indicate higher priority.
t_{im}: Lifetime of task. This characteristic defines the expected duration for the task's existence.
T_{im}^{max}: Maximum latency. Denoting the maximum allowable delay for each task, crucial for meeting performance expectations.
Δ_{im}^{max}: Maximum jitter. Describing the acceptable variation in latency for each task, ensuring consistent performance.
γ_{im}^{min}: Minimum reliability. Establishing the minimum reliability threshold that each task within a service must meet to ensure uninterrupted service availability.
E_{im}^{max}: Maximum energy consumption. This provides the maximum acceptable energy consumption for each task.

3.3.5.3 Relationship of Service and Tasks Characteristics

The relationship between the computational demand of each task and its corresponding service can be characterized by

$$c_{im} = \alpha_{im} c_m \quad (5)$$

where c_m is the computational demand of service m, c_{im} is the computational demand of task i of service m, and α_{im} describes their relationship. Subject to:

$$0 \leq \alpha_{im} \leq 1 \quad (6)$$

The sum of the computational demand of all tasks within a service should equate to the computational demand of that service:

$$\sum_i \alpha_{im} = 1 \quad (7)$$

Similarly, the relationship between the volume of each task and its corresponding service can be expressed as:

$$b_{im} = \beta_{im} b_m \quad (8)$$

where b_m is the volume of service m, b_{im} is the volume of task i of service m, and β_{im} describes their relationship. Subject to:

$$0 \leq \beta_{im} \leq 1 \quad (9)$$

The cumulative volume of all tasks within a service should correspond to the total volume of that service:

$$\sum_i \beta_{im} = 1 \quad (10)$$

Considering the lifetime and the maximum latency of each task and its corresponding service, various scenarios, including parallel or sequential task management, are analysed. In parallel task management scenarios, the relationship between the lifetime and the maximum latency of each task and its corresponding service are explored as

$$max(t_{im}) \leq t_m \quad (11)$$

$$max(T_{im}^{max}) \leq \mathrm{T}_m^{max} \quad (12)$$

and for sequential scenarios they are expressed by

$$\sum_i t_{im} \leq t_m \quad (13)$$

$$\sum_i T_{im}^{max} \leq \mathrm{T}_m^{max} \quad (14)$$

Task requirements in terms of maximum jitter and minimum reliability are constrained by the requirements of corresponding service and the relationship between them are expressed by

$$max(\Delta_{im}^{max}) \leq \Delta_m^{max} \quad (15)$$

$$min(\gamma_{im}^{min}) \geq \gamma_m^{min} \quad (16)$$

By quantifying these relationships through mathematical formulas, valuable insights into the interplay between task characteristics and service requirements are gained.

3.3.5.4 Computational Elements Characteristics

From the perspective of computational elements within the system model, resources are categorized into three groups. Local resources including J units:

$$U^L = \{u_1^L, u_2^L, \ldots, u_J^L\} \quad (17)$$

edge computing resources including Q units:

$$U^E = \{u_1^E, u_2^E, \ldots, u_Q^E\} \quad (18)$$

and a single cloud computing resource: U^C.

The current system modelling characterizes each category of computational elements include local, edge and cloud, i.e., $\{u_j^L, u_q^E, u^C\}$ with the following parameters. Local computational elements, edge computational elements and cloud computing centre characteristics can be expressed as follows:

$$u_j^L = \{C_j^{min}, C_j^{max}, C_j(t), C_j(t, t+\Delta t), E_j^{max}, F_j\} \quad (19)$$

$$u_q^E = \{C_q^{min}, C_q^{max}, C_q(t), C_q(t, t+\Delta t), E_q^{max}, F_q\} \quad (20)$$

$$u^C = \{C_c^{min}, C_c^{max}, C_c(t), C_c(t, t+\Delta t), E_c^{max}, F_c\} \quad (21)$$

C_j^{min}: Minimum Computational Usage. This parameter signifies the minimum level of computational activity expected from each computational element. It ensures that computational elements remain active and engaged, preventing instances of idleness to ensure balanced workload distribution.
C_j^{max}: Maximum Computational Capacity. Reflecting the upper limit of computational cores that each computational element can provide. It defines the peak capability of each computational element.
$C_j(t)$: Instantaneous Computational Capacity. Reflecting the number of processing cores that each computational element can provide at this moment.
$C_j(t, t+\Delta t)$: Computational capacity prediction by the prediction unit within the designated time window. The framework uses the prediction unit connected to the CCF to forecast the computational capacities of all computational elements over an upcoming time window. This approach enables effective resource allocation and task distribution based on computational capacity predictions.
F_j: Processing Rate. The processing rate indicates the rate of each computational element in managing computing tasks within a specified duration, typically measured in units per second. It provides insights into the unit's ability to handle workload efficiently.
E_j^{max}: Energy Level. Signifying the maximum energy capacity of each computational element, influencing its operational capabilities and longevity.

3.3.5.5 Communication Links Characteristics

The system modelling also covers the communication links between different elements utilizing various characteristics such as:

$$L_l^{(k)} = \left\{R_l^{(k)}, r_l^{(k)}(t), r_l^{(k)}(t, t+\Delta t), T_l^{(k)}, E_l^{(k)}, P_l^{(k)}\right\} \quad (22)$$

$R_l^{(k)}$: Capacity of the link. This parameter denotes the peak data transfer capability of the link l by using communication resource k.
$r_l^{(k)}(t)$: Instantaneous achievable data rate: Data rate of the link l by using communication resource k in this moment.
$r_l^{(k)}(t, t+\Delta t)$: Achievable data rate within the designated time window. The framework, working in coordination with the prediction unit connected to CCE, forecasts data rate of the link l by using communication resource k over a specified duration.

$T_l^{(k)}$: Latency of the link. Measuring the average delay experienced during data transmission over the link l by using communication resource k, this metric is important for assessing the responsiveness of the communication link and ensuring timely delivery of information.

$E_l^{(k)}$: Energy consumption for the utilization of a link. Quantifying the energy expended during data transmission over the link l by using communication resource k. This represents the power multiplied by the time it takes to send data over the link.

$P_l^{(k)}$: Transmission power on the link. This feature signifies the level of transmitting power utilized for sending information over the link l by using communication resource k, impacting the signal strength and reliability of data transmission.

3.3.6 PROBLEM DEFINITION

Based on the proposed system model (Section 3.3.5), three objective functions aimed at optimizing the joint task offloading and communications scheduling, which are executed to support the services, are introduced. The three defined objective functions seek to effectively support the task offloading process considering strict service level requirements from deterministic services, i.e., latency and jitter and reliability. In particular, a first objective function has been defined that focuses on guaranteeing that services are completed within a bounded delay with low jitter, independently of where the services' tasks are executed across the local or end-to-end continuum. The second objective function, which focuses on achieving a balanced distribution of the workload among different computational elements when they have the same efficiency of communication resource utilization, and to offload more workload to the computational elements with better efficiency of communication resource utilization for the distribution of the services' tasks across the local or end-to-end continuum. The third objective function seeks to maximize the services reliability by ensuring service delay remains below the upper band delay. Further details about these optimization functions are presented in the following subsection.

