6G Standardization Potential of the ORIGAMI Novel Architectures and Use Cases

Livia Elena Chatzieleftheriou, David de Andres Hernandez, Simone Bizzarri, Marco Fiore, Maurizio Fodrini, Andres Garcia-Saavedra, Marco Gramaglia, Esteban Municio, Dimitris Tsolkas

***Abstract*—The transition to 6G presents many barriers to be overcome, as well as opportunities for innovation. The integration of Network Intelligence (NI) is pivotal in optimizing network performance, enhancing security, and improving resource allocation. The ORIGAMI project identifies 8 critical barriers to 6G deployment and proposes both architectural and NI innovations to overcome them. This paper discusses the standardization potential of such innovations, which respond to 10 different use cases, each with diverse necessity and impact on the associated components, and across multiple network domains. We present in detail three key architectural innovations, namely the Compute Continuum Layer (CCL), the Zero Trust Layer (ZTL), and the Global Service-Based Architecture (GSBA), are leveraged to ensure dynamic adaptation and zero-trust business models; we then discuss two representative NI innovations targeting energy efficiency and infrastructure management. Overall, this paper shows how ORIGAMI's comprehensive approach to innovation aligns with and impacts ongoing standardization efforts.**

***Index Terms*—6G, mobile network architectures, network intelligence, global mobile services, compute continuum**

I. Introduction

The advancement of mobile network technologies towards 6G introduces a set of unprecedented challenges and opportunities. As the complexity of network architectures grows, the integration of Network Intelligence (NI) has emerged as the fundamental strategy to ensure secure, efficient, and adaptive network operations. NI refers to using data-driven insights, artificial intelligence (AI), machine learning (ML), and advanced analytics to optimize network performance, enhance security, and improve resource allocation. By enabling dynamic adaptation to change demands and fostering zero-trust business models, NI is a cornerstone of the 6G paradigm.

The ORIGAMI project identifies eight critical barriers to the successful deployment of 6G networks. These include unsustainable RAN virtualization, poor interoperability of RAN components, high latency in NI, underutilized programmable transport, the absence of global service APIs, an obsolete trust model, inadequate networking data representation, and high volumes of control plane signaling. Addressing these barriers requires innovative architectural models and NI solutions tailored to the unique demands of 6G.

In addition to solving immediate barriers, the use cases proposed in ORIGAMI are associated with novel architectural components, ensuring that the ORIGAMI project not only addresses present challenges but also aligns with evolving standards and global research efforts.

In this paper, we present a discussion of the standardization potential of 10 use cases designed to overcome these barriers. For each use case, we (i) motivate its necessity in the context of the associated barrier, and (ii) discuss its potential impact on the relevant architectural component. These solutions span multiple network domains—Radio Access Network (RAN), transport network, and core network—and leverage key architectural innovations such as the Compute Continuum Layer (CCL), Zero Trust Layer (ZTL), and Global Service-Based Architecture (GSBA).

For example, in the RAN domain, solutions such as data-driven task offloading for vRAN acceleration address unsustainable virtualization by optimizing workload distribution in real-time, leveraging the complementarity of CPUs and hardware accelerators. Similarly, in the transport domain, innovative distributed ML models improve programmable user-plane performance, while in the core network, anomaly detection and privacy-preserving analytics enhance global operator functionality.

This paper is structured as follows: In §II we discuss 3GPP Topics of Interest, in §III our architectural components, in §IV our results, and in §V we conclude the paper.

II. 3GPP Topics of Interest

In this section, we discuss the 3GPP topics of interest, as well as the ORIGAMI innovations and tackled barriers.

A. Topics in 3GPP standardization

The 3rd Generation Partnership Project (3GPP) is a central entity in the global standardization of mobile network technologies. Its work is crucial in shaping the future of advanced 5G systems and next-generation 6G networks. Within the 3GPP framework, a number of key research topics are of paramount importance, as they define the technological directions and solutions for the forthcoming evolution of telecommunication networks. These topics encompass a broad range of innovations, including network resource management, AI/ML integration, and energy efficiency, among others. Below, we outline 10 main topics, termed T1-10, highlighting their relevance to 5G advanced and 6G, as well as their corresponding 3GPP Working Groups (WGs).

