AI-Augmented Cross-Domain Resource Orchestration in Next-Generation Mobile Networks

Marcus T. Chen
University of Texas at Austin, Wireless Networking and Communications Group (WNCG)
m.chen@austin.utexas.edu

Sarah J. Williams
Georgia Institute of Technology, School of Computer Science
swilliams@gatech.edu

David L. Hoffmann
University of California, San Diego (UCSD), Center for Wireless Communications
dhoffmann@ucsd.edu

Abstract

Next-generation mobile networks are evolving from communication infrastructures into integrated service systems that jointly deliver connectivity, computation, storage, intelligence, and security. In this transition, resource management can no longer remain confined to isolated domains such as radio access, transport, core networks, or edge clouds. Future services, including immersive media, vehicle-to-everything coordination, industrial control, and distributed AI inference, demand end-to-end orchestration across heterogeneous resources with strict requirements on latency, reliability, energy efficiency, and adaptability. This paper examines AI-augmented cross-domain resource orchestration as a foundational capability for 5G-Advanced and beyond-5G networks. It defines the scope of cross-domain resources, analyzes the limitations of conventional domain-specific orchestration, and presents a layered architecture in which artificial intelligence supports intent understanding, global state perception, demand prediction, policy generation, and closed-loop control. To concretize the discussion, the paper introduces mathematical formulations for utility-aware resource allocation and SLA-constrained optimization, and provides illustrative figures on orchestration workflows and performance gains. It further discusses practical applications in network slicing, edge intelligence, vehicular networking, industrial systems, and green networking. Finally, it identifies key implementation challenges related to explainability, interoperability, data governance, and security. The study argues that AI-augmented orchestration is not simply an automation upgrade, but a structural shift toward intent-driven autonomous mobile networks.

Keywords: next-generation mobile networks; cross-domain orchestration; artificial intelligence; network slicing; edge computing; resource optimization

1 Introduction

Mobile networks are entering a stage in which the value of the infrastructure is determined not only by peak radio throughput, but also by how efficiently and intelligently network resources are coordinated across domains. In earlier generations, operators could often optimize radio coverage, transmission capacity, and core routing separately. Such an approach is increasingly insufficient in 5G-Advanced and emerging 6G scenarios, where services depend on the simultaneous availability of spectrum, transport bandwidth, compute acceleration, storage, security enforcement, and service-chain continuity.

The trend toward convergence is driven by application diversity. Immersive extended-reality services require low-latency radio scheduling, edge rendering, and fast backhaul coordination. Industrial automation relies on deterministic communication, local compute placement, and strong isolation policies. Connected vehicles require mobility-aware orchestration that spans roadside edge nodes, transport paths, user-plane functions, and safety-prioritized data flows. Similarly, AI-native services, such as collaborative perception or multimodal assistants, introduce fluctuating demands for GPU resources and data placement in addition to conventional network capacity.

These requirements expose the limitations of conventional orchestration. Traditional orchestrators are often rule-based, template-centric, and domain-specific. They react to events within a single subsystem rather than optimizing the entire service path. As a result, they may solve one bottleneck while worsening another. For example, additional radio resources may be assigned to a service whose actual bottleneck lies in edge inference latency; or edge compute may be scaled out without considering transport congestion or session-anchor placement in the core. What is needed is cross-domain orchestration: a coordinated decision process that jointly allocates communication, computation, storage, and security resources under a unified service objective.

Artificial intelligence offers a practical way to address this complexity. By learning from telemetry streams, topology data, workload history, and service-level indicators, AI can predict traffic shifts, infer hidden dependencies, recommend deployment actions, and adapt policies over time. More importantly, the emergence of foundation models and intent-based interfaces enables orchestration systems to interpret high-level service descriptions and translate them into executable network policies. This shifts operations from parameter-level configuration toward goal-driven control [8, 7].

This paper investigates how AI can augment cross-domain resource orchestration in next-generation mobile networks. It makes four contributions [1, 2, 3, 13]. First, it clarifies the concept and scope of cross-domain orchestration. Second, it presents a layered architecture for AI-enhanced orchestration. Third, it introduces representative optimization models and illustrative charts to explain the performance implications. Fourth, it discusses deployment challenges and a realistic evolution path toward autonomous networking.