3.3.6.1 Latency and Jitter Optimization Objective Function

This first objective function focuses on satisfying the services' latency and jitter requirements. One common approach in the literature is to minimize the services' latency. However, this approach might result in unnecessarily high utilization of communication and computational resources and system load when the services can tolerate higher delays. Relaxing the service's latency to near the maximum latency requirement eliminates the option for performing countermeasures, such as retransmissions or task redistribution. At the same time, many critical (deterministic) services benefit from maintaining low jitter. These three aspects have been addressed in the definition of the objective function, as graphically represented in Figure 33. This cost function $(K(x))$ penalizes latencies that are close to zero or close to the upper bound, aiming to maintain latency within an acceptable range. The cost function is defined as a truncated normal distribution between zero and the upper bound of latency. It assigns infinite value to latencies exceeding the upper bound, emphasizing the high penalty for exceeding latency constraints. The variance of the cost function (σ) can be adjusted to control jitter. Lower variance values result in lower jitter, and vice versa. One critical aspect in the design of the objective function is setting the value of T', which represents the target latency value that must be within the service latency limit.

$$min \sum_m \sum_i K\left(\frac{T_{im}}{T^{max}}\right) \quad (23)$$

$$K(x) = \begin{cases} -Ae^{-\frac{(x-T')}{2\sigma^2}} + M_2 & 0 \le x \le 1 \\ +\infty & x \ge 1 \end{cases} \quad (24)$$

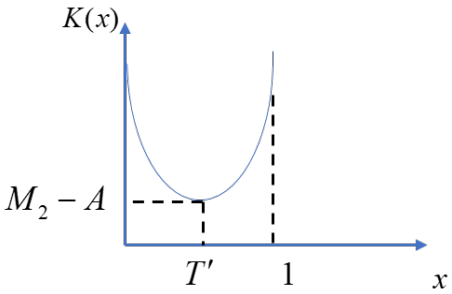

Figure 33 Latency and jitter optimization cost function

3.3.6.2 Distribution of Workload Based on Efficiency of Communication Resource Utilization

The second objective function aims to achieve a balanced distribution of the workload among different computational elements when they have the same efficiency of communication resource utilization, and to offload more workload to the computational elements with better efficiency of communication resource utilization for the distribution of the services' tasks across the local or end-to-end continuum, focusing on three key aspects. Firstly, ensuring that allocated resources for each task are sufficient to maintain its desired QoS. Secondly, considering the priority of tasks to allocate communication resources for task offloading. Thirdly, considering the distance of computational elements from the point where the task is generated. Thus, the following objective function is defined:

$$max\ \Phi = \prod_{j=1}^{J} \left(\frac{W_j}{CP_j}\right)^{\psi_j} \times \prod_{q=1}^{Q} \left(\frac{W_q}{CP_q}\right)^{\psi_q} \times \left(\frac{W_c}{CP_c}\right)^{\psi_c} \quad (25)$$

In this equation for $x \in \{j, q, c\}$, W_x is the workload of each computational element, CP_x is the processing power of each computational element, and $\psi_x$is defined as:

$$\psi_x = \sum_{m=1}^{M} \sum_{i=1}^{I_m} \frac{d}{p_{im}} a_{im}^{(x)} \sum_{l_{im}=1}^{L_{im}} \frac{\sum_{k=1}^{K} b_{l_{im}}^{(k)} \times r_{l_{im}}^{(k)}(t, t+\Delta t)}{\sum_{k=1}^{K} b_{l_{im}}^{(k)}} \quad (26)$$

In this equation, p_{im} denotes the priority of each task, and the coefficients $d \in \{d_l, d_e, d_c\}$ within the objective function illustrate the significance of the efficiency of communications resource utilization for task offloading to local computational elements, edge computational elements, and cloud, respectively. Considering that offloading tasks to edge computational elements and the cloud typically congest more communication resources, the coefficients are chosen as follows:

$$d_l \ge d_e \ge d_c \quad (27)$$

In addition, $a_{im}^{(x)} = 1$ when task i of service m is offloading to the computational element x, and $b_{l_{im}}^{(k)} = 1$ when communication resource k is allocated to link l_{im} for offloading task i of service m, and $r_{l_{im}}^{(k)}(t, t+\Delta t)$ is the predicted achievable data rate by using communication resource k in link l_{im}. L_{im}

shows the number of links in the route for offloading task i of service m, I_m is the number of tasks for completing service m, M is the number of services, and k is the number of communication resources.

3.3.6.3 Maximizing Reliability of Services

The third objective function aims to maximize the reliability of services. Acknowledging the importance of deterministic service provisioning, the objective function could be defined as:

$$max \sum_m \sum_i p_{im}\, \theta(f_{im}) \tag{28}$$

where p_{im}denotes the priority of each task, and one potential objective function for considering service reliability can be expressed as follows:

$$\theta(f_{im}) = -max\left(0, Y\left(\gamma_{im}^{min} - \gamma_{im}\right)\right) \tag{29}$$

where $\gamma_{im} = Prob\{T_{im} \leq T_{max}\}$. In this equation, Y is a large number, and γ_{im}^{min} is the minimum acceptable reliability of task i in service m, and γ_{im} is the reliability of task i in service m.

3.3.6.4 Mixing Different Objectives

Finally, a general objective function is proposed to achieve a balance and optimize all objective functions:

$$max\, \lambda_1 \Phi + \lambda_2 \sum_m \sum_i p_{im}\, \theta(f_{im}) - \lambda_3 \sum_m \sum_i K\left(\frac{T_{im}}{T^{max}}\right) \tag{30}$$

By adjusting and altering the values of the coefficients include $\{\lambda_1, \lambda_2, \lambda_3\}$ within the objective function, the significance of each term can be adjusted to address specific performance requirements and priorities. This flexibility enables fine-tuning of the optimization process to enhance joint task offloading and communication scheduling in support of services.

3.3.6.5 Constraints

Below, the constraints for solving the optimization problems defined above are detailed.

- Task scheduling: Every task is only assigned to one computational element.

$$\sum_{j=1}^{J} a_{im}^{(j)} + \sum_{q=1}^{Q} a_{im}^{(q)} + a_{im}^{(c)} = 1 \tag{31}$$

  where $a_{im}^{(x)} = 1$ when task i of service m is offloading to the computational element x for $x \in \{j, q, c\}$.

- Communication resource scheduling: Every communication resource is only assigned to one communication link.

$$\sum_{m=1}^{M} \sum_{i=1}^{I_m} \sum_{l_{im}=1}^{L_{im}} b_{l_{im}}^{(k)} = 1 \tag{32}$$

where $b_{l_{im}}^{(k)} = 1$ when communication resource k is allocated to link l_{im} for offloading task i of service m, L_{im} shows the number of links in the route for offloading task i of service m, I_m is the number of tasks for completing service m, M is the number of services.

- Latency limit: There should be an upper bound on the latency for each task to ensure timely execution.

$$T_{im} \leq T_{im}^{max} \tag{33}$$

where T_{im} is the delay for execution of task i of service m, and T_{im}^{max} is the maximum allowable delay for task i of service m.

- Latency limit: In addition, the latency for executing each task should be less than the designated time window of prediction units.

$$T_{im} \leq \Delta t \tag{34}$$

where Δt is the length of designated time window for prediction units.

