On the Deployment of an AI-driven Power Control Mechanism for D-MIMO in Beyond 5G Scenarios

Jordi Pérez-Romero[1], Ömer Faruk Tuna[2], Oriol Sallent[1], Nikolaos Bartzoudis[3], Elli Kartsakli[4], Oluwatayo Kolawole[5], Irene Vilà[1], Maria A. Serrano[6], Utku Gülen[2]

[1]*Universitat Politècnica de Catalunya (UPC), Barcelona, Spain;* [2]*Ericsson Research, Istanbul, Turkey;* [3]*Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Barcelona, Spain;* [4]*Barcelona Supercomputing Center, Barcelona, Spain;* [5]*Samsung R&D Institute, Staines, United Kingdom;* [6]*Nearby Computing, Barcelona, Spain.*

***Abstract*—The evolution beyond the Fifth Generation (5G) of mobile communications systems is expected to exploit different advanced radio access techniques, such as distributed multiple input multiple output (D-MIMO), together with an extensive support of Artificial Intelligence (AI) and Machine Learning (ML) solutions for different aspects of network optimization. In addition, these systems will also benefit from the evolution of edge computing technologies through the availability of computing and storage capabilities that will span from the radio access network nodes to the cloud, forming an edge-to-cloud compute continuum. The availability of these features will provide the beyond 5G and 6G networks with an increased flexibility for deploying the different AI/ML-based optimization functionalities, trading-off aspects such as the delay requirements, the computational complexity or the availability of data for feeding the AI/ML models. In this context, this paper takes as a reference the evolved edge computing architecture of the VERGE project and focuses on the possibilities to deploy on top it an AI/ML model for D-MIMO power control. The paper discusses several options to host the associated AI/ML model lifecycle management functionalities on the edge-to-cloud compute continuum and elaborates on the implications of the considered approach in terms of standardization.**

Keywords— Beyond 5G, 6G, edge computing, AI/ML-based optimization, distributed MIMO, edge-to-cloud compute continuum

I. INTRODUCTION

The evolution of communication networks beyond 5G (B5G) and towards 6G is expected to deal with the diverse and challenging requirements of innovative, immersive, and real-time vertical services [1][2]. Emerging architectural designs supporting network function virtualization (NFV) and Radio Access Network (RAN) disaggregation offer enhanced flexibility and scalability, whereas Artificial Intelligence (AI) and Machine Learning (ML)-enabled solutions leverage monitoring data to optimize both network and application performance, achieving zero-touch closed-loop automation. Moreover, advanced radio access techniques, such as distributed multiple input multiple output (D-MIMO), are expected to play an important role in B5G/6G systems, enabling a more uniform user experience across the coverage area while meeting the service requirements [3].

Fostered by the growing capabilities of communication infrastructures, emerging applications such as eXtended Reality (XR), Virtual Reality (VR) or holographic communications, which involve computationally heavy functions, such as rendering or spatial computation [4], are expected to progressively enter the mobile ecosystem. Nevertheless, the limitations of end user devices – phones, VR glasses, etc. – in terms of size, energy, computational capacity, and cost, necessitate the offloading of heavier tasks to more powerful computing elements, typically residing at the cloud. However, cloud computing strategies are no longer capable of meeting the latency requirements of such applications. As such, edge computing has been rapidly evolving as a novel computing paradigm that brings computational power and resources closer to where the data is generated, thus considerably reducing response times with a much lower carbon footprint [5].

Hence, the synergy between B5G and edge computing can provide computing and storage capabilities for applications residing at the boundary of operators' networks [6]. This can be particularly relevant when deploying functionalities with varying needs in terms of communication and computing requirements as well as data consumption, such as those associated with the lifecycle management of AI/ML-based solutions (e.g. training, inference, or model monitoring), suggesting that these functionalities can be hosted at different points of the network. For example, an offline training stage can require massive amounts of data and large computational requirements to build an accurate AI/ML model, so it can run on a central cloud. In contrast, an inference stage that uses a previously trained model can mostly require a fast reaction time to quickly adapt to network variations, so it can be better hosted on an edge server.

