Enabling far-edge intelligent services with Network Applications: the automotive case

Konstantinos V. Katsaros, Eirini Liotou, Francesca Moscatelli, Theodoros Rokkas, Georgios Drainakis, Edoardo Bonetto, Daniele Brevi, Dimitris Klonidis, Ioannis Neokosmidis and Angelos Amditis

***Abstract*—The fifth generation of mobile networks (5G) is rapidly reaching deployment across the globe, promising a series of advances for vertical service providers, both in terms of performance and in terms of operational capabilities. In this context, the 5G-IANA Network Application platform focuses on the rapidly advancing domain of intelligent, data centric, Artificial Intelligence / Machine Learning (AI/ML)-enabled applications, with a particular focus on the automotive domain. In this paper, we present the key functional features designed for the support of such services including the integration of (mobile) far-edge resources, as well as ML-aware orchestration primitives. This includes novel features such as decision support for the optimal distribution of end-to-end ML pipelines, as well as run-time support for client selection in Federated Learning setups, far-edge failure handling and distribution drift aware lifecycle management. Such features come to address a series of limitations associated with legacy 5G management & orchestration systems, such as resource consumption of data centric services and privacy support. In this context, we further discuss the new opportunities arising for service provisioning and corresponding business models in the automotive ecosystem, with a particular emphasis on the implications of the emerging data and/or ML-model sharing schemes.**

***Index Terms*—5G, AI/ML, management, orchestration, CAM, vehicular, federated learning**

I. INTRODUCTION

5G is expected to provide the connectivity fabric for a multitude of novel services in a series of vertical domains e.g., automotive, manufacturing, etc., to name a few. In numerous cases, these services are characterized by the pivotal role of the vertical-oriented end devices e.g., fleet vehicles, factory robots/Automated Guided Vehicles (AGVs), etc., which, as opposed to general purpose terminals e.g., smartphones, are typically under the operational control of the vertical service provider. This role often manifests itself through the generation of valuable data capturing (vertical) operational conditions and further feeding decision making and eventually AI/ML-enabled intelligent automated services. In this paper we focus on the Automotive domain as a notable example domain, of particular economic interest. Indeed, the global connected car market is projected to reach a value of €200 billion by 2025, growing at a compound annual growth rate (CAGR) of 14.8% from 2018 to 2025. It is also forecast that more than 125 million passenger cars produced in the next four years will be equipped with embedded connectivity, out of a total of 1.2 billion motor vehicles in use worldwide[1]. This connectivity is expected to fuel a series of AI/ML-based services, such as Automated Driving, Trajectory Planning, Preventive Maintenance, etc., which rely on the rich data generated by vehicles, often (expected to be) operated in fleets e.g., in Mobility-as-a-Service (MaaS) schemes [4], transportation/logistics truck platoons [5]. Such data include mobility traces, sensor readings, communication data, etc., and can often reach high volumes e.g., up to TBs per vehicle per day [6].

While enhanced 5G connectivity is rightfully seen as a facilitator to such services, a series of constraints raise significant barriers in translating this valuable data to services. Privacy concerns, energy consumption / sustainability requirements, economic viability and performance reasons often challenge the well-established centralized model that largely builds on the aggregation of high volumes of data in centralized data lakes [7].

Admittedly, several solutions have emerged against the aforementioned barriers e.g., Federated Learning solutions against privacy restrictions [8], distributed computing systems for the support of the compute continuum e.g., Kubernetes[2], MicroK8s[3], KubeEdge[4], as well as mobile offloading optimization solutions tackling energy-performance tradeoffs

Paper submitted for review on September 24th, 2022. This paper is part of the 5G-IANA project, co-funded by the EU under the H2020 Research and Innovation Programme (grant agreement No 101016427). *(Corresponding author: Konstantinos V. Katsaros).* Konstantinos V. Katsaros, Eirini Liotou, Georgios Drainakis and Angelos Amditis are with the Institute of Communication and Computer Systems, Athens, Greece (e-mail: {k.katsaros, eirini.liotou, giorgos.drainakis, a.amditis}@iccs.gr). Francesca Moscatelli is with Nextworks, Pisa, Italy (e-mail: f.moscatelli@nextworks.it), Theodoros Rokkas and Ioannis Neokosmidis are with inCITES Consulting, Luxemburg (e-mail: {trokkas, i.neokosmidis}@incites.eu), Edoardo Bonetto and Daniele Brevi are with LINKS Foundation, Turin, Italy (e-mail: {edoardo.bonetto, daniele.brevi}@linksfoundation.com), Dimitris Klonidis is with UBITECH, Limassol, Cyprus (e-mail: dklonidis@ubitech.eu).