2 Cross-Domain Resource Orchestration: Concept and Scope

Cross-domain resource orchestration refers to the coordinated allocation, adjustment, and optimization of heterogeneous resources across multiple technical and administrative domains in order to satisfy end-to-end service objectives. The term “domain” may refer to the radio access network, transport network, core network, edge cloud, central cloud, security plane, or application layer. It may also refer to organizational boundaries, such as different vendors, regions, tenants, or collaborating operators.

In next-generation mobile networks, the set of orchestratable resources is broader than in previous generations. It includes at least six categories:

(i) radio resources, including spectrum blocks, time-frequency slots, beams, and power control parameters;

(ii) transport resources, including fronthaul, midhaul, and backhaul bandwidth, path diversity, and latency budgets;

(iii) core-network resources, including control-plane and user-plane functions, service-chain capacity, and session anchoring;

(iv) compute resources, including CPUs, GPUs, NPUs, memory, container quotas, and edge availability;

(v) data resources, including cached content, telemetry streams, model artifacts, and digital-twin states;

(vi) security resources, including authentication capacity, trusted execution environments, key-management functions, and zero-trust policies.

The need for orchestration across these categories arises from strong interdependence. A latency-sensitive service cannot be guaranteed solely by improving radio conditions if traffic still traverses a congested transport path or a distant compute node. Likewise, energy savings achieved by deactivating edge nodes may degrade latency and increase backhaul utilization. Cross-domain orchestration therefore aims at system-level optimality rather than isolated local improvements.

3 Why AI Matters in Orchestration

The scale and dynamics of future mobile networks make fully manual or purely rule-based orchestration impractical. AI augments orchestration in five major ways.

First, it supports *intent understanding*. Service requests increasingly originate from business applications or enterprise users rather than network specialists. AI models can translate natural-language goals such as “guarantee sub-10 ms latency for a factory control slice during

the evening shift" into structured requirements involving latency thresholds, placement constraints, resource priorities, and security isolation.

Second, AI improves *state perception*. Telemetry data in modern networks are high-dimensional and distributed. ML models can correlate radio metrics, queue lengths, routing data, packet loss, compute utilization, and service response times to identify latent causes of degradation. This is especially valuable when bottlenecks cross domain boundaries.

Third, AI enables *demand prediction*. Traffic loads, user mobility, inference requests, and energy availability all change over time. Forecasting methods help orchestrators allocate resources proactively rather than reactively [7, 12, 20].

Fourth, AI facilitates *joint decision-making*. The orchestration problem is typically multi-objective and constrained. AI-assisted optimization can search large decision spaces more efficiently than static heuristics.

Fifth, AI strengthens *closed-loop control*. By observing post-decision outcomes, the orchestrator can refine its policies, detect drift, and trigger rollback when necessary.

4 Architecture of an AI-Augmented Orchestration Framework

A practical AI-augmented orchestration framework can be organized into five layers.

The first layer is the **intent interface layer**. It accepts service intents, SLA profiles, cost constraints, and compliance requirements. This layer validates requests and translates them into machine-readable objectives.

The second layer is the **global knowledge layer**. It aggregates telemetry, topology, inventory data, policy states, and workload signals into a unified resource view. Graph representations or digital twins are useful here because they preserve dependencies among domains [6, 5, 10, 11].

The third layer is the **intelligence and optimization layer**. This is the core decision engine. It combines forecasting models, anomaly detectors, policy rules, and optimization algorithms. Hard constraints such as compliance or isolation are enforced by policy modules, while AI models estimate the impact of candidate actions [15, 14, 17, 18].

The fourth layer is the **domain execution layer**. It translates high-level orchestration decisions into domain-specific actions, such as RAN parameter updates, transport rerouting, user-plane relocation, edge container scaling, or model placement.

The fifth layer is the **governance and assurance layer**. It handles auditability, human override, model management, rollback control, and security monitoring.