- Data transfer rate: The rate of data transfer on each link must not exceed the capacity of that link.

$$\sum_{m=1}^{M} \sum_{i=1}^{I_m} r_{l_{im}}^{(k)}(t, t+\Delta t) b_{l_{im}}^{(k)} \leq R_{l_{im}}^{(k)} \tag{35}$$

where $r_{l_{im}}^{(k)}(t, t+\Delta t)$ is the achievable data rate by using communication resource k in link l_{im}, and I_m is the number of tasks for completing service m, M is the number of services, $R_{l_{im}}^{(k)}$ denotes the peak data transfer capability by using communication resource k in link l_{im}.

- Maximum computational capacity: Each computational element has a maximum capacity limit for computational tasks.

$$\frac{1}{\Delta t} \sum_{m=1}^{M} \sum_{i=1}^{I_m} c_{im}(t, t+\Delta t) a_{im}^{(x)} \leq C_x^{max} \times F_x \tag{36}$$

where $x \in \{j, q, c\}$, $c_{im}(t, t+\Delta t)$ is the computational demand within the designated time window for task i of service m, C_x^{max} is the maximum computational capacity of computational element x, F_x is the processing rate of computational element x, and Δt is the length of designated time window for prediction units.

- Minimum usage: There is a minimum usage requirement for each computational element to ensure balanced resource utilization.

$$C_x^{min} \times F_x \leq \frac{1}{\Delta t} \sum_{m=1}^{M} \sum_{i=1}^{I_m} c_{im}(t, t+\Delta t) a_{im}^{(x)} \tag{37}$$

where $x \in \{j, q, c\}$, $c_{im}(t, t+\Delta t)$ is computational demand within the designated time window for task i of service m, C_x^{min} is the minimum computational usage of computational element x, F_x is the processing rate of computational element x, and Δt is the length of designated time window for prediction units.

- Maximum energy capacity: Each computational element has a maximum energy capacity to ensure sustainable operation.

$$\sum_{m=1}^{M} \sum_{i=1}^{I_m} E_{im,c}\, a_{im}^{(x)} \leq E_x^{max} \tag{38}$$

where $x \in \{j, q, c\}$, $E_{im,c}$ is the energy consumption for computation of task i of service m, and E_x^{max} is the maximum energy capacity of computational element x.

- Latency of transmission: The latency of transmitting a computing task is equivalent to the latency experienced across the various links within a given path.
- Energy consumption of transmission: The energy consumption for transmitting a computing task is equal to the total energy consumption across the links within a given path.

3.3.7 NEXT STEPS

One of the main enablers of the proposed framework is using prediction units. The prediction units enable efficient service management, forecast availability of computing capacity in the computational elements and the communication resources. These units leverage historical data and deploy sophisticated AI/ML techniques to make precise predictions. A significant future step involves the adoption of advanced AI/ML methodologies within these prediction units to enhance the confidence in their forecasts. Additionally, the framework's CCRM node leverages the characteristics of services along with the availability of computing and communication resources, and information provided by the topology controller to facilitate task offloading decisions. Developing sophisticated policies for the joint management of communication and computational resources within the CCRM node is a critical area for future focus. AI/ML algorithms will allow the CCRM node to dynamically plan task offloading decisions, taking into account the changes in requirements of the services, the availability of communication and computing resources, and network topology. Offloading computational tasks across the end-to-end computing continuum offers a flexible approach to meet diverse service requirements. Effective task offloading for supporting deterministic service level provisioning, necessitates seamless integration and strategic planning of both communication and computing technologies, as outlined in the proposed framework. As a next step, we plan to implement the proposed framework to achieve deterministic service provisioning in specific use cases within the subnetwork's ecosystem, including in-vehicle networks and Electrical/Electronic (E/E) architecture.

4 DYNAMIC SPECTRUM SHARING BETWEEN 6G AND IN-X SUBNETWORK

4.1 INTRODUCTION

The realm of 6G in-X subnetworks, as extensively detailed in [10], mostly addresses applications and capabilities within the Ultra-Reliable Low Latency Communications context. Key examples include closed-loop control systems within factory and automotive environments, characterized by millisecond cycle times. A significant challenge for these wireless applications lies in adhering to the stringent latency requirements while maintaining extremely high reliability.

A crucial aspect in this domain is spectrum usage for in-X subnetworks. The dilemma of choosing between unlicensed and licensed bands presents distinct challenges. Unlicensed bands are fraught with coexistence complexities and regulatory constraints. In contrast, using licensed bands entails considerable cost implications, often determined through market-based auctions that vary from country to country. This raises a pertinent question: Is the additional financial burden of licensed spectrum justified, particularly for in-X subnetworks applications like cabling replacements, which are inherently low-cost?

In this context, there is potential for developing in-X subnetworks either by optimizing existing technologies (e.g., Wi-Fi, Bluetooth or Sidelink) to meet the envisioned requirements or by innovating new technologies tailored for specific subnetwork services. Each pathway brings its own set of regulatory implications, influenced by regional variations. An illustrative example of a commercial wireless solution akin to in-X subnetwork applications is ABB's Wireless Sensor-Actuator Network (WSAN)-FA. It leverages the 802.15.1 PHY-layer [53] with an enhanced MAC layer to improve latency.

Subsequent sections will delve into the challenges associated with using licensed or unlicensed spectrum, the implications for in-X subnetworks and potential fundamental regulatory constraints that might impact SN operation.

4.2 SPECTRUM SHARING CHALLENGES

4.2.1 Challenges for Licensed-based Subnetworks

Within homogeneous networks, devices primarily operate within licensed bands. Despite notable advancements from the 5G era, several intricate challenges may hinder the effective deployment of spectrum sharing algorithms in in-X subnetworks. These networks, anticipated to be highly dense, often comprise numerous ultra-small cells within a single macro cell, all sharing the same frequency bands [42]. Furthermore, the dynamic nature of in-X subnetworks, particularly evident in automotive contexts, demands a versatile approach to data transmission. The spectrum sharing strategy must accommodate varying data sizes, ranging from small status updates to large-scale video or radar streams, to ensure reliable transmission.

The evolution toward 6G subnetworks may witness the continuation of Dynamic Spectrum Sharing (DSS)-like techniques developed during the 5G era. However, relying solely on the sharing of existing bands may prove inadequate for delivering the expected 6G services. Additional bands with adequate bandwidth might be necessitated. The mobile industry's drive towards higher frequencies necessitates long-term investment and could encounter challenges due to potential market demand shortages. Spectrum harmonization [54] stands as a foundational element to instil confidence in future markets, providing visibility for long-term investments, economies of scale, efficient spectrum utilization, and improved cross-border conditions.

Another challenge revolves around the feasibility and cost-effectiveness of specific in-X subnetwork applications. Edge devices, such as Distributed Units and Central Units, traditionally defined in the 5G paradigm, are equipped to execute complex algorithms and relay timely feedback via fronthaul networks. However, the intricate processes involved in spectrum sharing—encompassing sensing, inference, and scheduling—can introduce excessive costs and complexity, potentially rendering some subnetwork-oriented applications unfeasible, particularly where energy and price constraints play significant roles. Additionally, the overall architecture significantly impacts costs and performance. If all subnetwork elements are connected to the 6G network, technology licensing, radio access technology performance, scalability, and security become critical factors. Conversely, if only the subnetwork controller is connected, this can influence the total subnetwork price and efficiency.

