Augmented Reality App with AI-based Pervasive Latency Monitoring of RAN and Programmable Metro Packet-Optical Networks

F. Alhamed[1,2], M. Guaitolini[3], P. González [4], R. Berozashvili[5], L. Ismail[1], H. Shakespear-Miles[4], S. Barzegar[4], L. Velasco[4], M. Ruiz[4], J.J. Olmos Vegas[2], A. Sgambelluri[1] and F. Paolucci[3]

[1] Scuola Superiore Sant'Anna, Italy; [2] NVIDIA, Denmark;

[3]Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy;

[4] Universitat Politècnica de Catalunya, Spain; [5] Accelleran, Belgium

**e-mail: andrea.sgambelluri@santannapisa.it*

ABSTRACT

Deep data plane programmability is exploited at different future 6G network technological segments to realize end-to-end application delay telemetry. For the first time, data analytics obtained by RAN controllers and metro network collectors are processed by a Multi Agent System running AI algorithms with the aim of detecting latency anomalies and their location in the network, suggesting the most appropriate recovery countermeasure. The demo is shown applied to Augmented Reality application with extreme low latency requirements.

Keywords: Programmable Data Plane, SDN-based Network Control, Multi-Agent System, 6G Pervasive Monitoring.

1. OVERVIEW

Future 6G Networks will be fed and driven by Artificial Intelligence (AI) / Machine Learning (ML) and will require collection of pervasive monitoring from infrastructures and services to enable fast network configurability adaptations to traffic patterns, as well as to provide verification and enforcement of tenant applications performance [1]. To these goals, data plane programmability and advanced Software Defined Networking (SDN) are investigated as enablers of end-to-end service, real time, and heterogeneous metadata telemetry [2], including edge-cloud continuum segments [3], the Radio Access Network [4] and the optical x-Haul comprising packet-optical equipment [5].

In this live demonstration, we will present the efficient and nearly real-time functionality of a 6G network including the Radio Access Network (RAN) and the metro packet-optical network, fed by a distributed, intelligent Multi-Agent System (MAS), applied to an end-to-end Augmented Reality service requiring low latency. The MAS seamlessly integrates two key components: i) pervasive In-band Network Telemetry (INT) agents, leveraging the capabilities of P4-based components [5] in different networks and technological domains; and ii) multi-flow routing agents to dynamically adapt the network resources based on observed traffic patterns. In the RAN domain, it exploits the near real time controller configurability, while in the packet-optical domain, multi-path flow routing policies are enforced within the packet nodes, aiming to ensure optimal Quality of Service (QoS) performance. As a result, the network re-optimization to satisfy end-to-end latency is orchestrated by a diverse set of agents, each receiving real-time telemetry data from both the RAN and the metro-optical network thanks to latency metadata offered by programmable P4 switches.

The underpinning systems featured in this demonstration are currently in the developmental phase as part of the Horizon Europe DESIRE6G project [6]. The demonstration is deployed in a distributed federated testbed encompassing the CNIT/SSSA (ARNO testbed) located in Pisa, Italy, and the UPC testbed located in Barcelona, Spain. The strategic collaboration between these testbeds allows for comprehensive testing and validation of the RAN and packet-optical network's capabilities, along with the MAS's dynamic control mechanisms.

.

2. INNOVATION

In this demonstration, we aim at exhibiting the groundbreaking capabilities and functionalities inherent to P4-based switched networks [5]. A central focus of this showcase relies on the innovative application of P4 INT collectors [7]. These collectors play a key role in measuring and aggregating Quality of Service (QoS) metrics on a per-flow or per-route basis, introducing a monitoring innovation that facilitates the implementation of dynamic multi-path routing strategies.

