Enabling Low-Latency 6G Architectures with a Distributed B5G Core Framework

Maida Islamagić*, Raúl Cuervo Bello*, Miguel Camelo Botero†, Johann Marquez-Barja*, and Nina Slamnik-Krieštorac*
* University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Belgium
† University of Antwerp - imec, IDLab - Department of Computer Science, Belgium

***Abstract*—With the stringent requirements of evolving vertical services, there is a growing need for more flexible, ultra low-latency Beyond 5G (B5G)/6G architectures. One way to achieve this is by disaggregating core network functions and deploying them closer to the end users. The distribution and optimal placement of both Control Plane (CP) and User Plane (UP) functions at the network edge, along with the collocated orchestrated Application Functions (AFs), can significantly improve End-to-End (E2E) latency and network reliability by avoiding unnecessary traffic flows to the centralized cloud servers. Furthermore, as 5G evolves into 6G, automated and intelligent network management solutions will be needed in the distributed communication compute continuum. This paper presents an early-stage PhD research direction that focuses on the optimal quality-aware disaggregation of 5G core functions, aiming to reduce E2E latency and improve the throughput of 6G vertical services. The focus is on the optimal User Plane Function (UPF) placement, as well as the other CP functions and AFs, which altogether interact with the UPF. As UPF is directly involved in user data forwarding, it is crucial for handling user traffic over 5G/B5G network and, as such, it significantly impacts E2E latency. This paper provides i) an overview of theoretical concepts and State of the Art (SotA) UPF placement methodologies, and ii) future directions for optimized placement of B5G Core and application functions within the edge cloud continuum, leveraging intelligent network and service orchestration solutions.**

***Index Terms*—distributed B5G Core, 6G, UPF**

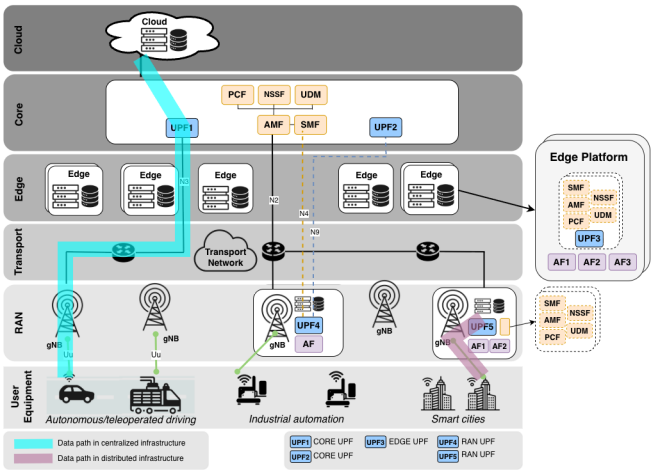

Fig. 1: Proposed distributed architecture

I. Introduction

The increasing demands of modern-day vertical services, particularly mission-critical ones requiring ultra-low latency, high reliability, and guaranteed Quality of Service (QoS), have triggered a shift from traditional, centralized Fifth-Generation (5G) networks towards distributed, adaptable architectures [1]. Centralized architectures suffer from increased E2E latency due to the computational and transmission latency caused by centralized, single instances of core components, as well as from scalability limitations and the risk of single points of failure.

In 5G, three main service categories: Ultra-Reliable Low-Latency Communication (URLLC), enhanced Mobile Broadband (eMBB), and massive Machine-Type Communication (mMTC), define the network capabilities that support a wide range of vertical services [2]. The mission-critical services have stringent latency requirements: for example, for remote vehicle operations and time-critical sensing, E2E latency needs to be lower than 30 ms [3]. However, achieving such latency requirements remains challenging in current 5G Standalone (SA) deployments. Based on the work of X. Limani et al. [4], 5G SA deployments were validated across many real-world settings, including industrial and urban areas with 5G SA coverage. The results showed that URLLC slice, under normal conditions, effectively maintained low-latency performance in the required range, while exhibiting notable fluctuations in high-stress conditions or when moving further from the base stations [4]. To reduce network latency, 5G implements fundamental modifications, including Network Function Virtualization (NFV), Multi-Access Edge Computing (MEC), Massive Multiple-Input Multiple-Output (MIMO), and modifications to the 5G radio, among others [5].

