AI-powered Edge Computing Evolution for Beyond 5G Communication Networks

Elli Kartsakli[1], Jordi Perez-Romero[2], Oriol Sallent[2], Nikolaos Bartzoudis[3], Valerio Frascolla[4], Swarup Mohalik[5], Thijs Metsch[4], Angelos Antonopoulos[6], Ömer Tuna[7], Yansha Deng[8], Xin Tao[6], Maria A. Serrano[5], Eduardo Quiñones[1]

[1]Barcelona Supercomputing Center, Barcelona, Spain; [2]Signal Theory and Communications Dpt., Technical University of Catalonia (UPC), Barcelona, Spain;[3]Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain; [4]Intel Intel Deutschland GmbH, Neubiberg, Germany; [5]Ericsson Research, Artificial Intelligence, Stockholm, Sweden;[6]Nearby Computing, Barcelona, Spain;[7]Ericsson Research, Istanbul, Turkey;[8]King's College London, United Kingdom

***Abstract*—Edge computing is a key enabling technology that is expected to play a crucial role in beyond 5G (B5G) and 6G communication networks. By bringing computation closer to where the data is generated, and leveraging Artificial Intelligence (AI) capabilities for advanced automation and orchestration, edge computing can enable a wide range of emerging applications with extreme requirements in terms of latency and computation, across multiple vertical domains. In this context, this paper first discusses the key technological challenges for the seamless integration of edge computing within B5G/6G and then presents a roadmap for the edge computing evolution, proposing a novel design approach for an open, intelligent, trustworthy, and distributed edge architecture.**

***Keywords*— Edge computing; AI/ML-based optimization; security and trustworthiness; B5G/6G evolution; edge-cloud compute continuum; closed-loop automation**

I. Introduction

Applications based on eXtended Reality (XR) and holographic representations are expected to become a key asset in a wide range of scenarios, fusing the digital and the real world to deliver new experiences to end users, such as metaverse environments, immersive online gaming, real-time 3D communications, etc. In addition, they allow for a more efficient implementation of digital twin models, remote assistance, training and collaborative design, thus realizing the much-sought digital transformation of several vertical industries, such as manufacturing and healthcare. At the same time, the wide penetration of the Internet of Things (IoT), involving massive numbers of sensing devices with different capabilities, ranging from low-complexity, low-cost sensors to smart cameras and actuators, is generating huge volumes of data. Leveraging Artificial Intelligence (AI) and big data technologies, such data is transformed into valuable and actionable knowledge, able to automate and optimize the decision-making process in multiple sectors, from smart city and autonomous driving, to energy and transportation.

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 AI-enabled solutions leverage monitoring data to optimize both network and application performance and achieve closed-loop automation.

As the capabilities of communication infrastructures grow, emerging applications are becoming increasingly demanding in terms of computation. Due to the limitations of end user devices – phones, Virtual Reality (VR) glasses, etc. –in terms of size, energy, computational capacity and cost, it becomes necessary to offload heavy tasks to more powerful computing elements, typically residing at the cloud. Moreover, cloud computing is no longer capable to meet the latency requirements of such applications, nor deal effectively with distributed and heterogeneous massive IoT deployments. In response to these needs, *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 [3]. Hence, the synergy between B5G and edge computing can provide computing and storage capabilities for applications residing at the boundary of operators' networks [4]. Nevertheless, many open challenges must be addressed in order to achieve an open, secure and distributed edge architecture that can be seamlessly integrated into an edge-cloud compute continuum, further boosting B5G capabilities enabling innovative next generation services [5].

In this context, the objective of this paper is twofold: i) elaborate on the technological challenges arising in B5G edge-enabled scenarios; and ii) provide insights on an integrated approach for driving the edge computing evolution. In this respect, this paper presents the vision of the project VERGE[1], which envisages a solution approach sustained on three main pillars: i) "*edge for AI*", namely a flexible, modular and converged edge platform design, unifying the lifecycle management and closed-loop automation for cloud-native applications, Multi-access Edge Computing (MEC) and network services across the edge-cloud compute continuum for ultra-high computational performance; ii) "*AI for edge*", namely an AI-powered portfolio of solutions leveraging the multitude of collected metrics for intelligent management and orchestration; and iii) "*security, privacy and trustworthiness of AI-based models at the edge*", providing a suite of methods to protect AI models against adversarial attacks, increase their explainability and reliability, and ensure data privacy.