Network Resource Model (T1). This is a fundamental framework that represents network elements and services in a

L.E. Chatzieleftheriou, D. de Andres Hernandez, and M. Fiore are with the IMDEA Networks Institute. S. Bizzarri and M. Fodrini are with FiberCop S.p.A. M. Gramaglia is with University Carlos III de Madrid (UC3M). A. Garcia-Saavedra is with NEC Laboratories Europe GmbH. E. Municio is with I2CAT. D. Tsolkas is with National and Kapodistrian University of Athens as well as with Fogus Innovations and services P.C.

technology-agnostic manner, facilitating efficient management in diverse network environments. Developed by the SA5 3GPP WG, it underpins network management processes, enabling interoperability and flexibility in all the management processes.

Orchestration (T2). It coordinates network management processes such as policy management, intent-based systems, and AI/ML lifecycle management. It ensures the automated and efficient management of resources in dynamic networks. The SA5 3GPP WG focuses on standardizing orchestration of the automatic processes for 5G and 6G networks, incorporating the results and outputs from the activities of other 3GPP Working Groups (e.g. RAN3 WG).

AI/ML (T3). It optimizes network performance, automates management, and enhances decision-making. It also supports AI/ML lifecycle management, data analysis, and energy efficiency. The integration of AI/ML is a cross-cutting activity involving all 3GPP Working Groups, as it has become a fundamental aspect of new networks, driving automation, optimization, and enhanced decision-making capabilities. To ensure alignment across the various groups on this topic, a specific work item [1] has been activated within TSG SA to coordinate the activities related to AI/ML standardization.

Energy Efficiency (T4). Energy efficiency focuses on reducing the energy consumption of networks, evaluating CO2 emissions, and assessing the overall environmental impact. It is essential as data demand grows. While SA5 plays an important role in this area, many other 3GPP Working Groups are also involved in advancing energy efficiency. Additionally, there are proposals to extend these efforts to include the evaluation and optimization of the OAM (Operations, Administration, and Maintenance) processes themselves.

Network Performance Management (T5). This involves defining KPIs and KQIs, as well as the processes for monitoring and optimizing network performance. The activity in 3GPP is led primarily by the SA5 WG, with the involvement of other WGs such as RAN1, RAN2, RAN3, and SA2.

RAN Architecture (T6). The evolution of RAN architecture, particularly with the potential introduction of Service-Based Architecture (SBA) in the RAN for 6G, could enhance flexibility and scalability in network design. The RAN3 Working Group is exploring this possibility as part of its research to optimize RAN systems for future networks.

Data Analytics (T7). It involves processing network data to derive insights for optimization and fault detection. The SA5 WG is focused on creating standards to enable data-driven network management, improving decision-making and operational efficiency.

Network Digital Twins (T8). They create virtual models of networks, simulating behavior and performance for optimization and troubleshooting. The SA5 WG is advancing the integration of digital twins in 5G and 6G networks.

Infrastructure Management (T9). It involves coordinating network functions with underlying technologies like NFV, cloud, and edge computing. The SA5 WG focuses on standardizing these processes to ensure efficient resource utilization in dynamic networks.

TABLE I
BARRIERS THAT ORIGAMI AIMS TO REMOVE (FROM [2]).

B#	Barrier	ORIGAMI Innovation #
1	Unsustainable RAN virtualization	3, 4, 5, 8
2	Poor interoperability of RAN components	2, 3, 8
3	High latency and unreliable Network Intelligence (NI) to process complex 6G network problems	3, 4, 5, 8
4	Under-utilized modern programmable transport	1, 6
5	Lack of global service APIs	1, 2, 6, 9
6	Obsolete trust model hinders performance	7, 9
7	Inadequate networking data representation	3, 5, 7
8	High volume of control plane signaling	3, 5, 7

Security Data Management (T10). It focuses on enhancing the collection and aggregation of security-related data, including metrics, alarms, and logs, from various network entities. This ensures automated and continuous monitoring of network security health. The SA3 WG aims to standardize these processes to address the complexity of 5G and beyond.

