Engineering 56 (2026) 87–103

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
6G: From Theory to Practice—Article

A Task-Driven Design Approach for 6G AI-Native Architecture

Xiaoyun Wang [a], Lu Lu [b], Qin Li [b], Qi Sun [b], Nanxiang Shi [b], Ziqi Chen [b], Tao Sun [b,*]

[a] China Mobile Communications Group Corporation, Beijing 100033, China
[b] China Mobile Research Institute, Beijing 100053, China

ARTICLE INFO

Article history:
Received 30 November 2024
Revised 17 July 2025
Accepted 8 September 2025
Available online 17 September 2025

Keywords:
Sixth-generation (6G) mobile network
Artificial intelligence
6G AI-native architecture
Task-driven approach

ABSTRACT

The deep integration of mobile networks with artificial intelligence (AI) has emerged as a pivotal driving force for the sixth-generation (6G) mobile network. AI-native 6G represents a paradigm shift for mobile networks, as it not only embeds AI into network components to enhance network intelligence and automation but also transforms 6G into a foundational infrastructure for enabling pervasive AI applications and services. This paper proposes a novel 6G AI-native architecture. The challenges and requirements for the AI-native 6G mobile network are first analyzed, followed by the development of a task-driven approach for architecture design based on insights from system theory. Then, a 6G AI-native architecture is proposed, featuring the integration of distributed AI data and computing components with layered centralized collaborative control and flexible on-demand deployment. Key components and procedures for the 6G AI-native architecture are also discussed in detail. Finally, standardization practices for the convergence of mobile networks and AI in fifth-generation (5G) networks are analyzed, and an outlook on the standardization of AI-native design in 6G is given. This paper aims to provide not only theoretical insights into AI-native architecture design methodology but also a comprehensive 6G AI-native architecture that lays a foundation for the transition from mobile communications toward mobile information services in the 6G era.

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The latest release of the IMT-2030 framework by the International Telecommunication Union Radiocommunication Sector (ITU-R) defines two categories of core sixth-generation (6G) scenarios. The first category builds upon three established fifth-generation (5G) scenarios: immersive communication, massive communication, and hyper-reliable low-latency communication. The second category encompasses three emerging fields: integrated sensing and communication, artificial intelligence (AI) and communication, and ubiquitous connectivity [1]. Additionally, in its latest work item proposal for 6G requirements, the 3rd Generation Partnership Project (3GPP) proposed that a fundamental requirement for 6G is the inherent integration of mobile communication networks with AI [2]. This integration, which began in 5G, aims to enhance network intelligence and automation by applying AI technologies to the radio access network (RAN), the core network (CN), and network management for analysis, prediction, and optimization [3,4]. To facilitate this goal, 3GPP specifies a network data analytics function (NWDAF) and a management data analytics function (MDAF) within the CN and management domains, respectively [5,6].

As society transitions from the Information Age to the Intelligence Age, ubiquitous AI applications and intelligent agents will impose higher demands on pervasive connectivity and intelligent computing capabilities [7]. With the vision of ubiquitous intelligence for 2030 and beyond, mobile communication networks will serve as a crucial foundational infrastructure in facilitating and supporting the development of AI services, thereby making AI a universal service at the societal level. Compared with that in 5G, the convergence of mobile networks with AI in the 6G era will further deepen integration, elevate intelligence level, and expand service scenarios, which means that AI will enable the 6G network to achieve native intelligence, while the 6G network itself will enable ubiquitous and inclusive AI services. Both aspects will mutually reinforce each other to fulfill the bidirectional integration and development of 6G and AI, finally achieving the ultimate goal of AI-native 6G.

