POLITECNICO DI TORINO
Repository ISTITUZIONALE

Demo: AIML-as-a-Service for SLA management of a Digital Twin Virtual Network Service

Original
Demo: AIML-as-a-Service for SLA management of a Digital Twin Virtual Network Service / Baranda, J.; Mangues-Bafalluy, J.; Zeydan, E.; Casetti, C.; Chiasserini, C. F.; Malinverno, M.; Puligheddu, C.; Groshev, M.; Guimaraes, C.; Tomakh, K.; Kucherenko, D.; Kolodiazhnyi, O.. - STAMPA. - (2021). (Intervento presentato al convegno IEEE INFOCOM 2021 - Demo Session tenutosi a Virtual conference due to COVID-19 nel May 2021) [10.1109/INFOCOMWKSHPS51825.2021.9484610].

Availability:
This version is available at: 11583/2871096 since: 2021-02-15T10:21:08Z

Publisher:
IEEE

Published
DOI:10.1109/INFOCOMWKSHPS51825.2021.9484610

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

25 April 2024

Demo: AIML-as-a-Service for SLA management of a Digital Twin Virtual Network Service

J. Baranda†, J. Mangues-Bafalluy†, E. Zeydan†, C. Casetti⋆, C. F. Chiasserini⋆, M. Malinverno⋆, C. Puligheddu⋆, M. Groshev‡, C. Guimarães‡, K. Tomakh§, D. Kucherenko§, O. Kolodiazhnyi§
†Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Spain, ⋆Politecnico di Torino, Italy,
‡Universidad Carlos III of Madrid, Spain, §Mirantis, Ukraine

***Abstract*—This demonstration presents an AI/ML platform that is offered as a service (AIMLaaS) and integrated in the management and orchestration (MANO) workflow defined in the project 5Growth following the recommendations of various standardization organizations. In such a system, SLA management decisions (scaling, in this demo) are taken at runtime by AI/ML models that are requested and downloaded by the MANO stack from the AI/ML platform at instantiation time, according to the service definition. Relevant metrics to be injected into the model are also automatically configured so that they are collected, ingested, and consumed along the deployed data engineering pipeline. The use case to which it is applied is a digital twin service, whose control and motion planning function has stringent latency constraints (directly linked to its CPU consumption), eventually determining the need for scaling out/in to fulfill the SLA.**

I. Introduction

Artificial Intelligence (AI) and Machine learning (ML) are increasingly pervasive and widely used for a fully-autonomous configuration and management of new-generation cellular networks, as confirmed by multiple ongoing initiatives ([1]–[3]). Along with this trend, the AIML-as-a-Service (AIMLaaS) paradigm has emerged to provide affordable access to AI/ML-based solutions on demand, in spite of the high computational and energy cost that such models may entail.

To enable the AIMLaaS paradigm, we have designed and developed a flexible, efficient AI/ML platform, which in this demo is exploited for Service Level Agreement (SLA) support in 5G networks, and more specifically, scaling. Such a platform can accommodate both supervised ML and reinforcement learning models, train them when needed, and create the files required during the inference phase for the model execution. The AI/ML platform has then been integrated in the 5G network NFV/SDN-based architecture defined within the EU 5Growth project [4] (5Gr) and used to implement ML-driven service scaling operations at the Service Orchestrator (5Gr-SO) whenever needed to meet the SLA requirements. In this direction, the 5Gr-SO, according to the network service descriptor (NSD), will automatically configure the metrics required and pass them through a complete data engineering pipeline (including the data collection, ingestion and consumption). The model is also automatically requested and downloaded to the 5Gr-SO that deploys it as part of its SLA management logic for the service at instantiation time. Finally, the model starts consuming real-time streamed data and provides decisions to take real-time scaling in/out decisions that are then applied by the 5Gr MANO stack.

Work supported in part by EU Commission H2020 5Growth project (Grant No. 856709) and H2020 Europe/Taiwan 5G-Dive project (Grant No. 859881).