While edge devices are equipped to handle such tasks, the timely implementation of policies is impeded by the complexity of spectrum sharing processes [37]. Potential solutions may include advancements in chip technology, innovative fronthaul architectures, streamlined sharing algorithms, or the utilization of AI-based predictive models [37]. Nevertheless, the feasibility and cost-effectiveness of these solutions for specific in-X subnetwork applications remain uncertain.

Furthermore, the dynamic nature of in-X subnetworks underscores the necessity for a flexible approach to data transmission, accommodating a wide range of data sizes. This aspect, coupled with the inherent short-range and low-power characteristics of in-X subnetworks, remains relatively underexplored in existing literature. Leveraging these attributes could potentially minimize interference between co-channel subnetworks, presenting an area ripe for research and exploration, as highlighted in Task 2.2 of the 6G-SHINE project.

4.2.2 Challenges for Unlicensed-based Subnetworks

In scenarios where unlicensed spectrum is utilized, subnetworks might often encounter significant challenges due to the decentralized nature of spectrum information. The absence of centralized coordination hinders their ability to access comprehensive spectrum data, complicating the development of effective strategies for radio resource management [37]. Interference from other SNs or external sources such as Wi-Fi or Bluetooth Access Points (Aps) presents significant research-worthy challenges [37]. These issues, addressed in Work Packages 4 and 5, underscore the need for centralized management or enhanced coordination among SNs and APs. From a legal standpoint, the utilization of unlicensed spectrum is anchored in medium access policies like the Listen-Before-Talk (LBT) mechanism. This policy facilitates the coexistence of devices by allowing them to check the channel before transmitting, making it suitable for delay-tolerant traffic. However, it falls short for applications requiring low or bounded latency. Consumer subnetwork use cases may need latency as low as 5 to 10 milliseconds, while industrial and automotive scenarios demand even more stringent requirements, with latencies as low as 100 microseconds. Consequently, the LBT mechanism, although effective for general traffic management, is not ideal for these high-performance applications.

However, the adoption of LBT introduces additional communication overhead and poses challenges in collaborative algorithm design, particularly due to SNs' reluctance from different providers to share information, driven by security and commercial concerns. Moreover, the dynamic nature of network topology in unlicensed bands, characterized by devices joining or leaving abruptly, undermines the efficacy of meticulously crafted sharing policies [39]. Despite the appeal of online learning algorithms, they are hindered by high computational demands and power consumption, making predicting, and managing spectrum allocation increasingly challenging in such environments.

Multiple companies suggest not to mandate LBT procedure, but to provide the designs for where they are needed by regulation, or if useful, for performance enhancements [55]. Consequently, 3GPP may

support both channel access mechanisms with and without LBT [55]. Various techniques are being considered to enhance LBT, including directional sensing or beam-based LBT, receiver-assisted LBT and the adaptation of sensing thresholds. Additionally, alternative coexistence mechanisms beyond LBT are proposed, such as Automatic Transmit Power Control (ATPC) [55], which operates autonomously based on good neighbour behaviour, and measurement/long-term sensing-based solutions like Dynamic Frequency Selection (DFS) and duty cycling. Furthermore, a mechanism for switching in and out of LBT mode may be adopted.

For the unlicensed 60 GHz band, 3GPP has adopted regulations defined by the ETSI Harmonized European Standard (EN) 302 567 as the baseline LBT mechanism. Enhancements such as adjustments to energy detection thresholds and contention window sizes based on sensing bandwidth may be considered. To facilitate harmonious coexistence in the 60 GHz band, 3GPP may support a channelization mode aligned with Wi-Gig (IEEE 802.11 ad) channels, with a bandwidth of 2.16 GHz, although smaller bandwidths may also be accommodated. Similar to NR-U operation below 52.6 GHz, NR-U operation in the 60 GHz band can be standalone, aggregated via carrier aggregation, or in dual connectivity with an anchor carrier.

4.3 Emerging Opportunities for 6G in-X Subnetworks

4.3.1 Terahertz Spectrum

The exploration of the terahertz (THz) spectrum, ranging from 0.1 to 10 THz, presents promising prospects for advancing existing wireless communication systems [40]. As illustrated in Figure 33, this spectrum realm offers opportunities for ultra-high-speed wireless links catering to virtual and augmented reality applications, robust backhaul and access links, and the realization of the Internet of Nano-Things [56]. Given the inherent characteristics of in-X subnetwork applications, which often operate within short ranges, require low power, and depend on line-of-sight connectivity [40], the THz spectrum emerges as a compelling alternative to mitigate frequency scarcity and substantially enhance communication network capacity. Nevertheless, challenges persist, particularly in the complexities of antenna array design and the energy constraints imposed by communication hardware, notably in battery-powered devices.

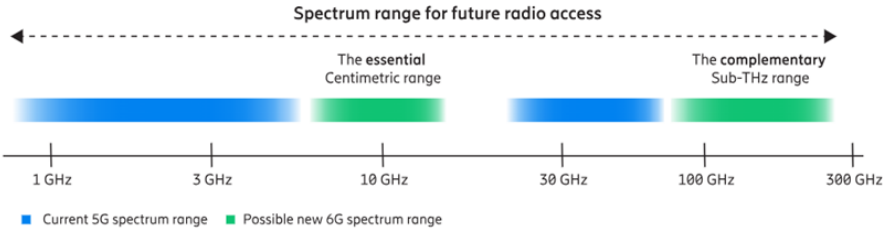

Figure 34 Possible new 6G spectrum range. Adapted from [57].

4.3.2 Blockchain-based Spectrum Sharing

Blockchain technology emerges as a revolutionary solution poised to enhance decentralized spectrum sharing [43]. By fostering trust among multiple wireless network operators, blockchain facilitates real-time spectrum management. Recent endorsements by government regulators, such as the FCC's advocacy for employing blockchain in spectrum market management, underscore its potential. Blockchain's capabilities in real-time processing, chronological data structuring, and prevention of double-spending render it particularly suitable for spectrum trading scenarios, effectively safeguarding against simultaneous spectrum allocation to multiple users and mitigating interference risks. For example, in the context of subnetworks, blockchain can be applied to channel access and resource

allocation by creating a decentralized spectrum ledger where each subnetwork records its spectrum usage and access requests. Smart contracts can automatically allocate spectrum slices based on real-time demand, ensuring no conflicts occur. The blockchain can resolve any access conflicts by referring to the chronological order of requests, granting access to the earliest valid request and mitigating interference by reallocating affected subnetworks. Additionally, regulators and network operators can audit the blockchain ledger to ensure compliance with spectrum policies, enhancing trust and reducing the need for manual oversight. Through these mechanisms, blockchain ensures that each subnetwork dynamically receives an appropriate spectrum slice without conflicts, thereby preventing simultaneous spectrum allocation to multiple users and mitigating interference risks, ultimately enhancing the efficiency and reliability of subnetwork operations in ultra-dense environments.

4.3.3 Dual-Band Design

The dual-band approach strategically exploits unlicensed bands while seamlessly transitioning to licensed bands when interference or resource contention jeopardizes reliable transmission. Initial data transmission involves a meticulous clear channel assessment (CCA) in the unlicensed band, triggering a switch to the licensed band upon detecting significant activity. While recent studies demonstrate the benefits of this approach for URLLC [38], implementing such a dual-mode operation in in-X subnetworks presents feasibility challenges, particularly for applications requiring consistent Physical/Medium Access Control operations.