[1] https://www.marketsandmarkets.com/PressReleases/connected-cars.asp

[2] Kubernetes, https://kubernetes.io/ (last accessed 01/11/2022)

[3] MicroK8s, https://microk8s.io/ (last accessed 01/11/2022)

[4] KubeEdge, https://kubeedge.io/en/ (last accessed 01/11/2022)

[9]. Nevertheless, as we discuss in this paper, such individual solutions lack a holistic service-oriented approach that would inherently integrate service management and provisioning considerations, taking into account the aforementioned limitations. Most importantly they lack visibility and control of vertical far-edge devices[5] and their data in a privacy aware, but unified, in terms of 5G-enabled service provisioning approach. In this context, this paper presents the 5G-IANA Network Application (5G) service provisioning platform, which aims to address the above limitations by supporting:

- The encapsulation of data generation, pre-processing and processing within service level re-usable and orchestrate-able application and network functions (AFs/NFs), organized in Network Application packages.
- A clear separation between vertical customer facing functionality and network side operations, including network slicing, as well as Management & Orchestration (M&O) operations.
- The explicit support of service provisioning primitives on top of far-edge resources, allowing to move any step of the data manipulation process across the extended compute continuum i.e., including the user equipment (UE).
- The explicit support of distributed ML schemes such as Federated Learning (FL).

Focusing on the automotive domain, we discuss how these features enable data centric services and novel use cases that were previously either impossible or inefficient to support, giving ground to new business models.

II. CHALLENGES

A. *Privacy*

Privacy concerns increase with the integration of devices that sense the environment of the terminal equipment user, and capture potentially sensitive / personal data e.g., daily trajectories, communication patterns, video footage or even Cooperative Awareness Messages (CAM) and Decentralized Environment Notification Messages (DENM) in Intelligent Transportation System (ITS) applications, etc. Subject to the enforcement of the European General Data Protection Regulation (GDPR)[6], these concerns are inherently tight with the aggregation of this data at centralized locations and their exposure to vertical service providers (or even third parties). As the 5G/6G ecosystem is evolving along the lines of digital transformation, including Private 5G deployments [10], more and more cases emerge where the sensing (end) device is owned by the vertical service operator e.g., vehicle fleets, smart city sensors (cameras). Nevertheless, privacy concerns do remain in many cases associated to the participation of vertical service customers e.g., vehicle fleet owner (smart mobility), smart city citizen. In view of the extend of digital transformation already taking place, explicit user consent, that often suffices to provide a legal basis for data exposure, may not be taken for granted.

B. *Resource efficiency – Sustainability*

Centralized data aggregation (and subsequent processing) may not be preferable, even in cases where privacy restrictions are not present or user consent if available. The potentially high volume of data generated by the far-edge devices may challenge a data lake model, subject to the associated resource consumption i.e., data aggregation may result in the excessive utilization of available bandwidth yielding a prohibitive energy and/or monetary cost footprint [2][3]. Obviously, this is only part of a broader optimization challenge which further includes aspects related to processing (energy) costs and performance tradeoffs. As discussed in [1] [2], the optimal dimensioning and placement of a data centric vertical service depends on the specificities of the vertical service in terms of data volume, as well as AI/ML aspects e.g., type/size of ML model used, hyperparameters, etc. The challenge becomes even more profound when further considering the continuous adaptation of the vertical service at hand to data distribution drifts, resulting in repetitive data aggregation and processing tasks [11].

C. *Obstacles to DML/FL Schemes*

The above challenges point to distributed solutions, where computation (i.e., ML training) takes place at the data source, avoiding the exposure of private data and the potentially excessive resource consumption associated to data aggregation. Obviously, these are functional features of FL [13], which already emerges as a strong candidate solution to the problem. FL implementations are currently largely provided as an over-the-top solution. Practical frameworks already exist[7] able to deploy an FL-enabled service over 5G. However, their operational scope does not go beyond the edge of the fixed network. Supporting such distributed ML (DML) schemes at their full extend i.e., over UE devices generating the data, currently requires an out-of-band mechanism i.e., decoupled to service provisioning inside the network. In view of the increased penetration of 5G enabled connectivity in IoT domains and digital transformation, this causes a problem as it overlooks the fact that UE-connected devices may well/often belong to the operational domain of the vertical service provider e.g., vehicle fleet, Industry 4.0 robots, etc., especially in view of Private 5G deployments.

