Advances on Unified Architecture for Open RAN-enabled Distributed and Scalable 6G Networks

Luis Blanco[1], Cristian J. Vaca-Rubio[1], Jorge Baranda[1], Farhana Javed [1], Selva Via[1], Engin Zeydan[1,], Luis M. Contreras[2], Javier Velazquez[2], Gil Kedar[3], Efi Dvir [3], Gianluca Rizzo Antonio[4], Berk Buzcu[4], Maria A. Serrano[5], Angelos Antonopoulos[5], Godfrey Kibalya[5], Jetmir Haxhibeqiri[6], Vasilis Maglogiannis [6], Dries Naudts [6], Jeroen Hoebeke [6], Tomaso de Cola[7], Balaji Kirubakaran [7], Sotirios Spantideas[8], Anastasios Giannopoulos[8], Panagiotis Trakadas[8], Harilaos Koumaras[9], Stefanos Plastras[9], Alice Piemonti[10], Vito Cianchini[10], Carla Amatetti[11], Bruno de Filippo[11], Zefu Gao[11], Pawel Kryszkiewicz[12], Marcin Pakula[12], Lukasz Kulacz[12], Paweł Płaczkiewicz[12], Adrian Kliks [12], Matteo Pagin[13], Hua Wang[13], Sihem Cherrared [14], Taki Djaidja[14]

[1]CTTC/CERCA, Spain, [2]Telefónica Innovación Digital, Spain, [3]Ceragon Networks, Romania, [4] HES-SO, Switzerland, [5]Nearby Computing, Spain, [6]IMEC, IDLab, Ghent University, Belgium, [7] DLR, Germany, [8]Four Dot Infinity, Greece [9]NCSR Demokritos, Greece, [10]Martel Innovate GmbH, Switzerland, [11]University of Bologna, Italy, [12]Poznan University of Technology (PUT), Poland, [13]Keysight Technologies, Denmark, [14]Orange France

Corresponding authors: {lblanco,ezeydan}@cttc.es

Abstract*—UNITY-6G introduces a AI-natively framework that unifies terrestrial (TN), non-terrestrial (NTN), and non-public networks (NPN), treating connectivity, computing, and intelligence as interdependent resources. The architecture utilizes an Inter-Domain Management Orchestrator (IDMO) based on Service-Based Management Architecture (SBMA) principles to coordinate services across heterogeneous domains. A core pillar of the framework is its AI-native design through autonomous agentic AI workflows following a standardized MS–AE–DE–ACT (Monitoring, Analytics, Decision, and Actuation) logical patterns. To enhance resource efficiency and sustainability, the architecture integrates Digital Twins (DT) for proactive system modeling and semantic communications to prioritize task-relevant information transfer. Security is addressed through a Trust Architecture leveraging Distributed Ledger Technology (DLT) for cross-domain auditability. The framework's utility is validated through proof-of-concepts targeting sustainable disaster handling, immersive XR/holographic communications, and time-sensitive services for Industry 4.0. The presented advances establish a foundation for the continuous development of high-performance, autonomous 6G systems.

***Keywords*—6G, NTN, O-RAN, Wi-Fi, SBA, Digital Twin, Sustainability**

I. INTRODUCTION

The rapid expansion of global connectivity demands architectures that are not only high-performance but also sustainable, resilient, and universally accessible. Conventional vertically integrated terrestrial networks are increasingly insufficient to meet coverage, energy efficiency, scalability, and deterministic service requirements, particularly in remote, industrial and underserved environments. The accelerating growth of Information and Communication Technology (ICT) infrastructures has raised significant sustainability concerns, with projections indicating that they could account for up to 21% of global electricity consumption by 2030 [2]. Hence, there is an urgent need for energy-efficient and environmentally sustainable network architectures. As 6G systems emerge, future networks must natively integrate renewable energy sources, energy-aware protocols and intelligent resource management, while supporting traffic demanding applications such as extended reality (XR). These services impose stringent requirements in terms of ultra-low latency, high bandwidth, reliability, synchronization accuracy and energy efficiency. Hence, necessitating new architectural foundations and resource orchestration paradigms.

Moreover, disaggregated Open RAN (O-RAN) architecture introduces open interfaces, virtualization and intelligent control loops, fostering multi-vendor interoperability. Beyond architectural disaggregation, conventional bit-centric communication models are insufficient to meet the efficiency [3]. Within this context, semantic communication paradigms have emerged, aiming to transmit only task-relevant and meaningful information rather than raw data, thereby reducing bandwidth consumption and energy expenditure. Recent works, such as SEM-NTN [4] demonstrate how semantic feature extraction can be embedded within unified satellite-terrestrial systems leveraging O-RAN principles, enabling context-aware and goal-driven transmission. Complementarily, the 6G-GOALS initiative [5] advances a semantic-aware, objective-driven architecture in which network behavior dynamically adapts to service intent.

The evolution toward fully autonomous networks is further accelerated by the emergence of agentic AI. These developments have led to a new paradigm of highly autonomous, self-directed systems referred to as agentic AI. In network management and automation, intelligent agents have long been employed for prediction, monitoring, optimization, and security. The emerging agentic paradigm introduces collaborative and multi-agent workflows in which agents not only act, but also evaluate outcomes, coordinate with other agents, continuously refining their strategies.