4.3.4 Big Data Processing

The utilization of detailed radio-service maps (RSMs) presents a sophisticated strategy for unlocking spectrum opportunities within 6G in-X subnetworks [41]. These maps, offering rich contextual information, play a pivotal role in dynamic spectrum management (DSM) within emerging dense and heterogeneous networks. They facilitate the integration and processing of extensive data with high geographical granularity, effectively addressing the inherent complexity and computational demands of such environments.

4.3.5 Application-Oriented Methodologies

In various applications, particularly wireless control systems, adhering strictly to fixed network requirements to accommodate worst-case scenarios often leads to inefficient spectrum utilization [44]. Tailoring spectrum allocation to specific application objectives or system states presents a more efficient approach [45]. For instance, implementing control system management strategies in factory environments could mitigate interference among subnetworks, a subject further explored in Task 4.1.

4.4 Relevant Regulatory Constraints for in-X Subnetworks

As explored earlier, developing in-X subnetworks can involve either optimizing existing technologies (Wi-Fi or Bluetooth) or innovating entirely new ones. Both approaches have distinct regulatory implications that vary across regions. This section delves into the regulatory landscape relevant to in-X subnetworks.

National Regulatory Bodies in most countries manage the radio spectrum, implementing policies based on agreements established by Regional Regulatory Bodies and International Regulatory Bodies. Nations report progress in applying decisions from the International Telecommunication Union and World Radiocommunication Conferences that aim to achieve global harmonization of spectrum usage practices.

Generally, each target market possesses its own regulatory framework for introducing new radio frequency band technologies or generic radio transceivers. This framework includes a dedicated

certification process that heavily impacts chip manufacturers, integrators, and IoT device importers. These players must invest time in understanding and ensuring they comply with local legal requirements for their devices. As illustrated in Figure 35, the process starts with manufacturing a system, integrating existing parts, or importing a product. Each product undergoes testing against a standardized test plan designed by the regulator. The certification step involves analysing test results along with additional technical documentation. Successful completion leads to a label issuance, allowing market access. Additionally, national authorities have a responsibility to perform post-market surveillance and monitoring.

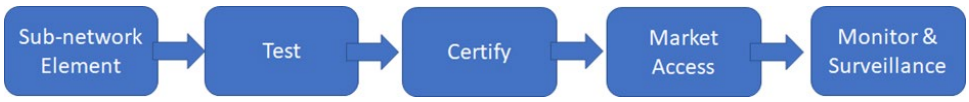

Figure 35 Approval process overview for a new SN radio equipment to gain access to the market.

While the core regulatory process exhibits similarities across the globe, certain regional variations exist. Table 5 provides an example of relevant details for the top GDP countries, showcasing how the certification process differs.

Table 5: Example of PHY Certification across the top GDP countries Worldwide. Adapted from [58].

	US	Europe	China	Japan	India	Brazil	Canada
Reference Standard (test)	47 CFR FCC Rules Part 15 subpart C §15.247	ETSI EN 300 220-2 EN 303 204	SRRC 423	Notification No.88 of MIC ARIB STD - T108	TEC2449:218	Resolution No. 242 Resolution No. 506	RSS-GEN RSS-247
Test Body	Recognized ISO 17025 Lab	Own / Other	Chinese ISO 17025 Lab	Recognized ISO 17025 Lab	Recognized ISO 17025 Lab	Brazilian Recognized ISO 17025 Lab	Recognized ISO 17025 Lab
In-country testing required	No	No	Yes	Notification No.88 of MIC ARIB STD - T108	Yes	Yes	No
Cerification Body	TCB	Own Producer / Notify Body (DoC if HS or NB UE type examination.)	MIIT	RCB	TEC	OCD	CB
Typical Lead Time (Test & Certification)	6 Weeks	4 Weeks	12 Weeks	9 Weeks	9 Weeks	9 Weeks	6 Weeks

Drawing from [58], this section further explores the key technical constraints that can significantly impact subnetwork performance based on local regulations. These constraints play a crucial role in Radio Resource Management (RRM) decisions, especially for ultra-dense deployments:

Frequency bands: This is a critical constraint, particularly considering the availability of new spectrum bands. For instance, it's expected that unlicensed spectrum can offer up to 1.2 GHz in the 6 GHz band, free from existing Wi-Fi devices (Figure 36). More than 70 countries are planning to adopt similar regulations. Regulators could also limit the number of channels used in multi-channel scenarios, which can directly impact, for example, the overall subnetwork performance. Additionally, if a subnetwork element is designed to be low-cost, it makes sense to limit the number of frequency bands it supports to reduce expenses. However, this limitation can restrict the device's ability to operate globally, affecting its versatility and roaming capabilities.

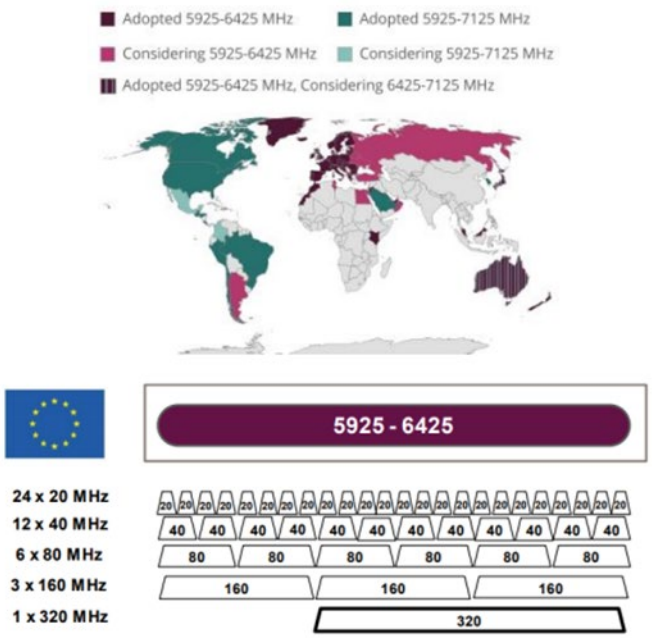

Figure 36 New 6 GHz channels for the United States and European Union shown above provide two to three times the available spectrum.

Maximum Transmit Power Limit: Regulators set a maximum transmit power limit for specific bands, which essentially determines the maximum coverage radius. This parameter varies significantly across countries. For example, sub-GHz IoT applications in the US allow roughly three times more transmit power compared to China and twice that of Japan. Additionally, regulators impose other constraints such as power spectrum density (PSD) and occupied bandwidth (OCB). These parameters further influence the performance and classification of subnetworks, playing a crucial role in defining them as low-power versions of Local Area Networks (LAN) or Personal Area Networks (PAN) according to the ETSI classification for the IoT landscape.

Maximum Spurious Emission Threshold: Regulations exist to ensure transmitters avoid generating spurious emissions that may interfere other users in the frequency domain. Different regions establish varying maximum spurious emission levels, which can directly impact subnetwork PHY layer design, particularly in ultra-dense deployments.

Medium Access Policy: Regulators might enforce specific spectrum access mechanisms like LBT or Adaptive Frequency Agility (AFA). In the case of LBT, regulations could specify a listening time window and the minimum signal strength threshold that differentiates signal from noise (Carrier Sense Level). Additionally, regulations might enforce polite policies such as maximum transmission time and minimum band Tx-OFF time to ensure fair medium access for other transmitters.