With all the above, this paper considers a B5G scenario built upon the following components: (1) a disaggregated RAN with advanced features such as D-MIMO, (2) an edge-to-cloud compute continuum consisting of a heterogeneous pool of connected edge and cloud resources, (3) extensive support of AI/ML-based techniques for the automated management of network and compute resources. To support this B5G scenario, this paper takes as a reference the system architecture of the VERGE project [7], which addresses the significant complexity that the overall B5G network management embraces due to the interplay of diverse functionalities and technologies from multiple domains. Then, the paper focuses on the deployment of an AI/ML model for D-MIMO power control from [8] on top of the VERGE architecture. The use of AI/ML models for this type of control mechanisms poses challenges on aspects such as the adequate placement of the inference and training, or the acquisition of the required data for these procedures. In this

This work is part of VERGE project, which 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 101096034. Samsung Research UK participants in Horizon Europe Project VERGE are supported by UKRI grant number 10071211. The work of Ericsson Research Turkey authors is also partly funded by The Scientific and Technological Research Council of Turkey (TUBITAK) through the 1515 Frontier Research and Development Laboratories Support Program under Project 5169902. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

direction, the contribution of the paper is the analysis of possible deployment options for the associated AI/ML model lifecycle management functionalities on the edge-to-cloud compute continuum. Moreover, the paper also discusses the implications of the considered approach in relation to standardization.

The paper is organized as follows. Section II describes the edge-enabled B5G architecture of the VERGE project. Then, Section III presents the specific applicability example related with AI-driven power control for D-MIMO for enhancing the operation of the RAN and Section IV discusses the different deployment options of this strategy using the components of the reference architecture and elaborating the corresponding workflows for lifecycle management. Then, Section V discusses the relationship with different standardization activities. Finally, conclusions are summarized in Section VI.

II. Reference Edge-enabled B5G Network Architecture

Edge computing involves an ecosystem of highly heterogeneous computing elements, spanning from embedded devices, intelligent base stations (BSs), edge and fog servers, to home gateways, and micro-datacenters, which may be located practically everywhere across the path between the end-devices, the access network and the central cloud [9]. At the same time, the current trend for disaggregated and softwarized RAN design for next generation mobile BSs, aligned with efforts such as the Open RAN (O-RAN) Alliance [10], splits RAN functionalities between distributed units (DUs) with radio capabilities and central units (CUs) that host other upper layer RAN functions, offering new opportunities for a flexible deployment and intelligent network management.

The computational and storage resources are highly heterogeneous and distributed across multiple layers, including edge sites close to the UEs, collocated with the BSs or relays (i.e., near edge), edge sites at aggregation points (i.e., far edge), cloud servers and even UEs with high computational capacities (e.g., equipped with onboard embedded processors). This pool of resources composes an *edge-to-cloud compute continuum* where virtualized services can be flexibly deployed and executed based on their requirements and available infrastructure. Such services may include cloud-native vertical applications, RAN and core virtual network functions (VNFs), as well as AI-enabled functions for network optimization and automation. The compute continuum management is carried out by a multi-site edge orchestration layer, coordinating the use of multiple edge sites and interfacing with the telco-cloud where operator-specific management operations are hosted, as well as external cloud infrastructures (public or private) where additional services may be deployed.

Fig. 1 depicts the system architecture for an edge-enabled B5G network considered by the VERGE project [7]. The architecture is based on three main pillars. The first one is “edge for AI” (Edge4AI), which is a flexible, modular and converged edge platform design that unifies the lifecycle management and closed-loop automation for cloud-native applications, multi-access edge computing and network services across a multi-domain edge-cloud continuum illustrated in the right-side of Fig. 1. The key components of Edge4AI are: (i) The Orchestration, Management and Control layer, which orchestrates services and infrastructures and enables the control of the B5G RAN elements. It includes functionalities at both multi-site and edge-site level. (ii) The Cognitive Framework, which enables the Lifecycle Management (LCM) of AI/ML solutions. (iii) The Distributed Knowledge Base (DKB), where all generated knowledge (e.g., trained AI/ML models, datasets, metadata) are registered. (iv) The Data Access layer, responsible of collecting all the relevant observability data. (v) The virtualization layer, providing a unified view of the communication and computational resources, forming an edge-to-cloud compute continuum that is tightly integrated with the B5G communication fabric.

The second pillar of the VERGE architecture is “AI for edge” (AI4Edge), namely a portfolio of AI-based solutions that leverage the multitude of information and metrics provided by the monitoring mechanisms to manage and orchestrate the computing and network resources. It encompasses functionalities for training and validating the AI models, for monitoring and managing them and for supporting the inference stage, in which trained models run on specific components of the Edge4AI, such as the orchestrators, RAN controllers, etc. The third pillar is "security, privacy and trustworthiness for AI" (SPT4AI), which provides a suite of methods and tools to ensure the security of the AI-based models against adversarial attacks, the privacy of data and models, the explainability capabilities of AI-driven decision making and the safe training and execution of such models. Further details on the VERGE architecture can be found in [7] and [11].