III. THE 5G-IANA NETWORK APPLICATION APPROACH

The 5G-IANA Network Application experimentation platform [15] high-level design is presented in Figure 1. The proposed architecture aims to tackle the aforementioned limitations, offering service providers mechanisms to easily design distributed intelligent services, which span from the remote cloud to the far-edge segment, and request their provisioning on top of 5G-enabled infrastructures. The platform is realized by four main building blocks:

- **Network Application Orchestration and Development**: the entry point for service providers. It exposes functionalities for designing distributed services composed

[5] By far-edge devices we refer to User Equipment (UE) connected to the remainder of the 5G System infrastructure through the Uu interface.

[6] https://gdpr-info.eu/

[7] E.g., FLOWER: https://flower.dev/ (last accessed: 01/11/2022)

by Network Applications. This layer hosts also a catalogue of available Network Applications that can be used and chained to realize the desired service.

- **Slice Management and Resource Orchestration**: this layer implements the functionalities for verifying the availability of a network slice instance suitable for supporting the operation of the vertical service. It also handles the orchestration of computational resources to be allocated to run the Network Applications.
- **Data Collection, Monitoring and Analytics**: it realizes the collection of data from distributed data sources (i.e., Network Application, infrastructure hosts etc.) and provides analytics based on service-level policies to optimize the Lifecycle Management (LCM) operations.
- **DML Orchestration**: provides explicit support for ML-oriented services, including FL primitives such as client selection and enhanced LCM e.g., drift management.

The platform provides service providers with the ability to wrap all data manipulation processes within Network Applications that can be re-used in broader service chains/graphs. The Network Application catalogue builds atop the Network Application Package construct. On the one side, the platform exposes this construct to service providers, allowing them to describe their service (see Section III.A). On the other side, the platform interfaces both the 5G System and the available far-edge/UE-side resources/nodes (Section III.B) so as to enable a series of Distributed Machine Learning Orchestration (DMLO) functional primitives (Section III.C). In the following, we describe the main concepts behind the platform.

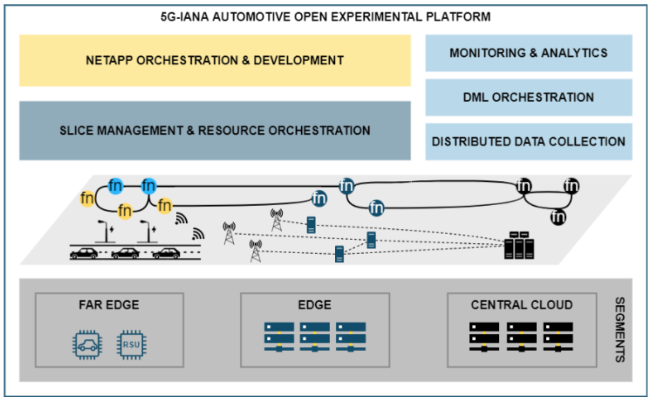
5G-IANA AUTOMOTIVE OPEN EXPERIMENTAL PLATFORM
NETAPP ORCHESTRATION & DEVELOPMENT
MONITORING & ANALYTICS
DML ORCHESTRATION
SLICE MANAGEMENT & RESOURCE ORCHESTRATION
DISTRIBUTED DATA COLLECTION
fn
FAR EDGE
EDGE
CENTRAL CLOUD
SEGMENTS
RSU

Figure 1: 5G-IANA Network Application (NetApp) Experimentation Platform

A. *Network Application Package*

In 5G-IANA, a Network Application is defined as a virtualized application that leverages 5G performance capabilities. Each Network Application implements and exposes a specific service. Network Applications constitute baseline components within more complex and distributed service-chains, depending on the specific logic that the service provider would like to implement. The Network Application Package is the construct that enables the Network Application chaining, orchestration and possible reuse in multiple vertical services. Indeed, the proposed Network Application Package information model has the primary objective of facilitating the re-usage of a Network Application by service providers to realize new 5G-enabled end-to-end services. To this end, the Network Application Package must include two different layers of information:

- **Service-level information**: it aims at simplifying the re-usage of a Network Application Package. Typical information includes the Network Application documentation, especially for what concerns the exposed services and interfaces and how the proper operation of the Network Application can be assessed. In addition, part of the information is structured in a Network Application Template that specifies high-level Quality of Service (QoS) parameters (e.g., the number of UEs to be accommodated). The Network Application Template includes also a high-level representation of the Network Application graph in terms of constituent components and related constraints e.g., features of vehicular (far-edge) resources required in service provisioning.
- **Orchestration-level information**: this type of information is used to actually perform the provisioning and LCM of a Network Application. It includes the expected 5G QoS indicators to be fulfilled formalized into a Network Slice Type (NEST) template. Network Application components packages and cloud-native service descriptors (i.e., Kubernetes compatible service templates) are included to enable the actual provisioning and orchestration of the Network Application.