The industry has a general recognition that 6G will deeply integrate with AI/machine learning (ML) to improve network

* Corresponding author.
E-mail address: suntao@chinamobile.com (T. Sun).

https://doi.org/10.1016/j.eng.2025.09.005
2095-8099/© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

self-management and performance metrics and act as a platform enabler for AI/ML to provide services. China's IMT-2030 (6G) Promotion Group has stated that the AI capabilities of 6G networks extend beyond self-operation and maintenance to include providing AI services for users or external third parties. With these AI-native functionalities, 6G networks can deliver on-demand customized AI services with high quality-of-service (QoS) guarantees through an orchestration and management system integrating AI with other connectivity elements [8]. Moreover, the European 6G Alliance Hexa-X has proposed an AI-native air–interface design that enables the native integration of AI and ML in 6G. This integration supports flexible, low-complexity, and reconfigurable networks (i.e., networks that can learn to communicate), as well as intrinsic in-network intelligence features (i.e., communicating for learning or 6G as an efficient AI/ML platform). The integration of AI in mobile networks for internal network operations (e.g., resource-allocation optimization) and external entities has been considered through an AI-as-a-service (AIaaS) approach [9–11]. The Next G Alliance, in North America, has proposed that the integrations of AI/ML into networks, computing architecture for AI/ML, datasets, and AI/ML security are key technology directions in AI-native wireless networks [12]. Ericsson defines AI-native as "the concept of embedding intrinsic trustworthy AI capabilities into systems, where AI is a natural part of the functionality, in terms of design, deployment, operation, and maintenance." This is achieved by planning AI-native architectures from the inception of new products. Key aspects of an AI-native architecture include ① pervasive intelligence, ② a distributed data infrastructure, ③ zero-touch automation, and ④ AIaaS [13]. Overall, there is a consensus in the industry regarding the necessity for deep AI/ML integration in 6G networks to improve network self-management and performance metrics and for the 6G mobile network to act as an efficient platform to provide AIaaS. With the progressive maturation of generative pre-trained transformers (GPTs) and large language models (LLMs) in recent years, burgeoning research efforts have emerged exploring their integration into network systems. The European Telecommunications Standards Institute (ETSI)'s ongoing research project, *Study on AI Agents-Based Next-Generation Network Slicing*, proposes an AI-CN, which utilizes multiple AI agents to manage and control the slicing network and flexibly process data for new services based on the dynamic requirements of various applications [14]. While leveraging foundation-model-powered AI agent technology to realize telecom AI-native systems represents a promising direction, significant challenges remain in constructing robust foundational models tailored for telecom-specific contexts. For instance, unreliable AI models could disrupt complex network interactions, and even marginal performance degradation remains unacceptable to communication service providers (CSPs) [15].

AI-native networks have also gained significant attention in academia. Wu et al. [16] proposed an AI-native network slicing framework for 6G, considering AI as a built-in component in the network slicing architecture. Hossain et al. [17] proposed an AI-native design for optimizing 6G CN configuration, significantly reducing execution time. Qiu et al. [18] discussed challenges and potential solutions for 6G network operation and management and proposed an AI-empowered network root cause analysis (Net-RCA) framework for 6G. Jiang et al. [19,20] proposed CommLLM, a multi-agent system for solving communication-related tasks using natural language, and validated its effectiveness by designing a semantic communication system. Xu et al. [21] proposed a framework for wireless network management based on LMMs. This framework integrates multimodal data fusion, retrieval-augmented generation (RAG), reinforcement learning (RL), and neuro-symbolic AI, aiming to enable intent-driven automation. Tarkoma et al. [22] proposed an AI-native interconnect framework for integrating LLMs into 6G systems. This framework leverages advanced AI/ML models coupled with data-driven analytics to dynamically optimize network resource allocation.

Current research primarily targets specific objectives, scenarios, and individual technologies, lacking a holistic framework addressing architectural-level challenges. The seamless deep integration of AI and communication requires architectural-level design, and an exploration of architecture design methodology is crucial to achieve this goal. To address these issues, this paper systematically analyzes the challenges and requirements in designing a 6G AI-native architecture, including capability generalization, quality assurance, efficiency optimization, and achieving a balance of the three. It then proposes a task-driven architecture design methodology grounded in system theory. The paper further details a comprehensive design for a 6G AI-native architecture.

The rest of this paper is organized as follows: Section 2 analyzes the evolution trends of mobile network architecture and identifies the requirements for designing a 6G AI-native architecture. A task-driven architecture design approach is presented in Section 3. Section 4 proposes an overall 6G AI-native architecture, analyzes its advantages, and provides experimental verification. Section 5 discusses the standardization progress and highlights standardization directions for the 6G AI-native architecture. Section 6 concludes the paper.