Within unified TN (terrestrial netwotk)-NTN (non-terrestrial network) and non-public network (NPN) environments, such agentic AI frameworks can orchestrate efficiently heterogeneous services across the edge-cloud continuum, dynamically allocate radio and computing resources and adapt network behavior according to high-level service goals and traffic demands. These capabilities are equally critical for NPNs, including private systems based on O-RAN and Wi-Fi, particularly when supporting Time-Sensitive Networking (TSN). By distributing intelligence across different domains, future unified 6G architecture can achieve sustainability, resilience and adaptive end-to-end control. Nevertheless, substantial challenges remain, including cross-domain interoperability, efficient spectrum and energy allocation, safe multi-agent coordination and the preservation of stringent Quality of Service (QoS) guarantees. Addressing these challenges through unified, disaggregated, semantic-aware and AI-native frameworks represents a pivotal step toward sustainable and intelligent 6G networks.

A. Paper Contributions

The key architectural and technological milestones are:

1. Integrated Multi-Domain Architecture: The paper introduces a natively convergent framework that unifies terrestrial (TN), non-terrestrial (NTN) and NPN into a single design paradigm.

2.Use of a Service-Based Management Architecture (SBMA) and an Inter-Domain Management Orchestrator (IDMO) to jointly orchestrate connectivity, computing and radio resources.

3. AI-Native framework based on Agentic AI Framework: A core contribution is the definition of a distributed AI/ML subsystem providing AI as a Service (AIaaS). The architecture moves toward agentic AI, utilizing a standardized MS–AE–DE–ACT (Monitoring, Analytics, Decision, and Actuation) logical patterns that allows autonomous agents to reflect, learn and improve integrated network operations.

4. Digital Twin and Semantic Communication Enablers: The paper describes a Digital Twin (DT) architecture that supports "what-if" scenario analysis and proactive resource optimization. Additionally, it integrates semantic communications as a system-level capability to prioritize task-relevant information transfer, critical for spectrum-limited NTN

5. Trust and Security Architecture: It incorporates a Trust Architecture that extends zero-trust principles to cross-domain orchestration. It utilizes DLT and smart contracts to ensure verifiable operations and transparent auditability.

6. Network Exposure: This paper defines a two-layer network exposure framework. Layer-1 provides telco-specific APIs for direct domain access (e.g., O-RAN, Core), while Layer-2 offers developer-oriented APIs based on CAMARA initiatives to simplify complexity for vertical services.

7. Validation through Proof-of-Concepts (PoCs): The architecture is demonstrated through 2 primary PoCs covering 6 experimental scenarios, including sustainable disaster handling via TN/NTN resource sharing, immersive XR/holographic communications and time-sensitive services for Industry 4.0 using multi-RAT (cellular and Wi-Fi).

II. ARCHITECTURE AND DESIGN PRINCIPLES

The UNITY 6G architecture is an AI-native framework that unifies terrestrial, non-terrestrial and non-public networks with compute continuum infrastructures and service orchestration layers under a single design paradigm. Moving beyond traditional network silos, it places programmability, openness, and multi-domain integration at its core to operate as a distributed, intelligent, and self-optimizing fabric spanning access, transport, core, edge, and cloud. This approach ensures that connectivity, computing, and intelligence are treated as interdependent resources, orchestrated jointly and dynamically in response to service intent, network state, and application requirements.

A central design principle is end-to-end harmonious integration achieved through standardized interfaces, cross-domain abstraction, and AI-native control loops. This architectural view emphasizes modularity, zero-touch automation, and energy-aware resource management. Furthermore, UNITY 6G adopts a security-by-design philosophy, embedding trust anchors and verifiable operations at every layer. Together, these principles provide the foundation for a flexible, future-proof architecture capable of supporting heterogeneous services, including real-time digital twins and resilient connectivity.

A. Integrated architecture overview

UNITY-6G integrated architecture builds upon a seamless fusion of multi-domain network components, incorporating terrestrial access, non-terrestrial systems, transport networks (xHaul), the 6G Core, and distributed computing domains. In contrast to earlier generations that primarily focused on radio enhancements, UNITY-6G addresses the full vertical stack, ensuring that data, control, compute and intelligence can flow consistently across all domains. A unified orchestration layer abstracts underlying heterogeneity, enabling service- and intent-driven provisioning that spans satellites, aerial platforms, macro cells, small cells, fiber, packet transport, and cloud/edge resources. At the *access domain*, UNITY 6G integrates TN and NTN under a common radio and mobility management umbrella. Terrestrial RAN elements collaborate with LEO/MEO/GEO satellites, HAPS, and other aerial assets, allowing devices and services to maintain consistent performance across vastly different coverage environments. This convergence is mediated by shared spectrum frameworks, unified authentication and mobility procedures, and AI-assisted link management. The *xHaul transport layer*, including front, mid and backhaul, acts as a programmable fabric interconnecting disaggregated RAN elements, edge platforms, and the core network, ensuring that user data and control traffic are delivered with predictable latency, jitter, and reliability. Unified telemetry feeds into a cross-domain control system that dynamically adjusts routing, capacity, and transport slices based on service requirements and network conditions.

Within the *6G Core*, the architecture extends the service-based approach of 5G into a fully cloud-native, AI-orchestrated environment. Core functions become highly granular, stateless where possible, and deployable across different points of the compute continuum. The core also plays a central role in coordinating TN/NTN convergence, multi access management, and distributed compute placement decisions. Finally, the *Edge–Cloud continuum* binds all the technological domains together. Compute resources at radio sites, metro edges, regional clouds, and hyperscale data centers participate in a single, federated computational environment. AI-powered orchestration distributes workloads dynamically across the continuum depending on latency, energy, mobility, and resilience constraints. The continuum interacts tightly with the network fabric, allowing data and computational paths to be co-optimized in a unified way, reflecting 6G's communication-compute convergence. Fig. 1 summarizes this view.

