A Proof of Concept Implementation of an AI-assisted User-Centric 6G Network

N. Gkatzios
INFOLYSIS P.C.
Athens, Greece
ngkatzios@infolysis.gr

H. Koumaras
NCSR "DEMOKRITOS"
Agia Paraskevi, Greece
koumaras@iit.demokritos.gr

D. Fragkos
NCSR "DEMOKRITOS"
Agia Paraskevi, Greece
dfragkos@iit.demokritos.gr

V. Koumaras
INFOLYSIS P.C.
Athens, Greece
vkoumaras@infolysis.gr

***Abstract*— The design of the 6G system will be based on a user-centric paradigm, enabling users to be involved in the creation and the management of network services and also providing them with a highly customized network experience. This paper proposes a realization of this paradigm by proposing a core network redesign from "Network Function-focus" to "User-focus" in conjunction with an AI-assisted approach that self-organises the network according to the user-requirements. A proof of concept implementation is presented based on both simulated and physical deployments, that demonstrate an optimal User Plane Function (UPF) placement taking into consideration user's preferences, proving the validity of the proposed approach.**

Keywords—6G; user-centric; AI; QoS; validation.

I. Introduction

Today mobile communication networks follow an operator-centric approach, utilizing monolithic network functions (NFs), such as access management functions (AMFs) and user plane functions (UPFs) in 5G, which are expected to serve a large number of end users at the same time. The 5G core network is currently built using the per-network function paradigm. As a result, some monolithic network functions—such as AMF and the session management function (SMF) in 5G—serve a high number of users/UEs, while performing extremely particular tasks. So, it is acceptable to assume that a centralized architecture is a natural outcome of the NF-centric design.

From a functional standpoint, the current operator-centric and function-centric core network architecture controls each UE's or end user's state by maintaining consistent states across various network functions. This results in complex signaling message exchange, which may limit how much network performance may be enhanced (e.g. latency), as well as create more potential areas of attack. Monolithic network services (both physical and virtual) could become significant bottleneck sources as the number of linked devices/users rises.

One might argue that the currently supported "per-user" slicing mechanisms of the existing NF-centric 5G core could accomplish this user-centric vision of 6G [1]. However, this is not feasible, since the existing slicing techniques of the 5G system, is built on the "NF-focused" fundamental underlying architecture, where each user may feel like having his/her own dedicated network services, but in reality, some centrally located monolithic NFs serve multiple users and UEs creating this user-centricity "illusion", while in practice scaling and performance issues remain.

In order to overcome these obstacles, network services need to be recentered on users, following a user-centric approach, the distinguishing characteristic of the 6G architecture, enabling user-definition, user-configuration, and user-control [2]. The visioned user-centric architecture in 6G alters how users, network services, and apps communicate, having an impact on the provided mobility management. With a shared context and a modular design, the per-user network will minimize the overwhelming message exchange between conventional network functions, resolving in this way the performance limitations of an NF-centric network, while at the same time will allow the per-user NF-placement based on the performance requirements of each user.

Following this architectural evolution in 6G, the necessity for advanced user-centric placement and deployment of the NF has been raised. Considering that various distributed virtualization infrastructures have become available, spanning from edge to central cloud, the various placement options have a direct impact to the Quality of Experience (QoE) and Quality of Service (QoS) levels that are delivered at the user.

In order to achieve the optimal NF placement and deployment based on the user and/or service requirements, the use of Artificial Intelligence (AI) techniques has been introduced, which allow to the 6G network not only to self-configure its operation based on the user-expectations, but also to predict how a specific session is going to be evolved and respectively to be adapted accordingly [3]. By exploiting data that are provided by the core network 3GPP APIs, such as NEF, NWDAF [4], an AI model can be tightly integrated with the core network and be used as a cognitive AI coordinator deciding for each user the optimal NF placement and deployment, realizing a user-centric 6G service provision.

This paper describes the core network evolution from a NF-centric to a user-centric design, which together with a cognitive AI coordinator, realizes a novel architecture for user-centric 6G network. The proposed approach goes beyond the current state-of-the-art, because it does not simply rely on triggering the optimal NF placement based on some QoS-degradation threshold [5], but it exploits on the openness of the core network [6] in order to predict the optimal placement, based on both spatial and temporal user-data.

The value of the proposed intent-based 6G architecture is validated with an emulated mobility scenario, which confirms the advances of the proposed user-centric approach on the optimal UPF-placement based on the user context.