2. Requirements for a 6G AI-native architecture

2.1. The evolution of mobile network system architecture

According to system theory, a system is an integrated whole composed of interacting and independent elements with specific structures and functions, often existing within a broader system [23]. In the context of mobile communication systems, this principle underpins network architecture design, which integrates the user equipment (UE), RAN, CN, and transport network into a cohesive whole. Mobile communication systems feature large-scale deployment to efficiently support mobile communication tasks through their design elements and interconnections. These elements include both logical functions and physical resources, while interconnections refer to procedure, protocol, and distributed topology. The objectives of these tasks are to ensure that services are delivered consistently, efficiently, and sustainably.

The architecture of mobile networks plays a pivotal role in determining resource-utilization efficiency and supported service types for each generation. It defines how various network elements are connected, how end-to-end services are established, and how the network infrastructure is deployed. As a result, its development has become one of the primary indicators of the generational evolution of mobile networks [24]. Fig. 1 maps the generational leaps in mobile communications, tracing technological milestones from the first generation (1G)'s analog roots to 6G's AI-driven horizons. While both 1G and second-generation (2G) networks centered on voice service/short message service (SMS), their defining divergence emerged through 2G's digital modulation schemes, which replaced 1G's frequency-based analog transmissions with time-division multiplexing (TDM). The transition from 2G to third-generation (3G) networks marked the first evolutionary leap in mobile service paradigms, pivoting from circuit-switched (CS) domains dedicated to voice service/SMS to integrated systems supporting packet-switched (PS) domains for Internet access and multimedia delivery. On the architectural level, this transition established foundational control/user plane separation. The transition from 3G to fourth-generation (4G) networks retained the same service suite while engineering an architectural metamorphosis. Voice services transitioned to Internet protocol (IP) multimedia

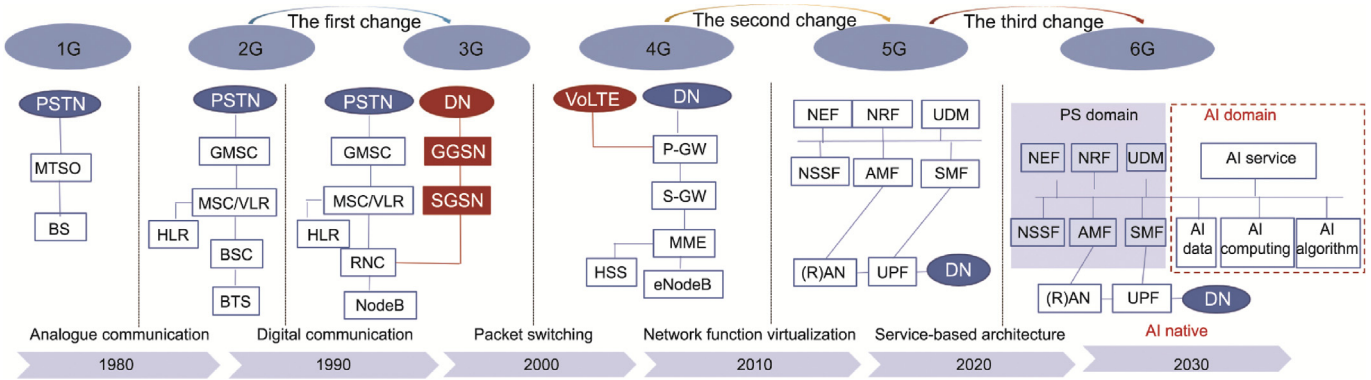

Fig. 1. Evolution of mobile communication systems. 1G: first-generation; 2G: second-generation; 3G: third-generation; 4G: fourth-generation; PSTN: public switched telephone network; MTSO: mobile telephone switching office; BS: base station; GMSC: gateway mobile switching center; MSC: mobile switching center; VLR: visitor location register; HLR: home location register; BSC: base station controller; BTS: base transceiver system; RNC: radio network controller; DN: data network; GGSN: gateway general packet radio service (GPRS) support node; SGSN: serving GPRS support node; VoLTE: voice over long-term evolution; P-GW: P-gateway; S-GW: S-gateway; MME: mobility management entity; HSS: home subscriber service; eNodeB: evolved NodeB; NEF: network exposure function; NRF: network repository function; NSSF: network slice selection function; AMF: access management function; UDM: unified data management; SMF: session management function; (R)AN: (radio) access network; UPF: user plane function; PS: packet-switched.