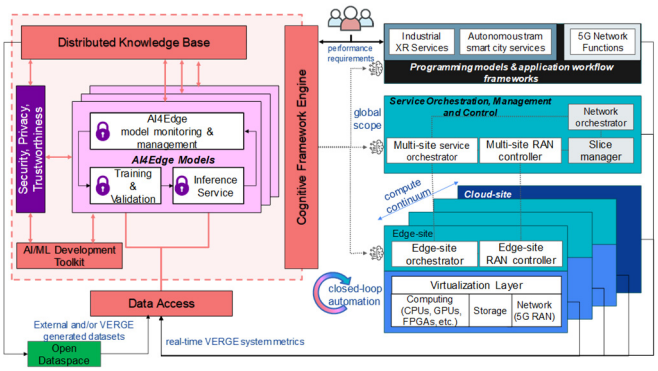

Fig. 1. VERGE system architecture.

III. AI-driven Power Control for D-MIMO

D-MIMO is a promising RAN technology to handle the complex propagation conditions that exist in certain environments (e.g., a factory plant), where providing a high service level reliability (e.g., >99.5% as defined in [11]) is challenging. D-MIMO consists of deploying multiple radio units (RUs), connected to a central processor via wireless/wired fronthaul links. The RUs are distributed in a region to increase the macro diversity and mitigate path-loss and shadowing effects [3]. D-MIMO is expected to be one of the main multi-antenna features to be used in deployment options in 6G. By its distributed nature, D-MIMO provides more uniform performance to users compared to collocated massive MIMO currently used in 5G, where many antenna elements are installed onto a large tower, although at the expense of additional deployment costs due to the increase in number of RUs.

Power control, which allocates the available power of RUs to the users according to channel conditions, is crucial to get the full benefit of D-MIMO as it helps to optimize both spectral and energy efficiency. There are actually three important benefits for applying power control techniques in D-MIMO: (i) Interference management, which allows to

optimize the spectral efficiency of the users by enhancing the desired signals and mitigating the interference signals. (ii) Energy efficiency, which should be optimized to make future networks more sustainable. If all UEs are close to the RU, the RU can decrease the total transmit power to minimize the power consumption. (iii) Connectivity management, to ensure reliability under varying channel conditions, which is especially important for critical applications. Even if the channel for a UE is good in general, there may be some moving blockers that make the channel conditions bad at an unknown time instant. In such conditions, the power level can be adjusted to increase the performance.

There are different strategies in the literature to control the power in a D-MIMO network, with Max-Min Fairness being one of the most preferred ones. The idea is to maximize the minimum spectral efficiency in the network by adjusting the power coefficients. This is actually a constrained optimization problem because for each RU there is a transmit power constraint (power amplifiers and antennas cannot transmit unlimited power). There is an exact analytical solution for this problem by using sequential second order cone programming. However, the computational complexity of the exact solution is extremely high, especially as the number of RUs and UEs in the network gets larger. That is why low complexity AI/ML based solutions are proposed to approximate the exact solution. Specifically, the AI regression model shown in Fig. 2 is used for the power allocation task. The input to this regression model is a vector of channel information coefficients for each pair of RU and UE. These coefficients are denoted as $\beta_{m,k}$, where m=1,..., M with M the number of RUs and k=1,...,K with K the number of UEs. The channel information coefficients can be estimated e.g. using uplink pilots, i.e. sounding reference signals (SRS), when Time Division Duplex (TDD) is used. The output is a vector of power control coefficients, denoted as $\eta_{m,k}$. Both the input and output vectors are of size $M \cdot K$.

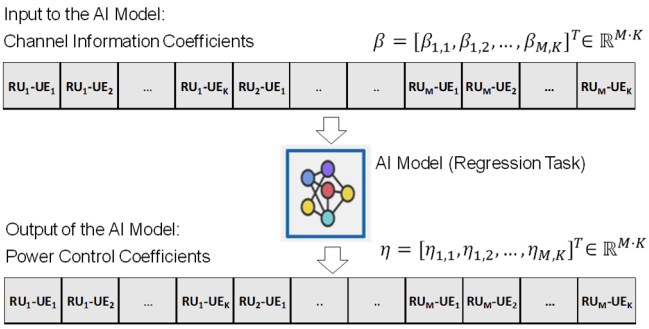

Fig. 2. Inputs and Outputs of the AI model in CP.

