AI@EDGE: A Secure and Reusable Artificial Intelligence Platform for Edge Computing

Roberto Riggio*, Estefanía Coronado†, Neiva Linder‡, Adzic Jovanka§, Gianpiero Mastinu∥, Leonardo Goratti**, Miguel Rosa††, Hans Schotten‡‡, and Marco Pistore‡‡

*RISE Research Institutes of Sweden AB, Stockholm, Sweden; roberto.riggio@ri.se
†i2CAT Foundation, Barcelona, Spain; Email: estefania.coronado@i2cat.net
‡Ericsson AB, Stockholm, Sweden; Email: neiva.linder@ericsson.com
§Telecom Italia S.p.A., Turin, Italy; Email: jovanka.adzic@telecomitalia.it
∥Politecnico di Milano, Milano, Italy; Email: gianpiero.mastinu@polimi.it
**Safran Passenger Innovations, Wessling, Germany; Email: leonardo.goratti@zii.aero
††Aerotools, Madrid, Spain; Email: miguel.rosa@aerotools-uav.es
‡‡German Research Center for Artificial Intelligence, Kaiserslautern, Germany; Email: schotten@dfki.uni-kl.de
§§Fondazione Bruno Kessler, Trento, Italy; Email: marco.pistore@fbk.eu

Abstract—Artificial Intelligence (AI) has become a major innovative force and a major pillar in the fourth industrial revolution. This trend has been acknowledged by the European Commission, who has pointed out how high-performance, intelligent, and secure networks are fundamental for the evolution of the multi-service Next Generation Internet (NGI). While great progress has been done in the accuracy and performance of AI-enabled platforms, their integration in autonomous decision-making and critical systems requires end-to-end quality assurance. AI@EDGE addresses these challenges harnessing the concept of “reusable, secure, and trustworthy AI for network automation”. To this end, AI@EDGE targets significant breakthroughs in two fields: (i) general-purpose frameworks for closed-loop network automation capable of supporting flexible and programmable pipelines for the creation, utilization, and adaptation of the secure, reusable, and trustworthy AI/ML models; and (ii) converged connect-compute platform for creating and managing resilient, elastic, and secure end-to-end slices supporting a diverse range of AI-enabled network applications. Cooperative perception for vehicular networks, secure, multi-stakeholder AI for Industrial Internet of Things, aerial infrastructure inspections, and in-flight entertainment are the uses cases targeted by AI@EDGE to maximise its commercial, societal, and environmental impact.

Index Terms—AI, 5G, MEC, automation, disaggregated RANs, ML-based security, hardware acceleration, serverless platforms

I. INTRODUCTION

Artificial Intelligence (AI) systems are irreversibly set on the evolutionary path of every future vertical as well as of every object and service we human will interact with in the near future. This trend is motivated by the need to support elastic and demanding real-world use cases such as automated mobility, e-health, gaming, etc. In this scenario, it is acknowledged that the operators will have the opportunity to fill a central role in providing innovative solutions for application and service developers that want to combine the advanced capabilities of 5G with the fluid cloud-based application development processes emerged in the last decade, such as, for example, the Platform as the Service (PaaS) and the microservice/serverless models. A significant example of this ecosystem are AI-enabled applications, which have become a major innovative force in almost any vertical, and are being foreseen as one of the pillars boosting the fourth industrial revolution.

This projection is supported by the fact that cloud and mobile networks are already converging at several technological levels. From one side, cloud technologies, such as cloud-native and virtualization, are making their way into the telecom operators’ domain. At the same time, networking use cases and requirements, such as access-aware operations and service function chaining, are influencing the evolution of cloud technologies. Despite such advances, the cloud operators’ approach remains centralized with few large data centers deployed at key locations, while telecom operators manage a distributed infrastructure based on a hybrid multi-cloud approach.

In this context, 5G is a paradigm shift: its high performance in terms of latency, bitrate, and reliability, call for a technological and business convergence between the cloud computing and the telecom worlds. 5G features like slicing, Multi-access Edge Computing (MEC), and more flexible radio connectivity can be used to support qualitatively different applications, and to deliver a richer user experience, faster interactions, large scale data processing, and machine to machine communications. Nevertheless, the challenges to be overcome to realize this connectivity/computing convergence are still notable. In particular, the increasing number of control and optimization dimensions of the end-to-end 5G infrastructure may result in an overly complex network that operators and vendors may find it difficult to operate, manage, and evolve [1].