subsystem (IMS) via voice over long-term evolution (VoLTE) implementation, while all traffic types—voice, video, and data streams—were consolidated over a unified IP transport layer through the evolved packet core (EPC). This architectural overhaul concurrently pioneered network function virtualization (NFV). The 4G-to-5G transition marks the second evolutionary leap in mobile service paradigms, pivoting from catering to individual users to serving industry users. This service expansion necessitated a fundamental architectural reinvention through 3GPP's service-based architecture (SBA), which deconstructs conventional network elements into cloud-native, service-oriented microservices [25].

Future mobile networks will not only act as bridges connecting every part of daily life but also serve as a comprehensive service platform integrating multidimensional functions such as connectivity, AI, and sensing, thus giving rise to the emergence of new elements and service demands that bring together communication, sensing, computing, intelligence, and security.

The 5G-to-6G transition represents the third evolutionary leap in mobile service paradigms, introducing an AI domain as a fundamental stratum alongside the PS domain. It will further expand connection-oriented services to include AI-driven services. This integration will address the demands of pervasive AI services in the future, making 6G not only a communication platform but also an intelligent service platform. To support these requirements, a novel 6G network architecture must be devised.

2.2. Design requirements for a 6G AI-native architecture

A 6G AI-native architecture should be designed to solve two categories of scenarios in a unified manner: AI-empowered networks and network-enabled AI. The core objective of AI-empowered networks is to leverage AI capabilities to improve network performance, optimize resource utilization, and enable the intelligent automation of network operations. This scenario positions AI as a productive element, driving innovations to overcome bottlenecks in efficiency and scalability. In this context, the role of AI is categorized as "production-oriented AI." In contrast, the core objective of network-enabled AI emphasizes the role of 6G networks as a foundational infrastructure for delivering high-quality, ubiquitous AI services. By addressing the demands of pervasive AI applications at the societal level, this scenario envisions the 6G network facilitating a paradigm shift in network services. The role of AI in this case is categorized as "service-oriented AI."

This paper proposes a 6G AI-native architecture with built-in AI elemental capabilities in computing, algorithms, and data. By embedding these fundamental AI enablers into the architectural fabric, the flexibility and efficiency of network AI services are significantly increased. In this regard, the four key challenges encountered in the design of a 6G AI-native architecture are analyzed below.

2.2.1. Capability generalization

The native AI in the 6G network should simultaneously support diversified AI services and meet extreme AI performance requirements. From the perspective of AI-empowered networks, system-level AI capabilities, such as spectrum efficiency, connection density, and channel information acquisition, require latency on the order of milliseconds or even microseconds. However, network-level AI capabilities, such as network energy conservation, resource allocation, and dynamic routing, require latency in the range of seconds. Similarly, service-level AI capabilities, such as resource scheduling optimization and service orchestration optimization, also require second-level latency. Thus, these production-oriented AI elements should focus on the highly efficient generation of AI production elements, such as AI data and models, and address the issue of adaptive AI task generation.

On the other hand, AI-native networks serving external applications encounter equally diverse requirements for their varied service objects—such as ubiquitous AI applications and AI agents—which exhibit significantly different service models and performance requirements [26]. For example, intelligent robots require an intelligent decision-making latency below 10 ms, while digital twins require a virtual reality (VR)-rendering latency below 15 ms and data rates exceeding 1 gigabits per second (Gbps). The complexity increases further for AI agents, which must integrate sensing, control, and execution. Depending on these diverse requirements, AI services may be delivered through service layers such as infrastructure as a service (IaaS), platform as a service (PaaS), or software as a service (SaaS). These service-oriented AI elements should focus on the pooling and flexible supply of computing, data, and AI model resources, the integration of services, and on-demand QoS. Therefore, the dual requirements of production-oriented and service-oriented AI create significant challenges for capability generalization in the design of the 6G AI-native architecture.

