ORIGAMI

Optimized Resource Integration And Global Architecture For Mobile Infrastructure For 6G

Project Introduction

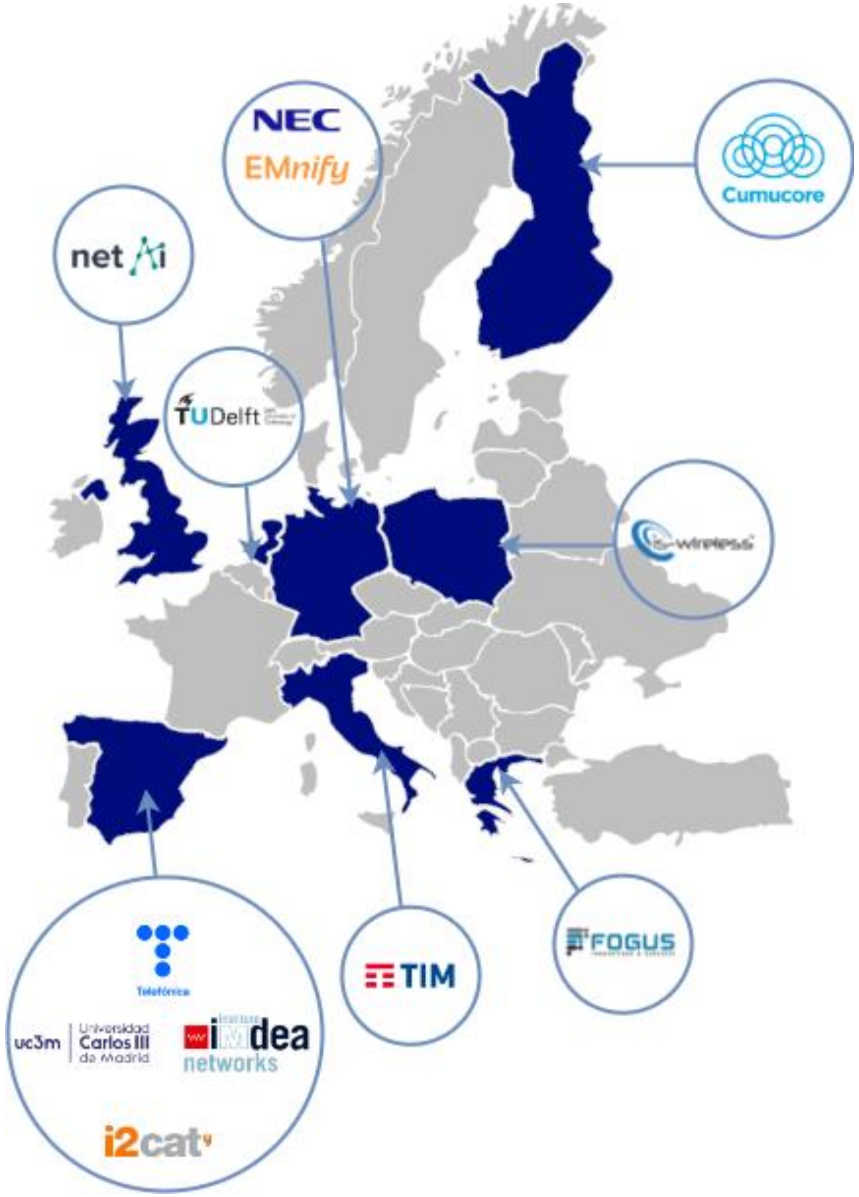
Andres Garcia-Saavedra
andres.garcia.saavedra@neclab.eu
Technical Manager, NEC

Co-funded by the European Union

6GSNS

GA:101139270
Call: HORIZON-JU-SNS-2023-STREAM-B-01-01

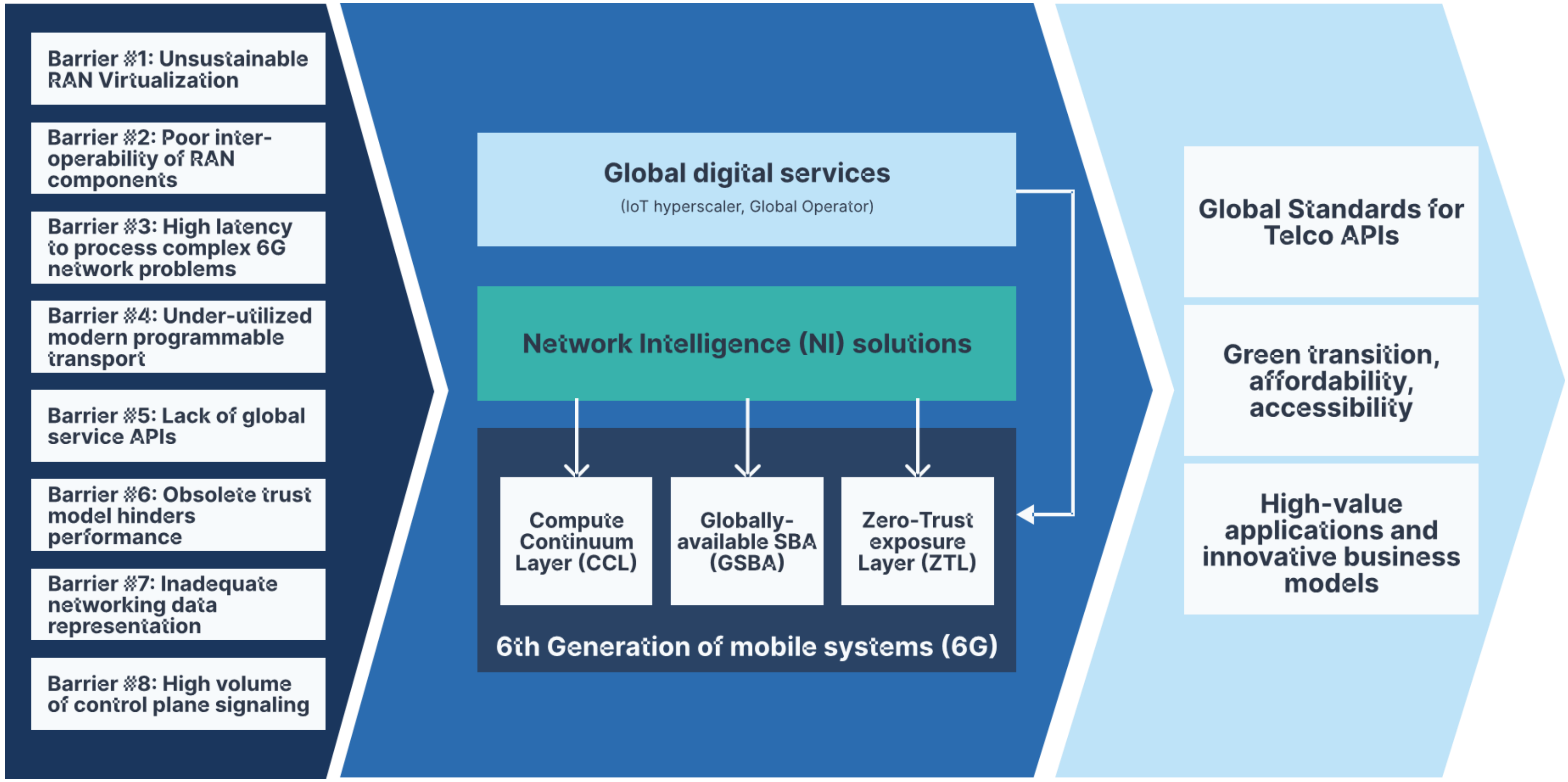
12 partners
9 countries
36 months
Kick-off on Jan. 2024
€3,9M (93% EU contr.)
466 PMs



Technical Objectives

ORIGAMI 6GSNS

- **Overall goal –** Developing a novel ***cross-plane architecture*** for 6G networks that supports original exposure and compute layers, which jointly
  - Remove practical barriers towards 6G,
  - Enable sustainable, energy-efficient, and affordable 6G systems, and
  - Promote new and disruptive 6G business models.