The rest of the paper is organized as follows: Section II introduces an architectural evolution of the core network, responding to the requirements set by the user/human-centric 6G vision. Section III describes the cognitive AI layer capable of taking decisions and providing reasoning on the user-centric instantiations of the core network. Section IV provides an experimental validation of the proposed user-centric 6G network, based on a cloud native implementation of the core network and the agility of the cognitive AI coordination. Section V provides future directions of the proposed approach as a technology enabler for trustworthiness provision in 6G. Section VI concludes the paper.

II. User-Centric Evolution of the Core Network

In order to offer users a highly personalized network environment, as ITU FG NET-2030 envisions, a user-centric concept will be used to build the 6G system architecture, allowing users to participate in network service development and operation. Architectural redesign of the core network is required, following the paradigm shift from “Network Function-focus” to “user-focus”, allowing users to participate in network service creation and operation, while also giving users full control over data ownership.

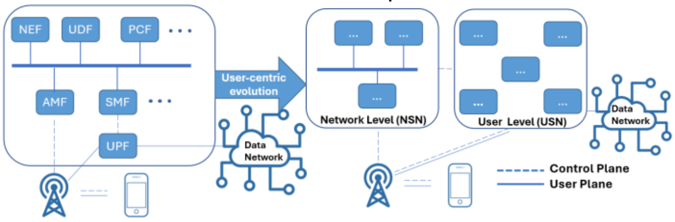

Fig. 1. Evolution of the current function-centric 5G core towards a user-centric evolution of the 6G core

To achieve this design, the network architecture should be segmented into user service nodes (USN) and network service nodes (NSN), and each user will receive a full instance of the 6G system properly distributed over the continuum based on her/his service requirements and level of trust. For instance, to satisfy a variety of service needs in the 6G era, USN will be composed of end-user-level network policies and tailored services. As a result, it may help develop user profiling as a representation of the physical world [7].

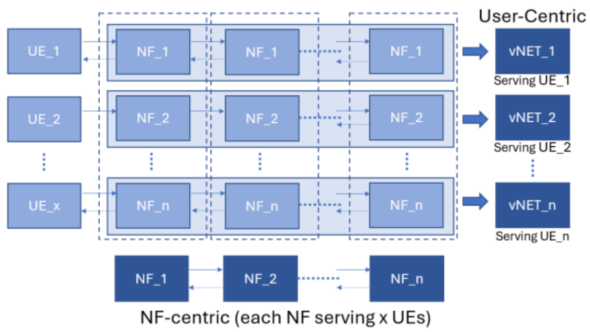

Fig. 2. The User-centric evolution of the core network vs. the current NF-centric approach.

In the envisioned user-centric evolution of the 6G core, each user will have a separate network that consolidates all necessary tasks for service delivery thanks to the user-centric architecture. According to this design approach, the User-Centric Network (UCN) is in charge of managing mobility, policies, sessions, and personal data. The much lower exchanged signals and resulting decreased latency are one of UCN's selling advantages. Thus, the proposed 6G Beyond Service Based Architecture (B-SBA) should allow the deployment of NSN components in one location, while aggregating multiple instances of USN in different locations but as part of the same communication platform. The B-SBA would allow the creation of network slices tailored for the users where the resources might be distributed or provided by different network providers.

III. Openess & AI Automation of the User-Centric 6G

The complexity imposed on the management and orchestration framework in user-centric 6G networks over the continuum requires evolving the orchestration concept through the massive integration of native AI as part of a distributed data-driven network architecture. Towards this the openness and the data exposure foreseen in B5G and 6G networks is acting as an enabler towards the envisioned AI-automation.

A. Openess and Data Exposure as enablers of AI-automation

The exposing of APIs, such as 3GPP TS 29.522 NEF/ 3GPP TS 33.521 NWDAF/ 3GPP 33.122 CAPIF from the NSN plane (i.e. network core and edge) allows to third party applications of the USN plane (i.e. service/application provisioning level) to integrate with the core network [8]. 3GPP SA6's work on vertical application enablers (VAEs) has emerged, with the goal of extending the NWDAF analytics APIs, so that decision and prediction AI models can be fed with data in order to support zero-touch management of the user-centric network.

