An Energy-Aware Multi-Domain Orchestration Architecture for Sustainable 6G Networks*

Berk Buzcu[a,d,*], Laura Finarelli[a,d], Engin Zeydan[b], Jorge Baranda[b], Nikolaos Tzanis[c], Stamatiou Nikolaos[c], Nikolaos Kanakaris[c] and Gianluca Rizzo[a]

[a]*Connected Experience Lab, HES-SO Valais/Wallis, , Sierre, Switzerland*
[b]*Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), , Castelldefels, Spain*
[c]*Department of Research, Technology & Development, Independent Power Transmission Operator (IPTO), , Athens, Greece*
[d]*Telecommunication Networks Group, Technical University of Berlin, , Berlin, Germany*

ARTICLE INFO

Keywords:
6G networks
multi-domain orchestration
energy-aware network management
AI-native automation

ABSTRACT

Next-generation 6G infrastructures are expected to behave as intelligent, sustainable, and self-optimizing systems that bring together heterogeneous communication and energy domains. This paper introduces an energy-aware, multi-domain orchestration framework for 6G networks that rests on a hierarchical control model composed of an Inter-Domain Management and Orchestration (IDMO) layer, per-domain Management and Orchestration (DMO) entities, and Infrastructure Domain Managers (IDMs), all coupled through a Service-Based Management Architecture (SBMA). The framework further embeds Virtual Power Plants (VPPs) and an inter-domain Energy Management System (EMS) that interface, both physically and logically, with the network fabric while preserving the operational independence of local grids. The EMS processes standardized, cross-domain energy data and forecasts into locality-aware reports for the IDMO, which drives cross-domain service decomposition, placement, and reconfiguration. The design is realized as an executable prototype in which the IDMO ranks candidate domains with a Multi-Criteria Decision Analysis (MCDA) score that weighs service capability, free capacity, latency slack, and an EMS-supplied greenness signal, and whose VPP–TSO interface solves a power flow on a CIGRE benchmark grid driven by hourly ENTSO-E data for two real national grids. On a two-domain proof of concept it reduces carbon by 31.2% (from 803 to 553 gCO_2eq/h) at 100% slice acceptance and within the slices' latency SLAs, and sustains 95% acceptance under load where a performance-first orchestrator retains only 72%. On the NOBEL-EU and GÉANT backbones (200 requests, 40 ms SLA) it reduces carbon by 19.9% and 16.0% at 95.5% and 90.0% acceptance.

1. Introduction

The emergence of Sixth Generation (6G) networks opens opportunities for intelligent, energy-efficient, and sustainable connectivity across highly heterogeneous infrastructures (Alhashimi, Hindia, Dimyati, Hanafi, Safie, Qamar, Azrin and Nguyen, 2023). Whereas earlier generations relied on single-domain management, 6G systems are expected to fuse terrestrial and non-terrestrial segments (Wang, Zhang, Yang and Quek, 2025), cloud-edge resources (SNS JU 6G Architecture Working Group, 2024), and vertical-specific infrastructures into a single, service-oriented fabric (Trichias, Kaloxylos and Willcock, 2024). This shift calls for orchestration frameworks that can master cross-domain complexity while treating energy awareness and sustainability as first-class design goals rather than afterthoughts.

Network management and orchestration research has matured considerably since the 5G era, moving beyond the original notions of Network Function Virtualization (NFV) and Software-Defined Networking (SDN) (Zeydan, Blanco, Batewela, Ranaweera and Liyanage, 2025). Those technologies delivered flexible, programmable infrastructures, but their orchestration remained largely confined to a single administrative domain. In contrast, 6G aims to unify multiple heterogeneous domains, spanning distinct technologies, ownership models, and geographic footprints, beneath a common, Artificial Intelligence (AI)-driven orchestration layer capable of dynamic service decomposition and cross-domain optimization (Alhashimi et al., 2023). This entails distributing Radio Access Network (RAN) and mobile-core functions across different points of presence (PoPs), co-located with vertical applications to satisfy stringent latency and throughput requirements, while also folding in non-terrestrial capabilities such as satellite access for ubiquitous coverage.

Recent "green orchestration" work has begun to weave carbon awareness into placement decisions, but energy intelligence, such as Virtual Power Plant (VPP) telemetry, remains largely unexplored inside inter-domain orchestration loops. Our design responds with a hierarchical control

*This work is supported by the UNITY-6G project, co-funded by the European Union's Horizon Europe Smart Networks and Services Joint Undertaking (SNS JU) research and innovation programme under Grant Agreement No 101192650, and by the Swiss State Secretariat for Education, Research and Innovation (SERI). During the preparation of this manuscript, the authors used ChatGPT to improve clarity, grammar, and sentence structure.

*Corresponding author
berk.buzcu@hevs.ch (B. Buzcu); laura.finarelli@hevs.ch (L. Finarelli); engin.zeydan@cttc.es (E. Zeydan); jorge.baranda@cttc.es (J. Baranda); nikostza@gmail.com (N. Tzanis); n.stamatiou@admie.gr (S. Nikolaos); n.kanakaris@admie.gr (N. Kanakaris); gianluca.rizzo@hevs.ch (G. Rizzo)
ORCID(s):

model: a global Inter-Domain Management and Orchestration (IDMO) layer coordinates several Domain Management and Orchestration (DMO) entities, each governing its own Infrastructure Domain Managers (IDMs). Within this hierarchy, our framework embeds energy-aware, multi-criteria decision logic, making sustainability a native ingredient of both service instantiation and runtime adaptation. Unlike prior approaches that treat energy efficiency as a post-deployment optimization step, our architecture exploits real-time energy data from distributed VPP systems to enable sustainability-aware placement, scaling, and migration across federated domains. This design targets carbon-conscious operation of federated 6G infrastructures. In particular, a single monolithic orchestrator cannot cope with (i) the heterogeneity and scale of administrative/technology domains, (ii) policy constraints that demand local autonomy, and (iii) end-to-end energy objectives that require global coordination of local trade-offs. Accordingly, our architecture is conceived around virtual entities and actors, rather than plain network functions.

Specifically, the IDMO acts as a global intelligence plane that orchestrates multiple administrative and technological domains; functioning as a master orchestrator, it coordinates end-to-end (E2E) network composition by negotiating deployment policies and combining inputs from individual domain managers (addressing (i) and (iii)). Each domain, in turn, is governed by its own DMO, which handles orchestration and management within its boundaries, resource allocation, lifecycle management, policy enforcement, and operational control of infrastructure and deployed network services (addressing (ii)). This paper therefore asks whether measured, per-domain grid telemetry can act as a placement criterion in its own right within a multi-domain orchestrator, and at what cost in end-to-end latency and other metrics.

Accordingly, we present (i) the chosen distributed future-networks architecture, built around the IDMO–DMO Blanco, Via, Vaca-Rubio, Zeydan, Murillo, Kedar, Antonio, Antonopoulos, Haxhibeqiri, Maglogiannis, Naudts, Hoebeke, de Cola, Giordano, Fontanesi, Spantideas, Koumaras, Neri, Vesco, Amatetti, Quijada, Dryjański, Kanakaris, Pamminger, Pagin and Ivanovic (2025) structure (See Section 3), together with a data model that exposes per-domain energy telemetry through an inter-domain Energy Management System (EMS) and VPPs; (ii) an Multi-Criteria Decision Analysis (MCDA) placement formulation that combines capacity, latency, and carbon intensity (Section 4); and (iii) a validation strategy for the resulting multi-domain, energy-aware orchestration framework, together with its implementation and evaluation on a two-domain prototype model and two SNDlib topologies (Section 5).

2. Related Work

The architecture proposed in this paper sits at the intersection of three research lines that have so far evolved largely independently: multi-domain orchestration of virtualized networks, the algorithmic placement of service function chains across federated infrastructures, and grid-side energy intelligence produced by energy management systems and virtual power plants. We review each in turn and then position our contribution with respect to the gap that separates them.

Multi-domain orchestration in future wireless networks
Network management and orchestration matured during the 5G era around NFV and SDN, but the resulting orchestration scope remained, for the most part, bound to a single administrative domain. The extension to federated, multi-domain settings has been addressed both architecturally and in terms of standardization. Afolabi, Taleb, Samdanis, Ksentini and Flinck (2018) survey the principles and enabling technologies of network slicing and softwarization, while Taleb, Afolabi, Samdanis and Yousaf (2019) propose a multi-domain slicing architecture with an explicit cross-domain coordination stratum operating over federated resources. Prototype multi-domain orchestrators have also been built: ESCAPE (Csoma, Sonkoly, Csikor, Németh, Gulyás, Tavernier and Sahhaf, 2014; Sonkoly, Czentye, Szabo, Jocha, Elek, Sahhaf, Tavernier and Risso, 2015) realises the orchestration layer of the UNIFY architecture and can discover, abstract, and manage heterogeneous technological and administrative domains behind a unified resource view, which is precisely the role of IDMO in our design.

On the standardization side, the ETSI Zero-touch Service Management framework (Benzaıd and Taleb, 2020) defines intent-based, closed-loop automation across management domains; the O-RAN Alliance (Polese, Bonati, D'Oro, Basagni and Melodia, 2023) specifies the near-real-time and non-real-time RAN Intelligent Controllers that expose the RAN to AI-based optimization; and the 3GPP Service-Based Management Architecture (Nokia Bell Labs, 2020) provides the modular service exposure and data-driven control plane on which our Service-Based Management Architecture (SBMA) bus is modeled. At project level, Hexa-X (Uusitalo, Farhadi, Boldi, Ericson, Fettweis, Khorsandi, Pavón, Mueck, Nimr, Pärssinen, Richerzhagen, Rugeland, Sabella and Wymeersch, 2025), ADROIT6G (Christophorou, Ioannou, Vassiliou, Christofi, Vardakas, Seder, Chiasserini, Iordache, Issaid, Markopoulos, Franzese, Järvet and Verikoukis, 2024) and UNITY-6G (Blanco et al., 2025) converge on the need for hierarchical, AI-native orchestration spanning semantic monitoring, intent translation and digital-twin feedback. What none of these frameworks provides is a normative mechanism for treating energy sustainability as a first-class decision parameter. In other words, energy appears as an observability metric rather than as an input to domain selection.