2.2.2. Quality assurance

The inherent uncertainty of AI poses a significant challenge to its integration into mobile networks, which demand high reliability and deterministic performance. This uncertainty stems from two primary factors: data and models. The effectiveness of AI inference and training heavily relies on the volume and quality of data. However, the limited availability of high-quality datasets and the challenges of storing and transmitting massive amounts of data make the widespread usage of AI models more challenging. Moreover, many AI models—especially those based on deep learning—operate as a "black box," making their predictions and decisions less interpretable and prone to inaccurate training datasets. Therefore, in the AI-empowered-network scenario, one of the challenges in designing a 6G AI-native architecture is to mitigate these uncertainties to meet definite quality and reliability standards in mobile networks. In the case of network-enabled AI, the mobile network can deliver real-time AI services with user-level QoS assurance, which is a distinct advantage over those offered by cloud providers. Thus, ensuring service quality for network-enabled AI services is another key challenge in designing a 6G AI-native architecture.

2.2.3. Efficiency optimization

AI technology is highly resource-demanding, requiring substantial computing resources for tasks such as AI/ML model training and inference. The data collection, processing, and transmission in AI tasks also require significant amounts of storage and transmission resources. Meanwhile, the primary goal of network development is sustainability through improved resource utilization, cost reduction, and lower energy consumption. Consequently, addressing the conflict between AI's high resource consumption and the goal of network sustainability is a crucial challenge in the design of a 6G AI-native architecture. A critical aspect of this design lies in integrating AI and network functionalities to share resources dynamically and reduce redundancies, paving the way for greener and more cost-effective operations.

2.2.4. Balancing capability, quality, and efficiency

The interplay between capability, quality, and efficiency in a 6G AI-native architecture introduces a complex optimization challenge. Capability generalization implies a wide range of capabilities and higher performance, which always necessitates more resources and complex functionalities, potentially leading to lower efficiency. Likewise, optimization for quality implies a higher level of reliability, which typically requires additional processing procedures and resource guarantees but could impact both efficiency and capability. Moreover, efficiency enhancement involves improving resource efficiency and reducing energy consumption, potentially compromising performance or quality. Thus, an inherent conflict exists among these three dimensions. The design of a 6G AI-native architecture must address the challenge of finding an optimal balance between capability, quality, and efficiency, thereby addressing the issue of global optimization.

3. A task-driven architecture design approach

3.1. Design principles for a 6G network architecture

In response to the requirements of a 6G network architecture and the challenges in its design, we first introduce the following design principles for designing a 6G AI-native network:

(1) **The principle of practicality: Prioritizing utility over perfection.** Following Pareto's law, this principle advocates for aiming to meet network requirements to cover 80% of service needs, with the remaining 20% of long-tail demands being optional. This principle prevents over-design [27] and aims to ensure that performance meets necessary requirements while enabling rapid, large-scale deployment.

(2) **The principle of simplicity: Favoring simplicity over complexity.** Following the concept of Occam's Razor, "if not necessary, not by entity" [28], this principle advocates for simplifying the structure and logic of the networks and avoiding unnecessary complexities. By simplifying processes, the architecture can achieve higher efficiency, lower costs, higher stability, and greater availability.

(3) **The principle of flexibility: Favoring flexibility over one-size-fits-all solutions.** Based on the Law of Flexibility in business, this principle recognizes that the speed of a process is correlated to its ability to adapt to changing requirements. Network functions (NFs) and services are designed with dynamic flexibility, based on the understanding that a flexible architecture is essential for quickly launching new services. The goal is to create an open, SBA that is both adaptable and scalable.

3.2. A task-driven 6G system architecture design approach

To guide the design of the 6G network architecture, a task-driven architecture design approach is proposed. Guided by systems theory, the proposed approach involves clarifying the system's target requirements, identifying the key tasks that need to be achieved, and proceeding with the architecture design based on task-driven principles. Specifically, the preliminary architecture design is completed through a four-step process that revolves around business scenarios: task definition, element definition, hierarchy definition, and connectivity definition. Finally, a comprehensive optimal architecture design plan is obtained through iterative optimization, as illustrated in Fig. 2.