However, putting this type of AI solution into practice involves some challenges. The first problem is related with the dynamic nature of the wireless domain and the inability of standard Deep Neural Networks (DNNs) to process variable input and output sizes. Most of the prior works in literature consider a simple scenario where the number of RUs and UEs in the network are fixed so that a traditional DNN architecture will work with fixed size input/output. However, in practice, the number of UEs connected to the live network is not fixed and dynamically changes in a short period of time. This makes standard DNN-based approaches not applicable. The second problem is that existing supervised learning solutions require excessive precomputation of training data. One needs to solve the analytical solution beforehand for the available training data. Solving the analytical solution is computationally expensive and this makes the supervised learning option practically not an ideal solution.

To effectively handle the issue of variable input/output size for DNN implementation, the solution considered here is to preprocess the channel input vector [8]. Instead of using a 1-d input vector, we suggest reshaping the input vector to a 2-d matrix with dimensions denoted as $\mathrm{R}^{\mathrm{M}\times\mathrm{K}}$. We then adjust the number of neurons in the network layers in such a way that keeps the output of all layers of the network as 2-d, by maintaining the second dimension always equal to K, which is the number of UEs. In this way, the effect of variable number of UEs in the network is eliminated. We will only need to consider the number of the RUs in the network, which might be fixed as M, and the variability of the number of UEs will be handled with the 2-d flow of data through the network as shown in Fig. 3. Finally, to satisfy the per-RU transmit power constraint for each RU, we suggest using sigmoid activation function at the DNN output layer and apply a normalization operation to output power coefficient values.

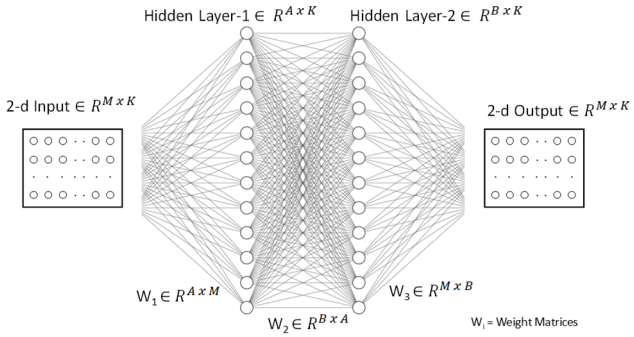

Fig. 3. Proposed operation with 2-d input/output.

For the second practical limitation, we propose the use of unsupervised learning. This way, we avoid the excessive precomputation of training data that supervised learning would necessitate. Our objective for this unsupervised learning task is to maximize the minimum user rate (spectral efficiency) in the network. Besides, applying unsupervised learning methodology to this kind of regression task will bring additional benefits of providing robustness to any kind of input variations in the training stage. Moreover, from a security perspective, the research work in [8] has demonstrated that using an unsupervised learning technique might lessen the influence of potential measurement mistakes or poisoning attempts on the training of AI models for these kinds of regression tasks. This is due to the fact that the whole purpose of a regression task is to approximate the underlying nonlinear function that maps the input-output pairs. Therefore, applying any kind of intentional or unintentional perturbation to the input vector will not change the model learning behavior in the unsupervised learning approach, as the model will still be trying to find the best output based on the given input data.

IV. Deployment of AI-driven Power Control in B5G Scenarios

This section discusses the deployment of the presented AI-driven power control for D-MIMO in B5G scenarios, which exploits the presented VERGE system architecture. To that end, a network managed by a Mobile Network Operator (MNO) is considered that provides service over a certain geographical area. The radio communication makes use of D-MIMO to facilitate a more homogeneous coverage and uniform service experience, avoiding coverage holes in specific areas. In particular, the geographical region where

the service is provided is organized in Service Areas (SA), each one covered by a number of RUs. The coverage in each SA is provided by one CU, one DU and multiple RUs when D-MIMO is used or just one RU if no D-MIMO is used. In this context, the functional components that constitute the workflow of the lifecycle of the D-MIMO power control solution are described first in Section IV.A. Then, in order to highlight the versatility of the VERGE architecture Section IV.B presents an example with different possibilities of deployment of the AI/ML functionalities in the considered architecture.