Figure 2 provides a visual overview of the 5G-IANA Package information model.

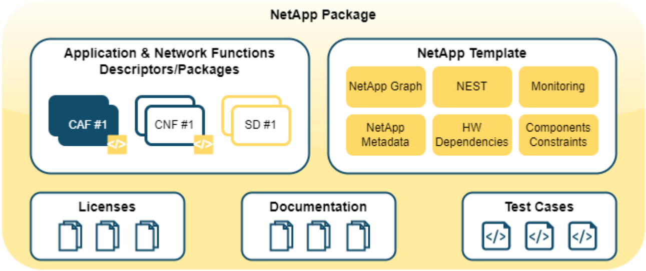

Figure 2: 5G-IANA Network Application (NetApp) Package

Network Application packages are baseline components for automotive verticals to create and operate in a facilitated manner innovative distributed service-chains that leverage 5G services. Indeed, one or more Network Application packages can be re-used and composed together by verticals: the service provider should understand which services a Network Application provides and how it can be chained with its own Network Applications, while the inner details related to the Network Application deployment are not exposed and its orchestration is handled in a transparent manner by the proposed platform.

Reusability is the “leitmotiv” behind the proposed information model. Examples of Network Application Packages to be used as baselines to compose new Automotive services include vehicles’ communication services (e.g., ETSI ITS long-distance communication), device-specific interfacing services (e.g., Advanced Driver Assistance Systems (ADAS) interfaces, sensors/actuators interfaces, cameras interfaces, etc.). Also, specific services related to manoeuvres coordination, infotainment etc. can be packaged as Network Applications and reused/chained with different Network Applications exposing compatible interfaces.

The described modelling is suitable to support the orchestration of AI/ML Network Applications, which indeed bring specific requirements in terms of data availability, data types, areas of service coverage, etc. As depicted in Figure 3, AI/ML pipelines composed by different functions can be packaged as DML Network Application service-chains and deployed on-demand to support the operation of vertical services. The possibility of packaging AI/ML pipelines into Network Applications and orchestrating them across remote cloud, edge and far-edge resources brings a great opportunity to verticals for exploiting the AI/ML potential in the Automotive segment.

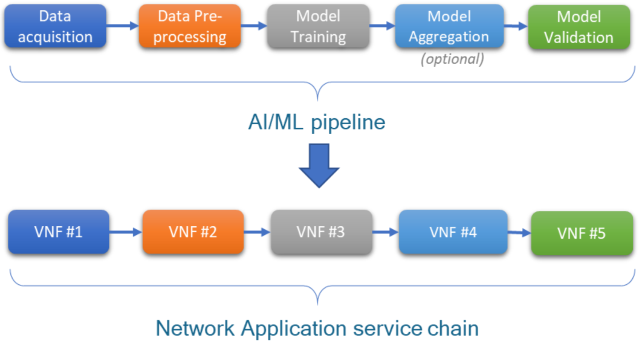

Figure 3: DML Service-chain Example

B. *UE Integration*

Given the focus on the automotive vertical, the far-edge devices in 5G-IANA, can be of the following types:

- **On-Board Unit (OBU)**: it is the main element on the vehicle that is responsible for the communication between the services on the vehicle and the edge and cloud services.
- **Road-Side Unit (RSU)**: it is used to collect data from roadside sensors, process them, and provide the raw or processed data to the relevant edge and cloud services.

The OBUs and RSUs, that are considered in 5G-IANA, are enhanced devices with respect to the currently commercially available ones, since they are devised: (i) to support AI applications (i.e., they are equipped with graphics processing units), and (ii) to be added as Points of Presence (PoPs) in the 5G-IANA service provisioning platform. These devices can indeed support the orchestration and the LCM of Network Applications through the deployment of virtualization and orchestration tools such as Kubernetes.

The service provisioning to OBUs and RSUs is critical from different points of view. First, the OBUs and RSUs are typically embedded devices, and they likely have a constrained availability of compute and storage resources. Second, the availability and quality of network connectivity of these far-edge devices may also change a lot depending on their physical location. Furthermore, the OBUs' connectivity varies over time as the vehicles move.