Barrier #1: Unsustainable RAN Virtualization
Barrier #2: Poor inter-operability of RAN components
Barrier #3: High latency to process complex 6G network problems
Barrier #4: Under-utilized modern programmable transport
Barrier #5: Lack of global service APIs
Barrier #6: Obsolete trust model hinders performance
Barrier #7: Inadequate networking data representation
Barrier #8: High volume of control plane signaling
Global digital services
(IoT hyperscaler, Global Operator)
Network Intelligence (NI) solutions
Compute Continuum Layer (CCL)
Globally-available SBA (GSBA)
Zero-Trust exposure Layer (ZTL)
6th Generation of mobile systems (6G)
Global Standards for Telco APIs
Green transition, affordability, accessibility
High-value applications and innovative business models

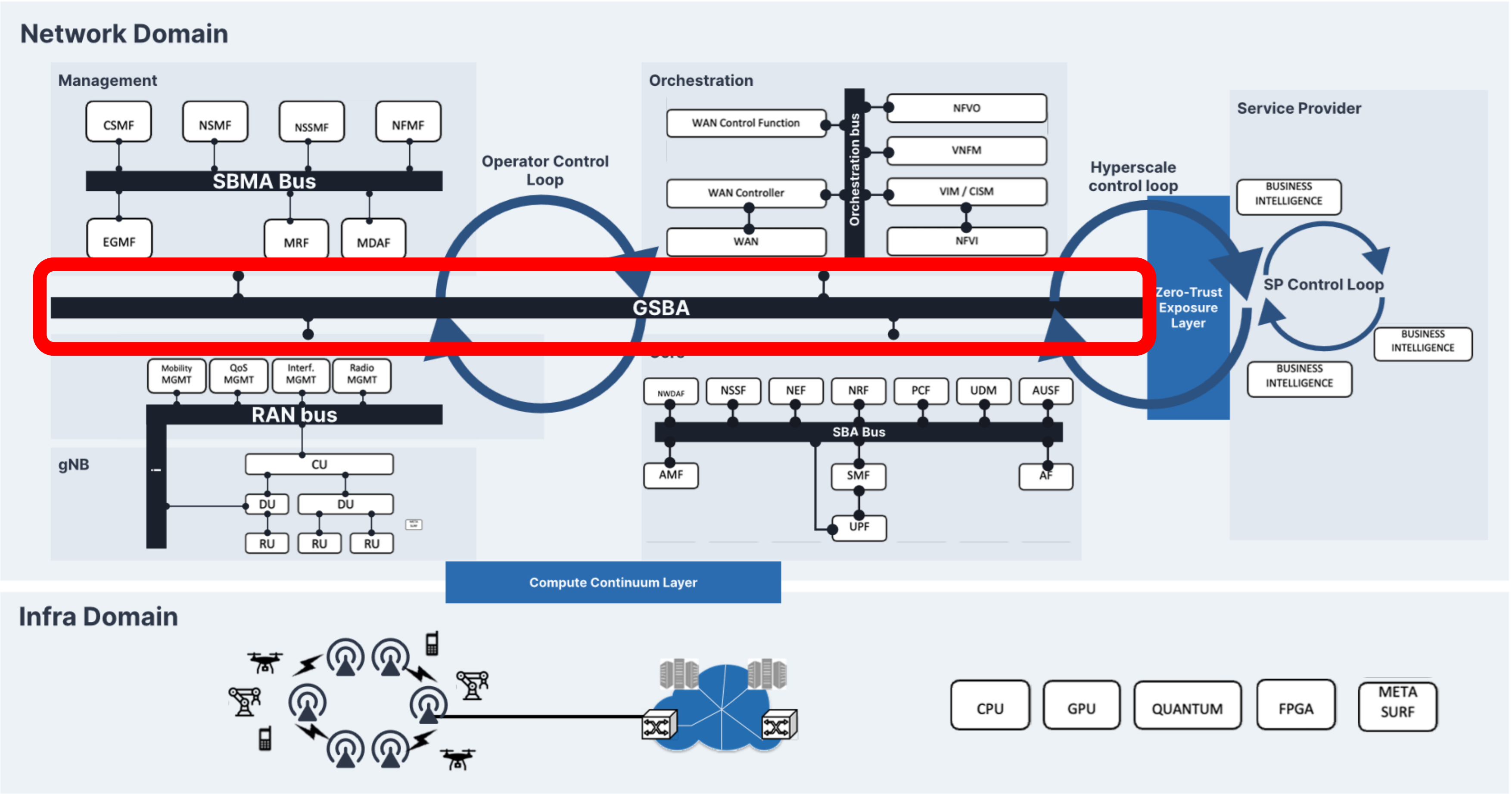
Co-funded by the European Union

Technical Objectives

- **Objective 1 –** Evolve the **architecture** of current mobile systems
  - **Sub-Objective 1.1 –** Address the challenges associated with deploying Network Intelligence (NI) solutions across multiple planes in next-generation mobile systems

- **Globally-available Service-Based Architecture (GSBA)**: inter-connects multiple planes or strata (NI plane, MANO plane, 3GPP/O-RAN planes, etc.).

- At least one experimental Proof-of-Concept (PoC)
  - Demonstrating the GSBA's ability to deploy NI that spans across multiple network functions or domains,
  - Achieving ultra-low-latency and at the same time having minimal overhead.



05/03/2024

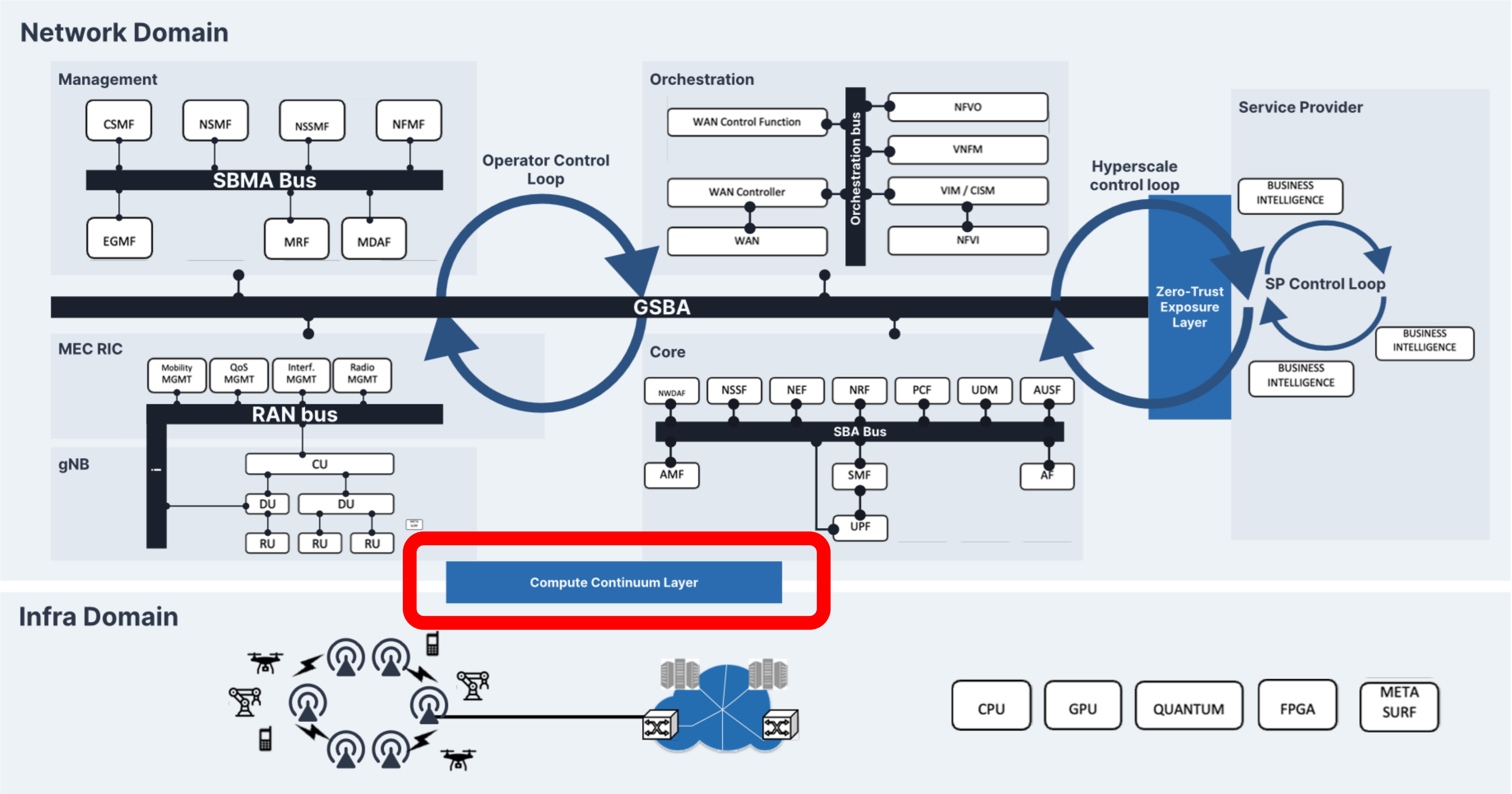
Co-funded by the European Union

Technical Objectives

- **Objective 1 –** Evolve the **architecture** of current mobile systems
  - **Sub-Objective 1.2 –** Compute Continuum Layer (CCL) that delivers enhanced processing power to ensure the dependable performance of increasingly complex operations in mobile systems while maintaining affordability and sustainability.