Based on the decomposition of use case scenarios and a requirement analysis, a "four-definition" approach is proposed to guide the design of a 6G AI-native architecture.

3.2.1. Task definition

To initiate the design process for a 6G AI-native architecture, the first step is task definition, driven by service scenarios. This involves identifying the services or tasks that the network must provide or complete and categorizing them based on shared characteristics. For instance, in the AI-empowered network scenario, typical tasks include intelligent channel estimation, handover management, and service perception. For both the AI-empowered network and network-enabled AI scenarios, this approach focuses on identifying and organizing use cases. Each use case is broken down into independently executable or callable tasks.

In the AI-empowered network scenario, tasks are categorized by service objectives across the RAN, the CN, and cross-domain applications. Typical use cases include: ① A UE triggers network-side AI model training for channel estimation; ② a UE triggers online federated training of a bilateral AI model for channel state information (CSI) compression feedback; ③ the RAN monitors the prediction accuracy of the AI-based CSI compression feedback model; ④ the CN employs AI-based service experience-evaluation models for real-time quality-of-experience (QoE) monitoring; and ⑤ the user plane function (UPF) conducts online training of AI-based traffic-classification models. Regarding the technical realization of these use cases, the tasks can be categorized as follows: UE-side AI training/inference, network-side AI training/inference, AI inference performance monitoring, and NF/UE access to network AI services, as illustrated in Fig. 3.

Conversely, in the network-enabled AI scenario, typical use cases include the following: ① provisioning graphics processing unit (GPU) computing resources with accelerated interconnect capabilities; ② computational offloading of AI-driven video analytics tasks to UE; ③ acceleration of distributed data storage through

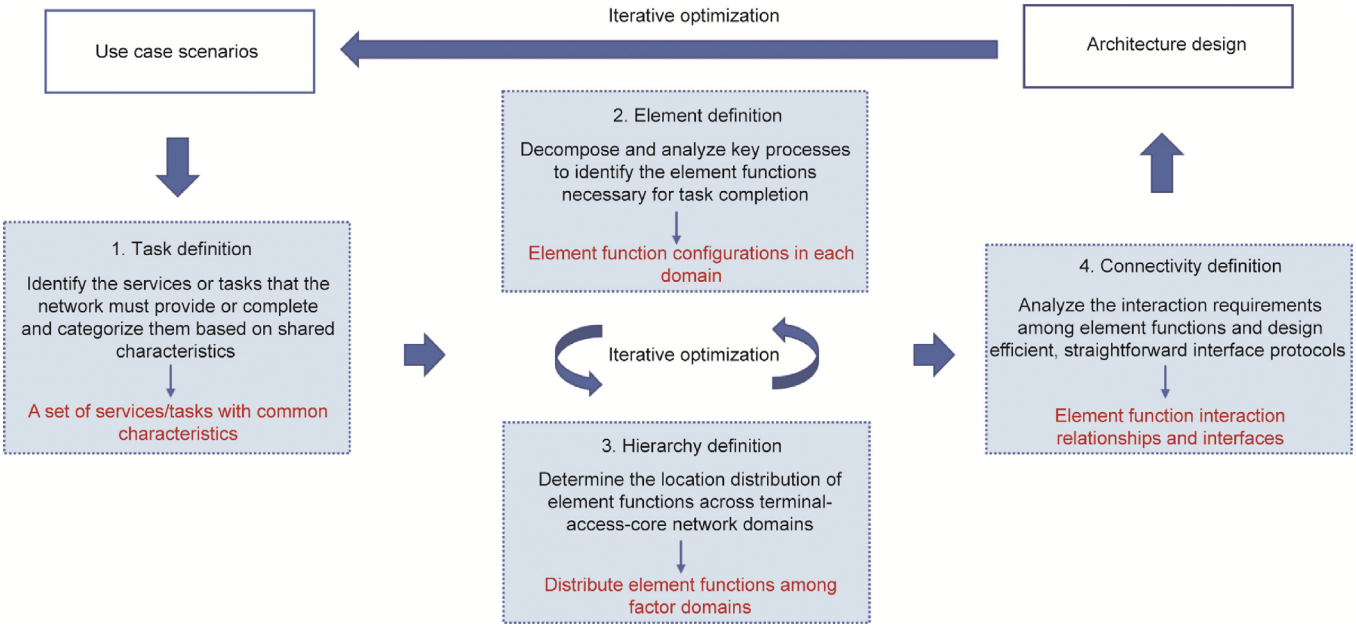

Fig. 2. The task-driven design method for a 6G system architecture.