Therefore, 6G will take network automation to an unprecedented level. More specifically, this tight integration between AI and 6G system (NSN and USN exposed data and APIs) will support the agile realizing of the user-centric provision. AI will transform network management into a cognitive process through which the network can self-adapt and self-react to changing conditions with minimal manual intervention (zero-touch). For example, Intelligent predictive orchestration will help to reduce OPEX, while reaching the relevant 6G KPIs/KVIs (i.e., optimal use of resources, decrease of energy consumption, service assurance) depending on the user preferences and context [9].

B. Cognitive AI system for user-centric 6G provision

This paper proposes a cognitive system that specializes in autonomous service and network operation by combining well-known AI technologies inside a flexible framework. The cognitive layer acts as an interface between tenants/users/3rd party apps and the network/environment via the 6G service exposure provider interface.

Such a cognitive coordination system consists of three major components: a knowledge base, a reasoning engine, and an agent architecture:

- The knowledge base includes an ontology of intents as well as domain-specific knowledge, such as the current state of the system.
- The domain-independent reasoning engine will use the knowledge graph as the primary coordinator function for locating actions, assessing their impact, and ordering their execution in order to realise the intent

which was requested by the tenant/user and/or third-party application.

- Finally, the agent design allows for the use of an unlimited number of models and services. The AI agents will incorporate machine learning models or rule-based policies, as well as providing services necessary in the cognitive reasoning process. For example, a machine learning model that can propose α mesh orchestration/configuration that improves reliability. This model is registered as an agent, which means as a "proposer" at the cognitive level, for configuration activities. Because of the discrete life cycle, the model may be replaced with an improved version as soon as it becomes available, regardless of the cognitive layer release cycles. Another example could be an agent assuming the role of "observer" that would monitor data sources and maintain the state's knowledge up to date.

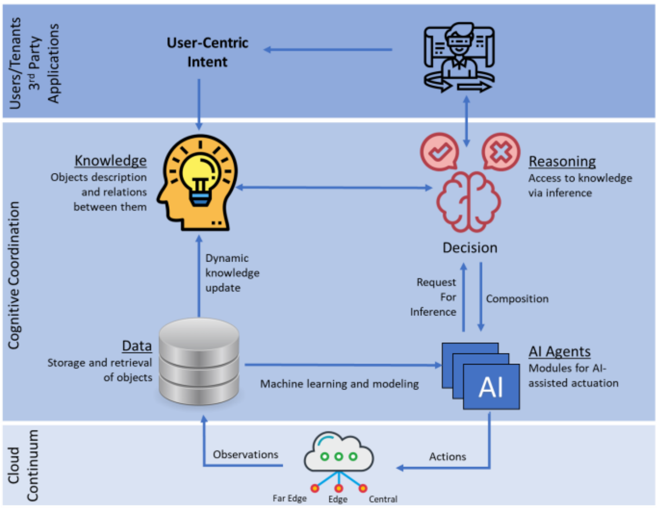

Fig. 3. The proposed Cognitive AI system for user-centric 6G provision

Smooth communication between the reasoning engine and knowledge base is essential for the cognitive coordinator to function properly. The intent-driven reasoning engine continually runs a process that looks for ways to take actions to bridge the gap between the desired user-centric state and the currently observed state. It gathers suggestions, receives forecasts of each proposal's impact, weighs benefit against risk and certainty, prioritizes its course of action, and implements its judgments. Every stage of the procedure makes extensive use of specialized agents. The ongoing cognitive process would continually seek activities for further optimization even in the absence of explicit problems. It may, for instance, aim to use less resources to provide the same services.

The proposed cognitive coordinator achieves a high level of dynamic adaptation to new contexts through its reasoning-based core mechanism. This stands in sharp contrast to systems that have been implemented using set workflows and rule-based policies, where every supported scenario requires consideration at the time of design through appropriate decision tree branches and varying rules. However, existing rule-based policies can still be implemented as agents on the cognitive coordinator and deployed there. This creates a method for upgrading old automation, adding AI-based models incrementally.

C. *Proposed User-centric architecture of 6G Networks*

The user-centric redesign of the core network will be driven by the distributed edge-cloud continuum and the openness of the 6G system, following the paradigm of the 5G core openness and its exposure capabilities with standardised 3GPP APIs, such as CAPIF/NEF, as well as the Network Apps paradigm driven by many European research projects and initiatives.