A. *D-MIMO power control model lifecycle workflow*

When using the D-MIMO power control solution explained in Section III, since the operating environment of each SA is different, the AI/ML models have to be trained and executed separately for each SA. Fig. 4 illustrates the different processes that constitute the lifecycle workflow of the AI/ML model for D-MIMO power control in one SA. The processes are grouped into two main stages, namely the training and the inference stage. The *training stage* makes use of RU-specific measurements that are pre-processed at the *data ingestion* process to build the training data to be used by the *model training* process, i.e. the unsupervised learning task mentioned in Section III that learns the adequate DNN weights for the SA. Specifically, the training data should consist of multiple realizations of the channel information coefficients $\beta_{m,k}$ for different users and RUs that are representative enough of the conditions experienced in a given SA (see Fig. 2). Since the training process can be executed offline, these measurements can be collected from the network, stored and provided to the data ingestion process whenever a training is required.

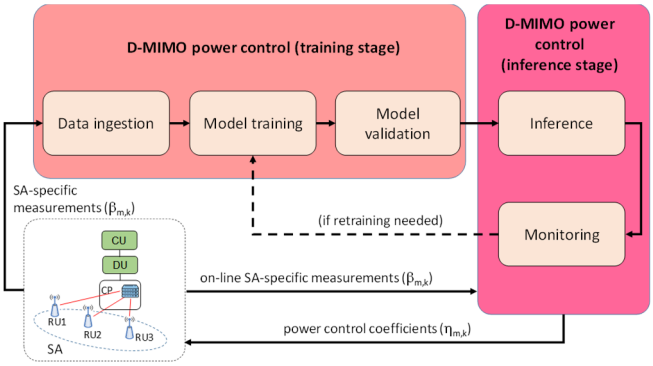

Fig. 4. Workflow of the D-MIMO power control model lifecycle for a cell.

Once trained, the resulting DNN can be validated against a separate dataset by the *model validation* process to assess its performance and make sure it is ready for deployment on the network. Afterwards, the *inference stage* uses the trained DNN model and takes as input real-time (on-line) measurements of the channel information coefficients $\beta_{m,k}$ of the different UEs and RUs to dynamically decide the power control coefficients $\eta_{m,k}$ to be allocated to each UE in each RU in the SA. Moreover, this stage also includes a *monitoring* process to assess the actual performance of the model. As a result of this assessment, it may be possible to request a retraining, e.g. if the conditions used in the training dataset are substantially different from the ones experienced currently by the SA.

B. *Example of D-MIMO power control model deployment*

Fig. 5 presents an example of deployment of the different stages of the AI-driven power control solution for D-MIMO in the network of an MNO that provides service across a wide area. The network is used to provide certain services with stringent latency requirements that involve the execution of computation tasks, e.g. rendering for XR services. Accordingly, the deployment includes a multiplicity of RUs organized in different SAs that ensure adequate coverage conditions when the users move across the network, facilitating smooth handovers between SAs. The figure also illustrates that, depending on their environmental characteristics in terms of geometry, served UEs, etc. the MNO can decide to deploy D-MIMO in some selected SAs, e.g. SAs #1, #2 and #6 in the example, while other ones, e.g. SAs #3, #4 and #5, may not require this feature.

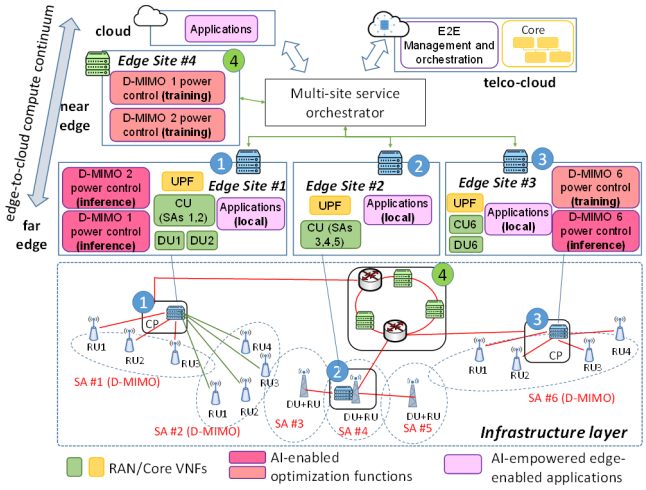

Fig. 5. Example of deployment of AI-driven power control for D-MIMO.