Category	Use cases for AI-empowered networks	Task
RAN physical layer	AI-based channel estimation, AI-based CSI compression feedback	UE-side AI training/inference Network-side AI training/inference
RAN high layer	AI-based mobility, AI-based energy saving	AI inference performance monitoring
CN	AI-based service experience evaluation, AI-based traffic classification	NF/UE access to network AI services
Cross-domain	AI-based end-to-end service QoS optimization policy	

Fig. 3. Task decomposition for AI-empowered networks.

network orchestration; and ④ real-time decision support for predictive path planning in autonomous driving systems. According to the intrinsic nature of these AI services, they are fundamentally categorized into three hierarchical classes: ① network AI infrastructure-as-a-service (NAIIaaS), providing essential AI resources (GPU/neural processing unit (NPU) clusters and storage pools) and their dedicated connectivity channels; ② network AI platform-as-a-service (NAIPaaS), delivering foundational platform-level capabilities including AI computing services and AI data services; and ③ network-native AI service (NNAIS), which synergistically combines network-hosted AI models, distributed data repositories, and cloud–edge computing platforms to enable communication-embedded intelligent services. Regarding the technical realization of these services, the tasks can be categorized as follows: network AI service exposure and discovery, full-stack implementation of network AI services, QoE-driven assurance for network AI services, seamless continuity-of-network AI services across mobility events, and end-to-end security assurance for network AI services, as shown in Fig. 4. To streamline task identification, the principle of practicality is applied; this ensures a focus on core scenarios and requirements, capturing a task set that prioritizes essential functionality while accommodating diverse service needs.

3.2.2. Element definition

The second step in designing a 6G AI-native architecture involves element definition, which entails extracting well-structured elements based on typical AI workflow and functional analysis.

3.2.2.1. Typical AI workflow and functionality analysis. In AI-empowered network scenarios, tasks must focus on acquiring network data, building AI models, and executing AI tasks with flexibility and efficiency. The design of processes and functions adheres to the principles of dynamic AI task scheduling, functional decoupling, system reuse, and on-demand deployment. For instance, the process design for "UE-triggered network-side AI training" and "NF-initiated AI inference service acquisition" is illustrated in Fig. 5.

In the UE-triggered network-side AI training procedure, upon receiving an AI model training request initiated by the UE, the RAN's AI service control unit performs AI task decomposition and issues data collection instructions to the UE based on the decomposed task requirements. After the UE uploads the collected data, the RAN AI control unit triggers data storage, selects appropriate computing nodes to initiate model training, stores the trained AI model upon completion, and distributes it to the requesting UE.

In the NF-initiated AI inference service acquisition procedure, upon receiving an AI service request from an NF consumer, the AI task-control unit triggers AI task decomposition. Based on the decomposed task requirements, the AI task-control unit dispatches data-collection tasks to designated data sources, requests relevant AI models from the model-management unit, selects appropriate AI inference nodes, executes the AI inference task utilizing the

Category	Use cases for network-enabled AI
Network AI resource service	The network provides GPU computing resources with accelerated interconnect capabilities
Network AI platform service	The network provides computational offloading of AI-driven video analytics tasks to UE
	The network provides acceleration of distributed data storage through orchestration
Network integrated AI service	The network provides real-time decision support for predictive path planning in autonomous driving systems

Task
Network AI service exposure and discovery
Full-stack implementation of network AI services
QoE-driven assurance for network AI services
Seamless continuity of network AI services across mobility events
End-to-end security assurance for network AI services

Fig. 4. Task decomposition for network-enabled AI.