The paper is structured as follows. Section II presents the envisioned B5G scenario for the edge computing evolution. Section III discusses the open issues for its materialization, whereas Section IV proposes a novel architectural approach to

[1] https://cordis.europa.eu/project/id/101096034

tackle these challenges, aligned with the VERGE project. The paper closes with some concluding remarks in Section V.

II. Envisioned Edge-Enabled Beyond 5G Scenario

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 [6]. 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 [7], 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.

Aligned with the 3GPP network architecture concepts for B5G, Fig. 1 illustrates an example of a virtualized network with evolved edge computing capabilities. The lower part depicts the network infrastructure layer, composed of heterogeneous nodes with different computational, storage and networking capabilities. A multi-access RAN, possibly including different radio access technologies (5G, LTE, Wi-Fi, etc.), provides wireless connectivity to mobile User Equipment (UE) devices through diverse types of radio nodes: BSs, disaggregated Radio Units (RUs), as well as fixed or moving relays. The computational and storage resources are highly heterogeneous and distributed across multiple layers, including edge sites closer to the UEs, collocated with the BSs or relays (i.e., far edge), edge sites at aggregation points (i.e., near edge), cloud servers and even UEs with high computational capacities (e.g., equipped with onboard embedded processors). This pool of resources composes an *edge-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, connecting the 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 also illustrates some use case examples enabled by this evolved edge computing architecture. At the leftmost part, an autonomous tram is depicted, equipped with a multitude of sensors that collect and process information on the surrounding area, implementing for example a collision detection application. While part of this process can be executed onboard the tram, additional services may run at the edge (e.g., offloading part of the computation or fusing input from additional sensors such as city cameras), or the cloud (e.g., providing a city-level mobility service to monitor the tram operation). Furthermore, the autonomous tram service could be deployed and horizontally distributed in more edge sites (e.g., site #2) to offer service continuity across the tram path. From the network perspective, in this example, the tram is connected, through a Radio Unit (RU) to the edge site #1, where the DU and the CU User Plane (CU-UP) of a gNB are hosted, together with a User Plane Function (UPF) of the 5G Core (5GC). This UPF enables a local break-out for delivering traffic offloaded by the tram control to be processed at the edge due to their strict real-time requirements latencies. The CU Control Plane (CU-CP) and the rest of 5GC functions (e.g. Access and Mobility Management Function (AMF), Session Management Function (SMF)) are centralized at the near-edge site #6, which could be a high performance data center.