Nevertheless, centralized core network procedures and long data paths contribute significantly to the overall E2E latency. Therefore, disaggregating core network functions and optimizing their distributed deployment at the network edge is a promising approach for shortening data paths to centralized cloud servers. It can also reduce signaling overhead, depending on the deployment and orchestration strategy.

In this paper, we present an early-stage PhD research direction, focusing on designing intelligent techniques for optimal distributed core network solutions to maintain required QoS levels for mission-critical services, especially focusing on reducing E2E latency. Existing approaches, analyzed later in Section 3, demonstrate that distributed UPF placement reduces E2E latency while addressing mobility, slicing, and edge deployment in limited, rather simplified scenarios. However, what is lacking are intelligent techniques to optimize the placement of UPF and related core functions across the large-scale compute-communication continuum, while accounting for collocation with edge-based application functions. To address these questions, we propose a distributed architecture, illustrated in Fig. 1. Our approach will include various combinations of CP and UP functions collocated on the same edge platform, as well as their (re)location, which will be optimally orchestrated relative to application function placement and the mobility of users. Placement and relocation decisions

ISBN 978-3-903176-82-9

TABLE I: Comparison of experimental studies on disaggregated UPF deployment

Work	Use case	Architecture type	KPIs evaluated	Main results	Key gap
5G Edge-Central Core Network Split [6]	Vehicle communication, smart city, multimedia	5G testbed with satellite and Ethernet backhaul	Median delay for CP procedures; RTT for UP	Autonomous Edge achieved minimal control plane delay	Limited set of components; software-emulated RAN
Smart farming testbed [7]	Smart farming (fire detection)	srsRAN + Open5GS (edge UPF)	RTT, throughput, jitter	∼60% latency reduction; up to 40% throughput improvement	Single UPF, no mobility or reliability analysis
Disaggregated Core (Smart City) [8]	Smart city services	Disaggregated core (edge + cloud UPF)	RTT, throughput, jitter	RTT reduced by ∼35% (factor 0.65); improved stability	Static UPF selection; ideal conditions (LoS, low load)
V2X Network Slicing [9]	V2X (e.g., remote driving)	Multi-UPF + network slicing (OAI + Open5GS)	RTT, throughput (UL/DL), packet loss	RTT reduced (17.53 ms to 15.43 ms); packet loss reduced (7.33% to 0.1%)	Static slice-based assignment; no dynamic steering

will be made by leveraging intelligent decision-makers based on Artificial Intelligence (AI) (Reinforcement Learning (RL), Federated Learning (FL), agentic AI).

In the next section, we provide a brief background description, focusing on 5G Core (5GC). Then we focus on the most relevant related work and SotA analysis. Finally, we close the paper by defining the research questions and the methodology we will adopt to address them.

II. Background

5G enables flexible solutions by adopting a Service-Based Architecture (SBA), in which Network Functions (NFs) are fully virtualized and communicate with each other in a service-based manner. Furthermore, separating CP functions from UP functions, the concept known as Control and User Plane Separation (CUPS), enables independent scalability and flexible deployments of the functions [2].

Core functionalities are traditionally located in the network's central infrastructure. This means that control and user data must traverse the backhaul network to reach the centralized servers, introducing additional latency. The 5GC network consists of several functions, responsible for establishing user session(s) securely and forwarding user data to and from mobile devices [10]. The Access and Mobility Management Function (AMF) can be considered a main function in the control plane, since it is the first interaction point between the Radio Access Network (RAN) and the Core Network (CN), when a user connects to a mobile network. It is responsible for user registration, authentication, and mobility management procedures [10], by interacting with other NFs. For user registration, the AMF interacts with the Authentication Server Function (AUSF) and the Unified Data Management (UDM) function to authenticate the User Equipment (UE) and retrieve subscription data. Upon successful registration, the AMF forwards all session management-related signaling messages to the Session Management Function (SMF). The SMF is responsible for establishing, modifying, and releasing individual sessions, and for allocating Internet Protocol (IP) addresses per session by assigning the appropriate UPF.

Since these core functions are affected by heavy traffic loads, SotA approaches adopt their disaggregation from the core network as a promising solution for E2E latency reduction (Section 3). With the optimal orchestration of these functions together with AFs deployed on edge platforms, a significant part of data processing could be moved to the network edge. Although core disaggregation improves latency, it introduces several trade-offs that need to be addressed. In particular, edge devices often have limited compute resources and less capacity to deploy applications, and are more vulnerable to cybersecurity attacks. Therefore, it is essential to consider these trade-offs when designing decentralized 5GC architectures.