To achieve an effective service provisioning, it is essential that the 5G-IANA platform considers all these aspects. The far-edge devices continuously have their compute and storage resources, as well as 5G connectivity status monitored (see *Data Collection, Monitoring and Analytics*). This information is provided to the 5G-IANA Network Applications provisioning platform in such a way that it can decide how and when to provision Network Applications to the OBUs and RSUs in order to fulfil the Service Level Agreement (SLA) of the different applications that should be deployed.

The Network Applications provisioning service can also be customized depending on the current location of OBUs. This information allows for implementing a service provisioning tailored to the needs of OBUs that are currently in a given geographic area and therefore have special needs (e.g., provisioning a Network Application related to a country-specific regulation). This functionality further gives ground to the *Client Selection* feature of the *Distributed ML Orchestrator* (DMLO) as described in the following.

The last aspect to be considered in the 5G-IANA context is that the OBUs and RSUs may lose the connection to the service provisioning platform. This event must be properly handled from both platform and far-edge device sides. Through monitoring, the platform identifies the disconnection, informs the relevant involved actors, and manages the reconnection event. The OBUs and RSUs should be able to manage the provisioned Network Applications without the support of the service provisioning platform for ensuring the continuity of the operations.

C. *Distributed ML Orchestration*

The pervasiveness of AI/ML in various vehicular applications goes hand in hand with the requirement to wrap UE Integration toolsets as the above, in functional, ready-to-use frameworks, where minimal domain expert knowledge and resources are required for configuration/optimization. ML-as-a-Service (MLaaS), which is currently limited to (centralized) cloud services, has emerged as a notion that aims to ease the process of ML training and deployment, by abstracting away challenges related to the intermediate steps of the ML process e.g., data acquisition, storage, feature engineering, data cleaning, tuning, training, monitoring, etc. The DMLO seeks to extend the traditional concept of MLaaS, facilitating AI/ML deployment in distributed (far-edge) environments. In that sense, the UEs (solely regarded as data sources in the past) can be part of the training process, therefore enabling the deployment of distributed ML schemes (see Figure 4) such as FL. In the following, we detail the primitive functionalities of the DMLO, which realize baseline M&O and LCM support.

ML-pipeline topology optimization: The selection of the most appropriate overlay service topology, corresponding to the placement of the various elements of a ML-pipeline on the compute continuum (underlying physical topology), depends on several factors ranging from privacy restrictions to resource consumption costs and energy footprint, as discussed in Section II, i.e., from fully distributed star topology in typical FL environments, to fully centralized ones (see Figure 4). While in some cases the decision is straightforward e.g., FL for privacy sensitive services, in other occasions it may require more complex optimization logic e.g., related to the resource footprint [1]. The DMLO aims to provide support for such cases by (i) realizing the corresponding decision support algorithms

[1], and by (ii) guiding the vertical service provider in the delivery of the required input information e.g., on the available sizes of the ML model at hand, the expected volume/rate of data, privacy requirements, desired UE features such as location, etc.

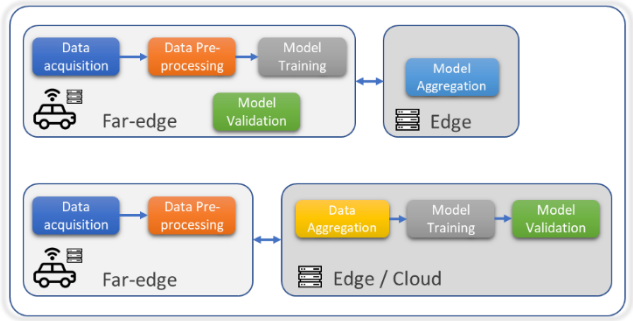

Figure 4: ML-as-a-Service examples: (top) fully distributed, FL-based deployment; (bottom) semi-distributed / centralized (hybrid) deployment

Client Selection: Choosing the clients (far-edge nodes/UEs) that will participate in each training round is a challenging problem in any distributed ML task [12]; a suboptimal selection may not only introduce delays in the training process, but also endanger the convergence of ML task. For that cause, the DMLO is designed to offer a multi-parameter selection scheme, based on both ML-specific (e.g., feature availability, data entropy, data freshness, etc.) and generic (data volume, connectivity, processing capacity, etc.) client criteria. On top of these criteria, client selection can be performed in various modes with respect to temporal granularity:

- One-off, where client selection occurs only once, during initialization (minimal overhead, low flexibility).
- Iterative, in a per training round basis (medium overhead, more flexibility).
- Event-based i.e., upon node change detection (maximum flexibility and overhead).