- CCL abstracts quantum computing, hardware accelerators and general-purpose processors:
  - Abstractions and architectural challenges
  - Network Intelligence (in Objective 3)

- At least one PoC improving performance in terms of:
  - cost-efficiency (network performance per dollar invested) and
  - energy-efficiency (network bits processed per energy joule invested).



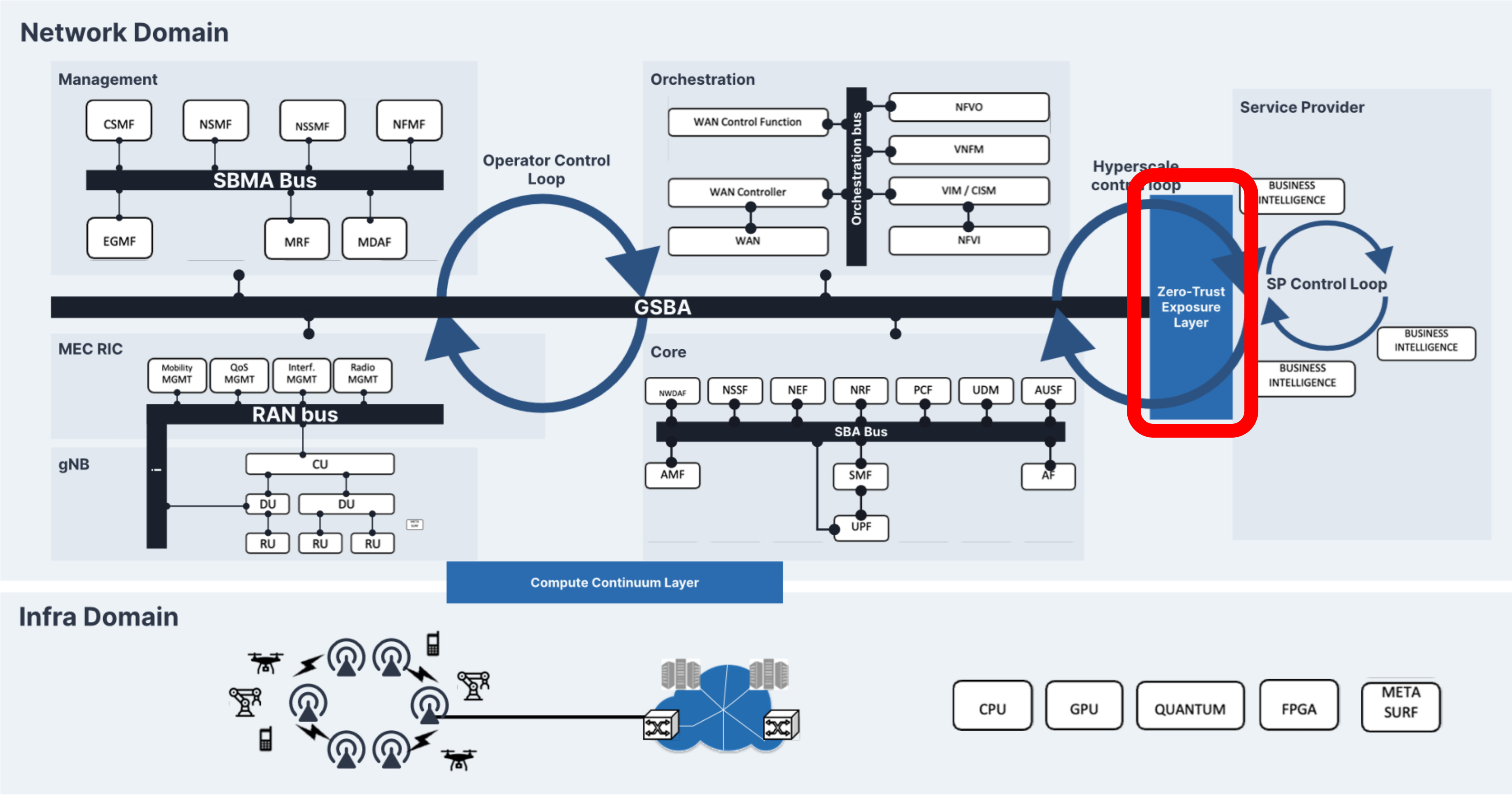
Co-funded by the European Union

Technical Objectives

- **Objective 1 –** Evolve the **architecture** of current mobile systems
    - **Sub-Objective 1.3 –** Zero-Trust exposure Layer (ZTL) that delivers cross-domain resource exploitation to ensure that third parties deliver optimal service to their end-users, in line with the latter’s diverse requirements.

- ZTL addresses challenges in terms of
    - Dynamics of business agreements and charging models (or lack thereof)
    - Open the path for new streams of revenues for resource owners
    - Dynamic interconnection of emerging players in the cellular ecosystem, such as IoT hyperscalers, with multiple resource providers within the ecosystem.

- At least one PoC demonstrating the global operator model (e.g., for IoT hyperscalers):
    - Tailored connectivity solutions for clients
    - Allowing users to connect to any cellular network

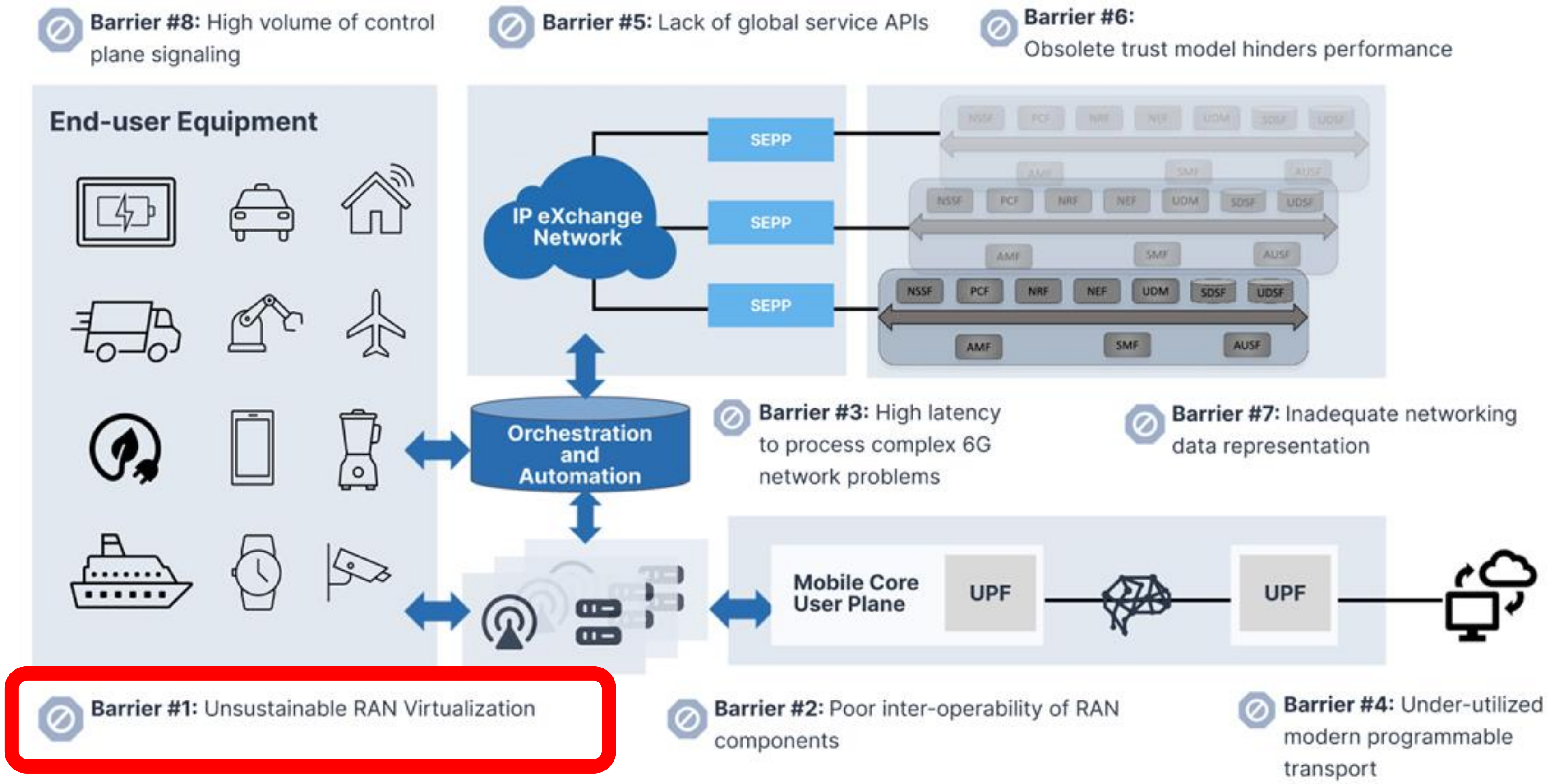


Technical Objectives

- **Objective 2 –** NI exploiting CCL and GSBA
  - **Sub-Objective 2.1 –** Infrastructure-aware virtualized RAN functions.