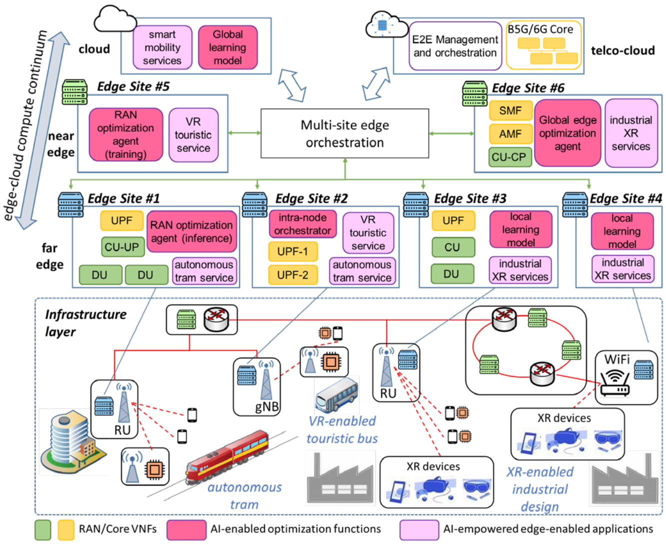

Fig. 1. Envisioned B5G scenario for the edge computing evolution

Another example in Fig. 1 shows a touristic bus, acting as communication relay for passenger connectivity and, at the same time, hosting an onboard server supporting VR applications that complement the sightseeing experience. Part of the computation could be offloaded at the edge site #2, whereas some centralized application component could be deployed at the near edge site #5, splitting the application vertically across the compute continuum. A third example considers an industrial application for collaborative product design enabled by real-time XR technologies, allowing two engineering teams to work on the same product design simultaneously from remote locations. In this case, the XR devices of one factory are connected to edge site #3, where a number of applications that process the data gathered by the devices (e.g., for rendering) are hosted. This data is shared with another factory, connected edge site #4, whereas other XR-supporting services with less stringent time constraints could also be centralized at the near-edge site #6.

In addition to the vertical applications and VNFs, this unified compute continuum can also host advanced AI-based applications for network management and optimization. Some examples include AI training and inference models for RAN optimization (e.g., dynamic CU-DU splitting), which could run at different locations (sites #1 and #5), or a distributed learning model running at sites #3 and #4, and feeding a global model hosted in the cloud (e.g., to optimize edge resource allocation for the XR application based on predicted user behavior and requirements). Different levels of optimization could also be considered, from both a local (intra-node) perspective, managing the multiple services running at the same node (e.g., two applications in site #2), or a global perspective across the edge domain (running in site #6 and supporting the decision-making process of the multi-site orchestrator).

III. Enabling Technologies and Challenges

Integrating all the elements of this envisioned scenario into an edge-cloud compute continuum, where cloud-native services and VNFs can be easily and flexibly deployed, scaled and orchestrated based on their specific requirements, is not trivial. In this context, this section aims to identify the most relevant challenges and technological gaps.

A. Intensive computation over heterogeneous architectures

Unlike cloud computing where the processing is spread over homogeneous computing and memory resource clusters, edge computing needs to contemplate a *heterogeneous panoply of highly distributed processing elements, with different performance and power consumption budgets, memory solutions and interconnectivity standards*. These processing elements range from micro-controllers, micro-processors, Graphics Processing Units (GPUs) and Field Programmable Gate Arrays (FPGA) accelerators, to complex multi-processor System on a Chip (SoC) devices, Application-Specific Integrated Circuits (ASICs) and AI-optimized processors. With virtualization solutions like containerization, software can be decoupled from the underlying hardware (HW), i.e., different applications can run over commercial off-the-shelf (COTS) HW. However, more targeted solutions are required in order to *fully leverage the massive parallelism capabilities of the HW-accelerated platforms*, which is fundamental in order to deliver the Key Performance Indicators (KPIs) for ultra-low latency processing required by computational-intensive B5G and 6G applications. On the other hand, the complexity of parallel programming techniques for such heterogeneous scenarios can be daunting [5], stressing the need for programming models capable of providing the *right level and granularity of abstraction* to express the performance capabilities of the underlying HW platforms, with the lowest possible overhead.

B. Network and service orchestration

One of the most radical changes brought by 5G is the high network programmability, driven by NFV and Software Defined Networking (SDN). The current trend for RAN disaggregation enables the flexible and dynamic placement of VNFs across the network infrastructure to meet specific performance requirements. At the same time, applications are adopting cloud-native design principles, breaking computation into microservices or even serverless functions, and enabling novel split and distributed computation execution models. Hence, it becomes evident that as we move towards 6G, there is a need to create an integrated communication and computation environment where network and application services can be seamlessly deployed, executed and orchestrated. This makes the *management and orchestration of services and resources* very challenging, further aggravated by the heterogeneous, multi-tenant and multi-site edge-enabled topologies. Additionally, for the *efficient lifecycle management of all network and application services*, it is necessary to integrate and monitor a multitude of underlying SW components (radio controllers, edge platforms, intra-node micro-orchestrators, etc.).