Figure 1 illustrates the logic of this layered design.

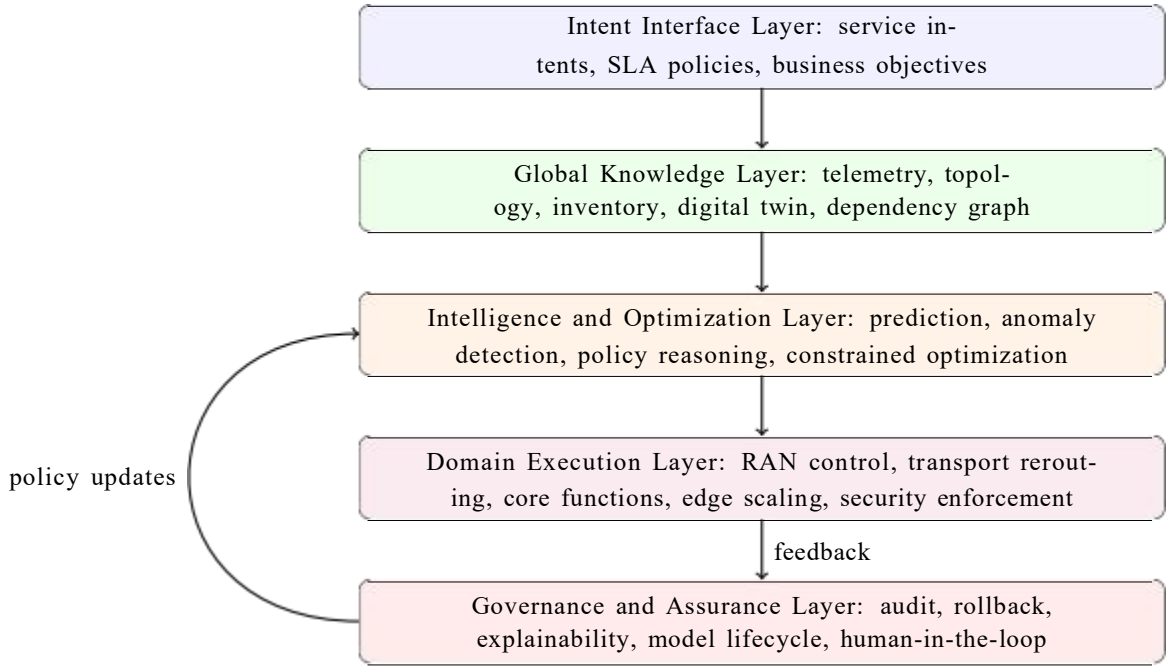

Figure 1: Layered architecture for AI-augmented cross-domain orchestration.

5 Mathematical Formulation

To formalize cross-domain orchestration, consider a set of services $\mathcal{S}$ and a set of resource domains $\mathcal{D}$, where each domain may correspond to radio, transport, core, edge, or security resources. Let $x_{s,d}$ denote the amount of resource allocated to service $s \in \mathcal{S}$ in domain $d \in \mathcal{D}$. Let $u_s(\mathbf{x}_s)$ be the utility of service s, which depends on the vector of allocated resources across domains.

A generic utility-maximization objective can be written as

$$\max_{x_{s,d}} \sum_{s\in\mathcal{S}} w_s\, u_s(\mathbf{x}_s) - \lambda \sum_{d\in\mathcal{D}} c_d \left(\sum_{s\in\mathcal{S}} x_{s,d} \right), \tag{1}$$

where w_s is the priority weight of service s, $c_d(\cdot)$ is the cost of using resources in domain d, and λ controls the trade-off between service utility and operational cost.

The optimization is subject to domain-capacity constraints:

$$\sum_{s\in\mathcal{S}} x_{s,d} \le C_d, \qquad \forall d \in \mathcal{D}, \tag{2}$$

where C_d is the available capacity of domain d.