Placement and embedding across federated domains
The decision the IDMO must actually make, i.e., mapping a logical service graph onto a substrate of PoPs and transport links under capability, capacity and latency constraints, is an instance of virtual network embedding, and of the closely

related Virtualized Network Function (VNF) forwarding-graph embedding problem, for which Schardong, Nunes and Schaeffer-Filho (2021) give a systematic taxonomy. Exact formulations are typically integer linear programs because these do not scale, the literature relies on heuristics, meta-heuristics and, increasingly, learning-based policies. Santos, Bezerra, Rocha, Ferreira, Moreira, Gonçalves, Marquezini, Recse, Mehta, Kelner, Sadok and Endo (2022) review the subset of this work that targets distributed scenarios, where no single entity holds a complete view of the substrate corresponding to the operative assumption in a federation of independently administered 6G domains. Game-theoretic formulations are also explored by SPEED (Battisti, Delicato and Muchaluat-Saade, 2026), which casts multi-domain service function chain (SFC) placement as a singleton congestion game and computes placements without centralized, complete infrastructure knowledge, which is why we adopt it as the distributed matching component of our implementation.

Energy has entered this literature, but narrowly. Lin, He, Luo and Zukerman (2023) embed service function chains in NFV networks under an energy-aware objective, and Sun, Li, Yu, Vasilakos, Du and Guizani (2019) combine energy and traffic awareness in multi-domain SFC orchestration. In both cases energy is modeled as a static power cost of activating nodes or links; the energy state of the underlying electrical system, which varies hour to hour, differs across countries, and is what actually determines whether a placement is low-carbon, does not enter the model.

Metaheuristic search has also been applied to this problem: Krishnamohan and Harvey (2026) evolve neural encodings of the chain-composition, function-embedding and virtual-link-embedding sub-problems jointly, evaluating candidates through an emulated multi-access edge deployment in order to minimise energy. Their objective is energy at the edge rather than carbon across administrative domains, and the grid state does not enter it. We adopt the genetic-algorithm idea as a search-based baseline in Section 5.4.

Energy management systems, VPPs, and grid-side intelligence On the power-system side, the virtual power plant is the established abstraction for aggregating distributed energy resources behind a single dispatchable interface (Pudjianto, Ramsay and Strbac, 2007).A substantial body of work addresses how a VPP reports upstream and how its flexibility is exposed at the TSO/DSO boundary: Riaz and Mancarella (2019) characterize the feasible operating and flexibility regions of VPPs at that interface, and Givisiez, Petrou and Ochoa (2020) review TSO-DSO coordination models more broadly. The communication layer that carries these exchanges has itself been standardized and evaluated: Zajc, Kolenc and Suljanovic (2019) describe the reference VPP communication system architecture, including the periodic upstream reporting of aggregated pool measurements that our VPP client reproduces, and Kolenc, Nemček, Gutschi, Suljanović and Zajc (2017) evaluate its performance when providing ancillary services. For grid-side evaluation, open reference models are available: the CIGRE benchmark systems for the integration of renewable and distributed energy resources (CIGRE Task Force C6.04, 2014) provide standardized high-, medium-, and low-voltage test networks, and `pandapower` (Thurner, Scheidler, Schäfer, Menke, Dollichon, Meier, Meinecke and Braun, 2018) provides an open-source solver for the corresponding power flows. This literature is mature, but it addresses the grid, not the network. The compute load of a telecommunication domain appears, when it appears at all, as an exogenous and inflexible demand; there is no interface through which an orchestrator could consume a normalized, comparable statement of "how green this domain is right now."

Energy- and carbon-aware networks and computing Two communities have approached the remaining gap from opposite sides. In mobile networks, energy efficiency has been pursued chiefly at the radio layer: López-Pérez, De Domenico, Piovesan, Xinli, Bao, Qitao and Debbah (2022) survey 5G RAN energy efficiency across massive MIMO, lean carrier design and sleep modes, building on a longer line of base-station sleep-mode research (Wu, Zhang, Zukerman and Yung, 2015); more recently, O-RAN's advanced sleep modes have been exploited for real-time radio unit control under QoS guarantees (Lozano, Ayala-Romero, Garcia-Saavedra and Costa-Perez, 2025). These techniques reduce the energy drawn by a given deployment; they do not choose *where* to deploy.

In cloud and edge computing, conversely, carbon-aware scheduling exploits the spatial and temporal variability of grid carbon intensity to shift workloads in time or across regions (Acun, Lee, Kazhamiaka, Maeng, Gupta, Chakkaravarthy, Brooks and Wu, 2023; Hanafy, Liang, Bashir, Irwin and Shenoy, 2023). Sukprasert, Souza, Bashir, Irwin and Shenoy (2024) provide a useful corrective: using carbon-intensity traces from a large number of regions, they show that the achievable gains from spatio-temporal shifting are bounded, that simple policies capture most of them, and that latency and regulatory constraints erode them further — an argument for coupling carbon signals to an orchestrator that already reasons about latency SLAs, rather than to a scheduler that does not. Within networking, renewable-aware VNF placement (Liang, Huang, Qiu, Liu, Aziz and Chen, 2024) and predictive solar-driven load management for data centers (Floridia, Laganà, Mastroianni, Meo and Renga, 2020) demonstrate the promise of sustainability-driven placement, but operate at the resource-management level of a single domain and consume renewable-availability signals that are synthetic or coarsely national rather than measured at the site.

Table 1 summarizes the comparison. Three gaps persist across the reviewed work. First, energy awareness and multi-domain orchestration are addressed separately: the orchestration frameworks that span domains do not carry an energy objective, and the placement algorithms that carry an energy objective do not span domains. Second, where renewable

Table 1
Positioning with respect to representative prior work. ✓: supported; ✗: not supported; (✓): partial.

Work	Multi-domain	Energy in the objective	Latency/QoS constraint	Real grid data in the loop	Implemented testbed
Multi-domain slicing (Taleb et al., 2019)	✓	✗	✓	✗	(✓)
ESCAPE (Sonkoly et al., 2015)	✓	✗	✓	✗	✓
ETSI ZSM (Benzaıd and Taleb, 2020)	✓	✗	✓	✗	✗
SPEED (Battisti et al., 2026)	✓	✗	✓	✗	✓
Energy-aware SFC (Lin et al., 2023)	✗	✓	✓	✗	✗
Multi-domain energy-aware SFC (Sun et al., 2019)	✓	✓	✓	✗	✗
Renewable-aware VNF (Liang et al., 2024)	✗	✓	✓	✗	✗
O-RAN sleep modes (Lozano et al., 2025)	✗	✓	✓	✗	✓
Carbon-aware shifting (Sukprasert et al., 2024)	(✓)	✓	(✓)	(✓)	✗
VPP–TSO reporting (Zajc et al., 2019)	✗	—	✗	✓	✓
Our work	✓	✓	✓	✓	✓

availability does enter a placement model, it enters as a static coefficient or a synthetic profile, not as a signal derived from the operational state of a real transmission system. Third, energy is predominantly treated as a post-deployment optimization rather than as a term inside the embedding objective that determines the deployment in the first place.

Our contribution addresses these three points jointly. We embed energy awareness as a native criterion in the IDMO-DMO-IDM hierarchy; we source that awareness from an inter-domain EMS fed by domain-level VPPs, which report upstream to a TSO service following the established VPP reporting architecture (Zajc et al., 2019) and receive back a normalized greenness index derived from a power flow solved on a CIGRE benchmark grid (CIGRE Task Force C6.04, 2014; Thurner et al., 2018) driven by real national SCADA data; and we carry that index into the embedding objective itself, as a weighted term traded off against end-to-end latency, so that sustainability shapes placement rather than merely correcting it afterwards.

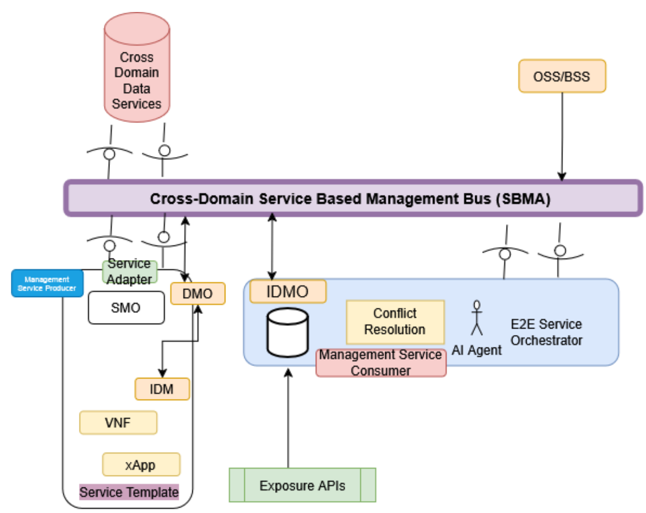

Figure 1: Cross-domain orchestration architecture for end-to-end 6G service deployment.

3. Background: Inter-Domain Orchestration Architecture

Future wireless networks are expected to be heterogeneous, multi-vendor, multi-domain ecosystems that combine diverse access technologies (e.g., 5G, 6G, Wi-Fi) with complex service chains stretching from the edge to the central cloud. To manage this scale and complexity, the proposed architecture adopts a hierarchical, modular design that realizes a distributed orchestration paradigm across multiple administrative and technological domains. Figure 1 depicts the overall structure and the interrelations among the relevant components.