B. Innovations and tackled barriers

As explained in [2], despite the advancements of 5G, mobile networks still face 8 inherent architectural barriers, **B1-8**. Removing them requires fundamental changes to pave the way for 6G. These barriers, ranging from infrastructure limitations to inadequate global operation support, are detailed in Table I.

6G demands efficient resource control across diverse providers. Barriers 1-4 emphasize the need to tackle complexity, interoperability, latency, and resource utilization. A unified management framework is lacking, especially with novel radio technologies (B1) and poor open-RAN interoperability (B2). High latency in complex network problems (B3) necessitates solutions like asynchronous functions and quantum computing, while under-utilized programmable transport (B4) hinders innovation.

Enabling global network operations for tenants and external MNOs is key. Barriers 5-8 highlight limitations: hindering smaller providers and global services (B5), obsolete trust models (B6), lack of common network data representation (B7), and excessive control traffic (B8). ORIGAMI addresses these barriers with nine innovations, termed **I1-9**:

Resource Abstraction (I1): Creates a generalized data model for interoperability and simplified management (B4,5).

Orchestration (I2): Enables automated network management, including resource allocation and service provisioning (B2,5).

AI/ML (I3): Transforms network management through intelligent automation, predictive analytics, and optimized decision-making (B1-3,7,8).

Energy Efficiency (I4): Focuses on monitoring and reducing energy consumption (B1,3).

Service & Network Performance (I5): Ensures optimal performance via monitoring/proactive issue resolution (B1,3,7,8).

Network Architecture (I6): Addresses the network architectures' evolution, including the transition to SBA for 6G (B4,5).

Data Analytics (I7): Extracts insights from network data for informed decision-making (B6,7,8).

Infrastructure Interworking (I8): Enables seamless interop-

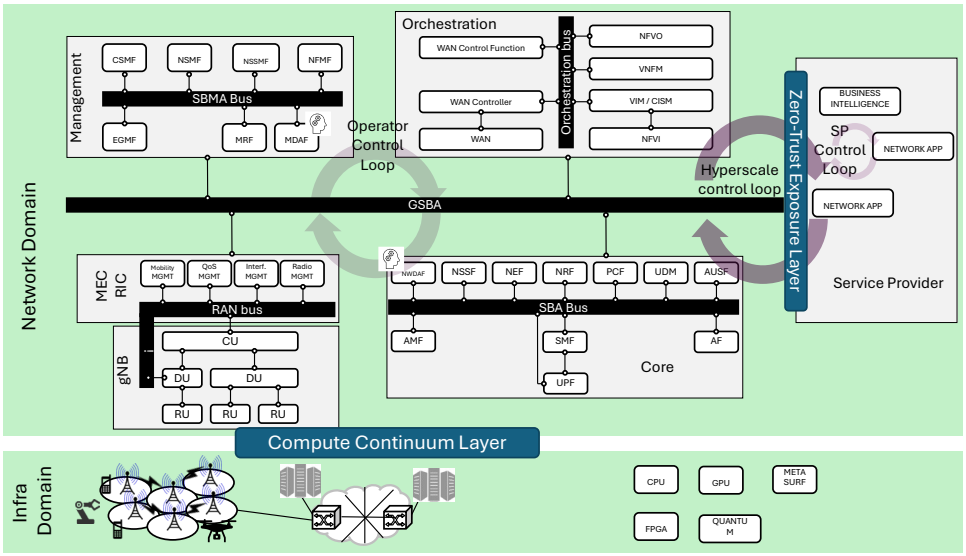

Fig. 1. Overall architecture envisioned by ORIGAMI.

erability between different infrastructure domains (B1,2,3).
Security (I9): Emphasizes advanced security technologies and strategies, including Zero Trust Architecture (B5,6).

III. Architectural Components and Impact over the 3GPP standards

In this section, we start by presenting the three key architectural innovations of ORIGAMI and their potential impact on 3GPP standards through ten use cases.