AI and Machine Learning (ML) technologies will be crucial in the cloud-network convergence process and will help operators achieve a higher level of automation, increase network performance, and decrease the time-to-market of new features. Early attempts at applying AI/ML in the cellular domain can be found in several academic works [2], [3], [4], [5]. Nevertheless, it cannot be expected that each and every subsystem of future access, edge, core, and cloud segments will
employ distinct and separated AI tools and datasets. Such an approach would lead to AI-silos, slowing down advances vital to achieving sustainable networking and ultra-scale complex services relying on distributed compute-connect fabric.

The approach of AI@EDGE to answer the above-mentioned challenges has two lines of action. First, we will design, prototype, and validate a network and service automation platform able to support flexible and programmable pipelines for the creation, utilization, and adaptation of secure and privacy-aware AI/ML models. Second, we will use this platform to orchestrate AI-enabled end-to-end applications. Here, we introduce the novel concept of Artificial Intelligence Functions (AIFs) to refer to the AI-enabled end-to-end applications sub-components that can be deployed across the AI@EDGE platform. Finally, the AI@EDGE platform will be validated using four well-chosen use cases with specific requirements that cannot be satisfied by current 5G networks according to 3GPP R15 and 3GPP R16, in particular in terms of support for latency-sensitive and highly dynamic AI-enabled applications.

The rest of the paper is structured as follows. In Sec. II we discuss the challenges addressed by the AI@EDGE Project. Section III covers the AI@EDGE concept for beyond 5G networks. The four reference use cases are described in Sec. IV. Finally, Sec. V concludes this paper.

II. CHALLENGES

The main objective of AI@EDGE is to build a secure connect-compute platform capable of enabling the automated roll-out and management of large scale heterogeneous edge and cloud computing infrastructures. To this end, the platform encompasses the required APIs to enable the deployment of large-scale virtual compute overlays (e.g., containers and serverless instances, etc.) across a multi-connected heterogeneous infrastructure able to support a range of future critical applications. This is illustrated in Fig. 1, which represents the functional overview of the platform across two dimensions: (i) AI/ML-driven multi-connected applications at massive scale; and (ii) AI/ML-centered security. However, this ambitious objective involves several important challenges. Such challenges are described in detail in the following subsections.

A. Network automation platform leveraging flexible and reusable AI pipelines

5G is a full paradigm shift where high performance in terms of latency, bitrate, and reliability, calls for a technological and business convergence between cloud computing and networking. The increasing number of control and optimization dimensions of the 5G infrastructure may lead to an overly complex network that operators and vendors may find it difficult to operate, manage, and evolve. AI technologies will be key in this roadmap and, although early attempts to address this issue can be already found [6], [3], [7], [8], they focus on specific problems that cannot be extrapolated to other network segments. However, in a full automated system, having distinct and independent AI tools and datasets would make impossible the sustainable management of networking and highly-scalable services on a distributed connect-compute platform.

Therefore, the challenge is to implement a general-purpose network automation framework capable of supporting flexible and reusable end-to-end AI pipelines. Scalable AI/ML models, fast data pipelines and effective data dissemination models are crucial to realize automation at scale. Some of the main pillars of AI@EDGE are the potential of multi-access edge computing and the powerful mechanisms of scalable distributed and federated learning in the 5G context. Based on this, AI@EDGE addresses this challenge by developing a platform for closed-loop automation that allows the deployment of AI/ML compute infrastructures over the edge, which also accounts for secure isolation of co-located AI/ML algorithms by multiple stakeholders running on shared MEC resources. The progress on this challenge will enable two main results:

- Scaling of AI/ML distributed algorithms to ensure application performance and model reliability under varying resource availability.
- Zero-touch end-to-end network and service management including the creation, utilization, and adaptation of reusable AI/ML pipelines in a connect-compute platform.
B. Secure and resilient ML for multi-stakeholder environments

From a security perspective, there are several relevant aspects for the success of 5G and beyond systems. In production networks, the potential risks for tangible assets, such as servers and human beings victims of attacks, is significantly increasing. Consequently, intrusion detection systems are a native part of the current 5G security architecture. However, they are usually under proprietary licenses, which highlights the need to enable an open exchange of models and parameters for intrusion detection, especially in multi-stakeholder environments. It must be noted that this issue worsens when the platforms are driven by AI/ML models, since a new and dangerous attack surface is added. The ability to achieve resilience and service continuity requires simple and information-effective data-driven models, suitably designed for running on Internet of Things and MEC devices with limited resources. Therefore, the challenge is to provide light-weight, secure and resilient ML systems that are robust to evasion and poisoning attacks.