C. Edge learning techniques

Edge intelligence, defined as the application of AI and Machine Learning (ML) at the network edge, has been identified as a key element in B5G, leveraging the multitude of data generated by the network and applications to learn and predict the runtime conditions and service requirements, thus enabling the proactive solutions for reduced latency and overall better performance and automation. However, the constraints on the training of such complex AI models over limited edge resources, the absence of sufficient local context information or data interpretation and the privacy concerns regarding the transfer of sensitive user datasets to the cloud, stress the need for decentralized leaning solutions. In this context, Federated Learning (FL), where a shared model is trained across multiple learning agents using local datasets, while the global model is updated by a central entity, is an emerging technique with many applications [8]. Another recent approach, known as split learning (SL), divides the ML model into several sub-models, distributing them to different entities. For instance, a few layers of a Deep Neural Network (DNN) could run at the UE, reducing the computational complexity compared to a FL approach, whereas the remaining layers could be deployed at an edge server. Exploring the tradeoffs between FL, being more efficient with small models and large datasets, and SL that scales better for high number of users, and leveraging the advantages for both approaches is a very relevant open issue.

D. Data-driven resource optimization

With the increased complexity of B5G systems, the use of the above mentioned edge learning techniques becomes key to design smart resource optimization solutions that achieve an efficient operation of the system. AI at the edge has been used for proactive network management [9], such as channel modelling and prediction, traffic and mobility forecasting, network resource allocation, task offloading, etc. However, while most AI solutions have been designed addressing stand-alone problems, the development of unified frameworks capable of overseeing end-to-end operations are not well investigated. In this direction, efficient conflict-free AI solutions will be needed, able to work in a harmonized way when dealing with closely related problems, such end-to-end slicing, optimally splitting DU and CU functions across the edge-cloud compute continuum, or resource allocation of network and computation resources under edge constraints. Besides, AI techniques are expected to exploit the huge amount of data collected at the edge for achieving an optimized operation. However, due to the limited computation, storage, and communication resources of edge nodes, and the low-latency, and reliability requirements of the applications, the design of these AI techniques needs to bear in mind the reduction of the magnitude of the data to be processed towards achieving scalable, sustainable and energy efficient solutions.

E. Trustworthiness of AI solutions

The successful adoption of AI-enabled orchestration as well as AI-driven B5G/6G use cases depend heavily on ensuring trustworthiness of AI. Trustworthy AI covers a wide range of concerns such as robustness, safety, security, privacy, explainability, fairness and reliability that are being increasingly mandated [10], some of which are discussed below.

1) Security and privacy concerns: Given the major role of AI in the decision-making process of edge-enabled B5G networks, ensuring the trustworthiness of AI models is extremely important to minimize any potential detrimental effects of AI decisions on these critical systems. Security and privacy are two key characteristics of trustworthy AI. In the past few years, we have witnessed extensive research showing the vulnerability of AI-driven systems in different domains. Despite the distributed nature of the communication domain and the network heterogeneity, we still have the risk of adversarial attacks in a telco environment [12], targeting the robustness of AI models [11] by compromising the decision-making process (evasion attacks) or undermining the training data, thus leading to performance deterioration (poisoning attacks). Because of their small footprints, adversarial ML-based attacks are more covert and difficult to detect, calling for novel solutions to ensure that the AI models' decisions are robust in the face of minor changes to the input data.

Another major factor is the privacy risks associated with the use of AI/ML in these networks. AI-powered systems are fueled by vast amounts of data, mostly containing private information on users, service providers, and infrastructure suppliers. However, AI models have the potential to leak private information. Attack scenarios have been demonstrated where an attacker could exploit weaknesses in the target system to obtain sensitive information about users' data and even extract deployed models. The privacy related threats against AI-driven systems may infer whether a particular data sample has been used during training or not (i.e., membership inference [13]), reconstruct private datasets used in model training (i.e., model inversion) or obtain AI model parameters by observing the outputs (i.e., model extraction).