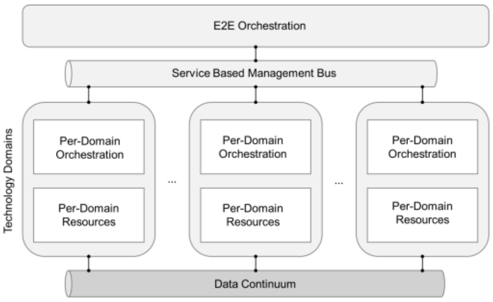

Figure 1. Data Continuum role across heterogeneous domains

B. Service-Based and AI-native design principles

To perform multi-domain integration, UNITY-6G architecture relies on ETSI ZSM architectural principles and the SBMA described in 3GPP TS 28.533 [1]. In that sense, the main idea is that management services (MnS) are exchanged between consumers and producers. In UNITY 6G architecture, the role of consumers is mainly played by the entity called IDMO. IDMO collects service requests realized through intents to decompose it based on the MnS provided by the different registered domains and coordinate such different management domains, which act as the service producers. The management domains, grouped technology-domain wise, e.g., O-RAN, Core Network, perform the lifecycle management of network services or configure transport flows interconnecting the domains, so mobile network functions entities are treated as virtual network functions (VNFs) and on-demand instances of such network functions can be created and interconnected through the edge-cloud continuum built with the interconnection of the different management domains. The communication between IDMO and the different domains is provided by the cross-domain SBMA bus substrate, similar to the cross-domain integration fabric defined in ETSI ZSM architecture. Within the UNITY-6G architecture, AI principles are natively reflected in multiple aspects and at different levels of architecture. At the top level (IDMO), the intent manager module oversees collecting natural language intents to translate them into deployment configurations. At the management domains, closed-loops for the different instances of running MnS are realized through the interaction of tuples, namely monitoring system, analytic engine, decision engine and actuator, thus providing service level specification (SLS) assurance. These tuples are supported by a specific management domain called the AI/ML subsystem based on an AIMLaaS concept, using a data continuum substrate between management domains to perform data collection, model training and distribution (details in Sect. IV).

C. Sustainability

The proposed network architecture is envisioned to be natively sustainability aware at the orchestration level by aligning network installations at the IDMO level with the availability of clean energy at the time of decision. By embedding Virtual Power Plants (VPPs) directly within individual network domains, the system respects the federated structure while locally managing renewable resources like solar and wind. These decentralized energy insights are coordinated to allow the network to act as a "carbon-aware" entity: it can autonomously shift energy-intensive data processing to locations with a power surplus or reduce activity when the local grid relies on high-carbon sources. To operationally realize this vision, the architecture is designed around a hierarchical control plane where the IDMO entity functions as the central decision-maker utilizing the SBMA [1]. In this proposed workflow, the intra-domain VPP telemetry is aggregated by a super-ordinate Inter-domain Energy Management System (EMS), which is tasked with synthesizing real-time data into probabilistic forecasts regarding generation availability and carbon intensity. This intelligence, both predictive and real-time is formatted for direct ingestion by the IDMO. Ultimately, the system is designed to algorithmically balance sustainability metrics with stringent QoS constraints (e.g., latency) to ensure that network performance is not compromised.

D. IEEE 802.11 domain

UNITY-6G aims to integrate not only the 3-GPP radio access technologies, but also non-3GPP RATs to enable seamless operation across multi-RAT deployments. As 3-GPP RAT domain will be controlled following O-RAN principles (as specified in section II B), the goal for IEEE 802.11 domain is to follow the same principles, allowing a unified control plane. At its core, the IEEE 802.11 domain consists of the Domain Management Orchestrator (DMO), which interfaces northbound with the IDMO while orchestrating the operation of the Infrastructure Domain Manager (IDM) in its southbound interface. The DMO hosts the Domain Data Fabric (DDF) that helps interaction between the domain telemetry database, domain topology database and domain scheduler. The telemetry database will be populated utilizing In-band Network Telemetry (INT) [7] and monitoring apps in IDM. The topology database will be populated by exposing association information from the access points (APs), with a dedicated app for topology monitoring. Both databases can exchange information with other end-to-end controlling entities at the level of IDMO. Similarly, the domain scheduler can interact via the northbound interface of the DMO with the end-to-end scheduler/orchestrator of the IDMO.

The IDM logically hosts the Radio Intelligent Controller (RIC) and the AP controller, which apply certain controlling functions in the IEEE 802.11 APs. RIC will activate different controlling APPs at different time scales (xApps, dApps) following O-RAN principles. Such apps can run at different locations (e.g. xApps at edge and dApps at APs), but logically they are part of RIC. Communication between RIC and APs is performed using E2 interface, by enhancing the IEEE 802.11 APs to be O-RAN compliant. For specific intra-AP logical control (e.g., association management, station control, etc.) custom APIs will be consumed. Fig. 2 shows IEEE 802.11 domain.