With the advent of ML, privacy techniques have been recently revisited to better accommodate the trade-off between privacy risk and data usefulness in the construction of ML pipelines. Federated Learning (FL) [9], [10], [11] and adversarial networks [12] have recently been used to cover the security aspects. FL allows assembling a common model combining local models built from edge devices without disclosing any data. However, this approach poses numerous problems such as local biases, temporal offsets, etc. Security is one of the cornerstones of the architecture of AI@EDGE. Consequently, AI@EDGE aims to define and implement an ML methodology following a distributed paradigm, allowing the development, implementation and evaluation of effective intrusion detection algorithmic framework, accounting for security and isolation of co-located AI/ML algorithms. Tackling this challenging research question will mainly lead to three key results:

- Increase in attack detection speed and intrusion resiliency, including early detection and automated configuration and speedup of countermeasures.
- Model propagation and computational efficiency to optimize the set of exchanged parameters, e.g., the weights of a neural network, and their aggregation methods.
- Privacy security of the model parameters exchanged between edge devices.

III. THE AI@EDGE CONNECT-COMPUTE FABRIC FOR BEYOND 5G NETWORKS

The design of the AI@EDGE platform envisions the automated roll-out of adaptive and secure compute overlays, and a new generation of AI-enabled end-to-end applications. Such applications are made possible by AI@EDGE through the introduction of the novel concept of AIFs, which refer to the AI-enabled applications sub-components that can be deployed and chained across the various levels of the architecture. This vision is presented in Fig. 2, in which AI@EDGE combines a set of cutting-edge cloud computing and 5G concepts with a reusable, secure, and privacy preserving AI/ML layer to enable an innovative network automation platform supporting all aspects of network and service management including the deployment and scaling of AIFs of different nature (i.e., latency-critical, low-latency, and latency-tolerant AIFs) over a distributed facility, and the various tasks needed to deploy such applications, e.g., the creation of a new network slice.

The remainder of this section describes the technological enablers on which AI@EDGE builds to convert the connect-compute platform into a reality, and that compose the functional blocks of the AI-driven platform, as sketched in Fig. 3.
A. Distributed and decentralized connect-compute platform

**Enabler.** $AI@EDGE$ combines the Function as a Service (FaaS) paradigm with serverless computing, hardware acceleration (GPU, FPGA and CPU), and a cross layer, multi-connectivity-enabled disaggregated Radio Access Network (RAN) into a single connect-compute platform to allow over-the-top providers to fully use the 5G capabilities though well-established cloud-native paradigms to develop and run applications. The serverless and FaaS approaches are gaining attention as cloud computing models in which the infrastructure provider manages on-demand infrastructure and resources, while the stakeholders (e.g., service providers) can focus only on their core activities. Building on this, $AI@EDGE$ intends to define a set of open APIs, by which network operators, vertical industries, service providers and users can interact with the network on a neutral host model. Moreover, the platform encompasses the path toward a hybrid multi-cloud-native deployment supporting Virtual Machines (VMs) and containers and their integration with the serverless paradigm.

**Innovation.** $AI@EDGE$ aims to account for this mixture by extending the current ETSI MEC/NFV architectures in which the infrastructure provider manages on-demand infrastructure and resources, while the stakeholders (e.g., service providers) can focus only on their core activities. Building on this, $AI@EDGE$ intends to define a set of open APIs, by which network operators, vertical industries, service providers and users can interact with the network on a neutral host model. Moreover, the platform encompasses the path toward a hybrid multi-cloud-native deployment supporting Virtual Machines (VMs) and containers and their integration with the serverless paradigm.

B. Orchestration of Artificial Intelligence Functions

**Enabler.** The provisioning of AI-enabled applications over a distributed computing platform requires reference models and standards, especially in heterogeneous and complex scenarios as edge computing platforms spanning across multiple domains. Defining these AI-enabled applications involves the representation of their AIFs (i.e., AI-enabled applications subcomponents). To this end, $AI@EDGE$ leverages standard knowledge representation languages and well-known state-of-the-art ontology engineering methodologies, to represent AIFs, as well as their relationships and status at the different levels of the technology stack. Conversely, $AI@EDGE$ considers “de-facto” standards for cloud and edge services orchestration, and from their emergent variants (e.g., FaaS) to for the end-to-end orchestration and chaining of AIFs.

**Innovation.** $AI@EDGE$ aims to build on the above ontologies to propose a reference model that provides the tools to describe AIFs, their requirements (e.g., storage, hardware acceleration, etc.), and the necessary metadata for their orchestration, which will also compose a catalogue of available AIFs. Furthermore, $AI@EDGE$ envisions innovative solutions for end-to-end orchestration to partition AIFs across different segments attending to their requirements (as shown in Fig. 2) considering the heterogeneity and complexity of the underlying edge computing platforms, and the collection of valuable quality of service indicators to create complex AI-enabled applications and detect abnormal situations.