This demonstration is conceived to captivate the interest of the ICTON audience, particularly those interested in innovative solutions in network intelligence, zero-touch autonomous network operation and interoperability between heterogeneous network segments - the three focal points of this scenario. The underlying objective is to showcase solutions that hold a potential in meeting the demands of emerging beyond 5G and 6G services. These services, such as ultra-low latency service including Augmented Reality-based applications, necessitate end-to-end bounded delay assurance even in the face of highly dynamic network conditions occurring at network segments with different time scales, patterns and distributions. The ability to make coordinated near-real-time decisions, assisted by AI/ML, is emerging as a key topic, representing a significant solution for reducing both capital and operational costs for network operators.

3. DEMO CONTENT & IMPLEMENTATION

A. Goals

The primary aim of the demonstration is to showcase a preliminary service-oriented end-to-end pervasive monitoring across various network segments. This involves the comprehensive monitoring of different network segments to assess their performance and identify potential bottlenecks or issues. Additionally, the demo aims to illustrate the concept of service re-optimization at runtime, leveraging the proposed MAS framework to dynamically adjust service parameters and configurations in response to changing network conditions or demands, thereby optimizing service delivery and resource utilization in near real-time.

By achieving these objectives, the demo seeks to highlight the feasibility and benefits of adopting a service-oriented approach to network control and optimization, paving the way for more efficient and adaptive network infrastructures.

B. Federated Testbed setup

The testbed hardware configuration (see Figure 1) is designed to facilitate advanced experimentation and analysis within the realm of beyond-5G network technology. It comprises various components seamlessly integrated to simulate real-world scenarios, ensuring robustness and flexibility in testing and development processes.

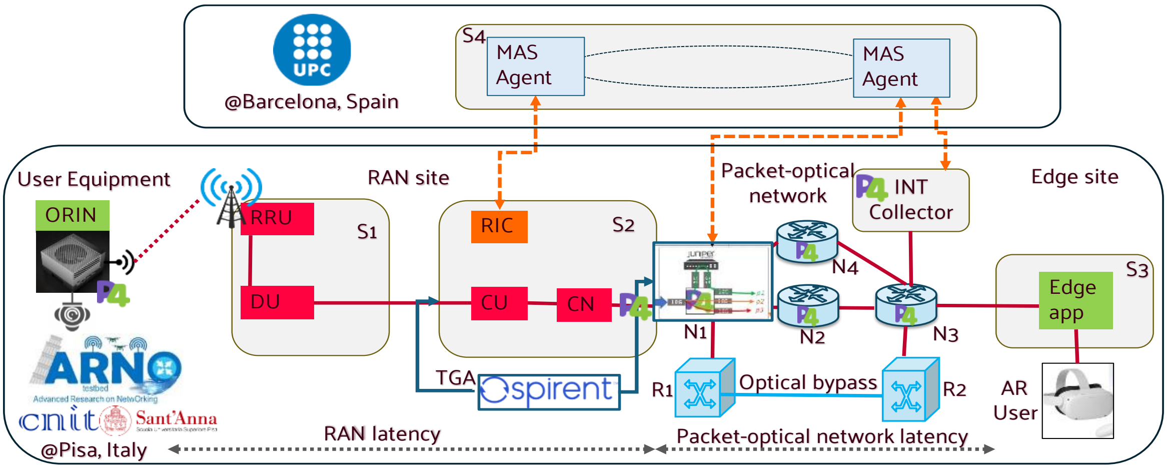

Figure 1: DESIRE6G Federated testbed including ARNO and UPC clusters

A 5G User Equipment (UE), consisting of a NVIDIA Jetson ORIN board is equipped with video camera and a Quectel 5G modem (see the 5G modem co-located with the ORIN in Figure 2). The ORIN and the camera enable visual data capture and transmission, while the Quectel modem provides high-speed connectivity, embodying the convergence of multimedia and communication functionalities.