A. The architectural components

Motivated by the above innovations, we point towards three architectural enablers to evolve the current 5G SBA. This new re-designed network architecture maintains the essence of the 5G architecture while enabling the transition to 6G by supporting novel NI functionalities and global services. As shown in Fig. 1, the proposed architectural innovations are the CCL, the ZTL and the GSBA.

Compute Continuum Layer (CCL). The CCL is a novel architectural enabler conceived to optimize operations according to the underlying network infrastructure. By allowing for the abstraction of heterogeneous computing resources through a CCL Broker, and a compute-aware control of Network Functions (NF) through a CCL controller, the CCL enables efficient use of the infrastructure in terms of improved energy efficiency and reduced Total Cost of Ownership (TCO). The CCL shall support a variety of computing resources such as CPUs, GPUs, FPGAs, SmartNICs, etc., as well as other emerging technologies such as quantum hardware or smart surfaces, to accelerate virtual NFs and ensure they are scaled efficiently while fulfilling their strict compute time guarantees.

Zero-Trust Layer (ZTL). ORIGAMI aims to enable zero-trust interactions through the ZTL between entities that support global services, while also tackling the limitations of the current trust model within the Mobile Network Operator (MNO) ecosystem. The ZTL brings two major changes to this ecosystem: i) *Horizontal Exposure:* enables MNOs to have a dynamic cooperation, enabling emerging operators' business models to establish relationships and perform financial clearing in almost real time. This is opposite to current trends, where the process using Data Clearing Houses is slow, and money moves slowly across geographies; ii) *Vertical Exposure:* enables the end-user to have control over its mobile connection. Currently, the end–user is locked-in only with one operator and migrating (though technically possible) is cumbersome and risky. These aspects allow for an architecture based on a new trust model that matches the internal operation of the service provider's business logic and the MNO's continuous optimization.

Global SBA Architecture (GSBA). ORIGAMI's architectural design aims to tear down the lack of trust and silo-based design of different network domains (core, orchestration, RAN, etc.), integrating them all into one single GSBA bus. The GSBA not only supports the CCL and ZTL architectural innovations but also other legacy domain buses such as the 3GPP SBA, enabling a multi-domain, holistic management that addresses trust issues between domains and ease and optimize resource sharing. Among the main highlights of the GSBA there is its extension towards the RAN, which means, splitting monolithic RAN modules into control and user plane functions (e.g., analogously to the ones in the 5G core). Introducing a SBA into the RAN brings several benefits to the network operations and management, such as improved scalability, enhanced modularity, improved automation, higher resilience and future-proofing for long-term evolving technologies.

B. Impacting standardization through ORIGAMI Use Cases

1) Use Cases: ORIGAMI seeks to address various barriers in the implementation of advanced network architectures by defining a set of 10 use cases, termed U1-10. These use cases are designed to overcome identified challenges, improve efficiency, and support sustainability in 6G network development. Each use case targets specific barriers and incorporates architectural innovations, non-functional requirements (NFRs), and KPIs. In the following, we summarize their respective