C. Hardware-accelerated serverless platform for AI/ML

**Enabler.** The most recent hardware acceleration solutions (FPGA, GPU, and CPU) and privacy preserving ML techniques allow and speedup the execution of sensitive and computing-intensive workloads over the same platform. The deployment of heterogeneous acceleration platforms at the edge enables advanced processing scenarios to be exploited in far more complex processing functions. In addition to GPUs, which are currently the prime solution for AI/ML processes acceleration, FPGA are gaining momentum for deployments at the edge due to their ability to ensure optimal performance to execute specialized functions (e.g., real-time network intensive processing), in an energy and cost efficient manner.

**Innovation.** $AI@EDGE$ makes resource-aware hardware-acceleration techniques a key point of its design with the goal of increasing resource efficiency across the computing
continuum. As depicted in Fig. 2, AI@EDGE aims to go a step further by exploring approaches to tame accelerators' heterogeneity and enable their integration (both GPUs and FPGAs) with the serverless computing concept, in order to offer a unified platform able to allocate resources and migrate functionality between accelerators on different edge devices or between edge and cloud infrastructures.

### D. Cross-layer, multi-connected, disaggregated radio access

**Enabler.** Supporting beyond 5G use cases requires relying on different communication technologies to increase reliability, as exposed in Rel15 and Rel16 through dual-connectivity techniques using data duplication at the PDCP layer. However, besides reliability, various use cases demand greater flexibility and openness in the RAN to implement more advance multi-connectivity layers and and to enable a higher degree of automation. This demand is being promoted in O-RAN specifications [13], which propose an open architecture where RAN control and management functions are divided into near-Real-Time (nRT) and non-Real-Time (nonRT) RAN Intelligent Controller (RIC). AI@EDGE aims to rely on both multi-connectivity options and O-RAN specifications to deliver a flexible, open and unified platform including 3GPP and non-3GPP radio access technologies.

**Innovation.** Building on current multi-connectivity options for increased reliability (i.e., PDCP data duplication), AI@EDGE seeks to investigate different approaches for user-plane data replication over a multi-path with non-3GPP interfaces in order to meet the requirements of highly demanding services, as illustrated in Fig. 2. Moreover, the current O-RAN architecture will be extended to account for both 3GPP and non-3GPP technologies. This will make it possible the collection of dual RAN telemetry, and the extension of the network automation platform, which will allow the interaction with the non-RT and nRT RICs and performing actions on path selection and switching, among others.

### IV. USE CASES

AI@EDGE will be validated using four high-impact use cases. This section provide a brief description of each of them with a particular focus on commercial relevance and KPIs.

#### A. UC1: Virtual validation of vehicle cooperative perception

The automotive use case is cooperative perception. Several vehicles exchange data related to their trajectories. The data are used to build a high-definition map of the surrounding environment that can be used to predict potential collisions. Today validation of vehicles’ cooperative perception is a challenge because cooperative perception deals with numerous vehicles that have to: detect in real time the surrounding traffic scenario; exchange their sensed data; and share their intended manoeuvres with other vehicles. Large tests are needed even to address one single traffic scenario.

Cooperative perception tests become even more complex when dealing with mixed traffic scenarios. To overcome the problem of simulating the human behaviour by means of a mathematical model, we plan to interconnect a dynamic driving simulator operated by a real human driver with a traffic simulator like Car Maker (or VI-Grade) and to design, implement, and test the digital twinning of a mix of real and emulated vehicles. The goal is to recreate the network-level data exchange required to build a cooperative perception between emulated vehicles and a virtual human-driven vehicle.

The AI-based digital twinning process will make use of the AI-enabled application features and of the distributed and centralized serverless AI@EDGE platform. A key role here will be the AI@EDGE network and service automation features allowing the digital twinning to cope with mutating radio network environment. The AI@EDGE platform will be interfaced with a 5G network emulator to allow testing a broader range of scenarios and network configuration and related 5G Key Performance Indicators (KPIs) will be measured.