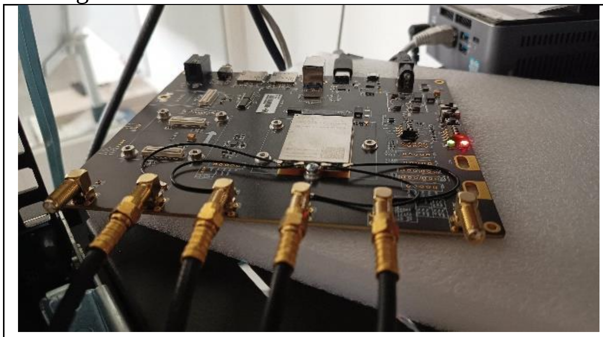	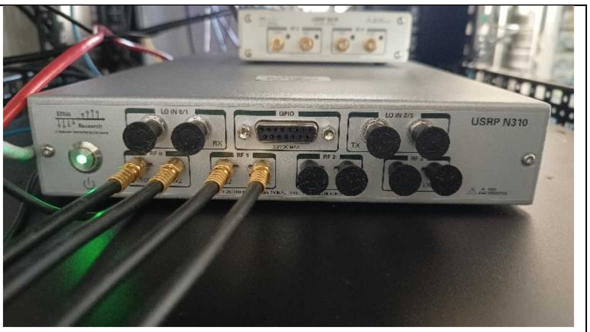
Figure 2: 5G User Equipment modem	Figure 3: Radio Unit USRP 310N

The implemented Radio Access Network (RAN) includes the following components. The utilized Radio Unit is the RRU USRP 310N (see the four radio channels using unconfined coaxial cables to avoid frequency interference with the environment, in Figure 3), connected to the DU through a 10GbE interface. The Accelleran Distributed

Unit (DU) software version 1.0.2 is deployed in server S1 (Dell Power Edge R760, Intel Xeon Gold 6438Y+ CPU with 128 cores, 256 GB RAM, Ubuntu Linux with low latency kernel v5.15). The Accelleran CU-CP/ CU-UP software version 5.0.6 is deployed in server S2 (same as S1 with standard Linux Ubuntu kernel version 5.15), connected to S1 with 100GbE interface. In order to complement the RAN, S2 hosts the Accelleran RIC version 7.2.0 and the open source Open5Gs Core Network (CN) version 2.7.1. The Programmable Packet-Optical network is powered by Nx P4 switches like BMv2 and NIKSS deployed on Dell servers and connected using 100GbE interfaces, facilitating programmable data forwarding and processing, and an optical bypass realized using a pair of Ericsson SPO ROADM R1 and R2 equipped with PM-QPSK 100Gb/s coherent cards with tunable wavelengths. A Two-Stage P4 Collector [7] enhances packet capture and analysis, offering deep insights into traffic patterns and behaviors. Additional traffic is injected by a Spirent N4U Traffic Generator and Analyzer. Finally, Python-based MAS Agents deployed within the UPC testbed environment oversee network orchestration and management (S3). The testbeds interconnection is realized through a dedicated GRE tunnel over the public Internet.

C. AR application

The AR application used in the demo is conceptually depicted in Figure 4. The source app runs in the ORIN and includes the camera video capture and a first object detection. Then, data are sent to the edge app for further data fusion and second object detection, including 3D reconstruction. Final data are streamed to the Oculus for 3D vision. In our current developmental phase, we are leveraging the YOLO (You Only Look Once) library to pioneer object detection capabilities [8]. This framework applies a single neural network to the entirety of an image, partitioning it into distinct regions and subsequently making predictions for each region.