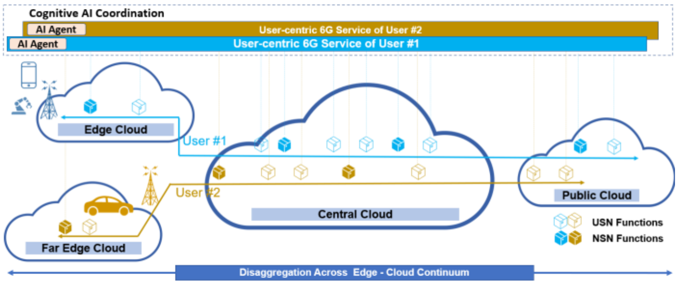

Fig. 4. Conceptual architecture of User-Centric 6G Core Network deployment with Cognitive AI Coordination over the Edge-Cloud Continuum

As Fig. 4 depicts, the cloud continuum in 6G will allow the realisation of the USN and NSN services in a distributed and adapted way closer to users, tenants, applications, data sources, and regulated processes. Knowledge of the continuum capabilities and of the user-context via the 3GPP APIs and ETSI APIs, can be used to optimize NSN and USN deployment and performance per user (i.e. user-centric approach), including aspects of security and robustness.

Another key benefit of the envisioned user-centric redesign of the distributed 6G core architecture is the ease with which NFs can be placed, subsequently scaled, and moved between the clouds of the continuum (i.e. far edge, near edge, central cloud on which the 6G core is realised), and the efficiency with which they can be executed depending on each user's requirements and preferences.

Similarly, the openness of the 6G system will enable the development of novel and innovative USN and NSN CNFs by third parties, creating even higher impact in the realisation of diverse use-cases with distributed intelligence.

IV. Proof of Concept Implementation of the Proposed AI-Assisted User-Centric 6G Network

For validating the proposed user-centric AI-enabled 6G architecture, this paper provides an experimental implementation, which interfaces an AI agent in the 6G core and consumes data from the 3GPP NEF API concerning the location of the user in order to identify usage patterns.

More specifically, for the experimental part of this paper a mobility scenario is considered where the location of a user in reported via the NEF API to the AI agent of the user-centric 6G network. Given that the user is requiring when is located at the work premises during the working hours to have very low latency access network (potentially because this allows to him/her to offload some computing tasks at the edge of the network), the proposed user-centric network performs the necessary AI analysis in order to predict the location of the user and respectively to place and deploy its functions in that way that low latency is reassured. For the experimental needs of the paper, the optimal placement and deployment of the UPF function in close proximity to the specific user is considered, allowing a local break-out and the low latency provision to that user and therefore validating the agility of the proposed user-centric and AI-driven 6G network. The proposed architecture allows users to have their own networks, while avoiding the "one-size-fits-all" philosophy in order to provide personalized services.

A. Scenario Description and Dataset Creation

For performing the experimental validation of the proposed user-centric network, a mobility scenario is simulated by using the NEF Simulator [10], which is an open-source software implementation to conduct core network simulations. More specifically, a map is provided in the GUI of the simulator where moving users and cells coverage can be placed in order to formulate a specific scenario and topology, as depicted in Figure 5. This user-friendly interface allows for the creation and storage (i.e., JSON or CSV format) of multiple scenarios across the map, serving as a valuable tool for generating the simulated datasets of the 3GPP APIs, such as NEF, addressing the critical data scarcity challenge for applying AI algorithms to B5G/6G networks. The produced simulated dataset produces GPS coordinates (i.e., latitude and longitude) of the mobile user alongside a timestamp.

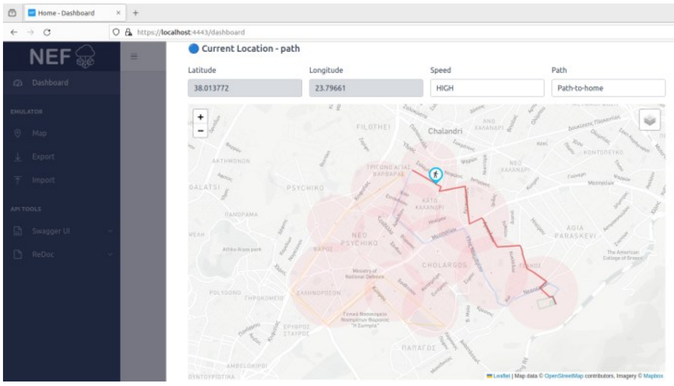

Fig. 5. The GUI of the NEF Simulator, depicting the scenarion under study

In this paper, we simulate a mobility scenario, where a user is moving across four distinct phases: Phase 1: routes from home to work, Phase 2: working hours spent at employer premises, Phase 3: routes from work back to home and Phase 4: random routes outside during no-working hours. The requirement from the user is the provision of low latency when he/she is located at the work, while this requirement is not valid when he/she is located at any other place.