The edge-to-cloud compute continuum of the VERGE architecture explained in Section II, is reflected in the example of Fig. 5 through the existence of four edge sites, numbered #1 to #4, and a cloud. They host computational, storage and networking resources where different applications or network functions can be executed. Edge sites #1, #2 and #3 correspond to the *far edge*, because they are co-located with the RAN equipment of different SAs. In turn, edge site #4 is located at an aggregation point in the transport network, so it corresponds to the *near edge* in the continuum. The different edge sites and the cloud are managed by the *multi-site service orchestrator*.

The availability of the multiple edge sites offers a lot of flexibility for placing the D-MIMO training and inference stages of the different SAs in accordance with the needs and the deployed resources. Regarding the inference, given that the power control needs to dynamically react to the fading variations and to the variations in the number of scheduled UEs at each time slot, it should run at the computational resources of the *far edge* co-located with the RAN equipment of each SA, and particularly with the DU, most likely being part of it because this will provide faster reactivity. Thus, in the example of Fig. 5, the edge site #1 includes the central processor that hosts the D-MIMO power control inference of SA #1, composed of three RUs, and SA #2, composed of four RUs. The two SAs are handled by two DUs associated with the same CU. The virtualized functions of these DUs and CU are also hosted at the edge site #1. Similarly, the inference of SA #6 is executed at the edge site #3, which is also hosts the virtualized DU and CU functions for this SA. Moreover, both edge sites #1 and #3 also host the applications with low latency requirements whose traffic needs to be processed locally at the edge. For that purpose, they include a local User

Plane Function (UPF) to transfer the packets of these applications to the local processor. This is similar to other edge sites deployed in the network, such as edge site #2, co-located with SA #4 that does not use D-MIMO. In this respect, the example illustrates the case in which SAs #3, #4 and #5 have a virtualized CU running at this edge site, together with the UPF and the applications, while the RU and DU are assumed to be implemented as physical network functions.

Regarding the training stage, on the one hand, it involves the processing of training data composed of multiple SA-specific measurements, thus having larger computational requirements than the inference. However, the training process does not have tight delay requirements, as it can be executed off-line. Thus, it provides more flexibility to be hosted at different locations of the edge-to-cloud continuum. Two different possibilities are shown in the example of Fig. 5. The first one corresponds to the D-MIMO training of SA #1 and #2, which is executed in the *near edge*, particularly at the edge site #4 (i.e. the *near edge*) co-located with an aggregation point of the transport network. In this way, the training process can benefit from the larger computational capabilities available at this edge site. Besides that, running this at an aggregation point avoids the need for the SA-specific measurements needed for the training, which can represent large data volumes, to be transferred deeper in the network (e.g. if the training was done at the cloud), thus enabling a more efficient use of transport resources only at the expense of some additional data storage at the edge site. The second possibility shown in Fig. 5 corresponds to SA #6, in which the training stage is executed locally at the same edge site #3 that hosts the inference. This could represent a situation in which the edge site has enough computational resources to support all its processes and/or the case of a remote site in which keeping the training data locally allows reducing the load in the transport links.

It is also worth mentioning that the example of deployment shown in Fig. 5 can be flexibly modified through the multi-site service orchestrator. The placement of the D-MIMO training function and the applications across the edge-to-cloud compute continuum can be decided based on criteria such as resource utilization efficiency, traffic load or energy consumption, following a human-based, rule-based, and/or AI-based orchestration action.

V. Relationship with Standardization

This section analyzes different aspects of the AI-driven D-MIMO power control deployment in the VERGE architecture from the perspective of their relationship with current standardization activities at different fora.

A. AI/ML model provisioning

The standardization efforts of 3rd Generation Partnership Project (3GPP) around the 5G system, with the introduction of network functions (NFs) for a service based architecture (SBA) [12], have paved the way for AI/ML-based network solutions. Regarding disaggregation, 3GPP supports deployment flexibility, allowing for multivendor deployments of CU/DU on the same site or across different sites [13]. In turn, the O-RAN alliance architecture has introduced an additional level of RAN disaggregation enabling a low layer functional split between a DU and an RU [14].