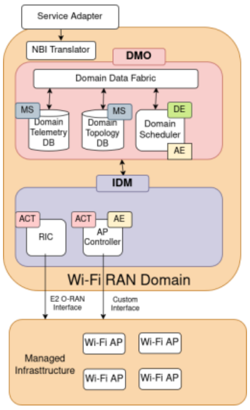

Fig 2. IEEE 802.11 domain architecture

III. DIGITAL TWIN FRAMEWORK

The DT represents real-time models of physical systems, capable of simulating, predicting, and analyzing behaviors under various conditions [8]. DTs are widely applied in what-if scenario analysis across multiple domains, enabling predictive insights and informed decision-making. In manufacturing systems, DTs support energy management by simulating energy consumption under varying operational conditions, allowing operators to optimize efficiency, reduce costs, and improve sustainability. In a integrated network, it is important to configure properly multiple services from different domains. In order to reuse DTs among various domains, DT repository will be connected to the service bus as depicted in Fig. 1. An example of DT in a UNITY 6G control loop is shown in Fig. 3.

A. DT internal structure

First, the DT repository architecture is discussed. In the literature there are different approaches to network DT creation varying from simple look-up tables, sophisticated AI approaches, up to full stack processing of user data considering realistic wireless channel models. While we treat a DT as a function that returns a given metric, e.g., Base Station load, based on a given configuration, each of the mentioned modules above support this functionality. Within this context, the selection of a proper DT structure depends on the task to be performed defined by the required output (e.g., load), its required accuracy (e.g., the required error) and the needed response time (e.g., full stack network emulation can take several seconds to return DT estimates). It is expected that for some applications, e.g., delay estimation in fibre link, look-up table can be enough, while this can be not accurate enough for over-the air wireless transmission modeling. Moreover, important aspect is sustainability of a DT, i.e., its computational compelxity. In an integrated network the aim should be to achieve a given KPI value from a DT using as small amount of resources as possible. These reasons support the repository-like structure of the DT proposed in Fig. 2. In the repository various DT functions are registered, each described by its capabilities, e.g., accuracy of the result, and cost, e.g., delay or computational complexity. For a given application, e.g., traffic steering in a 5G network, only a specific DT has to be used. The other reason supporting repository-based DT architecture is an integrated nature of UNITY network. If a network operator or vendor may be unwilling to share internal measurements or models. Exposing the DT via a dedicated API preserves confidentiality, while delivering full functionality.

B. Closed-Loop DT-Assisted Network Optimization

As presented in Fig. 3 the DT selected from the repository has to be connected to the integrated network, e.g., to update the DT model by adjusting to moving network users or removing estimation error of a neural network. It has to happen not only to active DTs, but also to DTs that can be shortly used. This requires proper interfaces to be used. Morover, the DT should be used by some Network Application, e.g., Traffic Steering performed in O-RAN using some xApp. By keeping the DT synchronized with the network conditions the Application can test various configurations first, before applying the best configuration to the real network. This can help in finding high quality solution in a complex, integrated network. While the DT can be performed inside a specific domain, an interesting aspect to be considered is building a DT for the whole service, facilitating fine-grained service optimization for IDMO.

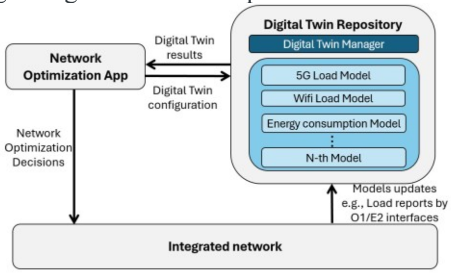

Figure 3: Network Digital Twin functional and internal model

IV. AI-NATIVE UNITY-6G NETWORK

The AI/ML subsystem provides intelligent services to the multiple domains, aligning with the concept of AI-native architecture with the following design principles: (i) The subsystem shall provide *AIaaS* to the 6G network in the form of interconnected tuples, i.e., training, inference and evaluation pipelines; (ii) *The intelligent services shall be available in all UNITY-6G domains* and at multiple locations, for instance in the RAN, Core and Management domains; (iii) The *ML pipelines shall be either deployed centrally or in a distributed manner* (e.g., ML models can be deployed at the location of the data), while supporting model management lifecycle, including data gathering and analysis, model training, testing, validation and performance monitoring; (iv) The framework shall *be able to gather data for model training from multiple domains* of the 6G system.

In this context, the internal AI/ML subsystem is shown in Fig. 4. Its components include, among others, an AI/ML Orchestrator that is used to interface with the rest of the domains, also coordinating the lifecycle management of intelligent services. In addition, a Data Management (DM) module is responsible for gathering data from the relevant network functions, as well as chaining the tuple components for building pipelines in the AI/ML Pipeline Subsystem. Other components of the system include typical MLOps functionalities (data processing, training services, model database, etc.), a Sandbox simulation test environment that can used to evaluate the performance of the ML models, along with a *Conflict Resolution* module and a *Trust Adapter* (details on how the latter components fit into the Unity-6G architecture are provided below). In addition, the AI/ML subsystem is designed to support closed-loop operation under uncertainty. To this end, the subsystem accommodates both deterministic and probabilistic learning approaches, allowing AI services to provide not only point predictions but also confidence-aware outputs, which has been shown to be critical for intelligent resource management [9], [10]. These capabilities are essential for time-critical and resource-constrained 6G scenarios, where control decisions must be taken ahead of time based on incomplete and evolving information. By exposing uncertainty-aware insights to decision-making entities, AI/ML subsystem forms a foundational enabler for resilient, risk-aware, and AI-native network operation.