In order for the simulation to be realistic, the duration of each simulation varies from 7 to 10 hours, while multiple days and routes were simulated. The generated dataset stored in a csv file and afterwards is processed by pattern extraction framework and the prediction AI model.

B. Pattern Extraction Framework

The dataset that was produced for the above-mentioned scenario includes the GPS coordinates along with their respective timestamps. The intention of this work is to extract possible routines of the user in order to be able to predict his/her movement, and by extend to identify the optimal place for the UPF deployment, with the aim to achieve a better quality of experience for the user.

The dataset includes both spatial (latitude, longitude) and temporal (timestamp) information of the user's movement. The developed framework is trying to exploit these two different dimensions and combine both these aspects in order to reach a better result.

Clustering is an unsupervised learning technique which discovers inherent structure within a dataset and enables the identification of patterns present in the data by grouping similar data points together into distinct clusters. When dealing with GPS coordinates, the spatial distribution of data along with the density variations are important parameters that should be taken into consideration in order to choose the appropriate clustering algorithm.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was chosen to perform the clustering. DBSCAN is a density-based clustering algorithm which groups together data points that are closely packed together. That attribute is particularly useful as far as spatial data are concerned because it enables the identification of clusters which have arbitrary shape. For the temporal data, assuming that timestamps in close proximity represent continuous movement with high probability, DBSCAN can also group the timestamps into different segments based on the density.

The results for the clustering depending on GPS coordinates are depicted in Fig. 6. Two different clusters were identified. The first cluster (red circles) includes the movement of the user i) from home to work; ii) activity during work; and iii) returning from work to home, while the second cluster (green circles) contains the random movement after the working hours.

Fig. 6. DBSCAN clustering results on spatial data (2 clusters)

For the clustering depending on the temporal dimension, a preprocess of the data was deemed necessary. Initially the timestamps were converted into Unix timestamps and then they were standardised (removing the mean and scaling to unit variance). The number of clusters that were identified was 8. The first cluster contains the movement from home to work. Clusters 2-6 correspond to the activity during work while cluster 7 and 8 include the return from work to home and the random movement after the working hours respectively. The outcome of the temporal analysis is presented in Fig. 7.

Fig. 7. DBSCAN clustering results on temporal data (8 clusters)

The parameters used for both spatial and temporal clustering are presented in Table I.

TABLE I. CLUSTERING PARAMETERS AND RESULTS

Clustering Dimension	**DBSCAN parameters**		**Number of Clusters**
	eps	***min_samples***	
Spatial	0.0002	3	2
Temporal	0.05	5	8

C. Placement Decision and Results

The prediction of a person's movement is a very challenging task due to several factors. Human behaviour is highly complex, and it can be influenced by various factors such as environmental conditions, personal preferences, random events, etc. In this work, we try to exploit some temporal dynamics (e.g. daily routine) along with some contextual information (e.g. working hours) in order to predict the future location of the user.

When the user starts moving, the location along with the timestamp are fed to an AI agent, which classifies them to the respective spatial and temporal clusters that we have identified previously. The results from the two classification tasks are combined in order to predict, with higher probability, the path of the user. During user's movement, each time new data for his/her location is produced, they undergo the same procedure. The new results are then compared with the current path prediction leading to an increase, or a decrease, of the probability that the forecasted path is the correct one.

Knowing the trajectory of the user, the problem of the placement of the UPF depends only on the KPI/KVI that we want to optimise (in our scenario is latency/QoE). Another AI agent, using the spatial information from the extracted forecasted path, identifies the optimal placement (in terms of latency) for the cloud native UPF and proposes it to the reasoning engine. This proximity minimizes the latency and facilitates service delivery that meets applications' strict performance requirements and for that reason the reasoning engine decides to accept the AI agent's proposal and orders its execution.

For evaluation purposes, we consider two possible cloud locations that the UPF can be placed, one (i.e. the edge cloud) within the working premises of the user (i.e. clusters 2-6 of temporal clustering dimension) and another one (i.e. the central cloud) located in a central/remote location of the area that the specific user is moving.