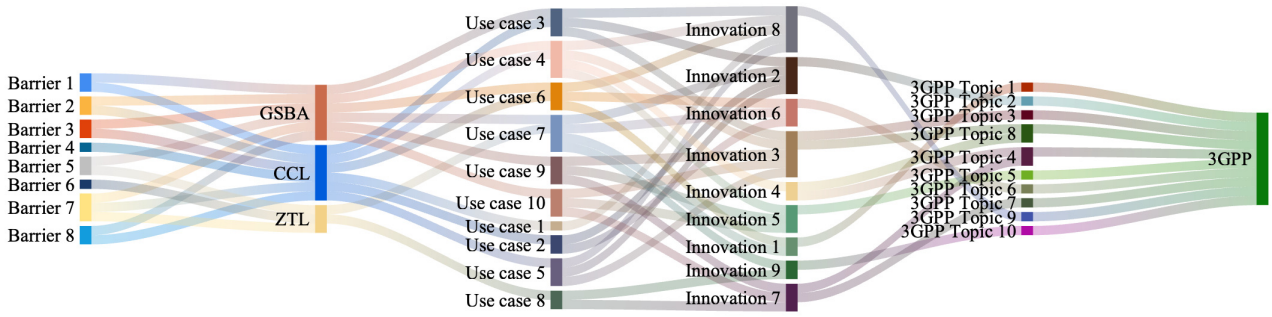

Fig. 2. Mapping of the standardization potential in the ORIGAMI project. The identified barriers (left) and the envisioned architectural components are mapped to the project use cases and the innovations analyzed in the project. They are then mapped to the 3GPP topics of interest for 5G Advanced and 6G.

architectural innovations and targeted KPIs:

Data-driven Task Offloading for Reliable vRAN Acceleration (UC1): Addresses unsustainable RAN virtualization by implementing opportunistic hardware accelerator (HA) offloading and processor pooling. Target KPIs: K2 (cost efficiency), K3 (reliability).

Conflict Mitigation of xApps and Interoperability of O-RAN Components (UC2): Focuses on conflict mitigation in xApps and ensuring E2 interface interoperability. Target KPIs: K2 (cost efficiency).

Enhancing Management and Stability in the 6G Architecture (UC3): Improves network management through scalable xApps. Target KPIs: K2 (cost efficiency).

Interoperable Machine Learning Models Improving RAN Energy Efficiency (UC4): Embeds interoperable ML models into RAN systems to enhance energy efficiency. Target KPIs: K3 (reliability), K4 (latency), K5 (accuracy).

Compute- and Fairness-Aware Radio Resource Allocation Algorithms (UC5): Designs resource allocation policies balancing energy savings and fairness. Target KPIs: K1 (energy efficiency).

Effective Access to U-Plane Computing Capabilities (UC6): Develops ML solutions for distributed user planes. Target KPIs: K4 (latency), K5 (accuracy), K6 (throughput).

Enabling a Global Operator Model (UC7): This key use case, proposes a novel architecture for global network interoperability and billing models. Target KPIs: K7 (CAPEX), K9 (control plane latency).

Limited Trust Network Analytics (UC8): Supports network-application integration under limited trust. Target KPIs: K10 (anomaly detection recall and sensitivity).

Anomaly Detection (UC9): Develops knowledge representations to detect IoT anomalies in global cellular networks. Target KPIs: K10 (recall and sensitivity), K11 (OPEX gains).

Network Core Traffic Analysis and Optimization (UC10): Optimizes network signaling traffic to improve scalability and efficiency. Target KPIs: K12 (control plane efficiency).

2) The architectural components in the standard: Integrating ORIGAMI's architectural components has the potential to impact the current and near-future standardization work in the context of 3GPP, as discussed in Section II-A. We now discuss how each of the three main architectural components can impact standardization, as graphically depicted in Fig. 2.

Global SBA Architecture (GSBA). The GSBA holds significant potential in various aspects of advanced Network Resource Management (NRM). By adopting the Service-Based Architecture (SBA) across all network domains, GSBA enables enhanced Resource Abstraction, allowing the NRM to integrate procedures spanning multiple parts of the network, including the RAN, Core, Management, and Orchestration. This capability is critical for supporting UC7—the Global Operator Model (GMNO). The GMNO leverages this functionality to enable seamless cross-operator procedures, particularly those related to authentication, charging, and the utilization of shared resources. Additionally, these features support other use cases, such as Anomaly Detection, which addresses challenges like B7 by utilizing data from all network domains, not just the Core.

GSBA also influences the design of AI/ML systems in network standards. The current foundation for Data Analytics and AI/ML frameworks is the Network Data Analytics Function (NWDAF) [3], which receives fine-grained data from Core Network functions and input data from the Management Data Analytics Function (MDAF). The MDAF provides Management Data Analytics Services (MDAS) within the management domain. While this coarse-grained interface facilitates data exchange across multiple domains, including the RAN and OSS/BSS frameworks, GSBA enables more direct and finer-granularity data access. This improvement enhances the precision of network analytics and AI-driven decision-making.