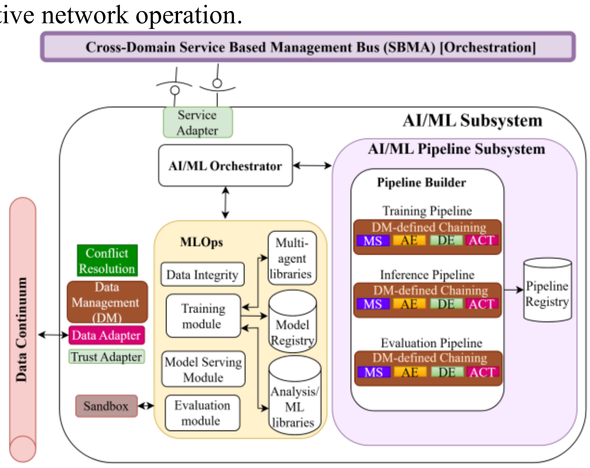

Figure 4: Architecture & components of the AI/ML subsystem

A. Agentic AI

The UNITY-6G framework introduces an agentic AI architecture that moves beyond rigid, manually engineered workflows towards autonomous, collaborative, and self-evaluative multi-agent orchestration. AI-Agents not only act but also reflect, learn, and improve over time [11]. Building on advances as general-purpose reasoning engines, the framework enables long-chain reasoning, multi-step planning, and direct interaction with network APIs. Rather than relying on isolated perception-reasoning-action loops, UNITY-6G establishes a closed-loop, intent-aware control paradigm in which agents decode natural-language service intents, generate orchestration strategies, coordinate actions, and continuously verify compliance with service-level objectives. Agents are characterized by two complementary dimensions: conceptual placement and physical placement. Conceptual placement defines the network domain an agent oversees (e.g., RAN, Core, or service layer), while physical placement specifies the HW tier on which it executes, ranging from base stations and edge nodes to regional data centers and central clouds.

Agents follow a standardized logical pattern denoted as *MS-AE-DE-ACT (Monitoring System, Analytics Engine, Decision Engine, Actuator): i) MS* performs data-engineering operations, collecting heterogeneous telemetry and enriching it with contextual metadata to produce structured statistical representations. *ii) AE* transforms this data into predictive intelligence through anomaly detection, forecasting, and performance modeling, combining telemetry with topology and historical knowledge. It employs validation-driven model selection and adaptive reasoning to identify the most suitable forecasting approach under current network conditions. *iii) DE* interprets analytical outputs within the constraints of system objectives and scheduling policies, reasoning over predicted demand and energy metrics to produce high-level orchestration decisions. It aligns predicted demand with available network resources, reasoning over allocation templates retrieved via an MCP server. It ultimately generates recommended actions that dictate the optimal scheduling policy. Finally, *iv) ACT* translates these decisions into executable resource-allocation and configuration commands. In Fig. 5, the DE reflects based on inputs, such as intents from the domains and insights from the AE Agent, including predictions of traffic or energy consumption and then determine actions by calling the ACT Agent. Architecture supports flexible agent configurations: some agents specialize in observation or analytics, while actuation may be delegated to shared controllers. This modularity enables scalable deployment across heterogeneous 6G environments and facilitates inter-agent coordination. A key application is proactive service orchestration for containerized applications and VNFs, where predicted workload is used to pre-allocate computational resources. By shifting from reactive scaling to proactive allocation, UNITY-6G reduces service degradation, improves performance stability and minimizes energy waste. Continuous learning and uncertainty awareness are integral to the framework. Agents can generate confidence-aware predictions, and an evaluation pipeline monitors inference quality to detect concept drift and trigger retraining.

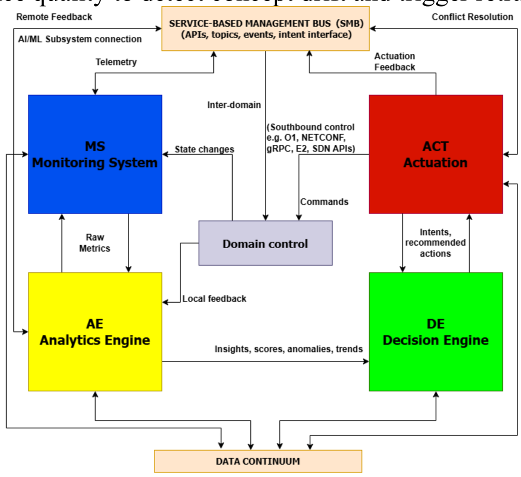

Figure 5. UNITY-6G Tuple inter-connectability

B. Conflict resolution in integrated O-RAN

In the unified architecture, a cross-domain orchestrator IDMO is introduced to coordinate business service realization and resources management across independent domains through interaction with individual DMOs. However, DMOs are not limited to intra-domain resource management and service deployment tasks; they must also handle operational challenges such as conflicts between concurrently operating applications. For instance, in Open RAN environments, *conflict management (CM)* becomes particularly challenging due to the presence of independently developed intelligent applications (xApps and rApps) that may generate conflicting control actions [13]. Consequently, multi-domain resource orchestration and management can naturally be viewed from a multi-agent perspective, where CM functionalities deployed in individual domains (RAN, CN, etc.) operate autonomously. To address this challenge, UNITY-6G proposes the concept of layered and hierarchical conflict management in virtualized networks. Specifically, we introduce a multi-agent CM framework coordinated by a supervisory entity operating across domains within the IDMO. When domain-specific CM mechanisms operate independently, their decisions may lead to globally suboptimal outcomes because actions taken in one domain can influence the performance of neighboring domains. Furthermore, once the IDMO layer is introduced, asynchronous supervisory decisions concerning E2E resource allocation must also be considered. For example, the IDMO may enforce cross-domain resource allocation to enable deployment of a new E2E service. If such decisions are not coordinated with underlying DMOs, conflicts and performance degradation may occur.