Furthermore, regulators can influence the use of specific modulation techniques, channelization (number and width of channels) on channelized bands, and the use of Frequency Hopping Spread Spectrum (FHSS) techniques. Duty cycles for sub-channels within a band can also be regulated, along with channel dwell time and channel duty cycle period. In multi-channel scenarios, regulators might also restrict the total number of channels used (Total Used Bandwidth). Understanding these regulatory constraints is crucial for the successful development and deployment of in-X subnetworks.

4.5 Next Steps

Building upon the challenges outlined for both licensed and unlicensed bands, coupled with the primary technical constraints associated with regulatory standards, our research aims to delve into and assess appropriate spectrum sharing mechanisms. Additionally, we intend to compare potential ultra-dense Subnetwork scenarios under licensed and unlicensed spectrums. Furthermore, we will explore mechanisms for dynamically allocating specific spectrum resources to a 6G subnetwork by its corresponding parent network. In conclusion, the road ahead for subnetworks involves several key goals: identifying and accessing the frequency bands to be supported, analysing and developing efficient PHY channel access mechanisms, and ensuring compliance with certification and regulatory standards. Achieving these goals will be crucial for the successful deployment and operation of subnetworks. Our research plans to thoroughly analyse these areas and deliver concrete results that offer practical solutions.

5 CONCLUSIONS

In this deliverable, investigations have been undertaken in the context of resource management within a subnetwork, among subnetworks and between the subnetwork and the 6G overlay network. In this context, aspects of routing data and control signaling have been investigated with the aim of enabling the formation of subnetworks as well as rendering their interaction with the overlay 6G network seamless. This seamless interaction with the overlay 6G network is crucial for providing network continuum. Moreover, the concept of computational resources offloading both locally within the subnetwork and to the remote edge-nodes via the subnetwork MgtN or via the 6G overlay network has been studied. Finally, a survey of the spectrum sharing access schemes with special focus on overcoming the diverse regulatory constraints imposed in different countries has been provided along with the envisioned next steps.

More specifically, in Chapter 2, the so-called virtual connections maintained between the 6G RAN-BS and the subnetwork devices have been proposed. These virtual connections constitute the enabler for the network continuum within subnetworks. In this context, the DL and UL flows have been defined from the 6G RAN-BS to the subnetwork nodes via the subnetwork management node. Additionally, the processes of joining and exiting the subnetwork while maintaining these virtual connections has also been provided with a special focus on dynamic topologies. Moreover, flexible subnetworks have been enabled by distributing the functionality of the UP and CP throughout the subnetwork nodes, in the form of the SN-CP and SN-UP, which boast a flexible deployment in terms of layer termination. These provide the architectural basis for the deployment of non-standalone UEs, which cannot establish a connection to the overlay network on their own but utilize the capabilities of other subnetwork nodes. In the next deliverable, the processes required for their configuration will be investigated and defined. Furthermore, the interaction between neighboring subnetworks will be investigated and new solutions will be presented benefiting from the inter-MgtN links, which provide an extra degree of coordination.

A preliminary approach for the routing of data and control signaling within the with a special focus on in-vehicle subnetworks has also been provided in Chapter 2. The proposal is aimed at exploiting predictability within the subnetwork (e.g. traffic correlations) for achieving a TSN-capable integration between the hybrid wireless and wired connections available in the end-to-end path communicating different in-vehicle subnetworks. This requires a tight coordination between the schedulers present in each subnetwork, as well as potentially the establishment of programmable and adaptive redundancy and multi-path links.

Additionally, the QoS framework that currently exists in 3GPP has been investigated in Chapter 2. This framework is to be used as baseline in the next step by analyzing QoS aspects for in-X subnetworks, addressing challenges and identifying areas for potential enhancements proposing new or improved solutions for QoS mechanisms related to selected use-cases, e.g., interactive gaming use case or any other use case that may be relevant.

Moving on to Chapter 3, the functionalities enabling dynamic computational resources offloading from subnetwork to the umbrella provide important aspect of the fully operational in-X subnetwork system with an overlay 6G network. There are three contributions related to these functionalities in this deliverable. Firstly, a framework-based approach is presented relevant for consumer subnetworks where XR traffic management mechanisms provide input information for offloading tasks. Secondly, an

approach that enables the convergence of communication and computation that adopts the introduction of new and enhanced CP and UP procedures to facilitate the distribution of computation among the network entities. Thirdly, a framework for leveraging coordinated planning between communication and computing technologies across subnetworks and towards the 6G overlay network that includes requirements such as deterministic service levels is also discussed.

In addition, the concept of local distributed computing has been investigated with a special focus on dynamic topologies. More specifically, a set of functional entities were introduced in line with [3]. Those functional entities were defined in such a way that they can be flexibly deployed anywhere in the subnetwork or the overlay 6G network, offering flexibility to the realization of the distributed compute concept in future cellular networks. In this context, the roles of the subnetwork nodes, which are required for enabling this convergence, have been described in more detail. As a next step, the current investigation will continue by examining and defining procedural aspects for local computational offloading.

Moreover, a preliminary framework for service-aware joint planning and allocation of communication and computational resources has been introduced that considers key communication metrics such as bandwidth, distance, and network congestion, which are indeed essential for effective task management. This framework focuses on coordinated planning across both subnetworks and a 6G parent network to support dynamic offloading and enhance service reliability, particularly for services with stringent deterministic requirements. Initially, the focus was on the subnetwork level, managing and allocating workloads across the local subnetwork continuum. This framework was then expanded to facilitate potential seamless integration and collaboration between subnetworks and the 6G parent network, forming a more integrated computing continuum. A system model and several objective functions have been proposed to ensure services are completed within a bounded delay with minimal jitter, to maximize the efficiency of communication resource utilization, and to enhance service reliability. A key future initiative is represented by the implementation of AI/ML algorithms to consider the established objective functions designed to optimize joint task offloading and communication scheduling.

In Chapter 4, contributions have been made towards understanding the challenges and opportunities of spectrum usage in both licensed and unlicensed bands in the context of in-X subnetworks. An outline of the intricacies of spectrum sharing, of the feasibility of various spectrum-based applications, and of the potential for technological innovations such as blockchain for decentralized spectrum management and big data processing for dynamic spectrum management is made. Moving forward, the next steps involve further investigating and refining spectrum-sharing mechanisms that accommodate the unique demands of ultra-dense subnetwork environments. This will include comparative analyses of subnetwork performance under different spectrum conditions and the development of more robust models for spectrum allocation, particularly in contexts where regulatory constraints play a pivotal role. Moreover, continuous advancements in technological solutions like chip technology and AI-based predictive models will be critical to overcoming the challenges of cost and complexity introduced by sophisticated spectrum-sharing processes. The findings aim to contribute towards scalable, efficient, and economically feasible 6G subnetwork solutions that can adapt to varying regulatory environments and market demands.