The IDMO is the top-level orchestrator that coordinates the different domains to build E2E mobile network instances. As the global orchestrator, it decomposes E2E service requests into domain-specific components (RAN, Transport, Core Network, AI Data Orchestration) and assigns them to the appropriate domains entities. Concretely, the IDMO splits each request into smaller tasks (e.g., the RAN part, the transport part, and the mobile core) and selects a suitable domain for each, thereby assembling the E2E view. The domain selection follows a multi-criteria decision process that evaluates every candidate domain against:

- *Service capability*: confirming that the domain can host the required network functions;
- *Resource availability*: ensuring that enough CPU, memory, and storage can support the part of the request assigned to the component running in this domain;
- *Latency and coverage suitability*: matching the deployment location to the target service area.

The DMO is the orchestrator present at a domain receiving domain-specific service components from the IDMO through the Service adapter component and interacts directly with its associated IDMs to manage the lifecycle of Network Services (NSs). These NSs are built from virtualized/containerized network functions (VNFs/CNFs) that realize the assigned component. Each DMO (e.g., DMO/Core, DMO/RAN, DMO/Transport, DMO/Data-Orch) instantiates and configures VNFs/CNFs at selected PoPs, performs intra-domain resource optimization, and start monitoring system instances to track the KPIs offered by the given NSs. IDMs are the managers that directly govern the infrastructure, servers, Kubernetes clusters, or virtualized PoPs, on which the network functions composing the NS instance run. A single domain may contain several IDMs; in the RAN segment, for example, PoPs are frequently distributed to guarantee coverage and honor tight latency constraints (e.g., for the fronthaul interface). Every IDM is responsible for deploying, monitoring, and managing the CNFs/VNFs requested by its controlling DMO. All orchestration layers communicate through the cross-domain service-based management bus (SBMA). Inspired by the 3GPP service-based architecture 3GPP TS 28.533 (2025), this component serves as lightweight communication, domain and service-discovery middleware that supports event-driven exchanges between the IDMO, the DMOs, and auxiliary components such as analytics engines and policy managers.

Taken together, the architecture forms a vertically integrated control hierarchy that unifies orchestration, analytics, and sustainability management: the IDMO provides global coordination and multi-domain decision-making, the DMOs perform local orchestration, and the IDMs enforce infrastructure-level control. Once the mapping has been determined, the IDMO decomposes the service into domain-specific components and dispatches orchestration intents through the SBMA to the corresponding domains. At each domain, the service adapter translates the received intent to concrete deployment instructions for the DMO, which trigger its underlying IDMs to instantiate the assigned VNFs or CNFs (e.g., gNB-DU/CU, UPF, or AMF) within their local PoPs using container orchestration platforms such as Kubernetes. After a successful deployment, the IDMs report operational status and endpoint identifiers to their DMOs, which aggregate this information and notify the IDMO of service readiness. Finally, the IDMO performs cross-domain configuration tasks, interconnecting network segments, establishing VPNs, and enforcing end-to-end policies, thereby completing service instantiation. Throughout this paper, the reported latency is the end-to-end *data-plane* latency of the provisioned slice, comprising propagation over the access and inter-domain links (including the VPN hop between domains) and per-function processing; it is the quantity constrained by the slice SLA. Control-plane exchanges over the SBMA, namely intent dispatch, status reporting, and cross-domain configuration, are a one-time provisioning cost that is not part of this per-slice budget and is therefore excluded from it; orchestration latency is instead a KPI of the physical testbed campaign outlined in Section 6, not of the evaluation reported here. Queueing and transmission delays lie likewise outside the steady-state model discussed in Section 5.6. We note that signalling delay does bound the reaction time of the runtime adaptation loop of Section 4.1, where it is a first-order concern. The resulting E2E instance is provisioned to balance performance, resource efficiency, and sustainability. The next section examines how energy awareness is integrated into this architecture.

4. Proposed Approach: Energy-Aware Orchestration

The operation of the proposed energy-aware multi-domain orchestration framework falls into two main phases: (i) a *provisioning and placement* phase, in which the system instantiates an E2E mobile service across heterogeneous domains, and (ii) a *runtime adaptation* phase, in which continuous monitoring and optimization sustain performance and energy efficiency. Both phases rely on hierarchical collaboration among the IDMO, the DMOs, the IDMs, and the inter-domain EMS through the SBMA.

To achieve energy-aware orchestration, our proposed framework integrates an inter-domain EMS. This system aggregates information from distributed power providers, renewable generation, battery energy storage, and other energy assets, and enables real-time monitoring and reporting of these resources. The resulting information is fed back into the orchestrator's decision process, letting it dynamically assess the energy mix available within each domain or site. Consequently, the orchestrator can favor greener locations when placing or migrating virtualized network functions, advancing sustainability goals while still respecting performance and Service Level Agreements (SLAs).

Figure 2 illustrates the proposed hierarchical *system-of-systems* architecture. At the local level, each domain hosts a VPP. The VPP acts as an aggregation entity that monitors and manages the collective energy status of all IDMs, including their relevant Distributed Energy Resources (DERs), storage systems, and flexible loads. It collects real-time energy telemetry from the domain's infrastructure and consolidates it into a unified representation of the domain's state (e.g., net-load profiles and local generation capacity/greenness). This VPP-level status report is then forwarded to the superordinate inter-domain EMS.

To incorporate energy awareness, each domain interfaces with a local VPP that aggregates information from its renewable sources and the local grid; these metrics are published to the local DMO and propagated by the inter-domain EMS to the IDMO via the SBMA. This provides real-time telemetry on: (i) the current percentage of renewable energy in the mix, (ii) carbon intensity (gCO_2/kWh), and (iii) available reserve power. Beyond these metrics, each domain can also expose an *index* value that quantifies its locally available green energy using real-time SCADA-derived power flows, together with the national (or inter-domain) renewable energy mix. This index provides a normalized, site-specific indicator of how favorable current conditions are for low-carbon computation. It reflects the local renewable surplus, ranked against that domain's own recent history rather than against other domains. A low index signals heavy import of nonrenewable energy into the domain, whereas a high index indicates a strong local renewable surplus. The green index is thus a *supply-side* indicator of renewable availability that complements, but does not replace, the grid carbon intensity: the orchestrator ranks domains for placement by their carbon intensity (Eq. 3), since a domain may present a high instantaneous renewable surplus yet still draw on a comparatively carbon-intensive average grid mix.

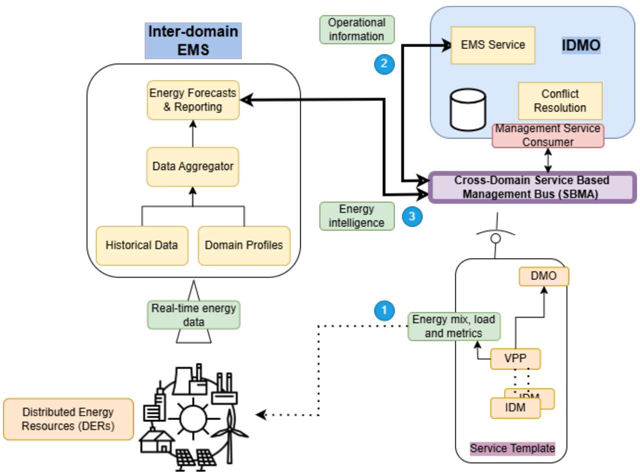

Figure 2: Proposed energy-aware unified architecture for 6G networks.

Operating at a higher hierarchical level, the inter-domain EMS is responsible for coordinating the coupled domains. A core capability of the EMS is advanced forecasting, which draws on both real-time VPP data streams and extensive historical datasets to produce probabilistic or deterministic forecasts of key quantities such as inter-domain power exchanges, net load, and generation availability. The processed and forecasted information is then aggregated into a standardized data structure, a formatting choice essential for interoperability and designed for ingestion by the IDMO, so as to provide a coherent, network-wide overview. Ultimately, the IDMO need not track the energy state of the network itself: all intelligence gathering is delegated to the inter-domain EMS, and the IDMO simply consumes the information it requires at decision time.

In parallel, the DMOs promote or deprioritize IDMs according to the energy mix advertised by local VPPs, which already monitor per-IDM energy information alongside resource abstractions. Whenever feasible, DMOs deploy VNFs at PoPs powered by renewable sources. This is especially relevant for distributed RAN domains, where a gNB VNF may be instantiated at a greener site as long as latency constraints are respected. IDMs regularly report their energy mix, resource load, and latency metrics to the DMO to enable localized, energy-aware decisions. Once the E2E request is established, local energy-efficiency operations, such as switching RUs and their associated processing VNFs (e.g., the CU) on or off during nighttime while preserving coverage, are handled by running tuples of monitoring, analytics, decision engines, and actuators. The actuator issues orders to the DMO, which may in turn inform the IDMO of the changes. These tuples must be orchestrated on demand for a given entity, since the underlying infrastructure may host multiple services; a "Mobile Core Domain," for instance, could run several mobile-core instances for different requesters. The domains are interconnected through VPN tunnels that carry both control and data traffic, for example, a site with a conventional power supply can connect over an encrypted VPN across the Internet to another site powered by sustainable sources. Throughout, the SBMA is responsible for consistent orchestration, lifecycle management, and policy synchronization between the IDMO and the distributed DMOs.

The provisioning workflow enables automated instantiation of a network that is not only compliant with Quality of Service (QoS) requirements but also optimized for carbon footprint and energy efficiency. It is triggered when the IDMO receives a high-level E2E service request or intent, and proceeds through four stages: Intent Translation, Candidate Discovery, Multi-Criteria Optimization, and Distributed Instantiation. At the heart of the workflow lies the placement engine within the IDMO, which matches resource advertisements from the DMOs against energy profiles from the EMS. The engine applies a MCDA algorithm to rank candidate domains via a weighted cost function. This enables informed trade-offs, for instance, selecting a more distant core network domain (with higher latency) when it is currently powered by surplus solar energy, provided the total latency stays within the network's tolerance margin.