- Smart RAN virtualization NI that exploit the CCL (heterogeneous and shared computing resources) with reliability guarantees

- At least one PoC (same of O1.2) improving performance in terms of:
  - cost-efficiency (network performance per dollar invested) and
  - energy-efficiency (network bits processed per energy joule invested).



Barrier #8: High volume of control plane signaling

Barrier #5: Lack of global service APIs

Barrier #6: Obsolete trust model hinders performance

Barrier #3: High latency to process complex 6G network problems

Barrier #7: Inadequate networking data representation

Barrier #1: Unsustainable RAN Virtualization

Barrier #2: Poor inter-operability of RAN components

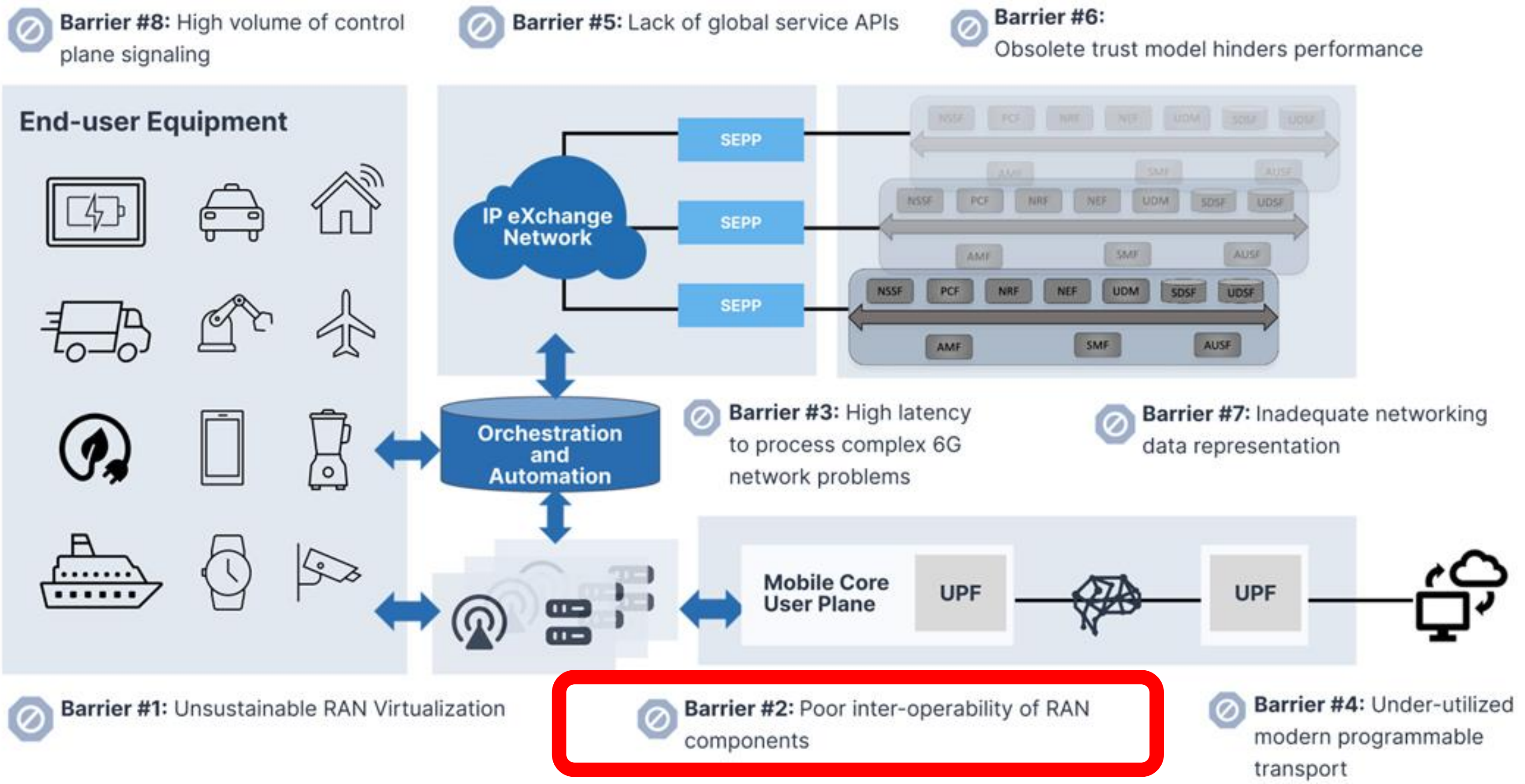
Barrier #4: Under-utilized modern programmable transport

Technical Objectives

- **Objective 2 –** NI exploiting CCL and GSBA
  - **Sub-Objective 2.2 –** Improve the inter-operability among open RAN components

- Build an inter-operable ecosystem that provides real-time control and effective collaboration between VNFs and RIC elements

- At least one PoC demonstrating seamless multi-vendor integration of open RAN components exploiting ORIGAMI's CCL