As a recap, a general architecture blueprint has emerged from this deliverable, mainly based on the contributions in Section 2 for the in-vehicle and consumer categories, with potential extensions to the

industrial category as well. In terms of the communication aspect related to Task 4.2a, the distribution of network functionality as well as the predictive scheduling mechanism presented in Section 2, lay the foundations for building the 6G in-X subnetwork system. In this front, the next step would be to define new methods and frameworks, particularly the subnetwork QoS framework, for harnessing the network functionality distribution achieved to satisfy the extreme requirements for all use case categories as well as seamless integration to the 6G overlay network. As for achieving computational offloading corresponding to Task 4.2b, an initial framework has been proposed in Section 3 for offloading to remote-edge. Additionally, some initial contributions were made towards enabling local and distributed compute offloading as well as a preliminary attempt to jointly orchestrate communication and computational resources. Moving forward, the procedural aspects of computational offloading within and across subnetworks as well as to remote-edge will be further investigated. The ultimate target is to achieve fully converged communication and computation in X subnetworks. Finally, an initial survey was presented in Section 4 for dynamic spectrum sharing for in-X subnetworks, addressing Task 4.2c. In the next deliverable, a further investigation and refinement of spectrum-sharing mechanisms will be made to accommodate the unique demands of ultra-dense in-X subnetwork environments.

REFERENCES

[1] B. Priyianto *et al.*, “D2.2 – Refined Definition of Scenarios, Use Cases and Service Requirements for in-X Subnetworks,” *6G-Shine*, February 2024.

[2] M. A. Uusitalo *et al*., "6G Vision, Value, Use Cases and Technologies From European 6G Flagship Project Hexa-X," in *IEEE Access,* vol. 9, pp. 160004-160020, 2021.

[3] O. Akgul et al., “Deliverable D3.2 Initial Architectural Enablers”, *Hexa-X II*, October 2023.

[4] F. Foukalas *et al.*, “D3.1 –Preliminary Results on PHY and MAC enablers for in-X Subnetworks,” *6G-Shine*, February 2024.

[5] D. Cavalcanti, J. Perez-Ramirez, M. M. Rashid, J. Fang, M. Galeev and K. B. Stanton, "Extending Accurate Time Distribution and Timeliness Capabilities Over the Air to Enable Future Wireless Industrial Automation Systems," in Proceedings of the IEEE, vol. 107, no. 6, pp. 1132-1152, June 2019, doi: 10.1109/JPROC.2019.2903414.

[6] 3GPP TS 38.300 v18.0.0, “NR; NR and NG-RAN Overall Description; Stage 2”, December 2023

[7] 3GPP TS 38.214 v18.0.0, “5G; NR; Physical layer procedures for data”, December 2023.

[8] 3GPP TS 38.323 v18.0.0, “5G; NR; Packet Data Convergence Protocol (PDCP) specification”, December 2023.

[9] 3GPP TS 38. 174 v18.4.0, “NR; Integrated Access and Backhaul (IAB) radio transmission and reception,” May 2024.

[10] 3GPP TS 22.859 v18.2.0, “Study on Personal Internet of Things (PIoT) networks”, December 2021.

[11] 3GPP TS 38.340 v18.0.0, “5G; NR; Backhaul Adaptation Protocol (BAP) specification,” Jan 2024.

[12] 3GPP TS 38.351 v18.1.0, “5G; NR; Sidelink Relay Adaptation Protocol (SRAP) Specification,” May 2024.

[13] B. Priyianto *et al.*, “D2.1. – Initial Definition of Scenarios, Use Cases and Service Requirements for in-X Subnetworks,” *6G-Shine*, October 2023.

[14] 3GPP TR 26.928 v18.0.0, “Extended Reality (XR) in 5G”, March 2023

[15] 3GPP TS 23.501 v18.4.0, “System architecture for the 5G System (5GS); Stage 2”, December 2023

[16] GP. Sharma, D. Patel, J. Sachs, M. De Andrade, J. Farkas, J. Harmatos, B. Varga, et al, "Towards deterministic communications in 6g networks: State of the art, open challenges and the way forward," in IEEE Access, vol. 11, pp. 106898-106923, 2023, doi: 10.1109/ACCESS.2023.3316605.

[17] The next step in E/E architectures, Whitepaper, August, 2023. Available at: link

[18] Y. Ju, Y. Chen, Z. Cao, L. Liu, Q. Pei, M. Xiao, K. Ota, M. Dong, and VCM Leung, "Joint secure offloading and resource allocation for vehicular edge computing network: A multi-agent deep reinforcement learning approach," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 5, pp. 5555-5569, 2023, doi: 10.1109/TITS.2023.3242997.

[19] W. Fan, J. Liu, M. Hua, F. Wu, and Y. Liu, "Joint task offloading and resource allocation for multi-access edge computing assisted by parked and moving vehicles," in IEEE Transactions on Vehicular Technology, vol. 71, no. 5, pp. 5314-5330, 2022, doi: 10.1109/TVT.2022.3149937.

[20] TX. Tran, and D. Pompili, "Joint task offloading and resource allocation for multi-server mobile-edge computing networks," in IEEE Transactions on Vehicular Technology, vol. 68, no. 1, pp. 856-868, 2018, doi: 10.1109/TVT.2018.2881191.

[21] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," in IEEE Transactions on Signal and Information Processing over Networks, vol. 1, no. 2, pp. 89-103, 2015, doi: 10.1109/TSIPN.2015.2448520.

[22] S. Mao, S. Leng, and Y. Zhang, "Joint communication and computation resource optimization for NOMA-assisted mobile edge computing," in 2019 IEEE International Conference on Communications (ICC 2019), pp. 1-6, May 2019, doi: 10.1109/ICC.2019.8761996.

[23] Y. Liu, J. Zhou, D. Tian, Z. Sheng, X. Duan, G. Qu, and VCM Leung, "Joint communication and computation resource scheduling of a UAV-assisted mobile edge computing system for platooning vehicles," In IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 8435-8450, 2021, doi: https://doi.org/10.1109/TITS.2021.3082539.

[24] Y. He, G. Yu, Y. Cai, and H. Luo, "Integrated sensing, computation, and communication: system framework and performance optimization", in IEEE Transactions on Wireless Communications , vol. 23, no. 1, pp. 1114-1128, June 2023, doi: 10.1109/TWC.2023.3285869.

[25] X. Mo, and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design," in Journal of Communications and Information Networks, vol. 6, no. 2, pp. 110-124, 2021, doi: 10.23919/JCIN.2021.9475121.

[26] LP. Qian, B. Shi, Y. Wu, B. Sun, and DHK. Tsang, "NOMA-enabled mobile edge computing for Internet of Things via joint communication and computation resource allocations," in IEEE Internet of Things Journal, vol. 7, no. 1, pp. 718-733, 2019, doi: 10.1109/JIOT.2019.2952647.

[27] J. Ren, Y. He, G. Yu, and GY Li, "Joint communication and computation resource allocation for cloud-edge collaborative system," in 2019 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6, 2019, doi: 10.1109/WCNC.2019.8885877.

[28] F. Liu, J. Huang, and X. Wang, "Joint task offloading and resource allocation for device-edge-cloud collaboration with subtask dependencies," in IEEE Transactions on Cloud Computing, vol. 11, no. 3, pp.3027-3039, 2023, doi: 10.1109/TCC.2023.3251561.

[29] G. Carvalho, B. Cabral, V. Pereira, and J. Bernardino, "Computation offloading in Edge Computing environments using Artificial Intelligence techniques," in Engineering Applications of Artificial Intelligence, vol. 95, pp. 103840, 2020, doi: 10.1016/j.engappai.2020.103840.