4.1. System Workflow and Interaction

The framework operates through a hierarchical interaction among the IDMO, the DMO entities, and their IDMs, with the VPPs and the inter-domain EMS providing the real-time energy metrics needed for sustainability-aware orchestration. As shown in Figure 3, the workflow proceeds in six main stages:

1. *Service Request Reception*: the IDMO receives a high-level E2E service instantiation request specifying the required network functions, performance indicators (such as latency), and sustainability preferences.

2. *Service Decomposition and Domain Selection*: the IDMO decomposes the E2E service into functional components (RAN, Transport, Core, etc.) and, based on the resource advertisements received over the Cross-Domain SBMA, selects appropriate domains using multi-criteria logic that balances performance and sustainability, namely: (a) service capability availability, whether a domain advertises the capability to instantiate the required VNFs; (b) resource availability, in terms of compute, memory, and storage; (c) coverage and latency constraints, ensuring placement meets radio and transport delay requirements; and (d) sustainability and energy mix, using VPP energy intelligence to prioritize greener domains (e.g., Domain A) when other performance conditions are met.

3. *Domain-Level Orchestration*: the selected DMOs (e.g., in Domain A or B) receive service components from

the IDMO and translate them into domain-specific deployments, instructing their local IDMs to instantiate the required CNFs or VNFs.

4. *Energy Intelligence from the EMS*: the EMS aggregates energy data from distributed sources (VPPs) and relays it to the IDMO, enabling informed placement decisions such as deploying high-load compute functions in domains currently powered by renewable sources.

5. *Runtime Monitoring and Energy-Aware Optimization*: once instantiated, services run under continuous supervision by Monitoring, Analytics, Decision, and Actuation (MADA) tuples. When conditions change, e.g., low traffic or reduced renewable availability in Domain A, the Decision Engine triggers adaptation actions such as scaling down low-load RUs or CUs in the RAN domain, migrating VNFs to a greener domain, or adjusting resource allocations to minimize carbon footprint while maintaining the SLA.

6. *Feedback and Closed-Loop Control*: adaptation decisions are executed by the respective DMOs, each of which then notifies the IDMO of the status change so that the IDMO maintains an accurate global E2E view of resource and energy states.

Through this workflow, the IDMO can effectively orchestrate sustainable services across federated infrastructures. Figure 3 shows the provisioning and placement messages for this E2E energy-aware placement.

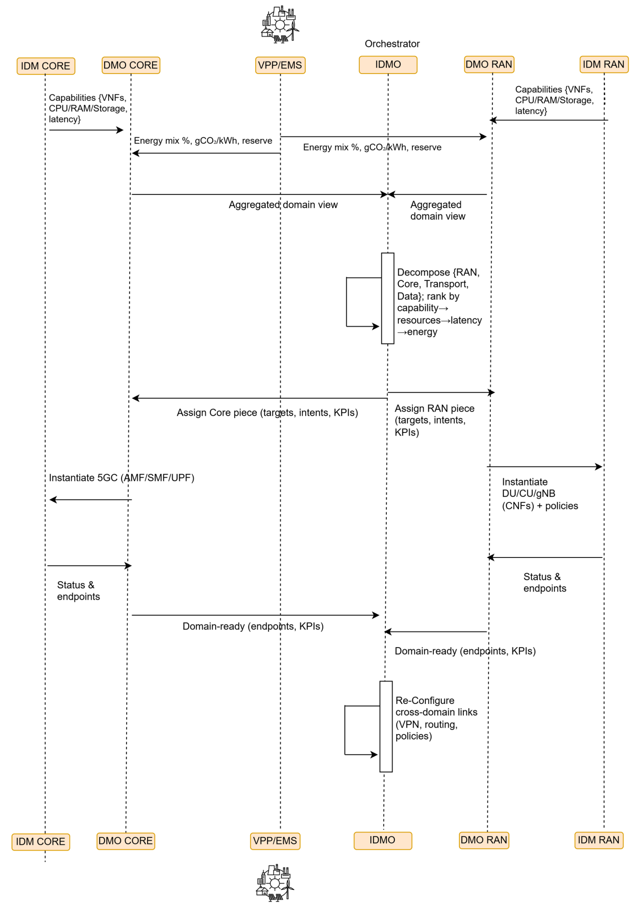

Figure 3: Workflow and provisioning/placement messages for energy-aware E2E 6G service deployment.

4.2. Green-Energy Index

The greenness signal exposed by each domain follows the Green Index methodology of the Greek TSO (IPTO) Vantzos, Skipis, Chassioti and Moraitis (2025), a *supply-side* measure of locally available green energy derived from grid telemetry. Let P_{net} denote the net power balance at the interface between the domain and the rest of the transmission grid, decomposed into its export and import parts, $P_{\text{net}} = P_{\text{exp}} - P_{\text{imp}}$ with $P_{\text{exp}}, P_{\text{imp}} \geq 0$. In combination with $r_{\text{RES}}(t) \in [0, 1]$, the share of renewable sources in the national energy mix at time t, yields the RES-adjusted power balance

$$R(t) := r_{\text{RES}}(t)\, P_{\text{exp}}(t) - \big(1 - r_{\text{RES}}(t)\big)\, P_{\text{imp}}(t), \qquad (1)$$

so that exporting power while the national mix is green scores high ($R \gg 0$), whereas importing power while the mix is carbon-intensive scores low ($R \ll 0$).

Since R is expressed in absolute power units, it cannot be compared directly across domains of different size. It is therefore normalized by its signed percentile rank $I(t; W) \in [-1, 1]$ over a sliding window W of one week, long enough to span day/night, weekday/weekend and weather-driven variability. Rescaled to the unit interval:

$$I_{\text{green}}(t; W) := \frac{1 + I(t; W)}{2} \in [0, 1]. \qquad (2)$$

The index admits a direct operational reading: values below 0.5 indicate a net importer and values above 0.5 a net exporter, with values near 1 corresponding to a local renewable surplus close to its weekly maximum. Because the percentile ranking makes the index dimensionless, indices reported by different domain instances share a common scale and a common operational reading: 0.5 marks the import/export break-even for every domain, and the extremes mark each domain's own weekly limits. What the index does *not* provide is an absolute comparison. Two domains reporting the same value are each at the same point relative to their own recent history, not at the same absolute level of renewable availability or carbon intensity. Cross-domain ranking for placement is therefore performed on carbon intensity (Section 4.3), with the Green Index serving as a supply-side indicator of local conditions.

Two aspects of this construction deserve comment. First, the rank is *signed* rather than a plain percentile. A plain percentile rank would place the neutral point at the median of the window, which is an arbitrary quantity that shifts from week to week; taking the sign from $R(t)$ and the magnitude from the rank of $|R(t)|$ instead anchors $I_{\text{green}} = 0.5$ at $R = 0$, the physical break-even between net import and net export. The importer/exporter reading above therefore holds in every window, independently of how green the week as a whole happened to be. Ranking is preferred to a z-score or a min–max rescaling because $R(t)$ is heavy-tailed: renewable surpluses arrive as infrequent large excursions, so a z-score would be dominated by a handful of extreme hours and a min–max rescaling would be fixed entirely by the single largest surplus in the window. A rank is invariant under any monotone transformation of R, and so is insensitive to both the scale and the shape of a domain's net-balance distribution.

Second, the window length is set to one week because 168 h contains a whole number of each dominant cycle in $R(t)$: seven diurnal cycles, driven by solar generation and the daily load curve, and exactly one weekly cycle, driven by the weekday–weekend swing in industrial demand. Every window therefore has the same phase composition, and the rank cannot be biased by which days happen to fall inside it, as would occur for a window of, say, three days, where a Sunday hour and a Wednesday hour would be ranked against different mixes of weekdays. A 24 h window would be too short: it would rank each hour only against the same day, removing the diurnal contrast that the index is meant to express, and would leave too few samples for a stable rank. A monthly or seasonal window, conversely, would readmit seasonal drift into the reference distribution. Because seasonality evolves on a scale of weeks to months, it is approximately constant within any one window and therefore cancels from the rank: the index measures how green a domain is *relative to its own recent behaviour*, not against an annual scale. This is what keeps it usable across seasons, since the index spans the full unit interval in a low-solar winter week just as it does in a sunny summer week, and the orchestrator retains its discriminating power year-round instead of seeing a signal pinned near zero for months. The price of this choice is that the index is a relative signal: a domain reporting 0.9 in December need not be cleaner in absolute terms than one reporting 0.4 in July. That is precisely why the sustainability score of Section 4.3 is derived from absolute carbon intensity rather than from the Green Index, and why a domain can exhibit a lower green index than its peer while running on a substantially cleaner grid, as is the case in Table 3. Each VPP exposes its per-domain index as part of the standardized energy telemetry introduced above, alongside carbon intensity, renewable share and reserve power, the inter-domain EMS ingests this telemetry and derives from the sustainability score consumed by the IDMO's multi-criteria domain selection. We refer the reader to for the full formulation.

4.3. EMS Sustainability Score

Beyond forwarding the raw per-domain telemetry, the key function of the inter-domain EMS is to *rank the coupled domains against one another* based on sustainability. Concretely, it converts the absolute carbon intensities into a normalized greenness score $\gamma_d \in [0, 1]$,

$$\gamma_d = \frac{\mathrm{CI}_{\max} - \mathrm{CI}_d}{\mathrm{CI}_{\max} - \mathrm{CI}_{\min}}, \tag{3}$$

where $\mathrm{CI}_{\max}$ and $\mathrm{CI}_{\min}$ are taken over the domains coupled by the EMS, so that the cleanest grid scores 1 and the dirtiest scores 0 (when all domains are equal, $\gamma_d = 1$). This dimensionless, self-normalizing signal makes greenness directly comparable across domains. It is computed from carbon intensity rather than directly from the Green Index of Section 4.2: the two are complementary but need not coincide, as the Green Index captures the instantaneous local renewable surplus whereas γ_d reflects the average emission intensity of the energy actually consumed.