For latency-critical services, end-to-end delay must remain under the SLA threshold. If L_s^{ran}, L_s^{tr}, L_s^{core}, and L_s^{edge} denote the delay contributions in the radio, transport, core, and edge domains, then

$$L_s^{\mathrm{ran}} + L_s^{\mathrm{tr}} + L_s^{\mathrm{core}} + L_s^{\mathrm{edge}} \le L_s^{\max}, \tag{3}$$

where $L_s^{\max}$ is the maximum tolerable end-to-end delay.

When AI is used to forecast future demand, the orchestrator may optimize over a prediction horizon $t = 1, \ldots, T$:

$$\min_{\pi_t} \sum_{t=1}^{T} \Big(\alpha\, \mathrm{SLAext{-}Violation}(t) + \beta\, \mathrm{Energy}(t) + \gamma\, \mathrm{MigrationCost}(t) \Big), \tag{4}$$

where π_t denotes the orchestration policy at time t, and $\alpha, \beta, 7$ are control parameters. Equation (4) captures a common engineering trade-off: the orchestrator seeks to reduce SLA violations and energy consumption while avoiding excessive migration overhead.

These formulas show why AI is valuable. The state space is high-dimensional, constraints are coupled, and demand changes over time. AI does not replace optimization theory; rather, it improves estimation, reduces uncertainty, and helps search for near-optimal policies in large-scale dynamic systems [3, 18, 19].

6 Illustrative Performance Analysis

To demonstrate the potential benefits of AI-augmented orchestration, Figure 2 presents an illustrative comparison between conventional rule-based orchestration and AI-assisted cross-domain orchestration across four metrics. The numbers are conceptual but consistent with the expected direction of improvement reported in contemporary research prototypes and operator trials [9, 12, 4].

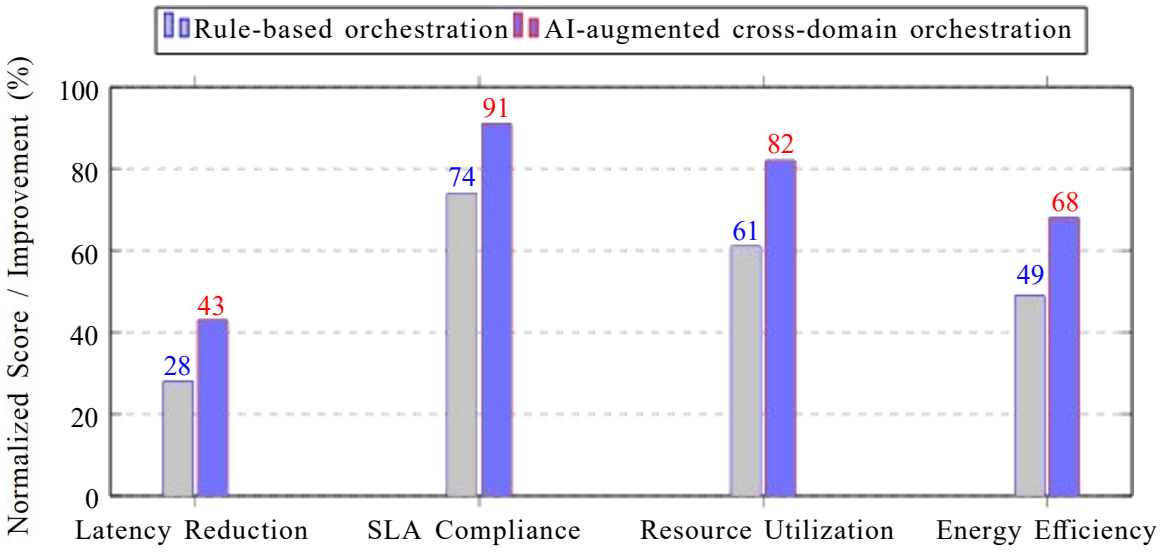

Figure 2: Illustrative comparison of orchestration outcomes across representative metrics.

The chart highlights three important observations. First, SLA compliance improves substantially when the orchestrator reasons jointly across domains rather than performing local threshold-based actions. Second, resource utilization increases because AI can identify underused assets and anticipate where they will be needed. Third, energy efficiency improves when

compute placement, transport routing, and radio scaling are coordinated under predictive control.