Based on the identified clusters, the prediction of the user's trajectory and the user's intent, the AI coordination layer decides what is the optimal UPF placement for the specific user in order to reassure that the service provision will continue to meet the low latency requirements.

Considering Round Trip Time in msec to be the time that the packet takes to travel from the mobile user to the application and back, we performed initially some baseline measurements in order to assess the Round-Trip Time (RTT) when the UPF has been placed at the edge or at the central cloud. For this reason a containerized implementation of the Open5GS core network was used, which has been properly configured to support dynamic migration of the UPF function between two placement options: i) The UPF to be deployed at a central cloud away from the working space of the user (in our case the working space is the lab), or ii) The UPF to be deployed locally at the edge cloud (within the lab/work area of the user).

Figure 8 shows the two RTT assessments for these two UPF placements, where a significant performance deviation in terms of latency is observed.

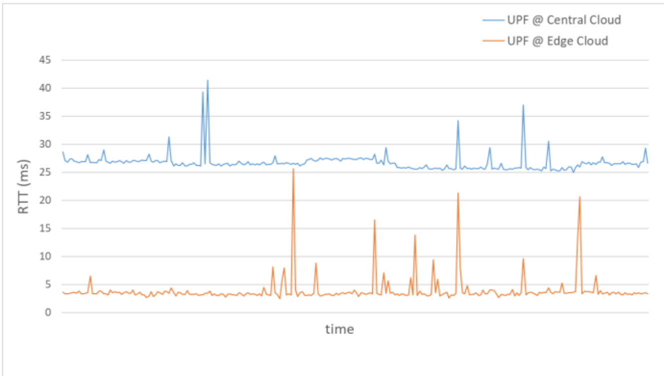

Fig. 8. RTT of the two UPF placements at the edge and the central cloud

Exploiting now the different granularity on the identified number of clusters of the two clustering methods (i.e. spatial and temporal), we apply the deployment decision algorithm once considering the spatial clustering dimension and afterwards considering the combined one.

As Fig. 9 depicts, in the case of the spatial clustering dimension, the UPF placement is decided to be at the edge cloud for the whole cluster #1 (i.e. the movement of the user: i) from home to work; ii) activity during work; and iii) returning from work to home), which results to over provision of the edge, since it includes also routes outside of the work premises that the UPF should not have been placed at the edge.

In the case of combined clustering dimension, higher granularity is achieved and the UPF placement at the edge cloud is performed only for the predicted clusters 2-6, which actually correspond to the routes that the user is located within the working area and the optimal user-centric core network provision is achieved.

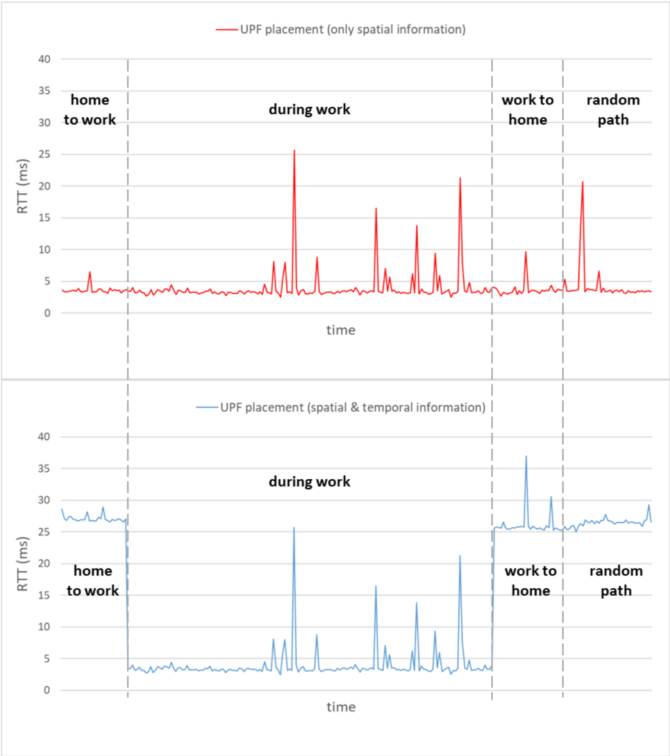

Fig. 9. UPF placement based on combined clustering

Therefore, it is deduced that the use of the 3GPP APIs as data source for feeding an AI agent that has been tightly integrated with the core for performing cognitive coordination is efficient and successfully manages to provide a user-centric core network placement according to the user's requirements.

