Edge4AI: Enabling Intelligent Edge Automation and AI Lifecycle Management for Beyond 5G Networks

Elli Kartsakli*, Nikolaos Bartzoudis†, Angelos Antonopoulos‡, Fred Buining§, Thijs Metsch¶, Joan Pujol Roig‖, Semiha Kosu**, Yansha Deng†† Jordi Pérez-Romero‡‡, Oriol Sallent‡‡,

*Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain; †Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain; ‡Nearby Computing S.L., Barcelona, Spain; §HIRO-MicroDataCenters BV Netherlands; ¶Intel Deutschland GmbH, Neubiberg, Germany;, ‖Samsung R&D Institute, Staines, United Kingdom; **Turkcell Technology, Istanbul, Turkey; ††King's College London, United Kingdom ‡‡Universitat Politècnica de Catalunya (UPC), Barcelona, Spain;

***Abstract*—Edge computing is a key enabling technology expected to play a crucial role in beyond 5G (B5G) and 6G communication networks. The EU-funded VERGE contributes to the evolution of edge computing by promoting an open, modular and distributed edge architecture that enhances B5G capabilities and supports next generation services. This paper highlights the core features of VERGE's "Edge for AI" pillar, which enables closed-loop automation and life cycle management (LCM) of cloud-native applications and network services across a unified edge-to-cloud compute continuum.**

***Index Terms*—Edge computing architecture; AI/ML life cycle management; B5G/6G evolution; edge-to-cloud compute continuum; closed-loop automation**

I. Introduction

The EU-funded research project VERGE introduces an evolved cloud-native edge architecture [1], powered by trustworthy, Artificial Intelligence (AI) solutions for the beyond 5G (B5G) and 6G communications. The VERGE architectural design is based upon three core conceptual pillars, namely the "Edge for AI (Edge4AI)", "AI for Edge (AI4Edge)" and "Security, Privacy and Trustworthiness for AI (SPT4AI)". The Edge4AI pillar provides mechanisms to develop, orchestrate, and monitor cloud-native applications and network services across a unified edge-to-cloud compute continuum, leveraging accelerated platforms. It also supports the life cycle management (LCM) of AI/Machine Learning (ML) solutions, developed within the AI4Edge pillar, for optimizing computing and network performance. Finally, SPT4AI focuses on ensuring the trustworthiness of the developed AI/ML models, addressing aspects on AI robustness, privacy, safety, and explainability. This short paper presents the key Edge4AI features and the relevant technological approaches that are adopted.

II. The Edge4AI pillar for edge computing automation, scalability and high performance

Fig. 1 outlines the overall Edge4AI pillar design. Its key components can be grouped as follows: 1) the **programming models and frameworks** (top of the figure), which are software assets that can be leveraged by developers during the design phase of application and network functions. 2) the **Orchestration, management and control (OMC)** layer

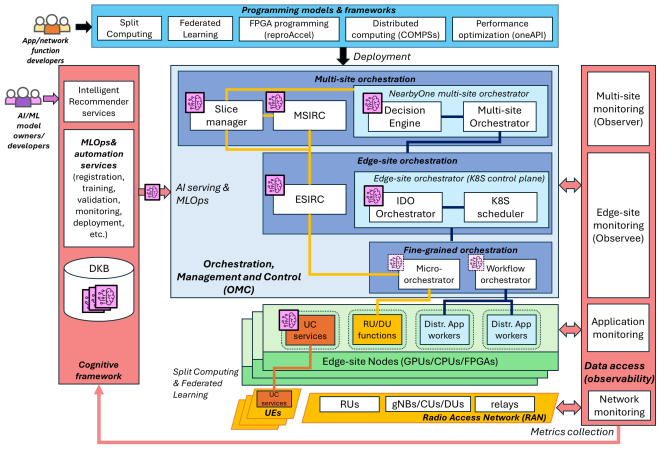

Fig. 1. Overview of the Edge4AI pillar building blocks

(center of the figure), handling the service LCM and the network intelligent control from a multi-site, edge-site and intra-node perspective, and 3) the **Cognitive Framework** (left side of the figure), handling the LCM of the AI/ML models and serving them to the OMC layer for intelligent decision-making, and the **data access layer** (right side of the figure) collecting data from the observability stacks and feeding them to the CF for AI model training and inference.

The key features of the Edge4AI design are described in continuation.

1) Edge4AI embraces a novel software abstraction layer to fully exploit the capabilities of Hardware (HW) accelerated platforms for ultra-high performance computation: Edge4AI supports a suite of programming models and frameworks that: 1) abstract the complexity of the underlying heterogeneous hardware infrastructure, facilitating developers in the implementation of network and application functions, thus improving programmability; 2) leverage the potential of diverse accelerated platforms, such as Graphics Processing Units (GPUs) and Field-Programmable Gate Arrays (FPGAs), by exploiting their parallelism and reprogrammability capabilities, thus improving the overall performance and addressing compute resources underutilisation.