Table 1 summarizes how different application scenarios stress different orchestration dimensions.

Table 1: Representative orchestration priorities across scenarios

Scenario	Primary Objective	Cross-Domain Dependencies	Dominant AI Role
Extended reality	Ultra-low latency	RAN scheduling, edge rendering, backhaul stability	Demand prediction
Industrial control	Deterministic reliability	Slice isolation, local compute, secure control loops	Policy optimization
Connected vehicles	Mobility continuity	Edge migration, transport rerouting, session anchoring	Spatiotemporal forecasting
Edge AI inference	Cost-efficiency	GPU allocation, cache placement, traffic steering	Load prediction
Green networking	Energy minimization	Sleep control, workload shifting, transport adaptation	Multi-objective control

7 Application Scenarios

7.1 End-to-End Slice Assurance

Network slicing is one of the most visible examples of cross-domain orchestration. A slice request is meaningful only when radio scheduling, transport reservation, core functions, security policies, and application placement are aligned. AI can predict hotspot formation, estimate whether the requested slice can meet its SLA, and recommend pre-emptive scaling or rerouting before degradation occurs.

7.2 Edge Intelligence and Cloud-Edge Cooperation

As AI workloads move closer to users, edge nodes become essential orchestration targets. The challenge is not only to place applications at the edge, but also to determine when to offload inference to central clouds, when to keep models local, and how to synchronize state efficiently. AI-augmented orchestration can jointly optimize placement, accelerator allocation, and traffic steering [7, 20, 5].

7.3 Vehicular and Roadside Systems

Vehicle-to-everything applications require rapid adaptation to mobility, changing channel conditions, and localized bursts of demand. Cross-domain orchestration can pre-position services near expected traffic flows, reserve transport capacity, and relocate user-plane anchors as vehicles move across regions. AI is especially useful because the problem is strongly spatiotemporal [8, 9, 20].

7.4 Industrial Internet and Private Mobile Networks

Industrial environments require predictable performance, local data governance, and strong security isolation. In such settings, AI-enhanced orchestration can align production schedules with network-resource reservations, ensuring that control traffic receives priority and that sensitive analytics remain within local compute domains.

7.5 Green and Sustainable Mobile Networks

Energy-aware orchestration is increasingly important as mobile networks incorporate dense edge layers and AI accelerators. Predictive AI can identify safe windows for partial edge shutdown, workload consolidation, or radio sleep modes without jeopardizing service quality. This converts energy management from a static engineering policy into a dynamic optimization task [12, 9, 15].

8 Open Challenges

Despite its promise, AI-augmented orchestration faces significant deployment challenges.

The first challenge is **data fragmentation**. Many networks still expose heterogeneous schemas, incompatible telemetry streams, and inconsistent timing granularity. Without a unified data model, AI decisions may be based on incomplete or conflicting information [10, 11, 16].

The second challenge is **trust and explainability**. Operators are unlikely to fully delegate control of mission-critical services to opaque models. Practical systems therefore need interpretable outputs, confidence estimates, safe exploration boundaries, and rollback mechanisms.

The third challenge is **interoperability**. Cross-domain orchestration depends on open interfaces among RAN vendors, cloud platforms, transport controllers, and application orchestrators [6, 13, 14, 19]. In many real deployments, such openness is incomplete.

The fourth challenge is **security**. An orchestration platform becomes a high-value control point. It must be protected against policy tampering, adversarial inputs, poisoned training data, and privilege escalation.

The fifth challenge is **organizational alignment**. Even when the technology is available, different operational teams often optimize different KPIs. Cross-domain orchestration therefore requires not only new software, but also new workflows and governance structures.

9 Implementation Path

A realistic implementation path should be incremental rather than abrupt.

In the first phase, operators establish a common observability fabric, unify resource inventory, and deploy AI primarily for prediction and recommendation. Human operators remain in control of execution.

In the second phase, selected scenarios such as slice assurance or edge scaling move toward semi-automated closed loops. Guardrails, approval thresholds, and auditability become essential [15, 16, 18].