III. State of the Art

Existing research on the topic of UPF distribution in the network edge can be classified into two categories: (i) analytical approaches focusing on optimal UPF placement, based on mathematical models and algorithms, and (ii) experimental testbed-based studies that evaluate the performance of UPF distribution. Table I provides a summary of related work. The research study cases span a variety of applications that require low latency and high reliability, including smart cities, Vehicle-to-Everything (V2X) communication, and remote healthcare.

Based on the work from M. Corici et al. [6], several 5G core network splits are possible. One solution is to place one UPF in the edge network, while the rest of the CP functions and additional UPFs remain in the core network. This way, the edge UPF can handle locally offloaded data traffic or forward it to the central side. Another solution is to deploy CP functions, AMF, and SMF locally at the network edge, next to the UPF. This approach additionally reduces the amount of signaling between the edge and the central entities. The experiments conducted on the testbed demonstrated the feasibility of decentralized networks, with reduced latency across various procedures. However, experiments were conducted in a testbed with a limited set of components and software-emulated RAN, resulting in optimistic results.

What needs to be further taken into account is the collocation with the radio network and the location of the applications. According to N. Makondo et al. [7], placing UPF in the network edge can reduce latency by 60-60.7%, and improve throughput by up to 40%. The solution is based on a testbed to support URLLC applications, such as quick-fire detection, and uses open-source 5G platforms, such as srsRAN and Open5GS. However, the experiment was limited to a single UPF instance and a single user. This is not a realistic scenario and, therefore, not sufficient for assessing large-scale distributed 5G architectures, where network congestion, queuing, and traffic load distribution impact E2E latency. Furthermore, the evaluation does not consider the impact of user mobility and handover on UPF placement and latency performance, or system resilience in the presence of failures. These two factors are important for orchestrated allocation and relocation of multiple UPFs, relative to the distance from the users.

Based on the work from P. Valente et al. [8], advanced architectures for a disaggregated mobile core are proposed for smart city applications. The work goes beyond comparing the deployment of UPF on the cloud and edge by evaluating different aspects of UPF disaggregation, including mobility scenarios, flow-level assessments for higher-priority services, and content distribution. The results show that the edge-based UPF deployment brings significant improvements, with the

Round-Trip Time (RTT) reduction by a factor of 0.65, for edge UPF [8]. It is worth noting that these results were obtained with a constant Line-of-Sight (LoS), when no other UE used the cell bandwidth. As mentioned previously, this is not a realistic scenario and does not reflect the behavior of disaggregated UPF deployments under real-life network load.

According to X. Limani et al. [9], the QoS requirements of V2X applications under network overload can be preserved by assigning different UPFs to different traffic slices. The goal is to design, implement, and validate an E2E framework that leverages distributed UPFs and RAN-core synergy. The results show that the average RTT for the critical slice decreased from 17.53 ms to 15.43 ms, while packet loss under congestion dropped from 7.33% to 0.1%. By applying policies to limit non-critical traffic, the network ensured sufficient uplink capacity for critical services such as remote driving [9]. The authors of the paper address that their current configuration lacks the direct allocation of dedicated radio resources for uplink traffic.

Overall, existing works demonstrate that distributing UPFs in the network edge can reduce E2E latency and improve network performance. However, they do not fully address dynamic, large-scale orchestration of distributed core functions in realistic scenarios, by including user mobility, and possibly mobility of edge nodes as well. This highlights the need for intelligent mechanisms for optimal core network collocation at the network edge.

IV. Research Methodology

Given the challenges posed by demanding verticals, current 5G architectures will have to evolve into more intelligent, self-managed Sixth-Generation (6G) networks. The disaggregation of both UP and CP functions and their optimal collocation with application functions at the network edge play a key role in maintaining the required QoS level, particularly in reducing the E2E latency. This is expected to be supported by advanced AI techniques for orchestration and adaptive network management.

In this section, we present an early-stage PhD research direction, formulated as Research Questions (RQ1, RQ2, RQ3 and RQ4), focusing on designing intelligent techniques for optimal distributed 6G core network solutions:

- **RQ1:** How does the orchestrated deployment of UPF and application functions in the network edge affect E2E latency in large-scale distributed 6G communication compute continuum? The collocation of other core functions that support UP communication (e.g. Policy Control Function (PCF), UDM) should also be considered to further reduce E2E latency.
- **RQ2:** How can user mobility and handover be incorporated into decision mechanisms for the optimal UPF placement on the edge-computing platforms, relative to the location of base stations?
- **RQ3:** How can dynamic and adaptive mechanisms, including intelligent solutions, such as agentic AI solutions, be applied for UPF selection and relocation?
- **RQ4:** Can the control plane functions (e.g., AMF, SMF) also be disaggregated to the network edge, in order to further improve latency?