Finally, the GSBA also directly impacts the RAN Architecture. Bringing this paradigm to the far edge of the network allows better interoperability of the different modules running in that domain. Through the GSBA it is possible to uniform two current state-of-the-art architectures such as O-RAN [4] and 3GPP. This is directly connected to the B2 in ORIGAMI.

Compute Continuum Layer (CCL). Through its integration, ORIGAMI aims to enhance the abstraction of resources across all network elements, enabling, through the GSBA, improved network resource modeling. Resource pooling facilitated by the CCL promotes energy efficiency and the seamless integration of AI/ML/Digital Twinning deep within the network.

As detailed in Sec. IV-A, an architectural module that abstracts the underlying computing capacity allows for its

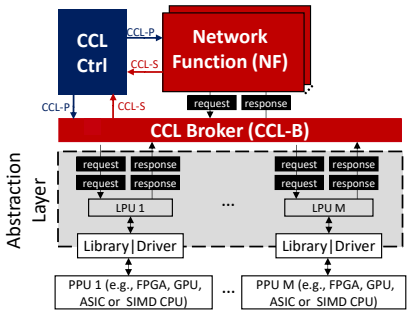

Fig. 3. ORIGAMI's Compute Continuum Layer (CCL).

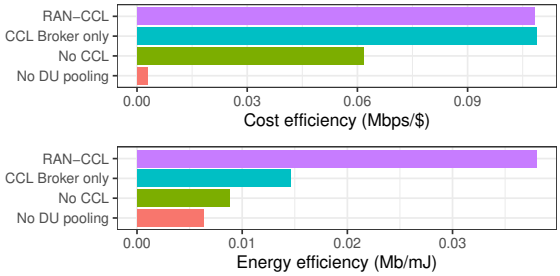

Fig. 4. RAN CCL vs benchmarks. 60 5G Distributed Units, 300 UEs, traffic traces from real-world cells in Madrid, Spain.

efficient on-demand utilization. Through standard interfaces, network functions—particularly those at the RAN—can dynamically select the execution setup (e.g., a Hardware Accelerator as in Sec. IV-A) to optimize energy efficiency metrics while maintaining carrier-grade network KPIs.

Another significant impact of the CCL lies in Infrastructure Management, as discussed in Sec. IV-B. The CCL enables the definition of a unified gateway for injecting AI/ML solutions directly into hardware components, such as the transport network (Sec. IV-B). Additionally, the CCL is foundational for integrating Digital Twins into the network, especially those modeling the behavior of specific functions onboarded on hardware platforms. This integration optimizes trade-offs between performance and resource utilization.

Zero Trust Layer (ZTL). Through the ZTL, various entities—such as other Mobile Network Operators (as envisioned in the GMNO framework) or third-party application providers—can interact with the network in a reliable and trustworthy manner. This innovation, combined with the GSBA, has the potential to improve access to external resources and deliver more customized services to end users. Crucially, it achieves this without exposing internal network information, thereby supporting business models such as those involving Non-Public Networks (NPNs) and the IoT.

The ZTL, integrated with the 3GPP Common API Framework (CAPIF) [5] and the service mesh paradigm, reduces the overall volume of control plane signaling (addressing B8) and offers features such as customizable analytics and the potential to revolutionize the roaming architecture. It enables simpler and more trustworthy interactions, for example, through the integration of distributed ledger technologies. Additionally, the ZTL facilitates efficient interworking of infrastructure across different tenants while ensuring security by design.

IV. Results

In this section, we present results for two representative NI innovations developed by ORIGAMI, which are related to energy efficiency and infrastructure management.