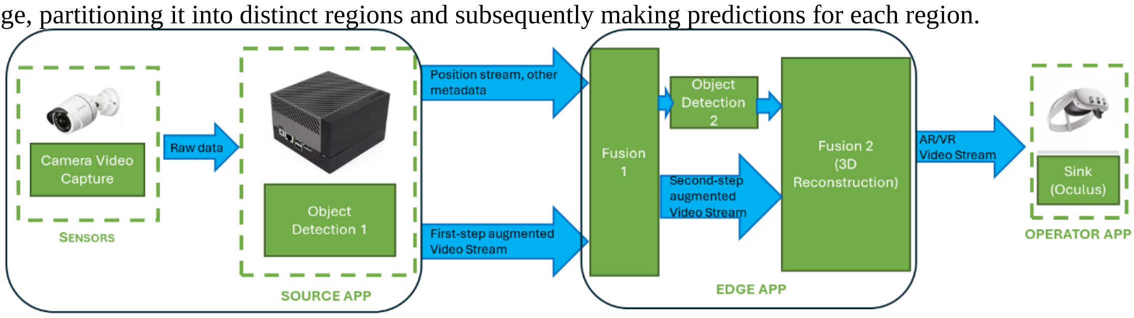

Figure 4: Distributed AR application

This approach makes YOLO faster than many traditional models, making it particularly suitable for real-time applications. Furthermore, YOLO could exploit Nvidia GPUs to run much faster than on CPUs. The model we're investigating is designed to run on a Nvidia Jetson Orin Module and leverages on CUDA Toolkit to exploit the acceleration granted by the GPU available onboard. Preliminary assessments have involved testing the model on sample video streams, as well as streaming from cameras demonstrating its viability for real-world deployment.

Our model is pre-trained, utilizing YOLOv7-tiny weights, and is proficient in detecting a diverse array of objects, spanning up to 80 categories, from which it’s possible to select the desired object to be detected. The number of object categories does not affect the performances in terms of time, but it may be possible to optimize the weights employed, using custom data tailored specifically on the final application. Notably, our current iteration of the model is capable of streaming video with detected markings at 60 frames per second (FPS), underscoring its suitability for real-time applications demanding swift and accurate object detection. Object detection can be performed on the source node or by transmitting images to the cloud node for processing; the device responsible for object detection runs the AI algorithm and sends data to an Oculus Quest 3 device. The Oculus, a virtual reality device configured in passthrough mode, displays the area surrounding the user captured by its stereo camera, then overlays information about detected objects behind obstacles. In fact, the camera captures images to detect objects occluded in the user's field of view. The app utilizes an environment based on Linux, Unreal Engine 5.3, and Nvidia Jetpack 5.

D. Implementation

The scenario of the demo is illustrated in Figure 5. The demo leverages the data-programmability offered by P4-capable switches in the network to perform two main functions related to delay measurements as demonstrated in Figure 5. The first function allows for accurately timestamping the data packets as they enter and leave the different nodes or segments in the network. This timestamping is inserted in the packet itself as INT and gives timely and accurate information regarding the delay that a packet suffers within every single network node or segment, allowing for precise diagnosis of the network latency performance and quick discovery of bottlenecks. The second function performs the collection and aggregation of per-packet INT data [7], making it possible to remove the overhead from the network, and allowing for line-rate calculations performed on the INT Report. For the purposes of this demo, the P4 switches are implemented in the Linux kernel using the eBPF technology [9], where the packet processing behavior is described using the P4 language in the portable switch architecture (PSA) and then it is compiled by the P4-eBPF compiler.

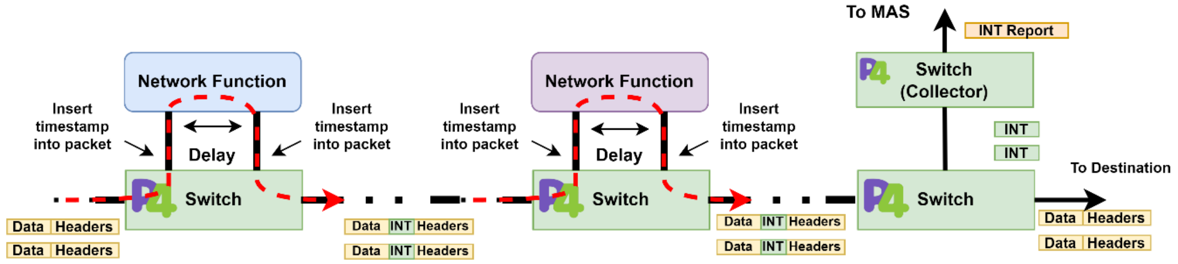

Figure 5: insertion and collection of INT in the system for Network Functions or chains/segments

E. *Proposed workflow*

The demo workflow aims at showing the monitoring and the countermeasures applied to different delay sources and events occurring at the RAN and the PDP.