C. AI/ML Subsytem Pipeline

The AI/ML subsystem is used to design pipelines in terms of connected tuples (MS, AE, DE, ACT) for three main purposes: *(i) training pipeline*, linking the functional tuple building blocks to achieve all the required operations for training of a ML model. This pipeline is extended to also incorporate distributed learning mechanisms, such as federated learning and multi-agent reinforcement learning; *(ii) inference pipeline*, serving an already trained model to an inference host to be used for online operation; *(iii) evaluation pipeline*, monitoring the online inference of a model and assessing its performance. This pipeline may trigger a re-train process in case of model degradation detection.

For instance, a *training pipeline* for an O-RAN control loop is initialized by an ML intent, i.e., a request that includes all the relevant information for performing the training of an ML model (optimization objective, required training data and algorithm, etc.). The AI/ML Orchestrator acknowledges this request and invokes the DM component to gather the raw data (corresponding to the MS component) and the requested AI/ML algorithm (ML code and base model) that will be used to perform the model training process. The raw data are then forwarded to the AE block (can be hosted in Near-RT RIC) that includes fundamental data integrity and pre-processing operations, along with data generative methods provided by the Sandbox to enhance the training dataset. In addition, the AE block prepares and initiates the training procedure (data loading, pre-processing, training, testing, and validation). Then, the DE component (also hosted in theNear-RT RIC) is responsible for exploring the training hyper-parameters and deciding upon the completion of the process. The trained model is stored centrally in a model registry database but can be also stored locally (at the location of the inference host as an xApp). Regarding *the inference pipeline*, the DM is responsible for identifying where the online data can be found for inferring the model and for properly configuring the tuples and the chaining of the subcomponents (MS, AE, DE, ACT) of the inference pipeline. The ML intent conveyed to the AI/ML Orchestrator in the inference pipeline includes this specification, along with the target optimization objective (e.g., forecasting of energy usage), providing the details for selecting the pre-trained model for deployment among different models in the ML catalogue.

In a functional scenario of the Performance Monitoring (PM) pipeline, the evaluation module is initialized based on the defined ML intent (assuming that the PM data have been gathered for accuracy of the ML model during the online inference phase). The DM in the evaluation process is responsible for pointing out where the PM data can be found and gathered, properly configuring the tuples and the chaining of the subcomponents (MS, AE, DE, ACT) of the evaluation pipeline. The ML intent requested from the AI/ML subsystem includes the PM specifications of an online operating ML model (e.g., accuracy of the model, latency, or defined KPIs).

The 6G network is expected to provide robust support for the AI/ML subsystem by guaranteeing uninterrupted interconnection among its constituent subcomponents. In the event of a disruption to the terrestrial network, for example, due to a catastrophic event affecting the links between subcomponents such as MS, AE, DE, and ACT, connectivity should be rapidly restored via the non-terrestrial segments of the 6G infrastructure. Furthermore, the AI/ML subsystem pipeline must be designed to tolerate increased latency under such conditions. Across the three pipelines, the AI/ML subsystem supports both deterministic and probabilistic learning paradigms, enabling the generation and exploitation of uncertainty-aware predictions within closed-loop control. Forecasting models deployed through the training and inference pipelines may output predictive distributions, quantiles, or confidence intervals in addition to point estimates, following recent advances in probabilistic time-series forecasting [12].

The *evaluation pipeline* plays a key role in sustaining the reliability of such probabilistic models over time. Beyond conventional performance indicators, it enables the assessment of uncertainty-related properties, such as prediction of interval coverage and stability, supporting the detection of concept drift or changes in network operating regimes. When degradation is observed, the pipeline can automatically trigger retraining or model adaptation.

D. Semantic Communications

Semantic communication enables the transfer of task-relevant meaning rather than transmitting source raw data. It prioritizes what needs to be conveyed, e.g., scene components for holographic and XR services. This capability is particularly important in integrated 6G deployments including NTN, where spectrum is limited, propagation delays are higher and link availability is intermittent. In UNITY-6G, semantic communication is treated as a cross-layer function spanning service endpoints, edge computing and RAN control. The architectural goal is to enable controlled data reduction and adaptive quality management based on semantic relevance.

The mapping of the semantic communication chain follows a clear separation between the Infrastructure Layer and the Management and Orchestration Layer. In the terrestrial domain, semantic processing and control functions are integrated into an O-RAN-based TN architecture, where semantic encoding is performed at the service source and semantic decoding at the external XR application. This placement ensures that semantic data generation and interpretation remain independent of the radio implementation. Semantic-aware control is realized through RIC: short-timescale adaptations, such as scheduling and traffic prioritization for semantic flows, are handled by semantic xApps in the near-real time (RT) RIC based on real-time feedback from the Distributed Unit (DU), while longer-timescale policy and configuration decisions are managed by semantic rApps in the Non-RT RIC within the SMO.