2) Need for safety and explainability: Among various AI technologies, reinforcement learning (RL) is widely used in 5G networks for complex decision-making problems, including network slicing optimization, power control, interference coordination and so on. Hence, imposing safety guarantees for RL agents is critical for their real-world application. A main risk stems from the fact that RL agents are trained through exploration of new actions in a given environment, but might be deployed in a different environment different from the ones used for training. In these situations, the RL agent may reach decisions that can potentially lead to dangerous situations (e.g., due to insufficient capacity allocated to a slice used to control an autonomous tram or for healthcare services) or just to unacceptable degradations in network performance (e.g., bad connectivity for a large number of subscribers), raising concerns to the operator on the trustworthiness of the AI solution. In edge computing, the distributed deployments further aggravate safety issues, resulting in asynchrony and time-delay in recommendation. Hence, developing scalable techniques for safe-training and safe-operation of a large number of RL-based agents in the edge can be a challenge.

On the other hand, understanding of how an AI model makes particular decisions and contributes to achieving the goals of the stakeholder is a key component of the trustworthiness of the AI system. Hence, the concept of explainable AI (XAI) has been gaining a lot of attention, with recent efforts focusing on the explainability for RL-based agents, to shed light in the decision-making process [14]. Although many studies have considered XAI, scalability and feature dependence are still an open issue. For example, providing explanation for each task of the model is challenging when the number of task increases, stressing the need for task-aware XAI models. Feature dependence also causes problems in explanation, especially when feature are correlated, making it very difficult to attribute the importance of each feature in the output of the XAI model.

IV. The VERGE Approach for the Edge Evolution

Motivated by the need to address the aforementioned challenges, the VERGE project will drive the edge computing evolution towards an *integrated, next-generation edge- cloud compute continuum*, as a key enabler for the most demanding applications and use cases in the road towards 6G. By adopting an interdisciplinary approach to converge techniques from multiple areas, including telecommunications, edge and cloud technologies, embedded and distributed computing, cybersecurity and AI, VERGE will achieve its vision by:

i) building an "*edge for AI*" framework (thereafter Edge4AI), to support the flexible and efficient deployment and execution of 6G AI-enabled applications across a jointly orchestrated compute and communication continuum, while fully exploiting the capabilities of multi-core/multi-accelerator platforms for ultra-high computational performance;

ii) leveraging "*AI for edge*" (thereafter AI4Edge), by employing cutting-edge AI methods to learn and optimize the network performance in the highly heterogeneous and rapidly changing B5G and eventually 6G environments, and

iii) providing the necessary Security, Privacy and Trustworthiness "*SPT for AI*" methods (thereafter SPT4AI), to address the relevant challenges that specifically emerge due to the decentralized edge computing environment and the extensive use of distributed AI methodologies in a dynamic and heterogeneous network structure.

Fig. 2 reflects the main components and concepts of the VERGE design towards this vision, which will be discussed in continuation. Emphasis is placed on providing some hints on the anticipated technical solution approaches and the expected advances with respect to the state-of-the-art.

A. Edge4AI: an integrated edge architecture enabling AI

The creation of the Edge4AI heterogeneous adaptive processing substrate encompasses and integrates a number of innovative agile computing techniques, resource orchestration frameworks and compute automation solutions. In this respect, the key features of the Edge4AI pillar will include:

1) Support for distributed and split computing over heterogeneous computation architectures: The evolved edge architecture should support different levels of distribution mechanisms for splitting the computation across the available edge and cloud resources: i) *in-node splitting* of computing tasks, i.e., within the same multi-accelerator platform, ii) *horizontal distribution* of computation among peer edge nodes, and iii) *vertical distribution*, between end users, edge and cloud. A key innovation towards this direction will be the design of an *adaptive virtualization layer* specifically targeting programmable accelerated HW platforms, enabling the dynamic reconfiguration of functions across embedded (AI) accelerators and general-purpose computing elements.