Regarding the support of AI/ML models in the RAN, the O-RAN alliance architecture includes specific functional elements for RAN control, such as the near-Real Time (RT) RAN Intelligent Controller (RIC) and the non-RT RIC [14]. These controllers can host software modules of inference/training stages of AI/ML models provided by third parties, e.g. in the form of the xApps hosted in the near-RT RIC, which have control loops between 10ms and 1s, or the rApps hosted in the non-RT RIC, which have control loops above 1s. Moreover, ongoing studies in the O-RAN next Generation Research Group (nGRG) - Research Stream 02 (RS02) focusing on Architecture towards 6G O-RAN [15] consider the distributed application (dApp) [16], which could form real-time control loops reaching even below 1 ms time scales. In general, the provision of AI/ML models by third parties as xApps, rApps or eventually dApps requires properly defined interfaces with the managed nodes to collect the necessary measurements and to provide the control commands to configure the involved nodes. In this respect, O-RAN architecture has already defined the E2 interface between near-RT RIC and DU and the O1 interface between non-RT RIC and CU/DU.

For the case of the D-MIMO power control considered in this paper, the decisions made by the inference stage modify the power control coefficients $\eta_{m,k}$ based on the channel information coefficients $\beta_{m,k}$. These decisions are made at the time scale of the scheduling algorithm that modifies the UEs that transmit in each time slot, thus being in the order of 1 ms. Thus, if the inference stage of the algorithm had to be implemented in the context of the O-RAN architecture, its timing constraints would only fit with the time scale of operation of a dApp. In such a case, the dApp should have access to Layer 1 (L1) data, e.g. to collect information from sounding signals that facilitates the determination of the channel information coefficients, and it should be tightly coupled with the scheduling algorithm. However, the interfaces currently defined by O-RAN are not able to provide such information. In this regard, third party dApp implementations would require the definition and standardization of new service models and the co-location of RAN and AI/ML functions on a shared hardware platform [17]. Given these limitations, the most feasible implementation of the inference would be a proprietary solution in which the AI/ML model is directly integrated in the DU and provided by the DU vendor.

In turn, the training stage, which does not have tight delay constraints and can be executed off-line, would accept an implementation e.g. as an rApp of the non-RT RIC. In such a case, a key aspect would be the provision of the channel information coefficients that will be used as training data. These coefficients should be collected by the DU and provided to the rApp. Given that current O1 interface does not include this type of data, a proprietary interface should be used in case.

B. Multi-site service orchestrator

The ETSI Zero-touch network and Service Management (ZSM) Industry Specification Group (ISG) addresses the challenges posed by the complexity and operational agility required to handle modern network infrastructures, by enabling end-to-end automation of network and service management [18]. To enable largely autonomous networks, an end-to-end architecture framework is being designed for closed-loop automation and optimized for data-driven AI algorithms.

The multi-site service orchestrator of the VERGE architecture (see Fig. 1) incorporates features that make it compliant with the ZSM architectural framework, like the

realization of end-to-end and domain specific closed loop service management thanks to a global visibility of the infrastructure and services metrics through support for deployment and interaction of distributed monitoring agents. The orchestrator also offers a number of open interfaces facilitating the integration of AI/ML-based algorithms on the closed loops for supporting service orchestration decisions, e.g. related to function placement. All these features allow an autonomous and dynamic decision-making process for the placement of the training service of the considered D-MIMO power control strategy.

The VERGE multi-site service orchestrator is also compliant with the specifications of the ETSI Multi-access Edge Computing (MEC) ISG that aims to define technical requirements, architecture, and interfaces that ensure interoperability and scalability of MEC solutions. As an example, it provides the functionalities of the MEC orchestrator or the MEC platform manager, as well as the interfaces Mm1 to Mm3 [19].

C. Hosting AI/ML functions at edge sites

As discussed in sections II and IV, the sites of the edge-to-cloud compute continuum host different types of applications that include the AI/ML training/inference functions, the end user applications (e.g. XR) and different virtualized network functions (e.g. CU, DU, UPF). Given that the requirements of these applications in terms of e.g. computing resources, virtualization approach, degree of isolation, latency or reliability can be very different, the ETSI MEC ISG has recently introduced the new concept of *MEC application slice*, defined as a logical MEC application service environment that provides specific MEC application functions and MEC services characteristics [20]. In this context, the D-MIMO power control training function can be regarded as a MEC application slice different from other MEC application slices that support the end-user applications.

The VERGE architecture of Fig. 1 includes a slice manager with the functionalities to jointly manage both the MEC application slices and the network slices that provide the network connectivity to transfer the data consumed by the applications. This joint management should ensure that the application requirements are met globally at both network and edge domains. Impacts of this joint management on standardization are currently being studied by the project, e.g. to identify network parameters and metrics that need to be exposed to the edge computing domain.