5. Implementation and Experimental Evaluation

To assess the feasibility of the architecture described in Sections 3-4, we implemented the IDMO/DMO/IDM control hierarchy together with the VPP and inter-domain EMS energy plane as an executable prototype and evaluated the energy-aware placement engine against non-sustainability baselines. The prototype is implemented in Python on top of the SPEED edge-cloud service-function-chain placement model[1]: the root aggregation zone acts as the IDMO, each compute zone is a DMO (a domain), and the compute nodes are the IDMs (PoPs). This section explains how the energy-intelligence plane operates, how the placement engine uses it, and the results obtained on a two-domain proof of concept and on two real European backbones.

At the top level, the IDMO acts as the global orchestrator, decomposing high-level service requests into domain-specific components (e.g., RAN, Transport, Core) and dispatching them through the Cross-Domain SBMA, which provides a unified orchestration and policy interface. Accordingly, Figure 4 shows the proposed architecture, comprising two distinct domains that operate under a unified orchestration framework and represent different provider capabilities and energy profiles:

- *Sustainability-Focused Domain*: represented by the blue area on the left, this domain includes a VPP that supplies green energy, for example from solar or hydro sources, to data center resources that primarily host core network functions.
- *Baseline/Performance-Focused Domain*: represented by the yellow area on the right, this domain hosts Transport, RAN, and data-management infrastructure and represents a typical deployment powered by the standard electric grid.

We offered a mix of eMBB, URLLC, mMTC, and core-only slices and compared the energy-aware IDMO (MCDA) against a performance-first orchestrator (latency only, akin to SPEED's shortest-delay placement), a carbon-greedy heuristic reusing SPEED's cost model, a genetic-algorithm search, and random placement. The genetic algorithm is

[1]SPEED, `https://github.com/anselmobattisti/speed`. We implemented a minimal, clean-room compatibility layer so that the SPEED zone hierarchy and its resource and cost aggregation execute unmodified.

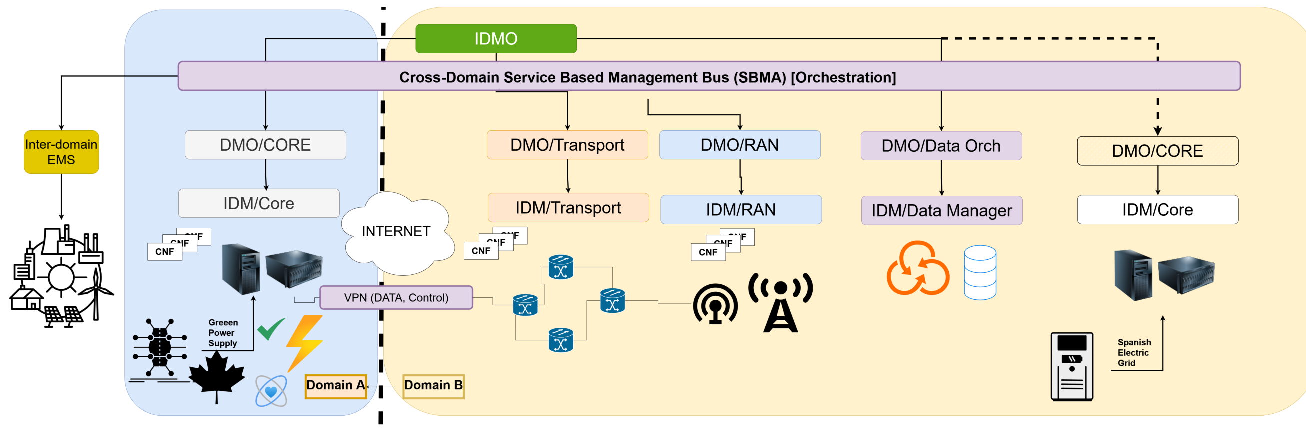

Figure 4: Cross-domain orchestration architecture for energy-aware end-to-end 6G service deployment.

included as a search-based reference point: service-function-chain embedding is NP-hard, so a metaheuristic stands in for the optimum that an exact solver cannot reach at this scale, and it tests whether the gains of the MCDA engine survive comparison with a method that actually explores the placement space rather than deciding in a single pass. Genetic algorithms are an established choice for energy-aware chain embedding (Krishnamohan and Harvey, 2026); our baseline is a distinct, simpler implementation of that idea, using a direct integer encoding with one gene per (request, function) slot whose value is the hosting domain, and a scalarised objective that penalises rejections, then minimises carbon, with energy and latency as tie-breakers. It uses a population of 60 over 100 generations, tournament selection ($k = 3$), uniform crossover, per-gene mutation at rate 0.12, and elitism of 2; its initial population is seeded with the greedy greenest-first and fastest-first assignments, so it starts from at least the quality of the greedy heuristics it is compared against. Throughout, the *acceptance* rate (equivalently *admission*; the **Acc.** column) denotes the percentage of offered end-to-end slice requests that the orchestrator successfully instantiates and allocates across the domains, out of all requests submitted; a request is rejected only when no placement satisfies its capability, capacity, and latency-SLA constraints.

The slice mix, the function chain of each slice type and its latency SLA are given in Table 2, and the compute, network and energy parameters of the two domains in Table 3. Together the domains provide 80 000 CU of compute against an expected demand of roughly 888 CU per request, so at the operating point of 40 requests the federation runs at about 44% utilization, and at the 60-request load used for the acceptance comparison at about 67%. The carbon and energy figures are reported over an operating horizon of $h = 1$ h. Transport links are assumed uncongested: a placement is constrained by the per-PoP CPU and memory capacity and by the end-to-end latency SLA, whereas the per-request bandwidth demand is carried through the model without being checked against a link capacity. The consequences of this assumption are examined in Section 5.6.

Table 2
Offered workload for the two-domain evaluation. Per-function demands are ru 120, du 200, cu 250, tn 100, upf 300, amf 200 and smf 150 CU. Requests are drawn from the mix with a fixed seed; the operating point is 40 requests, raised to 60 for the acceptance comparison.

Slice	Chain	Mix (%)	CPU (CU)	SLA (ms)
eMBB	ru, du, cu, tn, upf, amf	35	1170	120
URLLC	ru, du, upf	25	620	30
mMTC	ru, du, tn, upf, smf	20	870	100
core service	tn, upf, amf, smf	20	750	120

5.1. Green-Energy Index: Testbed Implementation

The energy testbed realizes the interface between the domain-level VPP and the local Transmission System Operator (TSO), through which each domain compoutes the index of Section 4.2 as illustrated in Figure 5. It consists of two services communicating over a REST API, reproducing the periodic upstream reporting of aggregated pool measurements of the reference VPP–TSO communication architecture (Zajc et al., 2019). In particular a VPP replays a daily consumption profile as timestamped pool measurements and a TSO service computes and returns the corresponding Green Index each report.

Internally, the TSO service maps the national grid state onto the CIGRE high-voltage benchmark network (CIGRE Task Force C6.04, 2014), solved with the open-source power-flow-engine pandapower (Thurner et al., 2018). The benchmark network acts as a reduced digital twin of the national grid, hourly national load and generation-mix time series, retrieved from the ENTSO-E transparency platform for the configured country. Background loads are scaled down, to fit the benchmark, by the ratio of the current national load to its peak and the wind, solar, and hydro generators so as to preserve the national generation mix, while the VPP pool measurement is injected as a controllable load. The net balance P_{net} of Eq. (1) is then read at the slack bus of the solved network (export-positive).

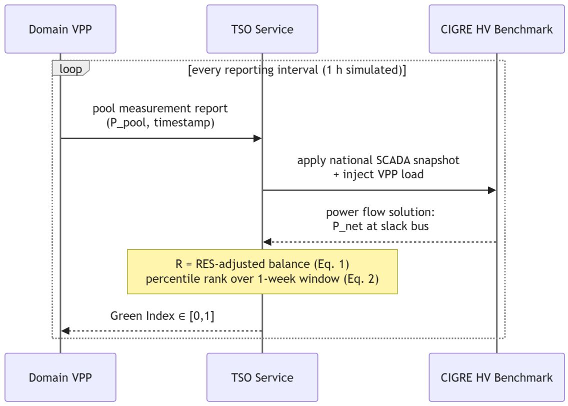

Figure 5: Message exchange of the energy testbed. The domain VPP periodically reports its aggregated pool measurement to the TSO service, following the reference VPP–TSO reporting architecture (Zajc et al., 2019). The TSO service drives the CIGRE HV benchmark with the national SCADA snapshot, solves the power flow and returns the normalized Green Index of Section 4.2.

Because the reduced twin does not model real cross-border interconnections, the exports–imports term of the original IPTO formulation is not observable in the testbed, we therefore approximate r_{RES} by the renewable plus the hydro share of total national generation. Since the index is percentile-ranked against its own one-week history rather than used in absolute terms, it is largely insensitive to this rescaling of the mix ratio. The rolling normalization of Eq. (2) is maintained over a one-week (168-hour) sliding window, and the simulation clock is accelerated so that months of hourly SCADA data can be replayed in hours.

Each running instance of the testbed represents one telco domain, deploying two instances side by side, each driven by the SCADA data of a different national grid, yields directly comparable per-domain greenness signals, which are the source of the per-domain green-index telemetry used in the evaluation of Section 5.4

5.2. Energy-Intelligence Plane: VPP and EMS

The energy-intelligence plane is the concrete implementation of the VPP and inter-domain EMS described in Section 4. It is deliberately structured so that the IDMO does not track any grid state itself: all energy intelligence is gathered by the EMS and used by the IDMO only when making decisions.

