

Augmented Reality in the DESIRE6G Cloud-native and Programmable Infrastructure with Multi-Agent System and Pervasive Monitoring

F. Paolucci, M. Guaitolini
CNIT, Italy

A. Sgambelluri, F. Alhamed,
D. Uomo, E. Paolini, M. Satler
Scuola Superiore Sant'Anna, Italy

P. Gonzalez, M. Ruiz, L. Velasco
Universitat Politècnica de Catalunya, Spain

S. Parker, S. Pryor
Accelleran, Belgium

G. Pongracz, A. Mihaly
Ericsson, Hungary

A. Dalgkitis, C. Papagianni
University of Amsterdam, NL

S. Laki, D. Kis
Eötvös Loránd University, Hungary

A. Nanos
NUBIS PC, Greece

V. Lefebvre, M. Angoustures
Tages Solidshield, France

J. J. Vegas Olmos
NVIDIA, Denmark

Abstract—6G networks promise ultra-low latency and adaptive digital services, including immersive experiences like Augmented Reality (AR). This demo showcases DESIRE6G, an integrated 6G architecture validated on the ARNO testbed in Pisa, Italy. The system combines cloud-native orchestration, programmable data plane infrastructure, secure multi-agent monitoring, and in-band telemetry to support latency-sensitive AR applications. A real-time use case involving camera-equipped drones and AR headsets showcases enhanced situational awareness. The architecture enables scalable service assurance, with latency recovery triggers ranging from microseconds to hundreds of milliseconds across multiple layers.

I. INTRODUCTION

The advent of 6G networks will enable ultra-low latency and highly adaptive digital services, including extended reality (XR) and autonomous systems. Achieving this vision requires the strict integration of cloud-native services, Artificial Intelligence (AI) into the network architecture, along with pervasive monitoring and dynamic reconfiguration across all the infrastructure layers. Data plane programmability and real-time in-band network telemetry (INT) will play a key role, supporting continuous optimization and ensuring strict service guarantees. End-to-end telemetry is made possible by combining data plane programmability, Software-Defined Networking (SDN) techniques and Multi-Agent Systems (MAS), spanning edge, cloud, RAN, and x-Haul segments. A unified telemetry system is expected to collect heterogeneous metadata, enabling coordinated control and adaptive management across domains.

In XR mission-critical or advanced surveillance applications, hybrid architectures—where application components are distributed across the user device, edge, and cloud—are increasingly used to balance latency, compute availability, and network load. This paper presents an integrated demonstrator of such an architecture using a 6G testbed equipped with innovative service orchestrators, infrastructure layers to distribute and enforce data plane programmability network functions

(NF), distributed and secure-attested multi-agents to monitor latency-sensitive Augmented Reality (AR) application and pervasive monitoring service and infrastructure frameworks, based on innovative in-band network telemetry and control.

Developed as framework of the Horizon Europe DESIRE6G project, this architecture is tested on a federated multi-site platform deployed in the ARNO testbed in Pisa, Italy. The demo details the architecture design, the AR application including drones and AR headsets, AI algorithms, and telemetry protocols, and reports deployment and service assurance workflows to guarantee the end-to-end latency budget in the case of network or service degradations (e.g., congestions, resource exhaustion) at the different layers of the DESIRE6G architecture, with particular focus on results including per-segment latency, MAS reaction times, and end-to-end performance. Compared to previous work, this demo highlights the full integration of cloud-native orchestration and control, data plane programmability orchestration, and AI-assisted telemetry and control mechanisms for latency-critical 6G applications.