A. Energy Efficiency

As introduced in §III, ORIGAMI's CCL is specifically designed to boost energy efficiency and TCO. The architecture of the CCL is depicted in Fig. 3. A diverse array of Physical Processing Units (PPUs), including hardware accelerators like Application-Specific Integrated Circuits (ASICs), Field-Programmable Gate Arrays (FPGAs), or GPUs, are accessed through an Abstraction Layer (CCL-AL). This layer homogenizes access to these heterogeneous resources by presenting them as Logical Processing Units (LPUs). Before real-time processing, some PPUs require the deployment of specific kernels, such as PIs for FPGAs or embeddings for quantum annealers. During this onboarding process, Network Functions (NFs) declare the Key Performance Indicators (KPIs) that the CCL must meet. NFs queue processing requests, specifying the required kernel and inputs. These requests are regulated by policies to ensure timely processing. The CCL Broker (CCL-B) routes requests between NFs and LPUs based on policies that prioritize factors like energy efficiency. In the control plane, the CCL Controller (CCL-C) mediates between NFs and the CCL-AL, computing policies to minimize cost and energy consumption while satisfying NF KPIs.

To demonstrate this, we implemented the CCL on the RAN domain (i.e., RAN-CCL) to: (i) virtualize 5G Distributed Units (DUs) in shared computing platforms (i.e., DU pooling), and (ii) balance the computing workload generated by DUs between high-performing but high-energy-consuming and low-perming yet low-energy-consuming L1 processors. Specifically, we virtualize 60 5G 100-MHz DUs and emulate the real behavior of 300 user equipment (UEs) by mimicking the workloads observed in real-world cells [6]. In Fig. 4, we compare the performance of ORIGAMI's CCL in this setting (RAN-CCL) with the industry-standard approach (i.e., no DU pooling, no heterogeneous computing) and two additional benchmarks: DU pooling without CCL Broker or CCL-C (No CCL) and DU pooling with CCL Broker only (no CCL Controller). Our results show that DU pooling achieves substantial gains in cost efficiency, but not necessarily in energy efficiency. Nevertheless, the CCL Broker is essential for maximizing cost efficiency (almost doubling that of a naive DU pooling mechanism without CCL), and the CCL Controller triples the energy efficiency of the system. These findings highlight the significant impact of the CCL on both the cost and energy efficiency of NFs such as DUs.

B. Infrastructure Management

Identifying traffic demands' requirements, such as latency or jitter, is crucial for efficient resource utilization in dynamic networks, particularly under stringent QoS constraints. These re-

TABLE II
PERFORMANCE COMPARISON OF ORIGAMI'S DISTRIBUTED INFERENCE

	Performance			Latency			
Dataset	Mousika	Jewel	Distributed	None	Mousika	Jewel	Distributed
UNSW	64.921%	65.718%	70.263%	852.54	871.64	981.47	1137.70
ToN-IoT	47.099%	61.718%	67.541%	852.54	869.18	1015.08	1183.61

quirements can be inferred by classifying the demands (flows), for instance, based on application or device type. Traditionally, deep packet inspection (DPI) has been the primary method for labeling traffic, relying on the analysis of upper-layer headers and payloads. However, the ubiquity of encrypted traffic and stringent latency budgets have rendered DPI-based solutions futile. For example, protocols like HTTPS or QUIC use native payload encryption, while at 100 Gbps, the packet processing budget can be as little as 6 nanoseconds, or a dozen CPU cycles. Under such constraints, even sophisticated deep-learning classifiers powered by dedicated hardware (GPUs) fail to meet the ultra-low latency requirements. Recent advancements demonstrate that deploying machine-learning models in the user plane is a promising approach to performing inference tasks under these conditions [7], [8]. This is possible thanks to techniques that allow implementing offline-trained models into programmable ASICs, allowing inference at line rate on transit traffic [9]. However, as models grow in complexity, implementing them in single network elements is not always efficient, or even possible, especially in the core domain, where they should co-exist with other essential functions.

We showcase how leveraging the computational capabilities of switches and middleboxes in the core domain, and recent advances, it is possible to distribute user-plane inference tasks across the available hardware—aligned with I3 and overcoming B3 and B4. Using ORIGAMI's proposed solution [10], we decompose high-performance, unconstrained models into sub-models optimized to balance accuracy and complexity, all through a fully automated process. This approach ensures efficient deployment while preserving inference performance. To validate our methodology, we employ two datasets: ToN-IoT for attack classification and UNSW for device identification. We evaluate our proposed distributed models using three programmable switches equipped with Intel Tofino BFN-T10-032Q chipsets. The performance of these models is benchmarked against two state-of-the-art monolithic user-plane inference solutions, Mousika [11] and Jewel [8]. Table II reports, for each solution, the achieved latency and F1 score—which are the target KPIs defined for UC6.