1-) Virtual Power Plant (VPP) One VPP is instantiated per domain. It exposes that domain's local energy telemetry as a standardized record containing (i) the grid carbon intensity CI_d ($\mathrm{gCO_2eq/kWh}$), (ii) the renewable share, (iii) the green index of Section 4.2, and (iv) the available reserve power. In our evaluation these quantities are sourced from real grid data (see Section 5.4): the carbon intensity is derived from the national generation mix by weighting each fuel by a standard emission factor, and the green index is read from the operators' logs. A VPP can report either a representative value for the operating period or the value at a specific hour of the underlying time series, which is what enables the day/night runtime adaptation of Section 4.1.

2-) Inter-domain EMS The EMS ingests telemetry from every VPP and compiles it into a single, network-wide report for the IDMO, in which each domain carries the greenness score γ_d of Section 4.3. The EMS additionally exposes a system-wide green index as the mean over all domains; because the per-domain indices are relative to each domain's own history, this aggregate is a coarse indicator of federation-wide conditions rather than an absolute one, and it is not used in any placement decision. The resulting locality-aware report, one entry per domain, each carrying the raw telemetry plus γ_d, is the only energy information the IDMO consumes.

5.3. Energy Model and MCDA Placement Engine

Each IDM follows the standard linear server power model, in which a node hosting no functions is suspended and draws no power, so consolidating functions onto fewer nodes saves energy:

$$P(u) = P_{\mathrm{idle}} + (P_{\mathrm{max}} - P_{\mathrm{idle}})\, u, \tag{4}$$

with $u \in [0, 1]$ the CPU utilization. The carbon emission of domain d over an operating horizon h combines its energy, its facility overhead PUE_d and the grid carbon intensity CI_d reported by its VPP:

$$C_d = \frac{P(u_d)\, h}{1000}\, \mathrm{PUE}_d\, \mathrm{CI}_d. \tag{5}$$

Following the provisioning workflow in Section 4.1, the IDMO decomposes each E2E slice into its functional components (RAN/Transport/Core) and, for each component, ranks the candidate domains with a MCDA score that weighs the four criteria in Section 4:

$$s(d) = w_{\mathrm{cap}}\, \kappa_d + w_{\mathrm{res}}\, \rho_d + w_{\mathrm{lat}}\, \lambda_d + w_{\mathrm{sus}}\, \gamma_d. \tag{6}$$

where κ_d is a hard capability filter (only domains advertising the required component are scored), ρ_d is the normalized free capacity of the domain, λ_d is the latency/coverage slack of hosting the component at d, and γ_d is the greenness score (3) supplied by the EMS. A latency guard subsequently pulls components back toward the closest domain whenever the greenest assignment would violate the end-to-end latency SLA, so that a slice is accepted only when it both fits and meets its SLA. Throughout the evaluation we used $(w_{\mathrm{cap}}, w_{\mathrm{res}}, w_{\mathrm{lat}}, w_{\mathrm{sus}}) = (0, 0.15, 0.25, 0.60)$, i.e. capability as a hard filter and a sustainability-leaning weighting of the remaining criteria.

5.4. Two-Domain Proof of Concept: Sustainable vs. Performance Domain

The PoC federates a sustainability-focused domain (the *Sustainable Domain*) and a baseline- and performance-focused domain (the *Performance Domain*), as shown in

Figure 4. The greenness of each domain is derived from real grid data for the period Nov. 2025-Feb. 2026. The VPP telemetry that the EMS forwards to the IDMO is summarized in Table 3, and its hourly evolution is shown in Figure 6. As noted in Section 5.3, the Sustainable Domain is the one with the lower grid carbon intensity (and hence the higher sustainability score γ_d), which need not be the domain with the higher instantaneous green index: here the Sustainable Domain runs on a cleaner average grid mix (117 vs. 301 gCO_2eq/kWh) despite a slightly lower green index. The Performance Domain hosts the RAN capability close to users, whereas the Sustainable Domain hosts only latency-tolerant Transport/Core functions.

Table 3
Domain (DMO) profiles. Compute, network and power parameters are the configured values of the two-domain scenario; carbon intensity, renewable share, and green index are computed from the real grid data and attached to the nodes via the VPPs; the EMS sustainability score follows (3). The green index is a period mean of the domain-relative index of Section 4.2, so the two values are not comparable in absolute terms; cross-domain ranking uses the carbon intensity.

	Sustainable *low-carbon*	**Performance** *low-latency*
Compute		
PoPs (IDMs)	8	6
CPU per PoP (CU)	4000	8000
Total CPU (CU)	32 000	48 000
Memory per PoP	4 GB	8 GB
Processing speed (CU/ms)	25	40
Network		
Capabilities	Transp., Core	RAN, Transp., Core
Delay to users	20 ms	5 ms
Inter-domain delay	25 ms	
Energy		
Carbon intensity (gCO_2eq/kWh)	117	301
Renewable share	60%	45%
Green index (period mean)	0.47	0.61
EMS sustainability score γ_d	1.00	0.00
PUE	1.15	1.50
Idle / peak node power (W)	90 / 200	120 / 350

Table 4 reports the outcome at an offered load of 40 slices, at which every feasible policy accepts all requests.

The energy-aware IDMO reduces carbon emissions by 31.2% (from 803 to 553 gCO_2eq/h) at 100% acceptance, paying only additional latency that remains within the slices' SLAs. As illustrated in Figure 7, it achieves this by offloading the latency-tolerant Core functions onto the clean Sustainable Domain grid (89 functions, 137 gCO_2) while the RAN stays in the Performance Domain near the users (101 functions, 416 gCO_2); the performance-first orchestrator instead concentrates all 190 functions in the Performance Domain (803 gCO_2). The carbon-greedy heuristic cannot split RAN-bearing slices and therefore captures almost none of the benefit, and the genetic algorithm, despite optimising carbon directly and starting from the greedy solutions, converges to 636 gCO_2eq/h against the MCDA engine's 553. The search does not reach the optimum of its own objective where capacity and SLA feasibility are hard constraints applied at decoding, so a single gene change can flip a chain from accepted to rejected and incur the rejection penalty, making the landscape a set of cliffs rather than a gradient, while uniform crossover recombines the per-function genes independently and readily splits a coherently placed chain across domains. The MCDA engine instead carries a graded latency-slack term and a latency guard, which steer placements away from the feasibility boundary while still leaning green. We therefore read this as evidence that a general-purpose search over a rugged, constraint-dominated space does not dominate a domain-informed scoring rule at this scale and search budget, rather than as a claim that the MCDA assignment is optimal. Random placement assigns functions to domains without regard to capability, so all but nine of the forty requests are rejected before capacity or latency are ever tested, and its apparently low absolute carbon merely reflects how little it embeds. Normalised per accepted slice it is the worst policy in the table by a wide margin, at 53.9 gCO_2 against 13.8 for the MCDA engine. The benefit is compounded by *acceptance* as the offered load approaches the Performance Domain's capacity, the IDMO relieves the Performance Domain by using the Sustainable Domain and keeps 95% of slices accepted at a load of 60 requests, against only 72% for performance-first.

5.5. Pan-European Evaluation: NOBEL-EU and GÉANT

To test the approach at scale, we replicated the experiment on two real European backbone topologies from SNDlib: NOBEL-EU (28 city PoPs) and GÉANT (22 PoPs). Each PoP becomes a candidate hosting domain whose grid carbon intensity is determined by its country; the two profiled domains (Sustainable and Performance) use the real telemetry described above, while the remaining countries use representative annual average intensities. The RAN part of every request is pinned to its source PoP, while the Core part may be hosted at any PoP reachable within the latency SLA, based on great-circle propagation delays. The IDMO's MCDA engine, driven by the same EMS greenness score (3), steers the Core toward clean-grid countries. Table 5 summarizes the outcome, and Figure 8 visualizes it the core load concentrates on low-carbon grids (France, the Nordics, Switzerland) and avoids coal-intensive ones (Poland, Serbia).

Across both backbones, the energy-aware IDMO reduces carbon emissions by 16-20% relative to the performance-first policy at comparable acceptance rates and sub-SLA latency. A purely carbon-greedy policy achieves marginally lower emissions but overcommits the few green PoPs, reducing acceptance to about 81%. In contrast, the MCDA formulation balances greenness against resource availability and latency, showing that sustainability can be treated as a native orchestration criterion without sacrificing service quality.

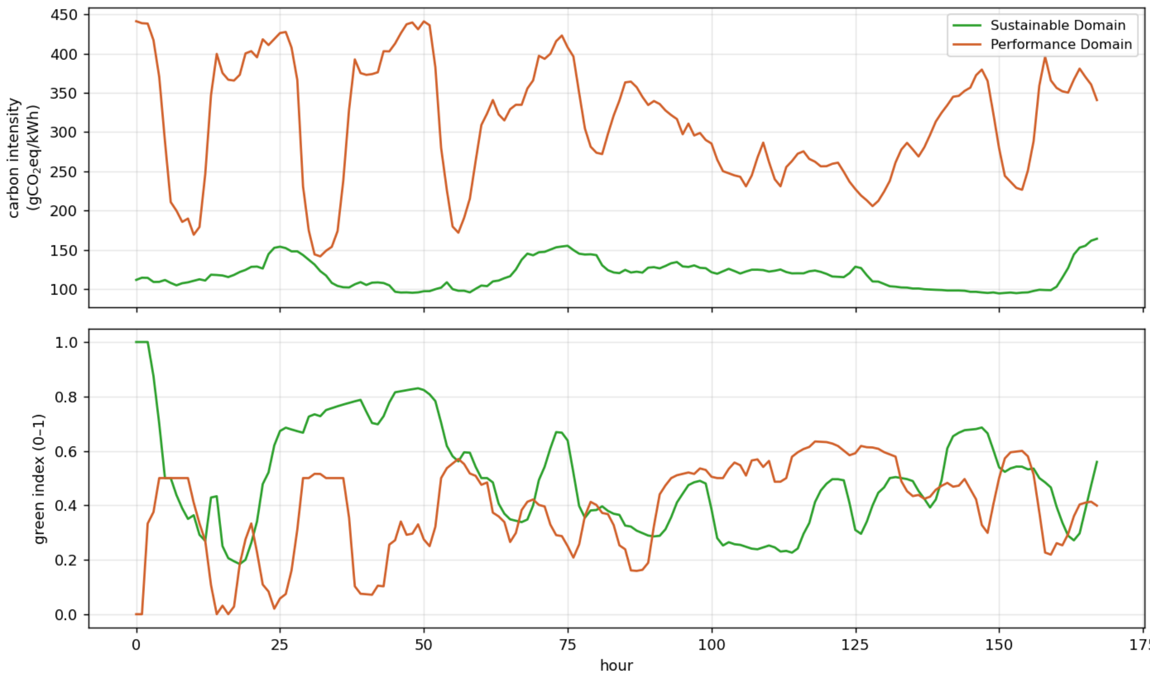

Figure 6: Real grid data driving the VPPs (first week): the Sustainable Domain's grid is consistently low-carbon, whereas the Performance Domain's grid is more carbon-intensive and volatile. The lower panel shows the corresponding local green indices.