In the NTN domain, the radio access path includes UE-NTN, satellite payload, and ground station components, with semantic encoding performed at the transmitting side and semantic decoding at the receiving side, optionally supporting split processing depending on service requirements. Semantic traffic control is implemented through RIC-based mechanisms, where near-RT RIC xApps apply short-timescale adaptations based on NTN-specific factors such as congestion, delay, and link availability, while non-RT RIC rApps handle semantic policy configuration and service parameter definition. Prior to transmission over the satellite segment, the NTN-DU enforces semantic-aware scheduling and prioritization decisions. UNITY-6G contributes to distributed AI-based optimization of semantic compression, traffic steering and resource allocation, as well as perception-aware Age of Information models and AI methods to balance semantic accuracy and energy efficiency.

V. TRUST MECHANISMS IN UNITY-6G

To complement the SBMA-based multi-domain integration, UNITY-6G introduces a dedicated *Trust Architecture* that makes trust enforceable across heterogeneous administrative domains rather than treating it as a set of isolated, domain-specific checks. This architecture addresses the explicit trust boundary requirements defined in 3GPP and ETSI [14] extends the *Zero Trust* principles studied beyond the network-centric model to encompass cross-domain orchestration scenarios. The architecture exposes *Trust Adapters* at the IDMO and DMO layer, so that identity management, attestation, reputation assessment, and trust-policy retrieval can be invoked across O-RAN, IoTs etc. DLT and its enabled DLT-based logs (e.g., operational measurements, compliance outputs, context signals, and authentication/authorization events) are collected via the Data Continuum and processed in a closed loop where trust assessment and enforcement actions remain aligned with orchestration decisions and user/tenant intents. To ensure cross-domain auditability and non-repudiation critical for inter-operator trust relationships; UNITY-6G anchors trust-relevant evidence, policies, and decisions on a distributed ledger, and uses smart contracts as one of the ways to encode programmable trust rules that can be automatically and in transparent way evaluated as part of the orchestration workflow. Concretely, this trust layer is realized through modular blocks based on DLT and non-DLT for access and credential control, enabling trust to be continuously monitored, verified across domains with transparent traceability.

VI. USE CASE IMPLEMENTATION AND POCS

UNITY-6G focuses on implementing four primary use cases: sustainable networks for disaster handling, immersive real-time XR and holographic communications, DTs for network evaluation, and multi-RAT O-RAN enabled NPNs for Industry 4.0. *PoC 1* targets network resource sharing and sustainability using an AI-driven orchestrator and a network evaluator enhanced by Generative AI data and DT. *PoC 2* focuses on end-to-end time-sensitive services across multi-RAT (cellular and IEEE 802.11) O-RAN architectures, demonstrating real-time AI/ML-based resource optimization and DT services for load balancing in Industry 4.0. A critical component of the PoCs is the exposure of infrastructure via open APIs.

A. Network Exposure

The architecture includes a network exposure framework designed to provide a unified, secure and interoperable interface between the integrated 6G network infrastructure and a diverse set of application functions, including vertical applications, service orchestrators, network controllers and digital twin platforms. Acting as a centralized exposure layer, the network exposure framework enables secure and standardized access to monitoring data and control mechanisms across multiple computing and network domains (RAN, transport, core). This framework serves as a key enabler for the PoCs. By exposing network information and fine-grained control mechanisms, it facilitates data-driven, AI-native optimization for service placement, resource management and end-to-end QoS/QoE satisfaction. In order to achieve this, UNITY-6G introduces technologies and APIs that expose network information and functionalities, in compliance with standardization and industrial initiatives (e.g., 3GPP NEF/CAPIF, CAMARA, and O-RAN). Exposure mechanisms are implemented across the different network segments:

1) Core network exposure: utilizes 3GPP CAPIF, NEF, and NWDAF to securely expose network events, analytics, QoS impact, and traffic management functionalities to application functions. *2) Transport/xHaul exposure*: leverages SDN-based interfaces to expose link- and path-level information, including topology status, latency, bandwidth availability, congestion indicators and flow-level metrics. The APIs allow external and internal UNITY-6G consumers to obtain a real-time view of the transport domain and to influence flow management decisions. *3) RAN* exposure (Multi-RAT NPN/NTN): encompasses both monitoring and control capabilities for multi-RAT terrestrial and non-terrestrial deployments. Leveraging O-RAN interfaces and in-band network telemetry, it delivers enhanced per-flow and per-device KPIs (e.g., CQI, resource utilization or UE throughput). Control mechanisms include parameters, scheduling policies selection and thresholds tuning.

In addition to domain-specific exposure, the network exposure framework distinguishes between 2 complementary API layers*: (i) Network exposure Layer-1*, provides telco-specific APIs facilitating direct access to domain functionalities such as xApps/rApps/dApps in the RAN, SDN controllers in the transport, and NEF/CAPIF in the core; and *(ii) Network exposure Layer-2*, for app-developer-oriented APIs founded on CAMARA initiatives, simplifying complexity for external developers and vertical service providers. Network exposure consumers, ranging from internal UNITY-6G architecture components such as orchestrators, to external application backends, can interact with the framework through direct API invocations. Exposed information is jointly exploited by vertical applications and internal network components to enable intelligent, cross-domain optimization decisions.

B. PoC 1

It focuses on resource sharing and sustainability and on integrated TN-NTN resource sharing during disaster scenarios where communication capacity is scarce. It utilizes policy-based DEs, Distributed AI, and DLT to ensure fair, multi-tenant access while maintaining traceability of all allocation decisions. A central objective is renewable energy awareness, which incentivizes tenants to use resources when solar or wind power is most abundant. Furthermore, the PoC incorporates a Semantic-Aware AE that extracts context-aware info rather than raw telemetry. It uses GenAI-enhanced DT to monitor sustainability metrics and evaluate the interplay between microgrid power and network energy consumption predictions.