Barrier #8: High volume of control plane signaling

Barrier #5: Lack of global service APIs

Barrier #6: Obsolete trust model hinders performance

Barrier #3: High latency to process complex 6G network problems

Barrier #7: Inadequate networking data representation

Barrier #1: Unsustainable RAN Virtualization

Barrier #2: Poor inter-operability of RAN components

Barrier #4: Under-utilized modern programmable transport

05/03/2024

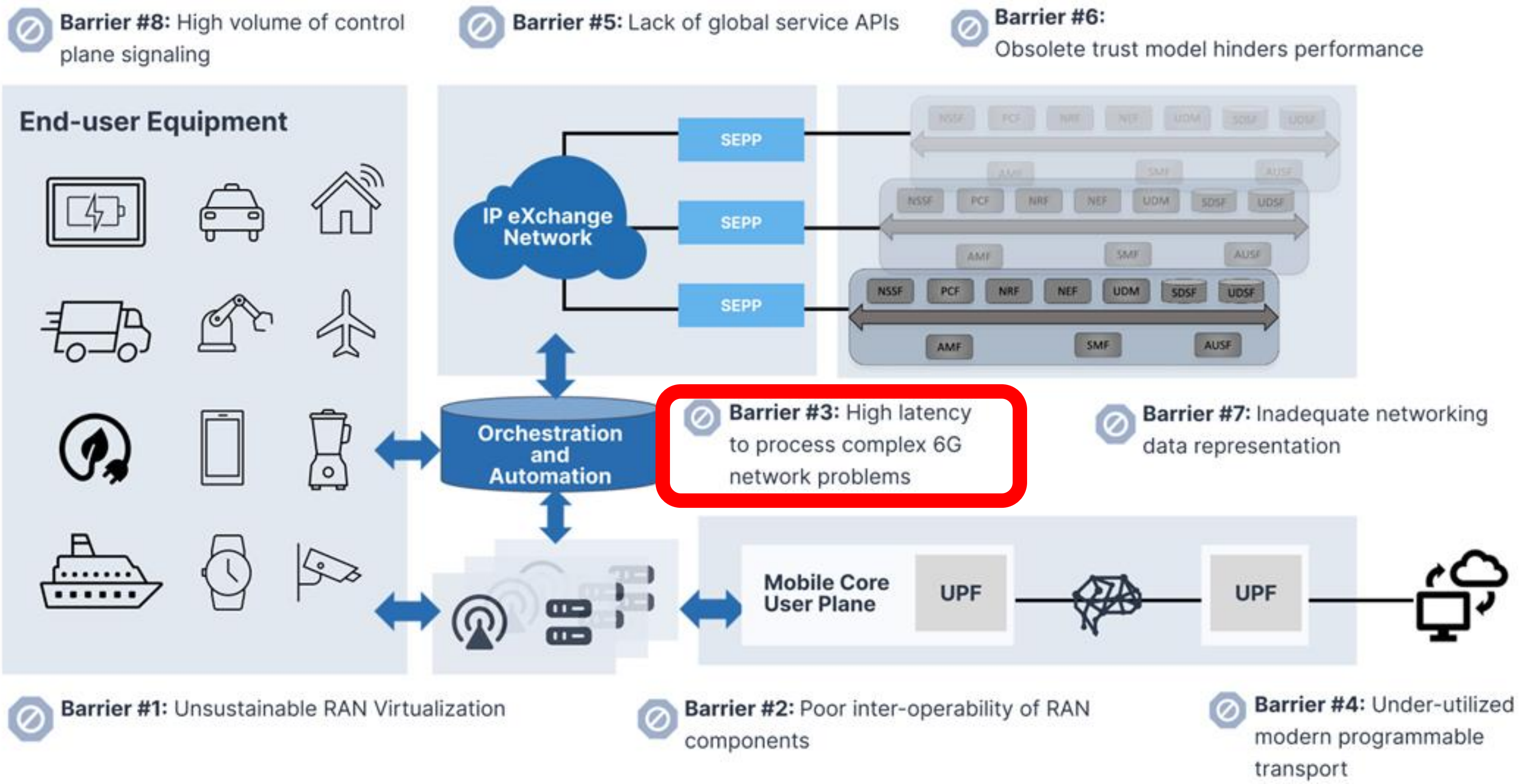
Co-funded by the European Union

Technical Objectives

- **Objective 2 –** NI exploiting CCL and GSBA
  - **Sub-Objective 2.3 –** Exploit novel compute paradigms to solve complex network processing problems with low latency

- optimize performance without compromising User Plane efficiency (?)
  - Asynchronous functions
  - binarized neural networks
  - scalable learning algorithms
  - quantum computing

- At least two PoCs demonstrating ultra-fast adaptability of UP functions:
  - RAN domain
  - Transport domain

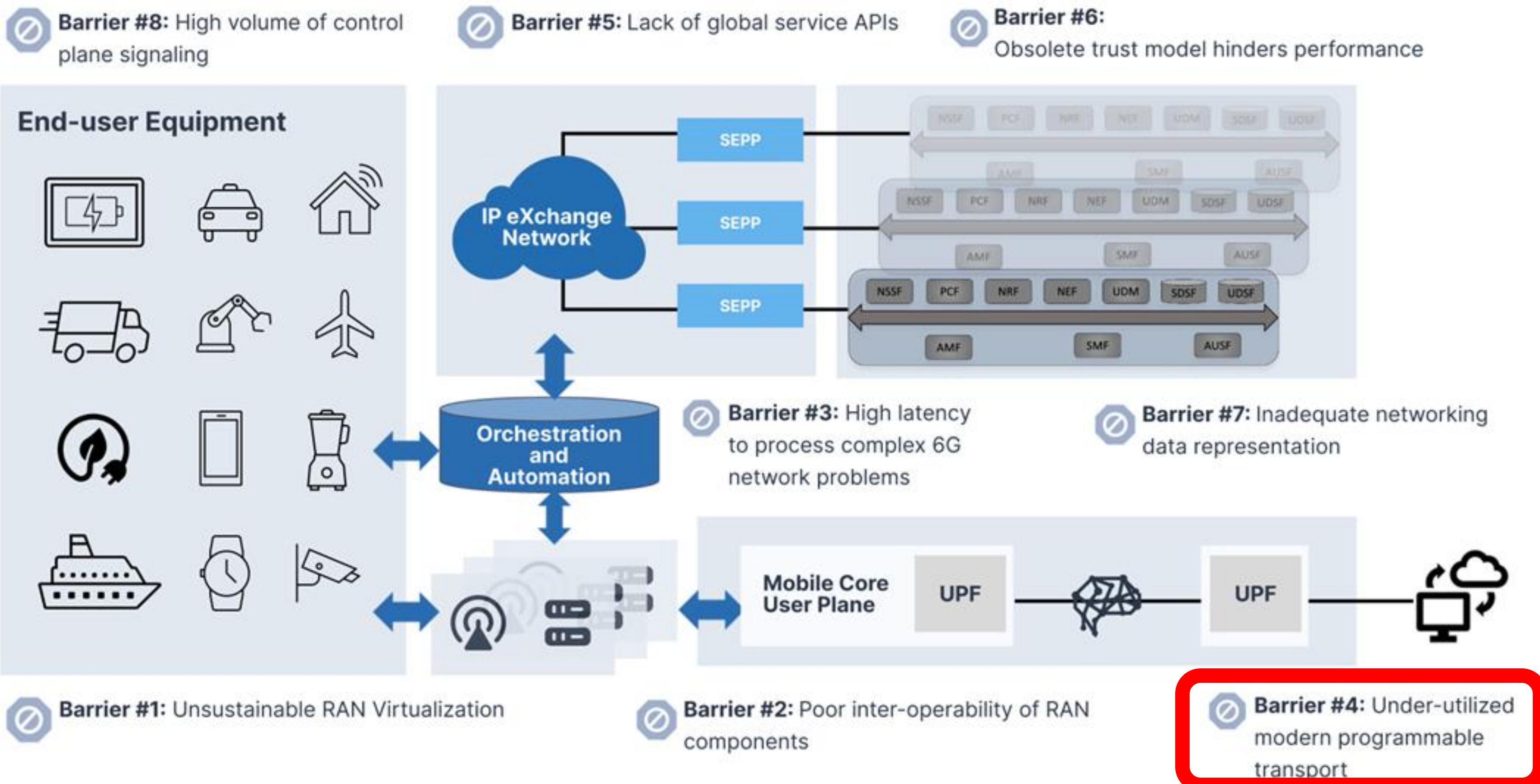

Barrier #8: High volume of control plane signaling
Barrier #5: Lack of global service APIs
Barrier #6: Obsolete trust model hinders performance
End-user Equipment
IP eXchange Network
SEPP
SEPP
SEPP
NSSF
PCF
NRF
NEF
UDM
SDSF
UDSF
AMF
SMF
AUSF
Orchestration and Automation
Barrier #3: High latency to process complex 6G network problems
Barrier #7: Inadequate networking data representation
Mobile Core User Plane
UPF
UPF
Barrier #1: Unsustainable RAN Virtualization
Barrier #2: Poor inter-operability of RAN components
Barrier #4: Under-utilized modern programmable transport

Technical Objectives

- **Objective 2 –** NI exploiting CCL and GSBA
  - **Sub-Objective 2.4 –** Design and build models and tools to exploit in-band computing capabilities in transport (backhaul, midhaul, fronthaul) and core domains

- NI for:
  - Ultra-low latency metadata collection
  - in-band data processing with privacy guarantees
  - real-time scheduling
  - decision-making policies under uncertainty for ML libraries

- At least one PoC demonstrating
  - sub-μs in-band AI/ML processing
  - with high performance
  - in constrained computing devices such as programmable switches and SmartNICs
  - preserving high network throughput (~100 Gbps)