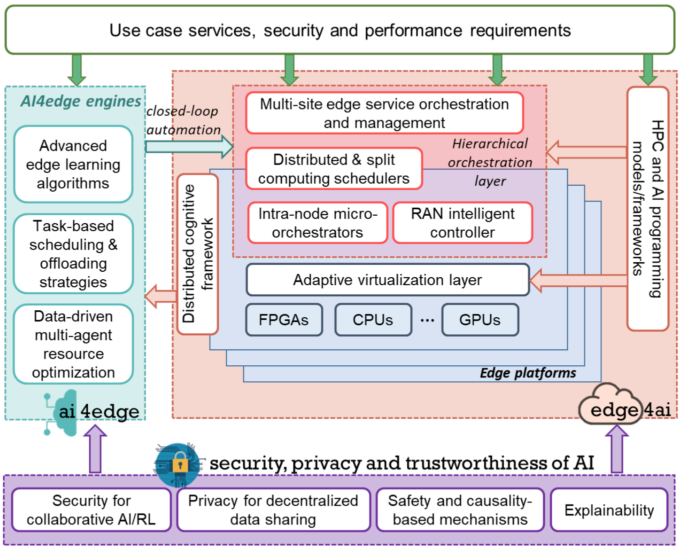
Use case services, security and performance requirements
AI4edge engines
Advanced edge learning algorithms
Task-based scheduling & offloading strategies
Data-driven multi-agent resource optimization
ai 4edge
closed-loop automation
Distributed cognitive framework
Multi-site edge service orchestration and management
Distributed & split computing schedulers
Hierarchical orchestration layer
Intra-node micro-orchestrators
RAN intelligent controller
Adaptive virtualization layer
FPGAs
CPUs
···
GPUs
Edge platforms
HPC and AI programming models/frameworks
edge 4ai
security, privacy and trustworthiness of AI
Security for collaborative AI/RL
Privacy for decentralized data sharing
Safety and causality-based mechanisms
Explainability

Fig. 2. The three pillars of the VERGE desingn for the edge evolution

Moreover, the most suitable programming models and practices from the embedded, High Performance Computing (HPC) and AI domains will be employed for the implementation of highly efficient AI-enabled application workflows at the edge- cloud compute continuum. On the one hand, HPC parallel programming models for shared memory (e.g., OpenMP[2]) and accelerator-specific programming environments (e.g., Intel® oneAPI[3], NVIDIA CUDA[4], Vitis AI development toolkit[5]) can be used to exploit the inner parallelism of multi- and many-core processor architectures. On the other hand, well-known frameworks such as Tensorflow and Pytorch, as well as open-source frameworks for distributed environments such as COMP Superscalar[6] (COMPSs) can facilitate the design of complex AI workflows, such as FL solutions. All these programming practices should be integrated into a common software architecture framework, ensuring interoperability between the different technologies, the underlying edge platform and the orchestration layer.

2) Multi-site hierarchical edge orchestration: VERGE will design an end-to-end orchestration framework for *multi-site service orchestration and management* of heterogeneous virtualized multi-edge infrastructures, extending the NearbyOne[7] solution. The orchestrator will jointly manage *network and application services*, by developing the necessary interfaces and communication pipelines with: i) relevant network management entities of the B5G network, such as network slicing managers and intelligent RAN controllers, to enable the implementation of joint computational and network resource management schemes, and ii) the underlying edge platform, built upon solutions such as the Kubernetes[8] based Intel Smart Edge Open Platform[9], as well as intra-node micro-orchestrators and split computing/distributed schedulers, to support hierarchical orchestration, enabling local autonomous decision-making and coordinated multi-site optimization.

3) Closed-loop automation framework: This framework targets the execution of zero-touch orchestration and optimization operations. This can be made possible through a *distributed cognitive framework*, responsible for collecting a wide range of parameters from all layers, RAN and core metrics, exposed by the B5G network, edge platform telemetry (e.g., CPU/storage/memory utilization, etc.), and application-related requirements, exposed by the employed programming models. The cognitive framework will fuel the AI4Edge layer (further described in the next section) with the necessary data to make intelligent decisions, which will, then, be enforced by the *multi-ste hierarchical edge orchestrator*, for closed-loop automation operations.