C. PoC 2

It is focused on Time-Sensitive services for Industry 4.0 validates a unified, multi-RAT (cellular and Wi-Fi) O-RAN architecture designed to maintain strict performance for industrial applications. Several Use Cases (UC) are considered: **(i)** *UC1 (Unified Multi-RAT Control)*: This scenario integrates Wi-Fi APs into the O-RAN framework, allowing industrial devices to maintain seamless time-sensitive connectivity across different radio technologies. **(ii)** *UC2 (AI/ML Resource Optimization)*: An AI/ML pipeline is deployed to optimize resources across both RATs in a unified manner. **(iii)** *UC3 (Digital Twin for Load Balancing)*: DT forecasts the impact of reconfigurations in a virtual environment before they are physically implemented. It enables disruption-free load balancing, ensuring the network meets the stringent latency requirements of TSN.

VII. LESONS LEARNED AND OPEN CHALLENGES

UNITY-6G architecture faces several challenges associated with its goals. The primary challenges identified include:

1. SBMA Heterogeneity: Integrating diverse domains through a SBMA is complex due to the inherent heterogeneity of the domains involved. This is particularly challenging for "autonomous-driven" standardized domains like O-RAN.

2. Conflict Management: In O-RAN, a major hurdle is managing conflicts between independent intelligent applications (xApps/rApps). Decisions made by a top-level orchestrator (IDMO) must be coordinated with the underlying DMOs to prevent decisions causing performance degradation.

3. Integration TN-NTN: Integration of NTN is still in its early stages. Current systems require further automation to move beyond simple satellite channel configuration toward on-demand creation of mobile network entities like terrestrial-based gNBs. This problem is even harder when we consider LEO satellites due to the non-continuous visibility of ground stations and the need for frequent handovers.

4. Sustainability vs performance: While architecture aims to be "carbon-aware", it faces the algorithmic challenge of weighing sustainability metrics against strict QoS constraints to ensure that shifting processing does not sacrifice performance.

VIII. CONCLUSIONS

UNITY-6G introduces an AI-native framework unifying terrestrial, NTN and NPN under an SBMA-based orchestration layer. By integrating agentic AI, DTs and semantic communications, the architecture enables proactive, intent-driven operations validated through scenarios like disaster recovery and Industry 4.0. The inclusion of DLT and energy-aware resource management ensures a secure and sustainable ecosystem. In conclusion, architecture establishes a flexible, future-proof foundation for autonomous 6G systems capable of supporting heterogeneous, high-performance services.

ACKNOWLEDGMENTS

This work is supported by UNITY-6G project, funded from European Union's Horizon Europe Smart Networks and Services Joint Undertaking (SNS JU) research and innovation program under the Grant Agreement No 101192650, and Swiss State Secretariat for Education, Research and Innovation.

REFERENCES

[1] 3GPP TS28.533, Management and Orchestration: Architecture Framework (Release 19), tech. Rep, 3GPP, 2025.

[2] IEA (2025), Energy and AI, IEA, Paris https://www.iea.org/reports/energy-and-ai, Licence: CC BY 4.0

[3] C. Chaccour et al. "Less Data, More Knowledge: Building Next-Generation Semantic Communication Networks," in *IEEE Communications Surveys & Tutorials*, vol. 27, no. 1, pp. 37-76, Feb. 2025

[4] E. Zeydan et al., "Semantic Non-Terrestrial Communications With Open RAN-Enabled 6G," in *IEEE Comm. Standards Mag*, vol. 9, no. 4, p. 72-78, 2025.

[5] 6G-GOALS Consortium, "Deliverable D2.5: 6G-GOALS Architecture and Operation Analysis," in 6G-GOALS project website, 2025.

[6] Buzcu, B., et al.,"Energy-Aware Multi-Domain Orchestration Architecture for Sustainable 6G Networks", in WONS 2026 (accepted) .

[7] Haxhibeqiri, J., et al., "In-band network monitoring technique to support SDN-based wireless networks". *IEEE TNSM*, 18.1 (2020): 627-641.

[8] I. Vilà, et al., "On the design of a network digital twin for the radio access network in 5g and beyond," Sensors, vol. 23, no. 3, p. 1197, 2023.

[9] V. Kasuluru, et al., "Cloud-Native Multivariate Probabilistic Forecasting for Next-Generation Resource Allocation in O-RAN". IEEE Access, 14 (2025): 2904-2923

[10] Vaca-Rubio, C., et al, "Probabilistic Forecasting for Network Resource Analysis in Integrated Terrestrial and Non-Terrestrial Networks". IEEE Comm. Standards Mag., 2025.

[11] Sapkota, R et al. "AI agents vs. agentic AI: A conceptual taxonomy, applications, and challenges". *Information Fusion,* 126, 103599, 2026.

[12] Vaca-Rubio, C. J., et al., A Primer on Kolmogorov-Arnold Networks (KANs) for Probabilistic Time Series Forecasting. arXiv:2510.16940.

[13] Giannopoulos, A., et al. "COMIX: Generalized Conflict Management in O-RAN xApps-Architecture, Workflow, and a Power Control case". *IEEE Access*, vol. 13, pp. 116684-116700, 2025.

[14] ETSI PDL ETSI GR PDL 030, "Permissioned Distributed Ledger (PDL); Trust in Telecom System", tech. rep., ETSI Industry Spec., May 2025.