Barrier #8: High volume of control plane signaling

Barrier #5: Lack of global service APIs

Barrier #6: Obsolete trust model hinders performance

Barrier #3: High latency to process complex 6G network problems

Barrier #7: Inadequate networking data representation

Barrier #1: Unsustainable RAN Virtualization

Barrier #2: Poor inter-operability of RAN components

Barrier #4: Under-utilized modern programmable transport

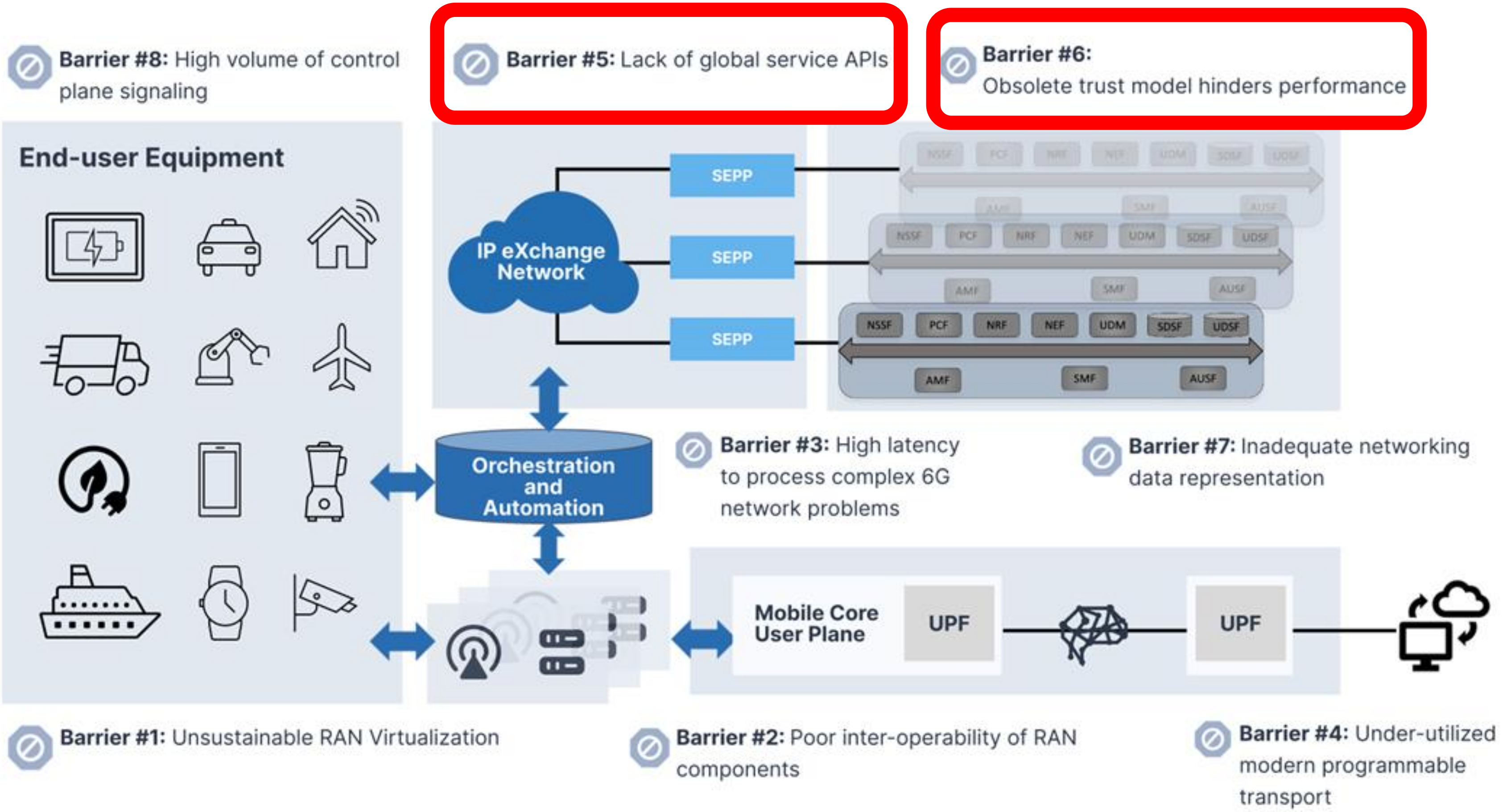
Co-funded by the European Union

Technical Objectives

- **Objective 3 –** NI and global APIs exploiting ZTL and GSBA
  - **Sub-Objective 3.1 –** Decouple authentication and billing from connectivity

- NI and APIs to enable entities interact **directly** and enable new commercial models (e.g. using DLT)
- New “Embassy” function for authentication” shall decouple end-user’s authentication from the infrastructure provider
- Facilitate dynamic business collaborations between MNOs and entities towards a truly global cellular provider model (enabling inter-MNO handovers)

- At least one PoC demonstrating
  - The global operator model is able to aggregate resources from >1 different providers able to perform inter-PLMN handovers
  - Latency reduction of frequent authentication procedures against home provider (from 1-2s to <300ms)



Barrier #8: High volume of control plane signaling

Barrier #5: Lack of global service APIs

Barrier #6: Obsolete trust model hinders performance

Barrier #3: High latency to process complex 6G network problems

Barrier #7: Inadequate networking data representation

Barrier #1: Unsustainable RAN Virtualization

Barrier #2: Poor inter-operability of RAN components

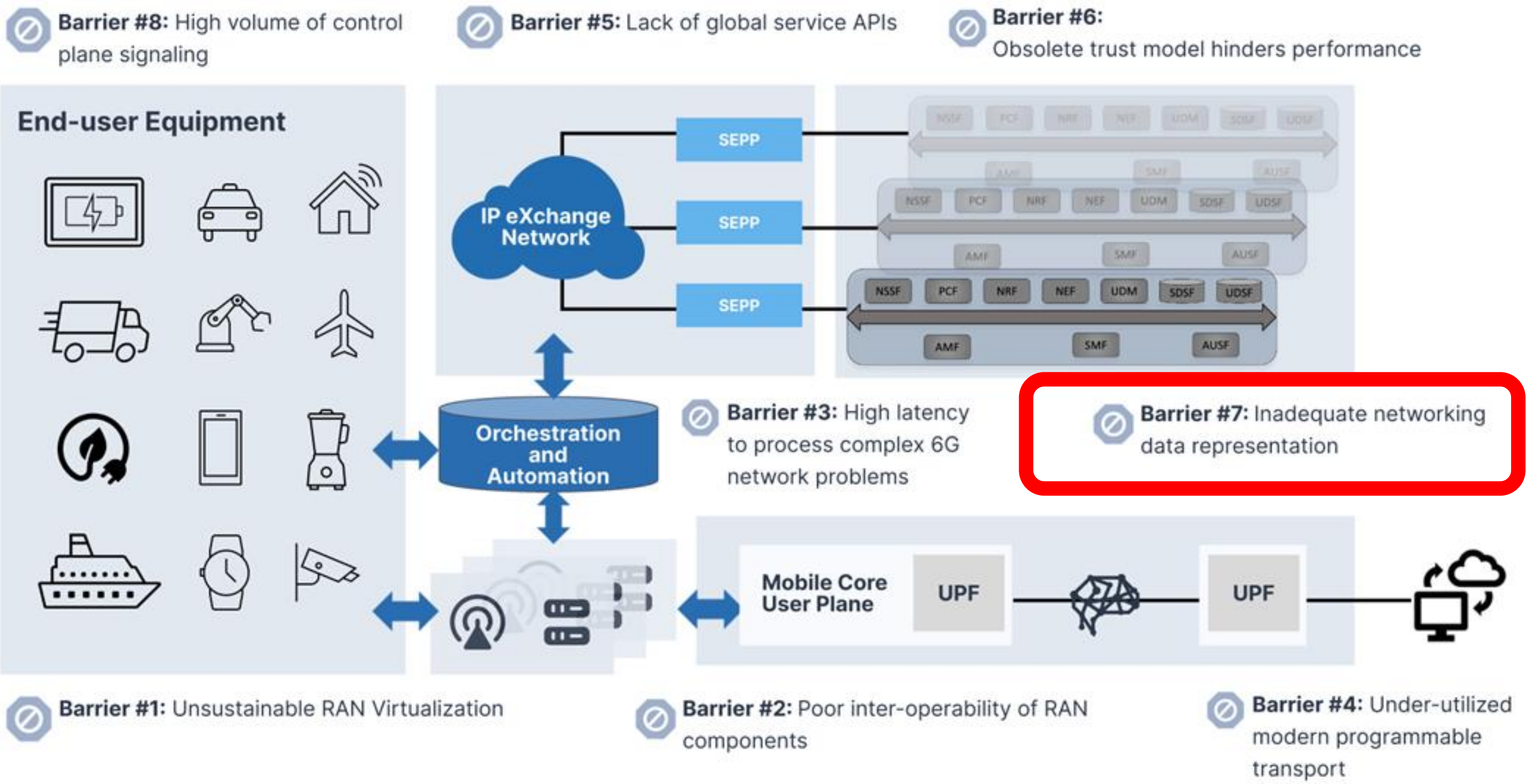
Barrier #4: Under-utilized modern programmable transport

Technical Objectives

- **Objective 3 –** NI and global APIs exploiting ZTL and GSBA
  - **Sub-Objective 3.2 –** Ensure reliable global operations

- Enable the design of orchestration policies of resources across multiple domains, while maintaining service quality guarantees
- Define network-specific data representation models that alleviate human troubleshooting time
- Integrate autonomous anomaly detection solutions that produce human readable explainable output