In this demo, we show the structure and the internal flow of the AI/ML platform (5Gr-AIMLP), its integration in the 5Gr architecture, and how it can be effectively used to enact service scaling taking a Digital Twin used in Industry 4.0 scenarios as a reference network service (NS). Specifically, the audience can look at: (i) the functional workflow for ML model (and dataset) uploading and model training, (ii) the service instantiation and automated configuration of the monitoring platform for later feeding real-time data to the AI/ML model, (iii) the triggering of the automated scaling in/out based on the AI/ML model output and the scaling operational flow.

II. System Architecture

Fig. 1 presents the setup under demonstration. This setup has been deployed in two different geographical sites interconnected through a virtual private network. The 5Gr MANO stack, the monitoring platform (5Gr-VoMS) and the infrastructure of the 5Gr framework are deployed in Barcelona (Spain), while the 5Gr-AIMLP is deployed in Turin (Italy).

The vertical user employs the 5Gr MANO stack to deploy the NS, which interacts with the 5Gr-AIMLP upon request in the NSD [5]. The 5Gr MANO stack consists of three building blocks. The Vertical Slicer (5Gr-VS) is the entry point for vertical users, providing a frontend that maps performance requirements of vertical service requests into network slices that are mapped to NSs. The Service Orchestrator (5Gr-SO) manages the end-to-end lifecycle of these NSs based upon the expressed requirements and the available resource at the underlying NFV infrastructure (NFVI) (compute, storage, network), managed by the infrastructure manager block, the Resource Layer (5Gr-RL). The 5Gr framework follows the architectural concepts of the O-RAN group [3], in which training and inference tasks are split into different blocks. The 5Gr-AIMLP supports multiple open source libraries, like Apache Spark (MLlib), Big DL, or Ray to train models using ML, Deep Learning, and Reinforcement Learning techniques, respectively. In this demonstration, the 5Gr-SO has been extended to perform the inference task and trigger possible scaling operations using Apache Spark streaming jobs. This inference is based on the monitoring data collected from the virtual network functions (VNFs) available in a dedicated

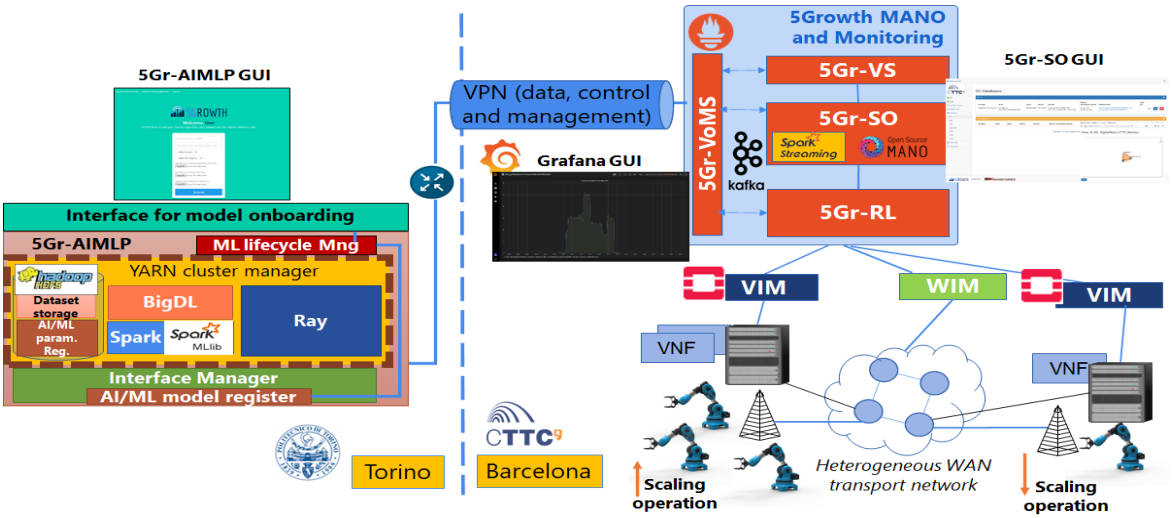

Fig. 1. Experimental setup under demonstration

Apache Kafka topic populated by the 5Gr-VoMs, which is based on Prometheus. Grafana dashboards are used to visualize the monitored data of the deployed NS. In this setup, the 5Gr-SO uses Open Source MANO (OSM) to handle the instantiation of the NSs and their VNFs at the underlying NFVI points of presence (NFVI-PoPs). The 5Gr-RL controls the NFVI-PoPs through their respective Virtual Infrastructure Managers (VIM), which are based on OpenStack DevStack.