B. AI4Edge: AI solutions for edge and network optimization

The AI4Edge pillar encompasses an AI-powered portfolio of solutions for the lifecycle management of computing, communication and networking resources and the intelligent decision-making across all layers of the system. It contains multiple AI models that can help power the network edge and their applications, in an efficient and scalable manner, exploiting the capabilities of the Edge4AI layer. The main components of AI4Edge pillar are detailed next:

1) Advanced edge learning algorithms: This category includes novel solutions for the efficient training of ML models at the evolved edge. Specific solutions should address the computational distribution of FL tasks, enabling an optimal selection of UEs as learning agents, and addressing the trade-off between computation and communication to optimally distribute ML models across the federation. Moreover, split computing approaches are also considered, dynamically splitting DNN models into head and tail portions (deployed at UEs and edge side, respectively), optimizing the splitting point on-the-fly based on varying environmental conditions (e.g., channel quality, throughput, UE battery level, edge server load). Similarly, *advanced distributed learning algorithms* to achieve edge-empowered applications, such as (multi-tier) asynchronous FL, transfer learning and robust FL under limited resources, will also be developed.

2) Task-based scheduling and computational offloading: Such schemes will offer intelligent solutions on a variety of *allocation challenges at the edge*, including the dynamic allocation of application tasks to the most suitable computing resources across the edge-cloud compute continuum, MEC selection strategies for computational offloading, or optimal computational splitting of AI models. Here, application-related metrics (e.g., sensor-to-actuator time) combined with real-time metrics from the edge platform (e.g., available computing resources), the 5G network (available bandwidth, communication latencies, etc.) and MEC services (e.g., UE location), provided by the distributed cognitive framework of the Edge4AI layer, can be jointly considered to design new scheduling and task distribution and offloading policies.

3) Data-driven resource optimization: The solutions belonging to this category should compose a *multi-level multi-agent end-to-end framework* that coherently and efficiently integrates the different AI-based solutions in a conflict-free and efficient manner for addressing different network optimization problems that account for energy,

[2] https://www.openmp.org/
[3] https://www.oneapi.io/
[4] https://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf
[5] https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html
[6] https://compss-doc.readthedocs.io/en/stable/
[7] https://www.nearbycomputing.com/nearbyone/
[8] https://kubernetes.io/
[9] https://smart-edge-open.github.io/

communication, and computation limitations. Examples of optimization problems and solutions considered here include the determination of the *optimum splitting point of the CU and DU functions* across the edge-cloud compute continuum during service run time based on RAN conditions (e.g., channel quality, load, physical and computing resources) and leveraging AI-based prediction methods (e.g., for channel condition estimates). Another relevant problem is the *capacity sharing for RAN slicing* in case of changes in the current deployment (e.g., due to the activation/deactivation of RAN nodes). The use of transfer learning can be useful here to accelerate the training of the model under the new conditions. Similarly *end-to-end slicing in vehicular scenarios* (e.g., for the autonomous tram), can also be considered, exploiting the predictability of the tram trajectory to timely prepare the slice allocation. Other optimization strategies can consider advanced RAN architectures, e.g. through the use of relay nodes that enable coverage extension and provide additional computational capabilities to the edge-cloud continuum. In this context, deep RL solutions can be used for dynamically deciding a relay activation or deactivation.

C. SPT4AI: Security, privacy and trustworthiness of AI

The third pillar of the proposed architecture will provide solutions to address the concerns for security, privacy and safety issues of systems with ML/RL-based models, specifically focusing on the threats arising in the dynamic, distributed and resource-constrained environment of the edge. Two main dimensions form part of the SPT4AI framework:

1) Security and privacy: Focusing specifically on collaborative AI and RL settings, the most probable adversarial attack scenarios relevant to edge-enabled B5G networks (including evasion and poison attacks) will be first identified, and targeted *efficient and robust mitigation methods* will be considered. To ensure the security and privacy of the sensitive information carried by the data and models against attacks such as membership inference, model inversion, and model extraction attacks, privacy preserving technologies such as *FL, differential privacy and homomorphic encryption* will be leveraged. Furthermore, distributed ledger technologies (DLT), and specifically blockchains, will be employed to provide secure, distributed and decentralized data sharing, under such heterogeneous communication environments.