Node Failure Handling: The DMLO is responsible for identifying a potential node failure e.g., due to loss of connectivity, and subsequently supporting graceful handling in the training process, by communicating related information to the AI/ML service. By leveraging the node (UE) health monitoring functionality (see also Figure 1 and Section III) provided by the platform, stragglers and drop-out nodes can be identified. Subsequently, the training process can be modified according to policies specified by the vertical service provider e.g., use event-based client selection primitives to remove a straggler node in a synchronous ML/FL scheme.

Termination Control and Model Monitoring: Refers to LCM handling i.e., controlling when the training process stops, subject to predefined criteria, related to e.g., ML task convergence, resource consumption restrictions, etc. Upon task completion, the (trained) deployed ML model performance is monitored at regular intervals, to account for model ageing i.e., degradation of model's accuracy, due to a (severe) change in the underlying data distribution. In the latter case, the DMLO will provide an option for the automated re-initialization of the ML training process.

IV. BUSINESS-LEVEL OPPORTUNITIES: THE AUTOMOTIVE CASE

5G promises to unlock new business models and enlarge the existing ecosystem. Going one step further, the creation of platforms like 5G-IANA will give the opportunity to third parties, that until now had no such role in the ecosystem, to create and/or offer their own Network Applications and services to users. By analyzing the rapid growth of private networks in the US, as well as the central role of cloud infrastructure providers like Amazon and Microsoft offering their services to users, either directly or through Mobile Network Operators (MNOs) (end-to-end solution), we expect that something similar will happen to platforms like 5G-IANA used for the development and provisioning of Network Applications. Especially in the automotive sector, there is a number of established stakeholders like the Vehicle Manufacturers (Original Equipment Manufacturers, OEMs) that are trying to identify their position in the new ecosystem.

Towards this direction, we initially identify the stakeholders that are expected to have (at different degrees) interactions with the 5G-IANA platform and group them into the following broad categories:

- Infrastructure providers who provide different types of resources such as MNOs, Cloud Providers and Road Infrastructure operators.
- Vendors who provide either H/W or S/W to all other actors.
- Vehicle manufacturers (OEMs) who provide the vehicles.
- Related to compliance, such as the different regulatory authorities, standardization bodies and policy makers.
- Service developers who include all types of service creators and research centers.
- Network Application-related players who include both Network Application developers and providers.
- Service providers that provide the service to users.
- Users that can be classified into Business in a B2B relation and customers (B2C).

Depending on the type of the developed Network Application, different stakeholders will interact with the 5G-IANA platform. Hereby we focus on the generic scenario where a stakeholder who owns a fleet of vehicles is in the position to collect data that can be used from the AI/ML models and then by different services. The following business models can then be defined based on the different levels of synergies that can be achieved between the stakeholders:

Single owner-user: the vertical user (fleet owner) is leveraging the data collected from its own fleet and uses it for its own benefit (for example predictive maintenance of the fleet vehicles). In this model, there can be several fleet owners that use their own data leading to the creation of different silos. It should be highlighted that the semi-distributed / centralized (hybrid) deployment of Figure 4 can be used in this case.

Multiple owner-users: The second model includes the collaboration among fleet owners with the exchange of the models among them in an effort to increase the flexibility and reliability of the services that use the collected data. (Cross-silo). In this case, a fully distributed FL-based deployment can be used. Each fleet uses the data to create a model which, combined with other models, can result in an aggregated model. That way, data are not exposed to other actors while at the same time all involved actors benefit from the aggregated model. Another option would be to exchange the models through a dedicated P2P network that will be formed by the interested parties.

Single – multiple owners – third parties as users: In this model, the vertical user(s) are providing the models to third parties that are interested to improve their services. Third parties can use the model from single users or combine the models from several users to create an aggregated model.

Single – multiple owners and MNO as a user: This model applies in the particular scenario where the AI/ML training process focuses on use cases with particular interest for the MNOs i.e., Predictive QoS [14]. Here, the data collected refer to the performance of the network as perceived by the vertical end user(s) and can be enhanced by network side data e.g., cell conditions. A synergy can be envisioned where the MNO can act as the eventual recipient and operator of the trained model i.e., in delivering Predictive QoS to automotive applications. The model can be trained on a multiplicity of sources, namely the vertical users themselves taking advantage of the high number of available data coming from their devices. It is noted, that alternatively, costly measurement campaigns are the only option for MNOs to collect their own data about network performance. MNOs can either exploit a single model or aggregate the models of different vertical users. Different agreements between MNOs and users on how to define the price for the models can be made.