VI. Conclusions

In the context of the evolution of B5G mobile communications systems towards the exploitation of advanced radio access techniques, the support of AI/ML solutions and the use of edge computing capabilities, this paper has focused on the deployment of an AI-based power control solution for D-MIMO on top of the VERGE project architecture. This architecture encompasses an edge-to-cloud compute continuum that provides a high level of flexibility for deploying AI/ML functions, user applications or network functions at different sites. In this respect, the paper has discussed different possibilities for deploying the AI model training and inference stages of the D-MIMO power control solution within this edge-to-cloud compute continuum depending on the geographical scope, the access to the required data or the required reactivity.

Moreover, the paper has also analyzed different aspects of the solution from the perspective of their relationship with current standardization activities. Specifically, the analysis has identified the limitations of current O-RAN architecture to support a third-party provision of the AI/ML model thus concluding that a proprietary implementation in which the inference is integrated in the DU becomes more plausible. Besides that, the paper has studied the features of the multi-site service orchestrator that would enable an autonomous and dynamic decision-making process for placing the training function of the solution in the context of the ETSI ZSM ISG. Finally, following recent trends of ETSI MEC ISG, the possibility that this training function is provided as part of a MEC application slice has been discussed.

Acknowledgement

The authors would like to thank Miguel Berg for his insightful comments that have helped to improve the discussions in the paper.

References

[1] K. Samdanis, T. Taleb, "The Road beyond 5G: A Vision and Insight of the Key Technologies", *IEEE Network*, vol. 34, March/April, 2020.

[2] W. Saad, M. Bennis, M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems", *IEEE Network*, vol.34, May/June, 2020.

[3] O. T. Demir, E. Björnson, L. Sanguinetti et al., "Foundations of user-centric cell-free massive MIMO", *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, 2021

[4] F. Alriksson, et al. "Future network requirements for extended reality applications", *Ericsson Technology Review*, April, 2023.

[5] AIOTI, "High Priority Edge Computing Standardisation Gaps and Relevant SDOs", April, 2022.

[6] W. John, et al. "The future of cloud computing: Highly distributed with heterogeneous hardware", *Ericsson Technology Review*, May, 2020.

[7] E. Kartsakli, et al. "An Evolutionary Edge Computing Architecture for the Beyond 5G Era", *IEEE Int. Workshop on Comp. Aided Modeling and Design of Comm. Links and Networks (CAMAD)*, Nov., 2023

[8] Ö. F. Tuna and F. E. Kadan, "A Flexible, Efficient and Robust Method for AI-driven Power Control in D-MIMO", *IEEE Int. Mediterranean Conf. on Comms. and Networking (MeditCom)*, 2023

[9] A. Yousefpour et al., "All one needs to know about fog computing and related edge computing paradigms: A complete survey", *Journal of Systems Architecture*, vol. 98, September 2019.

[10] A. Akman, et al., "O-RAN Minimum Viable Plan and Acceleration towards Commercialization", O-RAN Alliance, June 2021.

[11] G. Kalem, M. A. Durmaz, S. Kosu (editors), "Use cases, requirements and initial system architecture", Deliv. D1.1 of VERGE. July, 2023.

[12] 3GPP TS 23.501 v19.1.0, "System Architecture for the 5G system Stage 2 (Release 19) ", Sept, 2024.

[13] 3GPP TS 38.401 v18.3.0, "NG-RAN; Architecture description (Release 18)", Sept. 2024.

[14] O-RAN Alliance, "O-RAN Architecture Description", O-RAN.WG1.OAD-R003-v12.00, June, 2024.

[15] S. D'Oro, et al.,"dApps for Real-Time RAN Control: Use Cases and Requirements", Report RR-2024-10 of the O-RAN next Generation Research Group (nGRG), October, 2024.

[16] S. D'Oro, et al., "DApps: Distributed applications for real-time inference and control in O-RAN", *IEEE Comm. Mag.*, Nov. 2022.

[17] "Advancements in RAN through AI -AI for RAN", SoftBank, Feb. 26 2024, https://www.softbank.jp/en/corp/technology/research/story-event/040/, Accessed: October, 2024.

[18] ETSI GS ZSM 012 v1.1.1, "Zero-touch network and Service Management (ZSM); Enablers for Artificial Intelligence-based Network and Service Automation", Dec. 2022.

[19] ETSI GS MEC 003 v3.2.1, "Multi-access edge computing (MEC); framework and reference architecture", April, 2024.

[20] ETSI GR MEC 044 v3.1.1, "Multi-access Edge Computing (MEC); Study on MEC Application Slices", April, 2024.