TABLE I
KEY FEATURES OF THE EDGE4AI PROGRAMMING MODELS AND FRAMEWORKS

	Abstraction	Distributed computation
Split Computing Framework	Unified interface for implementing splittable DNN models while hiding the complexity of the underlying hardware heterogeneity and distribution mechanisms.	Split computation vertically, enabling to split complex DNNs in two segments, which can run at the UE and the edge.
Federated Learning (FL) framework	-	Federated learning across multiple UEs and aggregation at the edge, mitigating the wireless channel variability and node heterogeneity.
FPGA SoC reprogrammabiltiy (reproAccel)	Seamless reconfiguration for interconnected software and FPGA-accelerated functions running on FPGA SoC devices deployed at the far and extreme edge, by abstracting reconfiguration complexity and managing hardware-software interdependencies.	FPGA SoC resource management layer that enables the offloading of functions in different processing elements of FPGA SoC devices based on metrics and/or context events.
Distributed computing framework (COMPSs)	Application abstraction: facilitate the development of distributed workflows through simple task annotation Infrastructure abstraction: Application development is abstracted from infrastructure through the COMPSs runtime environment (which is deployed over Kubernetes).	Distributed computation of workflows horizontally, enabling execution of tasks across multiple computing nodes within the same edge cluster.
oneAPI programming model	Unified programming model for application optimization across different architectures	-

2) Edge4AI adopts and enhances split and distributed computing frameworks to develop and flexibly distribute computation tasks across the compute continuum: The Edge4AI programming models and frameworks also enable split and distributed computing. This is achieved by facilitating the decomposition of heavy workflows into smaller tasks that are deployed across different processing nodes of the compute continuum, including the User Equipment (UE) in some cases, so as to better match the computing capabilities of the available infrastructure and meet the application requirements.

Table I outlines the Edge4AI programming models and frameworks and their contribution to the first two objectives.

3) Edge4AI designs a unified compute continuum ecosystem, integrating the heterogeneous communication and networking components with the computing and storage resources that span from the edge to the cloud: Edge4AI builds a unified compute continuum ecosystem by: 1) leveraging cloud-native technologies such as Docker and Kubernetes to develop containerized network and application services, ensuring seamless interoperability across the compute continuum components; 2) enabling hierarchical service orchestration across the compute continuum, while also supporting intelligent network management and control. The modularity and flexibility of the OMC layer is reflected at the different granularity with which decisions are made. From a compute continuum perspective, Edge4AI supports orchestration both at multi-site and edge-site level, as well as fine-grained orchestration of distributed workflows, and intra-node micro-orchestration applied to FPGA System on Chip (SoC) devices. From a time-scale perspective, decisions may involve non real-time operations (>1000 ms), e.g. for multi-site infrastructure provisioning and service deployment, near real-time operations, e.g., for mobility control (10-1000 ms), or even decisions moving closer to the real-time domain (<10 ms) when function micro-orchestration is involved.

4) Edge4AI enables an AI-driven zero-touch closed-loop network automation: This is achieved by: 1) introducing the Cognitive Framework, a software asset developed in VERGE to facilitate the complete LCM of the AI4Edge AI/ML models by adopting and extending ML Operations (MLOps) principles and tools and providing intelligent AI model and dataset recommendations to the model developers; 2) creating a unified observability stack, collecting different metrics across the compute continuum and executed services and making them accessible to the AI/ML models through the cognitive framework interfaces; 3) integrating the different key components of the Edge4AI pillar through the necessary interfaces and adaptations, driven by requirements of the project use case demonstrations.

III. CONCLUSION

This paper discussed the features of the Edge4AI pillar design within the VERGE architecture, built to deliver the necessary building blocks for the flexible and efficient deployment and execution of B5G and 6G applications over an AI-powered edge computing infrastructure.

ACKNOWLEDGMENT

VERGE has received funding from the SNS JU under the EU's Horizon Europe R&I programme under Grant Agreement No 101096034. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the EU nor the granting authority can be held responsible for them. UK participants in VERGE are supported by UKRI grants 10071211 (Samsung Electronics (UK) Limited) and 10061781 (King's College London).

REFERENCES

[1] E. Kartsakli, et al., "Advanced Edge Computing Architecture for AI-Driven Automation and Slicing in Beyond 5G", 17th IEEE/ACM Int. Conf. on Utility and Cloud Computing (UCC), Sharjah, UAE, Dec. 2024.