V. CHARACTERISTIC USE CASE

In this section, we describe a characteristic use case (remote driving) that is enabled by the 5G-IANA Network Application (5G) service provisioning platform. This use case consists of three distinct and standalone services (Network Applications) that may be combined to run in sequence: a) an ML ***training service*** aimed to support the delivery of a trained ML model suitable to deliver QoS predictions [14], based on data generated by the vehicles, b) a network status monitoring and predictive QoS ***inference service***, and c) a vertical-/application-specific ***remote driving service***, that is tailored to the requirements and needs of any third-party automotive service provider. The *remote driving service* is meant to consume the *inference service*, i.e., to utilise the related predictive QoS information offered through the monitoring of the 5G network status from a UE/OBU perspective. Similarly, the *inference service* utilises the model artefact produced by the *training service*.

The training and inference services are enabled by means of the DMLO part of the 5G-IANA Network Application platform with the assistance of data generated by the far-edge devices (i.e., OBUs) (see Section III.B). All services are realized, by the respective vertical service provider(s), with the appropriate configuration of the Network Application Package service-level and orchestration-level information (see Section III.A) through the 5G-IANA Platform. The resulting service chain(s) is(are) composed of reusable AFs/NFs, offered by the Network Application "starter-kit" and any other proprietary VNFs at hand, stitched together in a meaningful order and with an appropriate configuration.

Delving into more details regarding this specific vertical service, and in particular the remote driving use case in the context of the 5G-IANA Network Application platform, this is envisaged to become reality as follows:

Training service (Network Application): The following functionality is realized (see Figure 5): (1) a model aggregator (here, lying in a Multi-access Edge Computing (MEC) server) conducts *Client Selection* for instance on node (vehicle) location, data availability, data volume, and data features criteria; this selection process is enabled by the DMLO which maintains a view of the available OBUs; (2) dispatches the currently available global model to the selected set of OBUs; (3) the OBUs perform local model training; (4) they upload the local models back to the aggregator, which then (5) aggregates the local models to produce a fresh version of the global model.

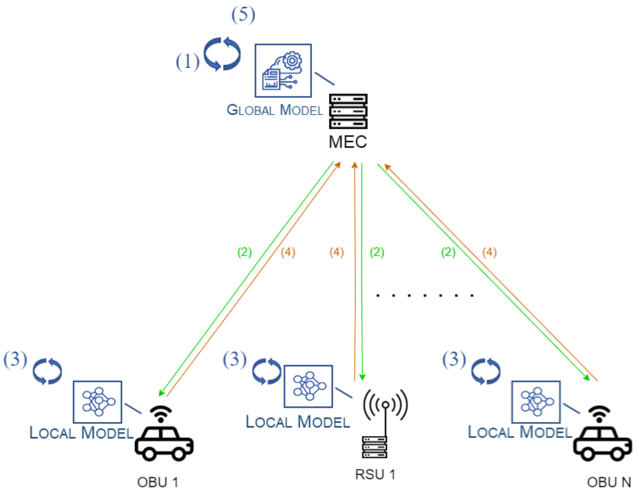

Figure 5:The DML-FL framework enabled through the 5G-IANA platform

Inference service (Network Application): Then, using the latest available version of this global model, each OBU would be in the position to conduct inference, namely, as examples, to conduct data traffic / QoS predictions, and to distinguish between normal and abnormal network behaviours as an attempt to predict network service deterioration. Alternatively, the global model can be deployed at the edge remotely supporting predictions for the vehicles; a decision subject to resource optimization (see Section III).

Remote driving service (Network Application): Then, the consumption of the previous services by the remote driving service is realized with the delivery of an In-advance QoS Notification (IQN) to the latter. The purpose of the IQN would be to help the remote driving application take a more informed decision in the context of expected QoS degradation e.g., a triggering of an “alarm”, making a remote operator take full control and drive the vehicle (see Figure 6); or/and the adjustment of the autonomous vehicle’s video stream quality so that fewer network resources are consumed; or/and instructing the vehicle to safely make a stop, among other potential options.