- At least one PoC demonstrating Explainable global anomaly detection pipelines



Barrier #8: High volume of control plane signaling

Barrier #5: Lack of global service APIs

Barrier #6: Obsolete trust model hinders performance

Barrier #3: High latency to process complex 6G network problems

Barrier #7: Inadequate networking data representation

Barrier #1: Unsustainable RAN Virtualization

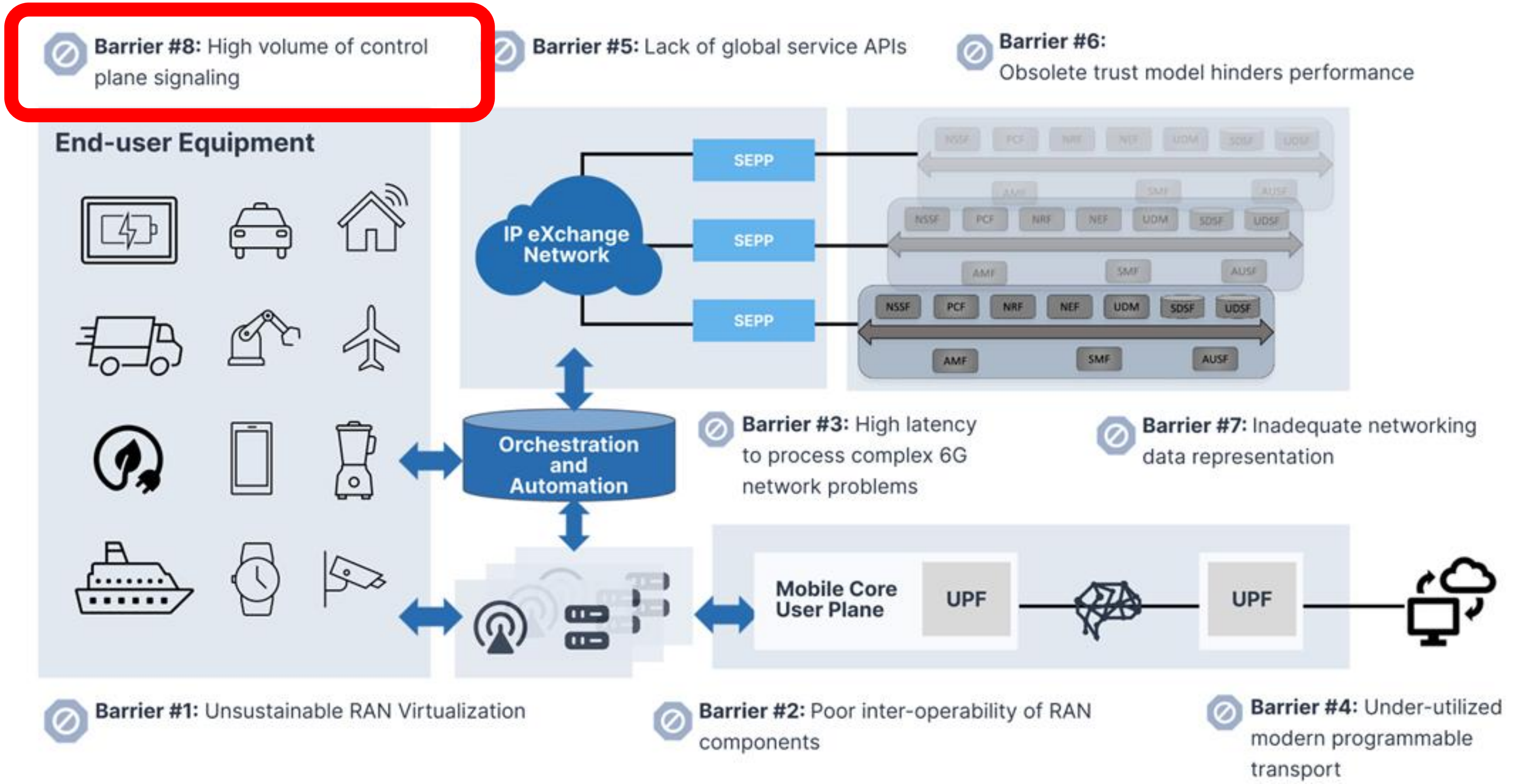
Barrier #2: Poor inter-operability of RAN components

Barrier #4: Under-utilized modern programmable transport

Technical Objectives

- **Objective 3 –** NI and global APIs exploiting ZTL and GSBA
  - **Sub-Objective 3.3 –** Realizing a global service mesh

- Exploring the benefits of an Service Communication Proxy (SCP)-enabled 6G network core operating as a full mesh
  - ML and graph theory to analyze and optimize traffic flows and network core signalling
- GSBA through a service mesh approach to provide a cloud-native implementation of the SCP functionality
- At least one PoC demonstrating the benefits of integrating the SCP functionality to reduce signalling traffic load



Barrier #8: High volume of control plane signaling

Barrier #5: Lack of global service APIs

Barrier #6: Obsolete trust model hinders performance

Barrier #3: High latency to process complex 6G network problems

Barrier #7: Inadequate networking data representation

Barrier #1: Unsustainable RAN Virtualization

Barrier #2: Poor inter-operability of RAN components

Barrier #4: Under-utilized modern programmable transport

Co-funded by the European Union

Performance Objectives

- **Objective 4 –** KPIs, demonstrations and PoCs
  - **Sub-Objective 4.1 –** KPIs

KPI	**Description**	**Network domain**	**Target**	**Evidence supporting the feasibility of the target and improvement over the baseline**
K1	Energy efficiency (bits-per-joule)	RAN	100% higher than today's vRANs	Today's RAN virtualization approaches rely upon energy-hungry hardware accelerators to process PHY operations. ORIGAMI's CCL will allow multiplexing a heterogeneous computing fabric comprised of energy-hungry yet powerful resources and energy-light yet capped resources opportunistically, using the former only when it is strictly required to preserve reliability.
K2	Cost efficiency (bps-per-$)	RAN	10x higher than today's vRANs	Today's RAN virtualization approaches rely upon expensive hardware accelerators to process PHY operations. ORIGAMI's CCL will allow sharing such expensive resource across multiple distributed units to amortize such expensive resource. Assuming a 5-μs/Km propagation delay, the area of radio units that can be connected to a single location is up to 400 Km^2 when using a 2D Manhattan tessellation model. Moreover, assuming PCIe v3.0+ bus and 100-GbE fronthaul interfaces, we could aggregate up to 3.8 GHz of radio spectrum to serve > 10 cells per hardware accelerator.
K3	Reliability (%)	RAN	99.999% probability of meeting deadlines	Attaining K1 and K2 target KPIs require relying upon inherently unreliably computing resources (energy-inexpensive processors or heavily shared accelerators). ORIGAMI's CCL will aid radio scheduling NI to meet K1 and K2 while guaranteeing that PHY processing deadlines are met with 99.9999% probability (5-nines reliability), which is the standard in the industry.
K4	In-band ML model inference latency (ms/µs)	RAN, Transport	Sub-ms (RAN), or sub-µs (transport)	Reducing the latency of inferences is currently hindered by the design of the employed ML models that are generic and agnostic to network hardware architecture and constraints. ORIGAMI's approach is to develop ultra-lightweight and scalable inference solutions, using online training-free algorithms that are natively designed and parametrized with the limitations of network hardware in mind, thus ensuring low-latency operation by-design.

Co-funded by
the European Union

Performance Objectives

- **Objective 4 –** KPIs, demonstrations and PoCs
  - **Sub-Objective 4.1 –** KPIs

KPI	Description	Network domain	Target	Evidence supporting the feasibility of the target and improvement over the baseline
K5	In-band ML model inference accuracy (%)	RAN and Transport	≥95%	The accuracy of in-band inferences is a key bottleneck for their exploitation. When confronted to non-trivial problems with decision space cardinality >10, current solutions struggle to achieve precision and recall above 0.85.23 ORIGAMI will remove this barrier and aim for 95% in large problems with cardinality >20 through (i) improved mappings of ML to hardware, which will allow accommodating larger models in the same equipment, and (ii) dynamic exploration at runtime of ML libraries configurations at per-flow granularity.
K6	In-band ML model inference throughput (Gbps)	Transport	100 Gbps	Current proposals for running ML models in programmable hardware have been only tested with low-throughput (e.g., order of Mbps) use cases, or, in the best scenario, with background traffic (which is not processed by the ML model but just forwarded) at 40-Gbps. Solutions to perform line-rate inference on order-of-Gbps traffic remain to be demonstrated. ORIGAMI aiming to demonstrate up to 100 Gbps of ML-processed traffic in the user plane, thanks to ML models that are natively designed to operate in production-grade programmable hardware (capable of achieving 100 Gbps throughput in legacy forwarding tasks) with near-zero overhead.
K7	Network CAPEX ($)	Core	50% reduction	The deployment of decentralized virtualized policy management (vPCF) and cloud-based user credentials provisioning (cUDM) will result in lower costs while enabling the usage of other infrastructure from different providers.
K8	Network energy consumption (KWh)	Core	35% less energy consumption	Seminal studies have repeatedly shown that substantial energy gains of 35% or more can be achieved by controlling RAN configurations. However, practical solutions adopted by operators today fall very short of that target as they adopt static policies to, e.g., switch off a fraction of manually identified base stations overnight. Thanks to the effective abstraction provided by the CCL, ORIGAMI will finally achieve a dynamic, automated, AI-driven optimization of the usage of RAN resources. In addition, the fully distributed core infrastructure enabled by GSBA and ZTL will allow a more efficient and energy-prudent allocation of network resources, based on traffic requirements and exploiting the available resources from different infrastructure providers.
K9	Control plane latency (ms)	Core	50% lower latency than current procedures	The usage of decentralized signaling and selecting the most optimal RAN and network resources from different providers will result in more efficient transport that will reduce latency.