II. AUGMENTED REALITY USE CASE: DRONE-OPERATOR COOPERATION FOR ADVANCED SURVEILLANCE

This demo showcases a latency-critical AR application designed to support cooperative surveillance and inspection tasks between a drone and a human operator. The use case aims to leverage the increased mobility of camera-equipped drones and the immersive capabilities of AR headsets to enhance situational awareness in real-time. The scenario involves a drone capturing high-resolution video streams of an area of interest that may not be accessible to surveillance operators directly (e.g., a forest, or a damaged building). Streams are transmitted via a low-latency wireless link to an edge computing node, where AI-based tasks are executed. These include object detection, classification, and AR augmentation, such as 3D overlays, annotations, or zoom highlights on relevant

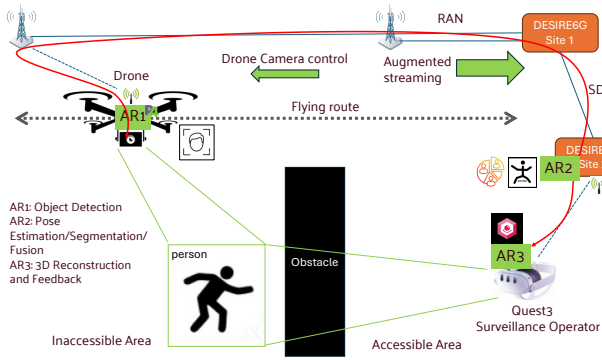


Fig. 1. Use case of augmented reality for immersive surveillance.

scene elements. The augmented video is then sent and 3D-transformed to the operator's AR headset, allowing interactive environment visualization.

The application follows a distributed design, with components deployed across three main layers:

- Source nodes (i.e., drones cameras) collect, perform preliminary detection and forward video streams.
- Edge nodes perform further real-time detections (e.g., pose estimation in the case of person) and data fusion.
- Operator devices (i.e., AR headsets) render the augmented visuals and allow immersive interaction.

The headset operator can influence the AR scene through a backward interaction loop. Using the headset's commands, may change camera angles, zoom into specific objects, or switch between different perspectives. These commands are transmitted back to the drone or other source nodes, enabling adaptive camera operations and creating the effect of natural perspective navigation, as if the operator were physically moving through the space. To maintain ultra-low latency and high reliability, the system requires key 6G features such as dynamic resource allocation (e.g., scaling and/or migration of containers) and real-time telemetry. Data plane programmability is employed both at the edge and within the drone's onboard network stack, enabling end-to-end in-band telemetry collection and latency monitoring.

III. THE DESIRE6G ARCHITECTURE

The AR use case is tailored in the DESIRE6G architecture, a flexible, programmable, and multi-site 6G-native system designed to support latency-critical, distributed applications in the 6G network [1]. The architecture comprises four key layers, each contributing to the deployment, orchestration, and performance assurance of complex services.

1) *Intent-based Orchestration Layer*: The Service Management Orchestrator (SMO) enables high-level, intent-driven control of network services. It translates user or SLA-driven intents into service templates, orchestrates service function deployment (both statically and dynamically), and manages multi-site coordination using a service catalog, machine learning function orchestrator (MLFO), and a federated data lake. The Optimization Engine (OE) maps the service chain and NFs

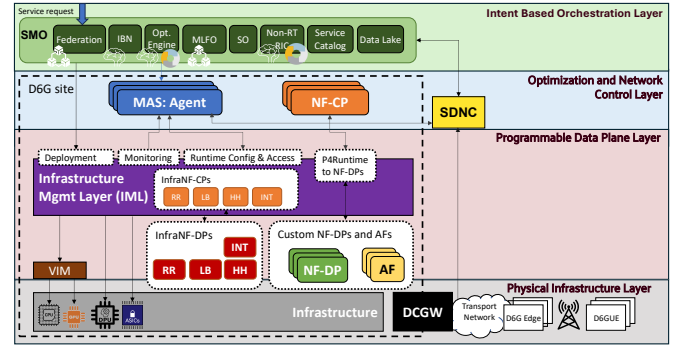


Fig. 2. Multi-layer architecture of DESIRE6G.

in the different sites based on SLA requirements. High-level service intents are then translated into concrete deployment templates that guide NFs configuration and activation.