Table 4
Two-domain results at 40 slice requests. ΔC is the carbon change relative to performance-first; it is not reported for random placement, whose acceptance differs, because a reduction obtained by rejecting requests is not comparable with one obtained by placing them better. The carbon per accepted slice normalises for this difference.

Policy	Acc. (%)	Energy (kWh)	Carbon (gCO_2/h)	Carbon/slice (gCO_2)	ΔC	Lat. (ms)
performance-first	100	2.67	803.1	20.08	(ref.)	33.1
carbon-greedy (SPEED)	100	2.84	792.7	19.82	-1.3%	38.2
GA (search-based)	100	2.59	636.0	15.90	-20.8%	64.9
MCDA (IDMO)	100	2.55	**552.9**	**13.82**	$\mathbf{-31.2\%}$	66.3
random	22.5	2.18	485.1	53.90	n/a	69.0

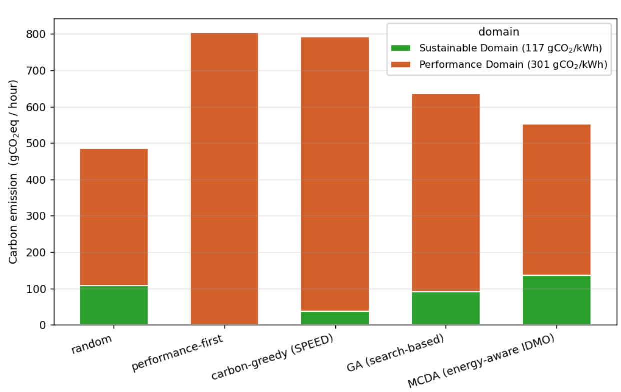

Figure 7: Carbon of the embedded functions, split by domain. The energy-aware IDMO shifts emission-heavy load off the Performance Domain grid onto the clean Sustainable Domain one.

Table 5
Pan-European results (200 requests, 40 ms SLA): the energy-aware IDMO versus the performance-first baseline.

Topology	Policy	Carbon (gCO_2/h)	Acc. (%)
NOBEL-EU	performance-first	2474	98.5
NOBEL-EU	**MCDA (IDMO)**	**1982** (-19.9%)	95.5
GÉANT	performance-first	2203	97.5
GÉANT	**MCDA (IDMO)**	**1851** (-16.0%)	90.0

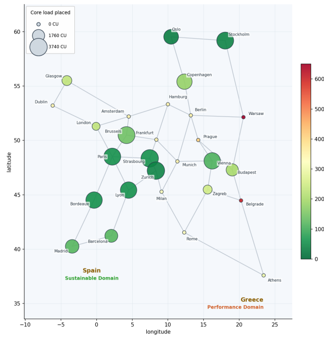

Figure 8: NOBEL-EU backbone: PoPs colored by grid carbon intensity (green = clean, red = coal-heavy); marker size shows the Core load placed by the energy-aware IDMO, which concentrates on the greenest grids.

5.6. Discussion and Limitations

The results support the central claim of the architecture: exposing VPP telemetry to the IDMO through the inter-domain EMS allows a multi-criteria placement engine to trade a bounded amount of latency for substantial carbon savings while simultaneously improving acceptance under load. The evaluation model considers steady-state embedding energy over an operating horizon rather than packet-level dynamics. Transport capacity is likewise outside the model: links carry a propagation delay but no bandwidth constraint, so the reported acceptance rates are governed by compute capacity and latency alone. The backbones considered here are provisioned well above the offered load, but the assumption is not neutral in the orchestrator's favour: the inter-domain hop that carries Core traffic to the clean domain is precisely the link that would saturate first, so a bandwidth-limited federation would realize smaller carbon savings than those reported above. Carbon intensity is derived from a fixed set of per-fuel emission factors and, by default, from the mean over the observed period, while the full hourly series remains available to drive the forecast-based day/night runtime adaptation loop of Section 4.1. Extending the study to time-varying, forecast-driven migration across the full federation is a natural next step.

6. Conclusion and Future Work

This paper has presented an energy-aware, multi-domain orchestration framework for sustainable 6G network operation. The architecture introduces a hierarchical management model comprising the IDMO, DMO, and IDM layers, augmented by integrated VPPs and an inter-domain EMS. Together, these elements enable cross-domain coordination of network functions and dynamic alignment of service placement with renewable-energy availability. Through a proof of concept driven by measured grid data from two national systems, and a pan-European study on the NOBEL-EU and GÉANT backbones, the framework shows that energy intelligence can be embedded into orchestration workflows at a bounded and quantified cost in latency and acceptance: the MCDA placement engine cut the carbon footprint of the embedded functions by 31.2% at full slice acceptance and within the latency SLAs, and by 19.9% and 16.0% on the two backbones, while relieving the performance domain under load and thereby raising acceptance from 72% to 95% at an offered load of 60 requests. Throughout the evaluation these two infrastructures are abstracted as the *Sustainable Domain* and the *Performance Domain*: the *Sustainable Domain* is the low-carbon, high-renewable-share infrastructure the orchestrator favours for green placement, whereas the *Performance Domain* is the latency-optimised infrastructure located close to end users that hosts the RAN and other delay-sensitive functions. This naming abstracts away the specific partner operators and grids, keeping the results dependent only on the two domains' energy and latency profiles rather than on their identities. By coupling energy management to orchestration across administrative boundaries, the proposed framework shows that grid-side telemetry can be carried into multi-domain placement decisions, and that doing so yields measurable carbon reductions at a quantified cost in latency and acceptance.

The most immediate next step is to implement the architecture on a federated physical testbed spanning two geographically distributed domains. In such a deployment, each domain runs its own DMO as a lightweight microservice suite exposing RESTful APIs for intra-domain orchestration, communicating with the IDMO over secure VPN tunnels and a publish–subscribe interface for telemetry exchange and intent delivery, while the IDMs manage the underlying compute through container orchestration platforms such as Kubernetes or OpenStack and the network functions (gNB DU/CU, UPF, AMF, and transport VNFs) are packaged as containerized workloads. A physical deployment would exercise precisely the parts of the framework that the steady-state evaluation of Section 5 leaves untested: end-to-end service instantiation across administrative boundaries, and the runtime adaptation loop under controlled traffic and energy variation, with synthetic user emulators driving the offered load and the VPP renewable share swept over day–night cycles, so that threshold-triggered VNF migration towards a greener domain and low-traffic RU power-down can be observed directly rather than assumed. It would also close the measurement gap on the indicators a placement study cannot report, namely orchestration latency, from intent submission to successful E2E service activation, and service continuity, in terms of packet loss and latency variation during migration or scaling; we plan to collect these automatically through a Prometheus monitoring stack and Grafana dashboards, alongside the resource-utilization, energy, and carbon figures already reported here.

Beyond this deployment, future research will extend the work along three complementary directions. First, we will explore the integration of AI-native decision engines to improve prediction accuracy and adaptability, enabling anticipatory orchestration that proactively responds to variations in traffic, energy supply, and service demand within a unified simulation environment. Second, we will embed trust and accountability mechanisms through decentralized verification and provenance tracking, so that the energy and resource claims exchanged between domains remain transparent and verifiable. Finally, interoperability and standardization efforts will focus on aligning the proposed architecture with ongoing activities in ETSI ZSM, 3GPP, and the O-RAN Alliance, so that the interfaces assumed here correspond to those emerging in practice.

CRediT authorship contribution statement

Berk Buzcu: Conceptualization, Methodology, Software, Writing – original draft. **Laura Finarelli:** Methodology, Writing – review & editing. **Engin Zeydan:** Conceptualization, Writing – review & editing. **Jorge Baranda:** Methodology, Writing – review & editing. **Nikolaos Tzanis:** Investigation, Data curation. **Stamatiou Nikolaos:** Investigation, Data curation. **Nikolaos Kanakaris:** Investigation, Writing – review & editing. **Gianluca Rizzo:** Supervision, Conceptualization, Writing – review & editing.

References

3GPP TS 28.533, 2025. Management and Orchestration; Architecture framework (Release 19). Technical Report. 3rd Generation Partnership Project (3GPP).

Acun, B., Lee, B., Kazhamiaka, F., Maeng, K., Gupta, U., Chakkaravarthy, M., Brooks, D., Wu, C.J., 2023. Carbon explorer: A holistic framework for designing carbon aware datacenters, in: Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 118–132.

Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A., Flinck, H., 2018. Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. IEEE Communications Surveys & Tutorials 20, 2429–2453.

Alhashimi, H.F., Hindia, M.N., Dimyati, K., Hanafi, E.B., Safie, N., Qamar, F., Azrin, K., Nguyen, Q.N., 2023. A survey on resource management for 6g heterogeneous networks: Current research, future trends, and challenges. Electronics 12. URL: https://www.mdpi.com/2079-9292/12/3/647, doi:10.3390/electronics12030647.