[30] J. Li, C. Zhang, Z. Liu, W. Sun, and Q. Li, "Joint communication and computational resource allocation for qoe-driven point cloud video streaming," in 2020 IEEE International Conference on Communications (ICC), pp. 1-6, 2020, doi: 10.1109/ICC40277.2020.9148922.

[31] J. Opadere, Q. Liu, N. Zhang, and T. Han, "Joint computation and communication resource allocation for energy-efficient mobile edge networks," in 2019 IEEE International Conference on Communications (ICC), pp. 1-6, 2019, doi: 10.1109/ICC.2019.8761886.

[32] F. Pervez, A. Sultana, C. Yang, and L. Zhao, "Energy and latency efficient joint communication and computation optimization in a multi-uav assisted mec network," in IEEE Transactions on Wireless Communications, vol. 23, no. 3, pp. 1728-1741, 2023, doi: 10.1109/TWC.2023.3291692.

[33] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4188-4200, 2018, doi: 10.1109/JIOT.2018.2875246.

[34] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Joint communication and computation resource allocation in fog-based vehicular networks," in IEEE Internet of Things Journal, vol. 9, no. 15, pp. 13195-13208, 2022, doi: 10.1109/JIOT.2022.3140811.

[35] J. Li, C. Zhang, Z. Liu, W. Sun, and Q. Li, "Joint communication and computational resource allocation for qoe-driven point cloud video streaming," in 2020 IEEE International Conference on Communications (ICC), pp. 1-6, 2020, doi: 10.1109/ICC40277.2020.9148922.

[36] J. Huang, J. Wan, B. Lv, Q. Ye, and Y. Chen, "Joint computation offloading and resource allocation for edge-cloud collaboration in internet of vehicles via deep reinforcement learning," in IEEE Systems Journal, vol. 17, no. 2, pp. 2500-2511, 2023, doi: 10.1109/JSYST.2023.3249217.

[37] Y. Gong, H. Yao, J. Wang, M. Li and S. Guo, "Edge Intelligence-driven Joint Offloading and Resource Allocation for Future 6G Industrial Internet of Things," in IEEE Transactions on Network Science and Engineering, 2022, doi: 10.1109/TNSE.2022.3141728.

[38] H. Guo, J. Liu, J. Ren, and Y. Zhang, "Intelligent task offloading in vehicular edge computing networks," in IEEE Wireless Communications, vol. 27, no. 4, pp.126-132, 2020, doi: 10.1109/MWC.001.1900489.

[39] Y. Cui, L. Du, H. Wang, D. Wu, and R. Wang, "Reinforcement learning for joint optimization of communication and computation in vehicular networks," in IEEE Transactions on Vehicular Technology, vol. 70, no. 12, pp. 13062-13072, 2021, doi: 10.1109/TVT.2021.3125109.

[40] P. Yang, L. Kong and G. Chen, "Spectrum Sharing for 5G/6G URLLC: Research Frontiers and Standards," in IEEE Communications Standards Magazine, vol. 5, no. 2, pp. 120-125, June 2021, doi: 10.1109/MCOMSTD.001.2000054.

[41] G. Hampel, C. Li and J. Li, "5G Ultra-Reliable Low-Latency Communications in Factory Automation Leveraging Licensed and Unlicensed Bands," in IEEE Communications Magazine, vol. 57, no. 5, pp. 117-123, May 2019, doi: 10.1109/MCOM.2019.1601220.

[42] R. Bajracharya, R. Shrestha, S. A. Hassan, H. Jung, R. I. Ansari and M. Guizani, "Unlocking Unlicensed Band Potential to Enable URLLC in Cloud Robotics for Ubiquitous IoT," in IEEE Network, vol. 35, no. 5, pp. 107-113, September/October 2021, doi: 10.1109/MNET.121.2100114.

[43] B. Hassan, S. Baig and M. Asif, "Key Technologies for Ultra-Reliable and Low-Latency Communication in 6G," in IEEE Communications Standards Magazine, vol. 5, no. 2, pp. 106-113, June 2021, doi: 10.1109/MCOMSTD.001.2000052.

[44] A. Kliks et al., "Beyond 5G: Big Data Processing for Better Spectrum Utilization," in IEEE Vehicular Technology Magazine, vol. 15, no. 3, pp. 40-50, Sept. 2020, doi: 10.1109/MVT.2020.2988415.

[45] R. Adeogun, G. Berardinelli and P. E. Mogensen, "Enhanced Interference Management for 6G in-X Subnetworks," in *IEEE Access*, vol. 10, pp. 45784-45798, 2022, doi: 10.1109/ACCESS.2022.3170694.

[46] S. Wang and C. Sun, "Blockchain Empowered Dynamic Spectrum Sharing: Standards, State of Research and Road Ahead," in IEEE Communications Standards Magazine, vol. 7, no. 3, pp. 72-80, September 2023, doi: 10.1109/MCOMSTD.0010.2200026.

[47] P. M. de Sant Ana, N. Marchenko, P. Popovski and B. Soret, "Wireless Control of Autonomous Guided Vehicle using Reinforcement Learning," *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, Taipei, Taiwan, 2020, pp. 1-7, doi: 10.1109/GLOBECOM42002.2020.9322156.

[48] P. M. de Sant Ana, N. Marchenko, P. Popovski and B. Soret, "Age of Loop for Wireless Networked Control Systems Optimization," *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Helsinki, Finland, 2021, pp. 1-7, doi: 10.1109/PIMRC50174.2021.9569366.

[49] P. Hande et al., "Extended Reality over 5G–Standards Evolution," *IEEE Journal on Selected Areas in Communications,* 2023.

[50] Webpage link: Home | AdvantEDGE (interdigitalinc.github.io)

[51] M. Satyanarayanan, .W. Gao, and B. Lucia,"The computing landscape of the 21st century," *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, 2019.

[52] C. Lee et al., "A scalable solution to the multi-resource QoS problem," *Proceedings 20th IEEE Real-Time Systems Symposium, No. 99CB37054*, 1999.

[53] V.-P. Bui, Van-Phuc, S. R. Pandey, F. Chiariotti, and P. Popovski, "Game Networking and its Evolution towards Supporting Metaverse through the 6G Wireless Systems," *arXiv preprint arXiv:2302.01672*, 2023.

[54] Huawei, "Cloud VR solution white paper," available at: https://www.huawei.com/en/news/2018/9/cloud-vrsolution-white-paper

[55] Extended Reality and 3GPP evolution, A 5G Americas White Paper, November 2022.

[56] Paymard, Pouria, et al. "PDU-set Scheduling Algorithm for XR Traffic in Multi-Service 5G-Advanced Networks." https://www.ieee802.org/15/pub/TG1.html

[57] R. Liu, Ruiqi, et al. "A vision and an evolutionary framework for 6G: Scenarios, capabilities and enablers." arXiv preprint arXiv:2305.13887, 2023.

[58] N. Rastegardoost, "3GPP: The Unlicensed Journey," 2020.

[59] M. Miraz *et al.*, "A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of nano things (IoNT)," 2015 Internet Technologies and Applications (ITA),pp. 219-224, 2015.

[60] E. Semaan et al. "6G spectrum-enabling the future mobile life beyond 2030," 2023.

[61] D. Castells-Rufas, A. Galin-Pons, and J. Carrabina, "The regulation of unlicensed sub-GHz bands: Are stronger restrictions required for LPWAN-based IoT success?." arXiv preprint arXiv:1812.00031, 2018.