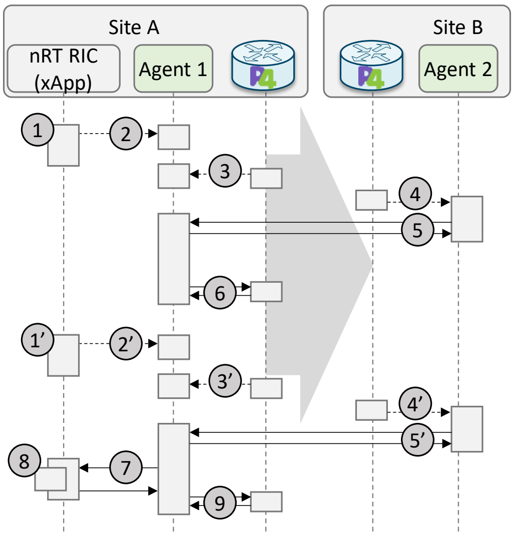

Figure 6: Proposed workflow

The proposed workflow is detailed in Figure 6. The near-Real Time (nRT) RAN Intelligent Controller (RIC), co-located at the CU site, monitors the quality of transmission in the RAN domain and the measurements are collected by a xApp (step 1 in Figure 6), which makes them available for telemetry consumers, in particular to the local MAS agent (2). Besides, Agent1 receives measurements of the input traffic to the PDP network (3), as well as telemetry measurements of the segments latency (i.e., RAN and PDP nodes) and the overall end-to-end delay collected in the egress of the PDP network (4), which are sent by Agent2 in the remote site (5). With these delay measurements, Agent1 can determine the optimal routing policy to be applied to the traffic in the PDP network (6). This process repeats if Agent1 is able to keep the end-to-end delay under the given maximum.

However, it might happen that because congestion in the PDP network, the end-to-end delay cannot be met. In that case (represented by messages 1'-5' in Figure 6), Agent1 requests the nRT RIC to reduce the delay in the RAN domain, so as to compensate the excessive delay in the PDP network (7).

In response, the nRT RIC identifies whether reoptimization of network slices can be carried out to better allocate resources, so as to improve performance (8). With such reoptimization, the delay budget for the PDP network is modified and new policies are obtained that satisfy the required end-to-end delay for the service (9).

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Smart Networks and Services Joint Undertaking under the European Union's Horizon Europe research and innovation program under G.A. No.101096466 (DESIRE6G).

REFERENCES

[1] F. Paolucci, F. Cugini, P. Castoldi, and T. Osinski, "Enhancing 5G SDN/NFV edge with P4 data plane programmability", IEEE Network, 35(3):154–160, 2021.

[2] D. Scano, F. Paolucci, K. Kondepu, A. Sgambelluri, L. Valcarenghi, and F. Cugini, "Extending P4 in-band telemetry to user equipment for latency- and localization-aware autonomous networking with AI forecasting", IEEE/Optica JOCN, 13(9) D103–D114, 2021.

[3] D. Scano et al., "Enabling P4 network telemetry in edge micro data centers with kubernetes orchestration", IEEE Access, 11:22637–22653, 2023.

[4] L. Velasco, M. Ruiz, P. Gonzalez, F. Paolucci, A. Sgambelluri, L. Valcarenghi, and C. Papagianni, "Pervasive monitoring and distributed intelligence for 6g near real-time operation", in EuCNC 2023.

[5] F. Cugini, C. Natalino, D. Scano, F. Paolucci, and P. Monti, "P4-based telemetry processing for fast soft failure recovery in packet-optical networks", OFC 2023.

[6] DESIRE6G Project, https://desire6g.eu/

[7] F. Alhamed et al., "P4 Telemetry Collector", Elsevier Computer Networks, Volume 227, 2023, 109727.

[8] YOLO, https://pjreddie.com/darknet/yolo/

[9] T. Osinski et al, "A novel programmable software datapath for Software-Defined Networking", CoNEXT 2022.