2) *Optimization and Network Control Layer*: It contains the control plane and the mechanisms for re-optimizing distributed services. It includes components such as the Multi-Agent System (MAS) [2], which ingests real-time telemetry and re-optimizes service quality metrics like latency and reliability. The MAS operates near real-time service assurance verification and recovery, with the aim of detecting degradation sources and suggesting the most appropriate reoptimization procedure. Being distributed per site and running AI for fast event detection, it requires secure attestation conferring trust between agents. This is done by D-MUTRA, a novel blockchain-based mutual remote attestation featuring a software-based root of trust anchored in SECaaS [3].

3) *E2E Programmable Data Plane (PDP) Layer*: This is the execution layer, where customized packet and application-level processing is performed. The data plane is composed of NFs implementing service logic, infrastructure services (e.g., service mapping, telemetry, routing), and application functions. The Infrastructure Management Layer (IML) abstracts hardware-specific platforms, supporting horizontal and vertical disaggregation, and offering a unified view to the control plane. The use of hardware acceleration, programmable switches, and telemetry-enabled NFs allows high performance and adaptability. For the AR case, end-to-end INT probes are deployed in the UE at the drone, in the RAN gateway and in the edge cluster to track per-segment latency values. Probes are reported to a P4 collector and exposed to the MAS.

4) *Physical Infrastructure Layer*: It comprises edge nodes, RAN, and network components. Backends include CPUs, GPUs, SmartNICs, FPGAs, and RAN-specific hardware, with site-specific configurations for roles like edge computing or core services. In the SDN domain, a specific INT is implemented, based on the in-band network control (INC) framework [4]. Using INC, peak delays experienced by small bursts of packets are detected at the Data Plane Active Collector (DPAC), placed inside a node, and novel INC messages are sent to upstream nodes to immediately steer traffic excluding the affected interface, without involving the SDNC.

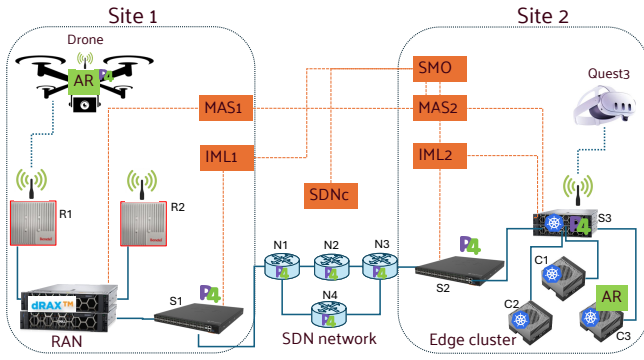


Fig. 3. Demonstrator testbed

IV. DEMONSTRATION

The demo is deployed in the ARNO testbed, Pisa, Italy. It spans two interconnected sites, showcasing a distributed 6G-enabled edge facility with programmable network and RAN. It integrates hardware and software P4 switches, edge nodes, and SDN control, supporting the AR application with the Quest3 headset and a drone equipped with camera and 6G User Equipment payloads. The testbed is shown in Fig. 3.

1) *Testbed*: In Site 1, drone connectivity is established through a UE enhanced with P4 programmability to support the D6G service mapping. The RAN segment is deployed in a disaggregated setup with two Benetel Radio Units connected via dual 10 Gb/s fiber links to a first server with vDU containers. A second server, linked to S1 by 10 Gb/s Ethernet, runs the Accelleran CU-CP/CU-UP v5.0.6, near real-time RIC v7.2, and the open-source 5G Core Network (Open5Gs). The CU, RIC, and CN deploy at bootstrap, while the vDUs and RUs are operator-deployed. The d-RAX dashboard manages independent deployment of all modules and displays the current topology with connected UEs, which use a Quectel 5G module registered via script. A P4 Tofino switch S1 links the RAN to the SDN network. Local orchestration and control are managed by MAS1 and IML1. The SDN domain interconnects the sites through a series of N_x P4 SDN switches running INC and 40GbE links. This network segment is managed by a centralized SDN controller (SDNc) and supports in-band telemetry for real-time monitoring and dynamic traffic steering. Site 2 hosts a Kubernetes-based edge cluster composed of three edge nodes ($C1$, $C2$, and $C3$) running Kubernetes workloads. A hardware P4 switch S2 connects the edge infrastructure to the SDN domain. A Quest3 AR/VR headset is connected via WiFi to the local cluster and edge services, enabling immersive, low-latency applications. The site's deployment and monitoring are managed by MAS2 and IML2, while SMO handles multi-site orchestration.