Co-funded by the European Union

Performance Objectives

- **Objective 4 –** KPIs, demonstrations and PoCs
  - **Sub-Objective 4.2 –** Evaluations (at least one PoC per barrier)

Barrier addressed (§ 1.1.5)	**Description of the solution to be demonstrated (§ 1.2.3)**	**ORIGAMI innovation exploited (§ 1.2.2)**	**Evaluation site (S) or Dataset (D) (§ 1.2.4)**	**Related KPIs (Table 1)**
#1* *Unsustainable RAN virtualization	Infrastructure awareness, enabled by ORIGAMI's CCL for 6G VNFs in the RAN enabling hardware accelerator pooling and opportunistic use of energy-efficient (yet slower) software processing opportunistically, while providing reliability guarantees.	Compute Continuum Layer (CCL)	(S) Distributed Infrastructure for Open RAN Experimentation	K1, K2, K3
#2* *Poor inter-operability of RAN components	Effective deployment of RAN bus and interoperability of RAN Intelligent Controllers (RIC).	Compute Continuum Layer (CCL), Global Service-based Architecture (GSBA)	(S) Distributed Infrastructure for Open RAN Experimentation	K1, K2, K3
#3* *High latency to process complex 6G network problems	Demonstrate solutions based on first-order scalable optimization algorithms; deep-learning-assisted optimization for mixed-integer problems; and quantum annealing and quantum approximate optimization for NP-hard problems.	Compute Continuum Layer (CCL)	(S) Distributed Infrastructure for Open RAN Experimentation; (S) MADQuantum-CM	K1, K4, K5
#4* *Under-utilized modern programmable transport	By enabling distributed, streamlined access to transport domain computing, hierarchical/federated models can operate in heterogeneous programmable user planes, utilizing efficient model design and hardware mappings. This approach improves speed, accuracy, and scalability compared to current solutions limited to single-device operation and not designed for user-plane functionality.	Compute Continuum Layer (CCL), Global Service-based Architecture (GSBA)	(S) Heterogeneous programmable network testbed	K4, K5, K6

Co-funded by the European Union

Performance Objectives

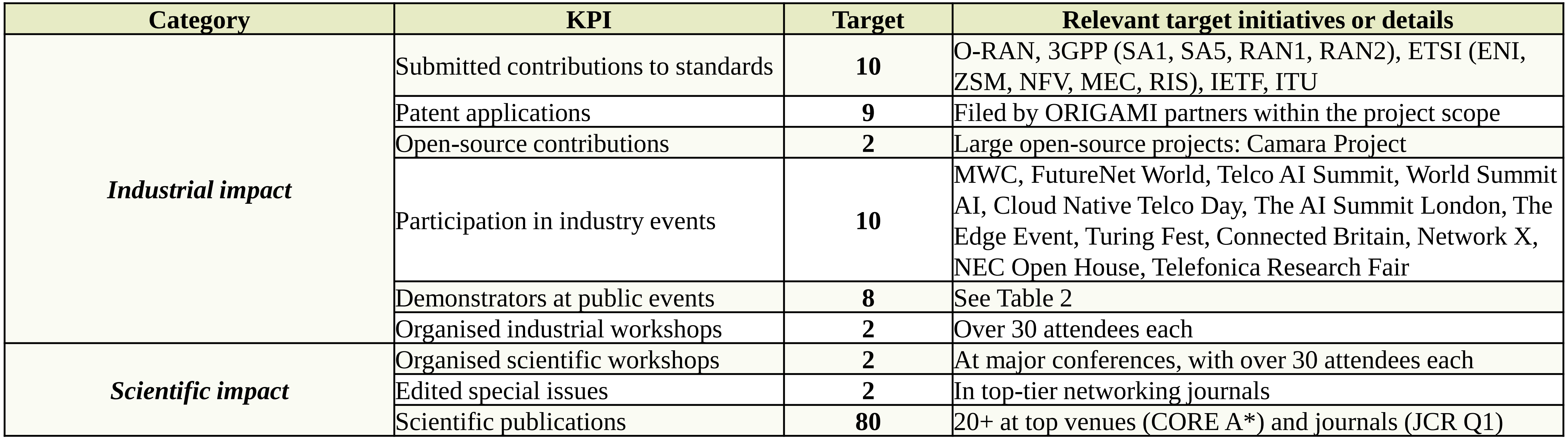
- **Objective 4 –** KPIs, demonstrations and PoCs
  - **Sub-Objective 4.2 –** Evaluations (at least one PoC per barrier)

Barrier addressed (§ 1.1.5)	**Description of the solution to be demonstrated (§ 1.2.3)**	**ORIGAMI innovation exploited (§ 1.2.2)**	**Evaluation site (S) or Dataset (D) (§ 1.2.4)**	**Related KPIs (Table 1)**
#5 ***Lack of global service APIs***	Global operator model (e.g., for IoT hyperscaler) enabling inter-MNO seamless mobility and handovers beyond the single provider boundaries to improve coverage and service quality for specific applications. The application-aware network for 6G integrates an Application Function to provide network external interface for service requests.	Zero-Trust exposure Layer (ZTL), virtualize PCF (vPCF) and cloud UDM (cUDM) to enable global credentials and cross service provider credentials	(S) 6G-SANDBOX	K7, K9
#6 ***Obsolete trust model hinders performance***	Decentralized identity model and dynamic charging solutions. New authentication function to enable on-demand authentication and authorization.	Zero-Trust exposure Layer (ZTL), cUDM to provide decentralized metaoperator authentication and charging models	(S) 6G-SANDBOX	K9
#7 ***Inadequate networking data representation***	Intelligent anomaly detection and Service-Level Agreement (SLA) monitoring in a global deployment model to enable proactive network troubleshooting across multiple domains. Application function to track and estimate available capacity to avoid congestion related problem.	Compute Continuum Layer (CCL), Global Service-based Architecture (GSBA), Zero-Trust exposure Layer (ZTL), vPCF and cUDM	(D) Large-scale monitoring datasets from commercial operators; (S) BatteryLab	K10, K11
#8 ***High control-plane signaling overhead***	Reshape network core functionality towards a cloud-native structure.	Global Service-based Architecture (GSBA)	(S) Service Mesh Platform; (D) Large-scale monitoring datasets from commercial operators	K12

Co-funded by the European Union

Productivity Objectives

- **Objective 5 –** Productivity



Category	KPI	Target	Relevant target initiatives or details
Industrial impact	Submitted contributions to standards	**10**	O-RAN, 3GPP (SA1, SA5, RAN1, RAN2), ETSI (ENI, ZSM, NFV, MEC, RIS), IETF, ITU
	Patent applications	**9**	Filed by ORIGAMI partners within the project scope
	Open-source contributions	**2**	Large open-source projects: Camara Project
	Participation in industry events	**10**	MWC, FutureNet World, Telco AI Summit, World Summit AI, Cloud Native Telco Day, The AI Summit London, The Edge Event, Turing Fest, Connected Britain, Network X, NEC Open House, Telefonica Research Fair
	Demonstrators at public events	**8**	See Table 2
	Organised industrial workshops	**2**	Over 30 attendees each
Scientific impact	Organised scientific workshops	**2**	At major conferences, with over 30 attendees each
	Edited special issues	**2**	In top-tier networking journals
	Scientific publications	**80**	20+ at top venues (CORE A*) and journals (JCR Q1)

Thanks!

The 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 programme under Grant Agreement No 101139270
www.smart-networks.europa.eu

Co-funded by
the European Union