#### B. UC2: Secure and resilient orchestration of large Industrial Internet of Things (IIoT) networks

Smart factories will be characterized by 5G connectivity using massive machine-type communications slices to interconnect both IIoT and IoT devices. Such deployments consider the interconnection of independent network segments, potentially managed by different stakeholders. Therefore, guarantee the confidentiality of proprietary information in a multi-stakeholder environment while exploiting as much information as possible in AI/ML detection and decision-making is one of the mandatory requirements and key challenges. Furthermore, IIoT environments are very sensitive to latency and transmission timing and are governed by strict access control policies, a constraint with which AI/ML solutions must cope. Conversely, local anomaly detection solutions introduce a significant reduction of the detection capabilities, as local/edge detection mechanisms are unable to detect system-wide events and correlations. Federated learning, detection model propagation and parameter exchange among edge devices can be applied to mitigate these issues and exploit the full potential of distributed architectures in terms of security and intrusion detection while enforcing data confidentiality between the stakeholders.

This use case envisions the design and validation of mechanisms for secure orchestration of large scale IIoT applications on the AI@EDGE platform with the aim of conducting autonomous workload management on a unified compute fabric. Flexible, intelligent, and secure management solutions will be developed with focus on AI-enabled multi-tier infrastructures. Data-driven management components will be designed to operate in synergy with security technologies, thereby implementing intelligent, protected, and trustworthy network services supporting advanced 5G applications.

#### C. UC3: Edge AI assisted monitoring of linear infrastructures using drones in BVLOS operation

Drones are gaining attention in the industrial world for capturing data in a flexible and innovative way, offering new operation and maintenance procedures. Complex scenarios such as monitoring large linear infrastructures can benefit from extensive and diverse data capture mechanisms, configuration in Beyond Visual Line of Sight (BVLOS) flight, low-cost
operation, and use of combined on-board sensor solutions. Existing studies estimate that a small drone fleet could easily create 150 terabytes of data per day. To manage this volume of data in a production setting drones require a combination of on board and on the ground edge-computing capabilities. By leveraging 5G fast data transmission and advanced AI-enabled edge image and video processing solution, AI@EDGE will boost the efficiency of drones in their tasks.

To optimize these operations and provide innovative functionalities a reliable communication technology supporting high-rate data transfer and multi-connectivity is required in BVLOS scenarios. The 5G network must therefore allow for dynamic AI-aided workload deployment in combined drone, edge, and cloud environments, considering the tight latency restrictions and diverse computational capacities, to process a constant dataflow. The coordinated workflow will allow reducing response time and quick decision taking. Moreover, AI models will also carry on automated incident detection. In-flight problem determination as well as the use of drone fleets will significantly benefit from this approach to speed up the monitoring process and ultimately support more rapid and frequent preventative actions for maintenance.

D. UC4: Smart content & data curation for in-flight entertainment services

In Flight Entertainment and Connectivity (IFEC) focuses on delivering entertainment and engagement to airline passengers. The traditional IFEC system made of a central content server and embedded seat screens was conceived to offer only static content over a resource-constrained daisy-chained wired network. Currently, it is relentlessly evolving toward a broadband connectivity platform through on-board Wi-Fi technology and ground networks via high throughput satellite connectivity. Recently, with the maturity of 5G, network softwarization and AI, the IFEC system has the potential to transform into a smart edge-cloud platform composed of servers, multi-radio access technologies (5G, Wi-Fi, etc.) and embedded seat screens. As the content consumption pattern in the IFEC market is changing rapidly, there is an expectation to provide dynamically curated content to the airline passengers, based on data about the flight routes and passenger demographics (e.g., frequent flyer subscriptions). These content sources include traditional options (major movie makers) and new sources such as live and near-live streams including news, viral clips, etc.

The clear challenge is thus to develop efficient algorithms and flexible protocols for managing both the content and the edge infrastructure. In this context, the lightweight AI@EDGE platform will equip the IFEC system with edge computing capabilities and dynamic workload distribution within the infrastructure. Moreover, the multi-connectivity environment will allow performing as many tasks as possible at the aircraft edge without costly satellite technologies which, together with AI-aided efficient content distribution, stands for a clear path forward for the IFEC industry.

V. Conclusions

In this paper we presented the challenges and conceptual architecture of the AI@EDGE project. The project aims converging and evolving AI/ML and 5G at the network edges with the goal of providing a flexible platform on top of which the next generation AI-enabled application and services can be deployed. The paper describes also the four reference use cases that will be used to validate the project concept, namely: cooperative perception for vehicular networks, secure, multi-stakeholder AI for Industrial Internet of Things, aerial infrastructure inspections, and in-flight entertainment.

Acknowledgments

This work has been performed in the framework of the European Union’s H2020 project AI@EDGE co-funded by the EU under grant agreement No 101015922. The views expressed are those of the authors and do not necessarily represent the project. The Commission is not liable for any use that may be made of any of the information contained therein. The authors would also like to acknowledge CERCA Programme / Generalitat de Catalunya for sponsoring part of this work.

References