Figure 6: Illustration of remote driving concept of an AGV vehicle (https://fivecomm.eu/)

This example use case can be mapped to several business models (see Section IV). One potential option is the “single owner-user” model, where the OEM in charge of the remote driving application uses the QoS prediction inference that is prepared exclusively by its own fleet. However, we cannot exclude cases where each one of the aforementioned three distinct services is owned and run by three different OEMs, who use SLAs in order to collaborate with each other and take advantage of each other’s offered services. Another viable business model is the “multiple owner-users” (i.e., cross-silo concept), where the training model is built through all involved OEMs’ fleet-generated data. Such an approach would be particularly valuable, as it would lead to the collection of more, in terms of quantity and of geospatial spread, real-time network-related data, and thus, to the training of a more accurate ML model. In turn, this would produce more reliable IQNs, handed over for the disposal and management of more than one fleet owners, enforcing mutual benefits for all involved parties. Privacy-related challenges are not to be neglected though in such kinds of scenarios, as discussed in previous sections (Section II-A).

VI. CONCLUSION

The 5G ecosystem is rapidly emerging with deployments across the globe promising a series of performance and operational advances to vertical service providers. The 5G-IANA Network Application platform aims to take a next step in this direction in supporting next generation intelligent, AI/ML-enabled automotive services. This is accomplished through a series of functional features including the LCM of AI/ML pipelines, including the optimized selection of the corresponding pipeline component placement across the compute continuum, which subsequently builds on innovative platform features such as the support for service provisioning over far-edge resources, client selection capabilities in federated learning deployments, failure handling and drift management. This is expected to facilitate a series of interesting service provisioning / business models, including cross-silo interactions.

REFERENCES

[1] I. Sartzetakis, P. Soumplis, P. Pantazopoulos, K. V. Katsaros, V. Sourlas and E. Varvarigos, “Resource Allocation for Distributed Machine Learning at the (Edge-Cloud) Continuum”, in Proc. of the IEEE International Conference on Communications (ICC): Communication, QoS, Reliability and Modeling Symposium (IEEE ICC'22 - CQRM Symposium), Seoul, Korea (South), May, 2022

[2] G. Drainakis, P. Pantazopoulos, K. V. Katsaros, V. Sourlas and A. Amditis, " On the distribution of ML workloads to the network edge and beyond," in Proc. of the First International INFOCOM Workshop on Distributed Machine Learning and Fog Networks (INFOCOM WKSHPS FOGML) 2021,,,

[3] G. Drainakis, K. V. Katsaros, P. Pantazopoulos, V. Sourlas and A. Amditis, “Federated vs. Centralized Machine Learning under Privacy-elastic Users: A Comparative Analysis,” in Proc. of 19th IEEE International Symposium on Network Computing and Applications, 2020 (IEEE NCA 2020)

[4] C. Expósito-Izquierdo, A. Expósito-Márquez, and J. Brito-Santana, “Mobility as a Service,” Smart cities: Foundations, principles, and applications, pp. 409–435, Wiley Online Library, 2017.

[5] A. Balador, et al. "A survey on vehicular communication for cooperative truck platooning application." Vehicular Communications (2022): 100460.

[6] Dell EMC, “PowerScale Deep Learning Infrastructure with NVIDIA DGX A100 Systems for Autonomous Driving,” December 2021 (White Paper)

[7] A. Cuzzocrea. “Big data lakes: models, frameworks, and techniques.” 2021 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2021.

[8] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, “A survey on security and privacy of federated learning,” Future Generation Computer Systems, vol. 115, pp. 619–640, 2021, doi: https://doi.org/10.1016/j.future.2020.10.007.

[9] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," in IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628-1656, thirdquarter 2017, doi: 10.1109/COMST.2017.2682318.

[10] A. Rostami, "Private 5G Networks for Vertical Industries: Deployment and Operation Models," 2019 IEEE 2nd 5G World Forum (5GWF), 2019, pp. 433-439, doi: 10.1109/5GWF.2019.8911687.

[11] Mehmood, H., Kostakos, P., Cortes, M., Anagnostopoulos, T., Pirttikangas, S., & Gilman, E. (2021). Concept drift adaptation techniques in distributed environment for real-world data streams. Smart Cities, 4(1), 349. doi:https://doi.org/10.3390/smartcities4010021

[12] Li, Tian, et al. "Federated learning: Challenges, methods, and future directions." IEEE Signal Processing Magazine 37.3 (2020): 50-60.

[13] M. Shaheenet, et al. “Applications of federated learning; Taxonomy, challenges, and research trends.” Electronics 11.4 (2022): 670.

[14] 5G Automotive Association (5GAA), “5GS Enhancements for Providing Predictive QoS in C-V2X,” White Paper, May, 2020.

[15] F. Moscatelli, et al. “The 5G-IANA platform: Bringing far-edge resources and ML potential to the disposal of automotive third parties,” Symposium on 5G for Connected and Automated Mobility, IEEE Future Networks World Forum, Oct. 2022.