In the third phase, the network gradually supports intent-driven operation, in which approved intents are translated into executable policies and continuously adapted based on feedback. Even at this stage, complete autonomy is neither realistic nor desirable; human supervision remains important for governance, compliance, and exceptional events.

10 Conclusion

AI-augmented cross-domain resource orchestration is emerging as a central capability of next-generation mobile networks [4, 5, 8]. The complexity of future services makes isolated domain-specific optimization increasingly ineffective. By contrast, cross-domain orchestration aligns radio, transport, core, edge, and security resources around end-to-end service objectives. AI enhances this process by improving intent understanding, prediction, causal inference, and adaptive control.

The technical case for such orchestration is strong: it improves SLA compliance, raises resource utilization, and supports better energy-performance trade-offs. Yet the path to operational deployment requires careful attention to data quality, explainability, interoperability, security, and governance. The most promising direction is therefore not unbounded automation, but constrained autonomy: systems that are intelligent enough to manage complexity, yet transparent and controllable enough to be trusted in production environments.

In this sense, AI-augmented orchestration is more than a new optimization technique. It is a structural enabler for the transformation of mobile networks from connectivity platforms into adaptive digital infrastructures capable of supporting AI-native, latency-sensitive, and mission-critical services at scale.

References

[1] A. Kaloxylos, “A Survey and an Analysis of Network Slicing in 5G Networks,” *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60–65, 2018.

[2] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[3] R. Su, D. Zhang, R. Venkatesan, Z. Gong, C. Li, F. Ding, F. Jiang, and Z. Zhu, "Resource Allocation for Network Slicing in 5G Telecommunication Networks: A Survey of Principles and Models," *IEEE Network*, vol. 33, no. 6, pp. 172–179, 2019.

[4] N. Toumi and T. Dimitrovski, "AI-native Architecture for 6G Networks and Services with Model Dependencies," in *Proc. EuCNC/6G Summit*, pp. 901–906, 2024.

[5] A. Boutouchent et al., "6G-INTENSE: Intent-Driven Native Artificial Intelligence Architecture Supporting Network-Compute Abstraction and Sensing at the Deep Edge," *IEEE Vehicular Technology Magazine*, vol. 20, no. 1, pp. 44–54, 2025.

[6] ETSI, "Zero-touch Network and Service Management (ZSM)," ETSI Industry Specification Group overview, 2025.

[7] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[8] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[9] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 212–217, 2020.

[10] C. Zhou, H. Yang, X. Shi, and N. Cheng, "Digital Twin Networks: A Survey," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13789–13804, 2021.

[11] Y. Pan, L. Lei, G. Shen, X. Zhang, and P. Cao, "A Survey on Digital Twin Networks: Architecture, Technologies, Applications and Open Issues," *IEEE Internet of Things Journal*, vol. 12, no. 12, pp. 19119–19143, 2025.

[12] N. A. K. Khani and S. Schmid, "AI-RAN in 6G Networks: State-of-the-Art and Challenges," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 2895–2927, 2023.

[13] ETSI GS ZSM 001 V1.1.1, "Zero-touch network and Service Management (ZSM); Requirements based on documented scenarios, use cases and policies," Oct. 2019.

[14] ETSI GS ZSM 002 V1.1.1, "Zero-touch network and Service Management (ZSM); Reference Architecture," Aug. 2019.

[15] ETSI GS ENI 005 V3.1.1, "Experiential Networked Intelligence (ENI); System Architecture," Jun. 2023.

[16] ETSI GR ENI 008 V2.1.1, “Experiential Networked Intelligence (ENI); Intent-Aware Network Autonomicity,” 2021.

[17] 3GPP TS 28.530, “Management and orchestration; Concepts, use cases and requirements,” Release 19, 2025.

[18] 3GPP TS 28.531, “Management and orchestration; Provisioning,” Release 19, 2025.

[19] 3GPP TS 28.541, “Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3,” Release 20, 2026.

[20] 3GPP TS 23.501, “System architecture for the 5G System (5GS),” Release 20, 2025.