Battisti, A.L.É., Delicato, F.C., Muchaluat-Saade, D.C., 2026. SPEED: A distributed approach for SFC placement in the edge-cloud continuum. Ann. Telecommun. doi:10.1007/s12243-026-01157-w.

Benzaıd, C., Taleb, T., 2020. Ai-driven zero touch network and service management in 5g and beyond: Challenges and research directions. IEEE Network 34, 186–194. doi:10.1109/MNET.001.1900252.

Blanco, L., Via, S., Vaca-Rubio, C.J., Zeydan, E., Murillo, L.M.C., Kedar, G., Antonio, G.R., Antonopoulos, A., Haxhibeqiri, J., Maglogiannis, V., Naudts, D., Hoebeke, J., de Cola, T., Giordano, L.G., Fontanesi, G., Spantideas, S., Koumaras, H., Neri, M., Vesco, A., Amatetti, C., Quijada, N.C.T., Dryjański, M., Kanakaris, N., Pamminger, C., Pagin, M., Ivanovic, F., 2025. Unity-6g: Unified architecture for open ranenabled distributed, scalable and sustainabilityenhanced 6g networks, in: 2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 440–445. doi:10.1109/EuCNC/6GSummit63408.2025.11036929.

Christophorou, C., Ioannou, I., Vassiliou, V., Christofi, L., Vardakas, J.S., Seder, E.E., Chiasserini, C.F., Iordache, M., Issaid, C.B., Markopoulos, I., Franzese, G., Järvet, T., Verikoukis, C., 2024. Adroit6g dai-driven open and programmable architecture for 6g networks. URL: https://arxiv.org/abs/2403.05277, arXiv:2403.05277.

CIGRE Task Force C6.04, 2014. Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources. Technical Brochure 575. CIGRE.

Csoma, A., Sonkoly, B., Csikor, L., Németh, F., Gulyás, A., Tavernier, W., Sahhaf, S., 2014. ESCAPE: Extensible service chain prototyping environment using mininet, click, NETCONF and POX, in: Proceedings of the 2014 ACM Conference on SIGCOMM (Demo), pp. 125–126. doi:10.1145/2619239.2631448.

Floridia, M., Laganà, D., Mastroianni, C., Meo, M., Renga, D., 2020. Load management with predictions of solar energy production for cloud data centers, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8986–8990. doi:10.1109/ICASSP40776.2020.9053186.

Givisiez, A.G., Petrou, K., Ochoa, L.F., 2020. A review on tso-dso coordination models and solution techniques. Electric Power Systems Research 189, 106659. URL: https://www.sciencedirect.com/science/article/pii/S0378779620304624, doi:https://doi.org/10.1016/j.epsr.2020.106659.

Hanafy, W.A., Liang, Q., Bashir, N., Irwin, D., Shenoy, P., 2023. Carbonscaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency. Proc. ACM Meas. Anal. Comput. Syst. 7. URL: https://doi.org/10.1145/3626788, doi:10.1145/3626788.

Kolenc, M., Nemček, P., Gutschi, C., Suljanović, N., Zajc, M., 2017. Performance evaluation of a virtual power plant communication system providing ancillary services. Electric Power Systems Research 149, 46–54.

Krishnamohan, T., Harvey, P., 2026. Energy-aware optimal service function chain embedding in emulated multi-access edge computing for internet of things, in: Proceedings of the 4th International Workshop on Testing Distributed Internet of Things Systems, Association for Computing Machinery, New York, NY, USA. p. 19–24. URL: https://doi.org/10.1145/3802513.3803485, doi:10.1145/3802513.3803485.

Liang, J., Huang, S., Qiu, Y., Liu, L., Aziz, F., Chen, M., 2024. Sustainable virtual network function placement and traffic routing for green mobile edge networks. IEEE Transactions on Green Communications and Networking 8, 1450–1465. doi:10.1109/TGCN.2024.3392813.

Lin, R., He, L., Luo, S., Zukerman, M., 2023. Energy-aware service function chaining embedding in nfv networks. IEEE Transactions on Services Computing 16, 1158–1171. doi:10.1109/TSC.2022.3162328.

López-Pérez, D., De Domenico, A., Piovesan, N., Xinli, G., Bao, H., Qitao, S., Debbah, M., 2022. A survey on 5G radio access network energy efficiency: Massive MIMO, lean carrier design, sleep modes, and machine learning. IEEE Communications Surveys & Tutorials 24, 653–697.

Lozano, J.X.S., Ayala-Romero, J.A., Garcia-Saavedra, A., Costa-Perez, X., 2025. Kairos: Energy-efficient radio unit control for o-ran via advanced sleep modes. URL: https://arxiv.org/abs/2501.15853, arXiv:2501.15853.

Nokia Bell Labs, 2020. The 3GPP-defined Service Based Management Architecture: A rich toolbox for 5G network and service automation solutions. White Paper. Nokia Bell Labs. Technical brief based on 3GPP Release 16.

Polese, M., Bonati, L., D'Oro, S., Basagni, S., Melodia, T., 2023. Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges. IEEE Communications Surveys & Tutorials 25, 1376–1411. doi:10.1109/COMST.2023.3239220.

Pudjianto, D., Ramsay, C., Strbac, G., 2007. Virtual power plant and system integration of distributed energy resources. IET Renewable Power Generation 1, 10–16.

Riaz, S., Mancarella, P., 2019. On feasibility and flexibility operating regions of virtual power plants and TSO/DSO interfaces, in: 2019 IEEE Milan PowerTech, pp. 1–6. doi:10.1109/PTC.2019.8810638.

Santos, G.L., Bezerra, D., Rocha, É., Ferreira, L., Moreira, A., Gonçalves, G., Marquezini, M., Recse, Á., Mehta, A., Kelner, J., Sadok, D., Endo, P., 2022. Service function chain placement in distributed scenarios: A systematic review. Journal of Network and Systems Management 30.

Schardong, F., Nunes, I., Schaeffer-Filho, A., 2021. NFV resource allocation: A systematic review and taxonomy of VNF forwarding graph embedding. Computer Networks 185, 107726.

SNS JU 6G Architecture Working Group, 2024. Towards 6G Architecture: Key Concepts, Challenges, and Building Blocks. White Paper. Smart

Networks and Services Joint Undertaking (SNS JU). 6G Architecture Working Group white paper on key concepts and enablers for 6G.

Sonkoly, B., Czentye, J., Szabo, R., Jocha, D., Elek, J., Sahhaf, S., Tavernier, W., Risso, F., 2015. Multi-domain service orchestration over networks and clouds: A unified approach, in: Proceedings of the 2015 ACM Conference on SIGCOMM (Demo). doi:10.1145/2785956.2790041.

Sukprasert, T., Souza, A., Bashir, N., Irwin, D., Shenoy, P., 2024. On the limitations of carbon-aware temporal and spatial workload shifting in the cloud, in: Proceedings of the Nineteenth European Conference on Computer Systems (EuroSys). doi:10.1145/3627703.3650079.

Sun, G., Li, Y., Yu, H., Vasilakos, A.V., Du, X., Guizani, M., 2019. Energy-efficient and traffic-aware service function chaining orchestration in multi-domain networks. Future Generation Computer Systems 91, 347–360. doi:https://doi.org/10.1016/j.future.2018.09.037.

Taleb, T., Afolabi, I., Samdanis, K., Yousaf, F.Z., 2019. On multi-domain network slicing orchestration architecture and federated resource control. IEEE Network 33, 242–252.

Thurner, L., Scheidler, A., Schäfer, F., Menke, J.H., Dollichon, J., Meier, F., Meinecke, S., Braun, M., 2018. pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. IEEE Transactions on Power Systems 33, 6510–6521. doi:10.1109/TPWRS.2018.2829021.

Trichias, K., Kaloxylos, A., Willcock, C., 2024. 6g global landscape: A comparative analysis of 6g targets and technological trends, in: 2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 1–6. doi:10.1109/EuCNC/6GSummit60053.2024.10597064.

Uusitalo, M.A., Farhadi, H., Boldi, M., Ericson, M., Fettweis, G., Khorsandi, B.M., Pavón, I.L., Mueck, M., Nimr, A., Pärssinen, A., Richerzhagen, B., Rugeland, P., Sabella, D., Wymeersch, H., 2025. Toward 6g — hexa-x project's key findings. IEEE Communications Magazine 63, 142–148. doi:10.1109/MCOM.003.2300735.

Vantzos, O., Skipis, D., Chassioti, E., Moraitis, I., 2025. Green index: A measure of locally available green energy for sustainable, grid-aware data center operations, in: 2025 IEEE International Conference on Environment and Electrical Engineering and 2025 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe). doi:10.1109/EEEIC/ICPSEUROPE64998.2025.11169146.

Wang, F., Zhang, S., Yang, H., Quek, T.Q., 2025. Non-terrestrial networking for 6g: Evolution, opportunities, and future directions. Engineering URL: https://www.sciencedirect.com/science/article/pii/S2095809925002917, doi:https://doi.org/10.1016/j.eng.2025.05.013.

Wu, J., Zhang, Y., Zukerman, M., Yung, E.K.N., 2015. Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey. IEEE Communications Surveys & Tutorials 17, 803–826. doi:10.1109/COMST.2015.2403395.

Zajc, M., Kolenc, M., Suljanovic, N., 2019. Virtual power plant communication system architecture, in: Smart Power Distribution Systems. Elsevier, pp. 231–250. doi:10.1016/B978-0-12-812154-2.00011-0.

Zeydan, E., Blanco, L., Batewela, S., Ranaweera, P., Liyanage, M., 2025. Advances in network orchestration for security, scalability, and sustainability, in: 2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 399–404. doi:10.1109/EuCNC/6GSummit63408.2025.11037064.