Our results demonstrate that ORIGAMI's distributed models consistently outperform the benchmarks in accuracy across all metrics for both application scenarios, except for the Weighted F1 score on the UNSW dataset, where Mousika slightly outperforms. Specifically, the distributed approach delivers accuracy gains of 4.5% and 5.8% in attack classification and device identification, respectively. On average, the gains come at a negligible cost (0.16-0.31 μs) of added latency compared to the no inference, Mousika, and Jewel cases. These findings highlight the potential of distributed ML techniques, combined with user-plane computing capabilities (UC6), to enable efficient traffic classification. This capability is critical for the effective management and operation of dynamic networks under stringent performance constraints.

V. CONCLUSION

The ORIGAMI project identifies eight critical barriers to be overcome for the deployment of 6G, and proposes ten innovative NI solutions that tackle specific use cases. We propose three key architectural advances, namely the Compute Continuum Layer (CCL), the Zero Trust Layer (ZTL), and the Global Service-Based Architecture (GSBA), and we discuss their potential impact over 3GPP standards through ten use cases. Finally, we present two specific use cases and their respective solutions to tackle energy efficiency and infrastructure management. The novel architectures ORIGAMI proposes and evaluates pave the way for the standardization of 6G.

ACKNOWLEDGEMENTS

ORIGAMI project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101139270. L.E. Chatzieleftheriou is a Juan de la Cierva awardee (JDC2022-050266-I), funded by MCIU/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR, and MADQuantum-CM project, funded by the Regional Government of Madrid and the EU "NextGenerationEU"/PRTR.

REFERENCES

[1] 3rd Generation Partnership Project (3GPP), "Technical specification group services and system aspects; study on 3gpp ai/ml consistency alignment (release 19)," tech. rep., 3GPP, 2024. version 0.2.0.

[2] L. E. Chatzieleftheriou *et al.*, "Towards 6G: Architectural Innovations and Challenges in the ORIGAMI Framework," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pp. 1139–1144, 2024.

[3] M.-A. Garcia-Martin *et al.*, "Network Automation and Data Analytics in 3GPP 5G Systems," *IEEE Network*, vol. 38, no. 4, pp. 182–189, 2024.

[4] A. Garcia-Saavedra and X. Costa-Pérez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," *IEEE Communications Standards Magazine*, vol. 5, no. 4, pp. 96–103, 2021.

[5] A.-S. Charismiadis *et al.*, "The 3GPP Common API framework: Open-source release and application use cases," in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pp. 472–477, 2023.

[6] L. L. Schiavo *et al.*, "CloudRIC: Open Radio Access Network (O-RAN) Virtualization with Shared Heterogeneous Computing," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 558–572, 2024.

[7] A. T.-J. Akem *et al.*, "Flowrest: Practical Flow-Level Inference in Programmable Switches with Random Forests," in *IEEE INFOCOM*, pp. 1–10, 2023.

[8] A. T.-J. Akem *et al.*, "Jewel: Resource-Efficient Joint Packet and Flow Level Inference in Programmable Switches," in *IEEE INFOCOM*, 2024.

[9] C. Zheng *et al.*, "Planter: Rapid Prototyping of In-Network Machine Learning Inference," *SIGCOMM Comput. Commun. Rev.*, vol. 54, p. 2–21, Aug. 2024.

[10] B. Bütün *et al.*, "DUNE: Distributed Inference in the User Plane," in *IEEE INFOCOM*, 2025.

[11] G. Xie *et al.*, "Empowering In-Network Classification in Programmable Switches by Binary Decision Tree and Knowledge Distillation," *IEEE/ACM Transactions on Networking*, vol. 32, no. 1, pp. 382–395, 2024.