V. Future Directions and Challenges

Currently the experimental validation of the proposed user-centric 6G network architecture relies on a simulated mobility scenario and not actually generated UE data. The authors are planning to expand the validation activities of the proposed user-centric 6G network with AI automation to a physical experimentation testbed by collecting real user-data and applying the AI-automation in the UPF selection in large scale field trials. Moreover, different decision models are planned to be used for benchmarking their performance and selecting the most optimal and efficient model for this task.

VI. Conclusions

This paper has presented a novel user-centric 6G network architecture that integrates an AI coordination layer in order to automatically decide the appropriate placement for the NFs according to the user requirements. The proposed architecture has been validated using simulated/synthetic data to train the AI model for NFs/UPF placement at the optimal location for low latency provision of a containerized Open5GS core network. The experimental validation showed that the proposed AI-assisted user-centric network implementation can successfully provide the requested low latency access to the user when she/he is located within a specific area (i.e. employer premises).

Acknowledgment

The work presented in this paper is partially supported by i) the SAFE-6G project that has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101139031, ii) the 6G-SANDBOX project that has received funding from the SNS JU under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101096328, iii) the ENVELOPE project that has received funding from the SNS JU under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101139048 and iv) the aerOS project that has received funding under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101069732.

References

[1] Bahare, M. K., Gavras, A., Gramaglia, M., Cosmas, J., Li, X., Bulakci, Ö., Rahman, A., Kostopoulos, A., Mesodiakaki, A., Tsolkas, D., Ericson, M., Boldi, M., Uusitalo, M., Ghoraishi, M., & Rugeland, P. (2023). The 6G Architecture Landscape - European perspective. Zenodo. https://doi.org/10.5281/zenodo.7313232

[2] X. Yan, X. An, W. Ye, M. Zhao, Y. Xi and J. Wu, "User-Centric Network Architecture Design for 6G Mobile Communication Systems," 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Gothenburg, Sweden, 2023, pp. 305-310, doi: 10.1109/EuCNC/6GSummit58263.2023.10188283.

[3] C. Li, W. Hua, A. Ming and S. Shaohui, "AI-native User-Centric Network for 6G," 2022 IEEE/CIC International Conference on Communications in China (ICCC Workshops), Sanshui, Foshan, China, 2022, pp. 494-499, doi: 10.1109/ICCCWorkshops55477.2022.9896699.

[4] D. Tsolkas and H. Koumaras, "On the Development and Provisioning of Vertical Applications in the Beyond 5G Era," in IEEE Networking Letters, vol. 4, no. 1, pp. 43-47, March 2022, doi: 10.1109/LNET.2022.3142088.

[5] I. L. Pupo, C. C. Pastor, "An intelligent scheduling for 5G user plane function placement and chaining reconfiguration", Computer Networks, Vol. 237, 2023, https://doi.org/10.1016/j.comnet.2023.110037

[6] G. Makropoulos, D. Fragkos, H. Koumaras, N. Alonistioti, A. Kaloxylos and F. Setaki, "Exploiting Core Openness as Native-AI Enabler for Optimised UAV Flight Path Selection," 2023 IEEE Conference on Standards for Communications and Networking (CSCN), Munich, Germany, 2023, pp. 254-258, doi: 10.1109/CSCN60443.2023.10453206.

[7] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou and W. Zhuang, "Holistic Network Virtualization and Pervasive Network Intelligence for 6G," in IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 1-30, Firstquarter 2022, doi: 10.1109/COMST.2021.3135829.

[8] N. Gkatzios et al., "A Chatbot Assistant for Optimizing the Fault Detection and Diagnostics of Industry 4.0 Equipment in the 6G era," 2023 IEEE Conference on Standards for Communications and Networking (CSCN), Munich, Germany, 2023, pp. 124-129, doi: 10.1109/CSCN60443.2023.10453129.

[9] C. -X. Wang et al., "On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds," in IEEE Communications Surveys & Tutorials, vol. 25, no. 2, pp. 905-974, Secondquarter 2023, doi: 10.1109/COMST.2023.3249835.

[10] D. Fragkos, G. Makropoulos, A. Gogos, H. Koumaras and A. Kaloxylos, "NEFSim: An open experimentation framework utilizing 3GPP's exposure services," 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Grenoble, France, 2022, pp. 303-308, doi: 10.1109/EuCNC/6GSummit54941.2022.9815829.