To address these questions, we will adopt new 6G architectures and algorithms for deploying 6G core functions, while taking into account QoS requirements, energy consumption, availability of scarce edge resources, and mobility of users and network nodes. The proposed architecture, illustrated in Fig. 1, forms the basis for the design and evaluation of the distributed core network solution. To validate the proposed solution, we will adopt an experimental approach using a nextG-Lab[1] testbed that integrates Open Air Interface (OAI)[2] and Open5GS[3]. We will evaluate and compare the performance of a centralized core architecture and a case in which UP and CP functions are deployed at different locations, closer to users and application functions. Latency is one of the main Key Performance Indicators (KPIs) for mission-critical URLLC applications, and we will evaluate the E2E, defined as the total time delay for data to travel from its origin to its destination, since it encompasses propagation, processing, and computational delay [11]. Furthermore, other important KPIs will be evaluated, including resource availability, throughput, reliability, and energy efficiency.

Finally, our work will be based on AI-driven decision-making techniques for optimal function placement, as a key enabler of scalable, adaptive, and latency-aware network architectures. In particular, we consider agentic AI-based approaches, leveraging techniques such as RL and FL, due to their ability to enable autonomous, adaptive, and distributed decision-making under dynamic network conditions. Therefore, we will place a specific focus on the trustworthiness of AI-driven decisions, by measuring AI accuracy and the quality of decisions.

Acknowledgements

This work has been funded by the imec.icon project MAGNOLIA, co-financed by imec and Flanders Innovation & Interpreneurship (VLAIO) under project nr. HBC.2025.0436.

References

[1] C.-X. Wang *et al.*, "On the road to 6g: Visions, requirements, key technologies and testbeds," 2023. [Online]. Available: https://arxiv.org/abs/2302.14536

[2] 3rd Generation Partnership Project (3GPP), "5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 version 19.7.0 Release 19)," Technical Specification TS 23.501, 2026, online [Available]: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144.

[3] 3rd Generation Partnership Project (3GPP), "5G; Service requirements for the 5G system (3GPP TS 22.261 version 19.13.0 Release 19))," Technical Specification TS 22.261, 2026, online [Available]: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3107.

[4] X. Limani *et al.*, "Optimizing 5g-based teleoperation: Synergy of vulnerable road user awareness and advanced traffic management systems," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2024, pp. 1067–1072, doi: https://doi.org/10.1109/EuCNC/6GSummit60053.2024.10596999.

[5] I. Parvez *et al.*, "A survey on low latency towards 5g: Ran, core network and caching solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018, doi: https://doi.org/10.1109/COMST.2018.2841349.

[6] M. Corici *et al.*, "A study of 5g edge-central core network split options," *Network*, vol. 1, pp. 354–368, 2021. [Online]. Available: https://www.mdpi.com/2673-8732/1/3/20

[7] N. Makondo *et al.*, "Implementing an efficient architecture for latency optimisation in smart farming," *IEEE Access*, vol. 12, pp. 140 502–140 526, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3466994.

[8] P. Valente *et al.*, "Disaggregated mobile core for edge city services," in *2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2023, pp. 157–166, doi: https://doi.org/10.1109/WoWMoM57956.2023.00030.

[9] X. Limani *et al.*, "Network slicing as the ultimate enabler of enhanced service quality in vehicular-to-everything (v2x) world," in *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, 2024, pp. 1–5, doi: https://doi.org/10.1109/VTC2024-Fall63153.2024.10757617.

[10] S. Rommer *et al.*, *The Core Network for 5G Advanced: Architecture overview*, 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780443291883000065

[11] J. R. Bhat and S. A. Alqahtani, "6g ecosystem: Current status and future perspective," *IEEE Access*, vol. 9, pp. 43 134–43 167, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3054833.

[1] nextG-Lab:https://www.uantwerpen.be/en/research-groups/idlab/infrastructure/nextg-lab/

[2] OAI: https://gitlab.eurecom.fr/oai/openairinterface5g

[3] OPEN5GS: https://open5gs.org/