III. Demonstration

We show the capabilities of the 5Gr-AIMLP and the data engineering pipeline integrated in the 5Gr platform to perform AI/ML-based scaling operations based on dynamic service conditions, hence fulfilling NS SLAs during run time operation. The dataset and the emulated Digital Twin NS used during this demonstration are derived from [6]. The NS initially consists of two VNFs, namely the Digital Twin app (DT_App) VNF and the Control and Motion Planning (C&MP) VNF. The DT_App VNF receives the position information of attached robots, and processes this information to create the digital twin model of the robot so the human user can control its movement. Then, the C&MP VNF processes these movement commands to plan and generate the required instructions for the robot and enact the desired movement. The critical VNF is the C&MP VNF, whose CPU load, hence processing latency, is directly related to the number of attached robots. Such VNF thus needs to be properly scaled out/in to ensure that instructions are delivered to robots on time while avoiding wasting CPU resources. For this reason, this NS exhibits multiple instantiation levels (ILs), each corresponding to a different number of C&MP VNF instances, over which the load generated by multiple robots can be balanced. The main steps of the demonstration are:

1) The dataset, the training algorithm, and inference classes are uploaded in the 5Gr-AIMLP. The training algorithm is used to process the dataset and generate a trained ML model. In this demonstration, a Random Forest classifier is generated using Apache Spark MLlib.

2) The Digital Twin NS is instantiated according to the procedure explained in [5], with its NSD requiring the use of AI/ML-based techniques to manage the scaling operation. During the instantiation process, the ML model trained in step 1) and the inference file to run the on-line classification are downloaded from the 5Gr-AIMLP to the 5Gr MANO platform to perform scaling decisions based on the monitored CPU load of the C&MP VNF instance/s and the current IL.

3) We stress the CPU load of the initial C&MP VNF instance simulating the addition of multiple robots. Then, the on-line classification job, based on the measured CPU load and the current IL, determines the new IL of the NS, which corresponds to adding a new instance of the C&MP VNF. The 5Gr-SO receives a scaling (out) request to change the IL of the Digital Twin NS instance. Then, the 5Gr-SO proceeds with the requested scaling (out) operation, as explained in [7].

4) When the CPU load decreases, the on-line classification job determines that the system can return to the initial IL with a single instance of the C&MP VNF and the 5Gr-SO receives a notification requesting to scale (in) the Digital Twin NS instance to the former IL.

During the demonstration, all steps are shown through the GUIs of the different building blocks of the 5Gr platform (e.g., ML model training, NS instantiation and information status after successive scaling operations, and time-profile of monitored metrics), as depicted in Fig. 1.

References

[1] ETSI Zero touch network & Service Management (ZSM), https://www.etsi.org/committee/zsm [Accessed in January 2021].

[2] ETSI Experiential Networked Intelligence (ENI), https://www.etsi.org/committee-activity/eni [Accessed in January 2021].

[3] O-RAN Working Group 2: AI/ML workflow description and requirements, *Tech. Rep. O-RAN.WG2.AIML-v01.01*.

[4] C. Papagiani, J. Mangues-Bafalluy, et. al., "5Growth: AI-driven 5G for Automation in Vertical Industries", in Proc. of EuCNC 2020, June 2020.

[5] J. Baranda et al., "On the Integration of AI/ML-based scaling operations in the 5Growth platform" , in Procs of the IEEE NFV-SDN 2020, November 2020, virtual event.

[6] L. Girletti at al, "An Intelligent Edge-based Digital Twin for Robotics", in IEEE Globecom Workshop - AT5Gp, Dec 2020, virtual event.

[7] X. Li et al, "Automating Vertical Services Deployments over the 5GT Platform", in IEEE Comms Mag., July 2020.