2) Safety and explainability: In order to increase the overall trustworthiness of the AI4Edge models, a portfolio of algorithms for building safe, explainable and causal RL solutions will be provided, following three interconnected approaches for the design of: i) Safe-RL techniques based on *formal verification*, which can be used to verify the correctness of AI models from the second pillar AI4Edge, both during training and during deployment, ii) *causality-based methodologies* to build causal graphs, identify causal variables and hidden confounders, and using them to build more trusted RL models. The effectiveness of the causal RL models will be demonstrated when applied in transfer learning scenarios where there could be inherent distributional shifts or noise in the data, and iii) techniques based on knowledge graphs, their maintenance and knowledge extraction, to pose queries to the RL-based agents and obtain *semantically rich explanations* for the decisions.

V. Conclusions

This paper has presented a roadmap for the edge evolution, as a key enabling technology to support innovative, distributed and secure AI-enabled processes across a unified edge-cloud compute continuum, integrated within the B5G/6G ecosystem. After describing the envisioned scenario and identifying the most relevant technical challenges, the paper presented a novel and holistic design approach for the edge evolution, aligned with the vision of the VERGE project. The proposed solution has introduced a flexible, modular and converged edge platform design, supporting and empowered by decentralized learning and intelligent automation agents, while also ensuring the security, privacy, safety and explainability of the employed AI models.

Acknowledgement

This work has been partially supported by the European Commission's Horizon Europe, Smart Networks and Services Joint Undertaking research and innovation program, under project VERGE (grant agreement #101096034). The authors would like to thank the whole VERGE consortium for their contribution to shape the concepts, research challenges and solution approaches considered in project VERGE.

References

[1] K. Samdanis and T. Taleb, "The Road beyond 5G: A Vision and Insight of the Key Technologies," in IEEE Network, vol. 34, no. 2, pp. 135-141, 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, no. 3, pp 134-142, May/June 2020.

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

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

[5] L. Kong, et al., "Edge-computing-driven Internet of Things: A Survey," in ACM Comput. Surv., vol. 55, no. 8, Article 174, Aug.2023.

[6] A. Yousefpour et al., "All one needs to know about fog computing and related edge computing paradigms: A complete survey", in *Journal of Systems Architecture*, vol. 98, pp. 289-330, Sept. 2019.

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

[8] S. Samarakoon, M. Bennis, W. Saad and M. Debbah, "Distributed Federated Learning for Ultra-Reliable Low-Latency Vehicular Communications," in *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146-1159, Feb. 2020.

[9] Y. Koda *et al*., "Communication-Efficient Multimodal Split Learning for mmWave Received Power Prediction," in *IEEE Communications Letters*, vol. 24, no. 6, pp. 1284-1288, June 2020.

[10] EC - High Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI", April 2019.

[11] M. Sadeghi and E. G. Larsson, "Adversarial Attacks on Deep-Learning Based Radio Signal Classification," in *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213-216, Feb. 2019.

[12] O. F. Tuna, F. E. Kadan, and L. Karacay, "Practical adversarial attacks against ai-driven power allocation in a distributed mimo network," 2023. [Online]. Available: https://arxiv.org/abs/2301.09305

[13] H. Hu, Z. Salcic, L. Sun, G.Dobbie, P. S. Yu, and X.Zhang, "Membership Inference Attacks on Machine Learning: A Survey,"in *ACM Comput. Surv.*, vol. 54, no. 11, Article 235, Jan. 2022

[14] S.Milani, N. Topin, M.Veloso, F. Fang, "A Survey of explainable reinforcement learning," 2022. [Online]. Available: https://arxiv.org/abs/2202.084