2) *MAS Attestation and Service Deployment*: The workflow begins with the SMO computing a service deployment descriptor. The AR service graph is submitted to the orchestrator, that retrieves service details from the catalog and the topology sub-modules, asks site decomposition to the OE, thus obtaining

per-site service sub-graphs. MAS agents are then deployed across the sites (MAS1 and MAS2), equipped with attestation software, and smart contracts are prepared. The system sets up accounts, registers smart contracts, and distributes credentials to the agents. Each agent undergoes attestation, verified through a root of trust to ensure integrity. Finally, verified agents securely receive encryption keys, completing the secure MAS pipeline configuration. Once inter-site SDN service connectivity is requested and deployed, sub-graphs are sent to IML instances of the sites, responsible for deploying, configuring and stitching NFs and application chain modules. In this case, the source app will be linked to NF chains at the drone P4 switch, and at nodes S1, S2, S3 and C3 (edge app), to reach the headset.

3) *Service Assurance Workflows*: Three workflows are envisioned to demonstrate service latency assurance at the different architecture layers. First, a congestion will occur at the SDN network (e.g., N1) and will be fully recovered at the data plane thanks to the INC functionality, activating a route bypass along N4. Second, resource exhaustion at the K8S cluster involving node C3 will trigger the infrastructure monitoring, that will scale/migrate the AR edge app thanks to the rapid intervention of local IML2. Finally, a degradation at the vDU causing latency increase will be discovered by the MAS thanks to reports provided by INT collector. Moreover, using xAPP of RAN metrics from the near real time RIC, the MAS will localize the degradation and will repotimize the end-to-end latency by forcing the Handover of the RAN from R1-DU1 to R2-DU2. These workflows demonstrate the ability of the DESIRE6G architecture to minimize the service assurance recovery time thanks to the different trigger mechanisms at different layers, from tens of microseconds (INC) up to hundreds of milliseconds (distributed MAS).

ACKNOWLEDGMENT

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 programme under G.A. No. 101096466 (DESIRE6G). Authors acknowledge the support of the BRIEF "Biorobotics Research and Innovation Engineering Facilities" project (Prj code IR0000036).

REFERENCES

- [1] G. Pongrácz, A. Mihály, I. Gódor, S. Laki, A. Nanos, and C. Papagianni, "Towards extreme network kpis with programmability in 6g," in *Proc. 34th International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, ser. MobiHoc '23. New York, NY, USA: ACM, 2023, p. 340–345.
- [2] P. González, F. Alhamed, S. Barzegar, F. Paolucci, J. J. Vegas Olmos, M. Ruiz, and L. Velasco, "Distributed multi-agent system fed with telemetry data for near-real-time service operation," in *2024 Optical Fiber Communications Conference and Exhibition (OFC)*, 2024, pp. 1–3.
- [3] P. González Pacheco, S. Barzegar, M. Ruiz Ramírez, and L. Velasco, "Deployment of multi-agent system pipelines for near-real-time operation of 6g network services," in *24th ICTON 2024, International Conference on Transparent Optical Networks*. IEEE, 2024.
- [4] F. Alhamed, A. Sgambelluri, C. Papagianni, and F. Paolucci, "In-band collection and control of end-to-end latency in programmable packet-optical networks," in *Optical Fiber Communication Conference (OFC) 2025*. Optica Publishing Group, 2025, p. M3H.4.