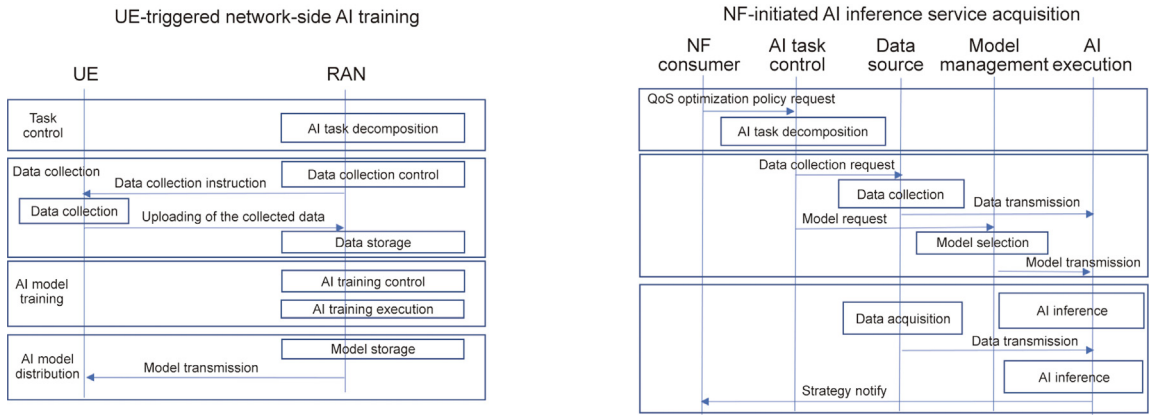

Fig. 5. Typical task process design for AI-empowered networks.

acquired data and model assets, and finally returns the inference results to the NF consumer.

In network-enabled AI scenarios, the task objectives are to match diverse services, utilize resources efficiently, and ensure QoS. Therefore, the principles of process and function design emphasize decoupling application logic from platform capabilities, separating control from execution, and coordinating distributed and centralized systems. The process of providing AI computing services for UEs is analyzed, as shown in Fig. 6. When a UE initiates an AI computing service request to the access and mobility management unit, after completing AI service access authentication, the AI task-control unit performs AI task decomposition. Based on the decomposed task requirements, it issues computing-control tasks to the computing-control unit and connection-control tasks to the connection-control unit. The computing-control unit generates computing strategies and delivers them to the computing-execution unit, while the connection-control unit generates connection strategies and delivers them the connection-execution unit. The computing-execution unit and the connection-execution unit execute AI inference computations and packet forwarding as mandated by their respective control units, subsequently reporting the execution results back to the control units upon task completion.

After comparing, breaking down, sorting, and integrating the key business processes, the main functionalities are identified. These include AI service access, authenticating, AI service orchestration, AI task control, AI model training, AI inference, AI model performance monitoring, model storage, model control, data control, data collection, data storage, data transmission, connection control, and package forwarding.

3.2.2.2. Element identification and functional characterization. Network elements are atomic functional entities that realize specific capabilities through physical/virtual resources in a defined architectural topology. Based on the functionalities abstracted from the preceding analysis of network AI service-provisioning processes, 6G networks require resources and capabilities pertaining to connectivity, data, and computing. Crucially, the realization of connectivity, data, and computing functionalities differ significantly in terms of resources, implementation mechanisms, and performance evaluation. Therefore, building on existing connectivity elements, the 6G AI-native architecture design expands to include data and computing elements. In each element, the control function and execution function are separated, allowing flexible control, distributed deployment, and independent evolution.

By analyzing key task processes, the functionalities within each element function can be identified and designed. For example, to complete network-assisted AI inference, computing-control functions and computing-execution functions are involved; moreover, within the computing-execution function, two functionalities are required—namely, AI/ML model training and AI/ML model inference.

The element functionality design should adhere to three core principles: self-containment, reusability, and independent management. Self-containment ensures the completeness and autonomy of atomic element functionalities, while reusability enables each element functionality to provision heterogeneous service classes. Independent management enables element functionalities to be deployed, upgraded, or migrated independently.

To ensure flexible and efficient utilization of element functionalities, the design should incorporate a combined invocation of

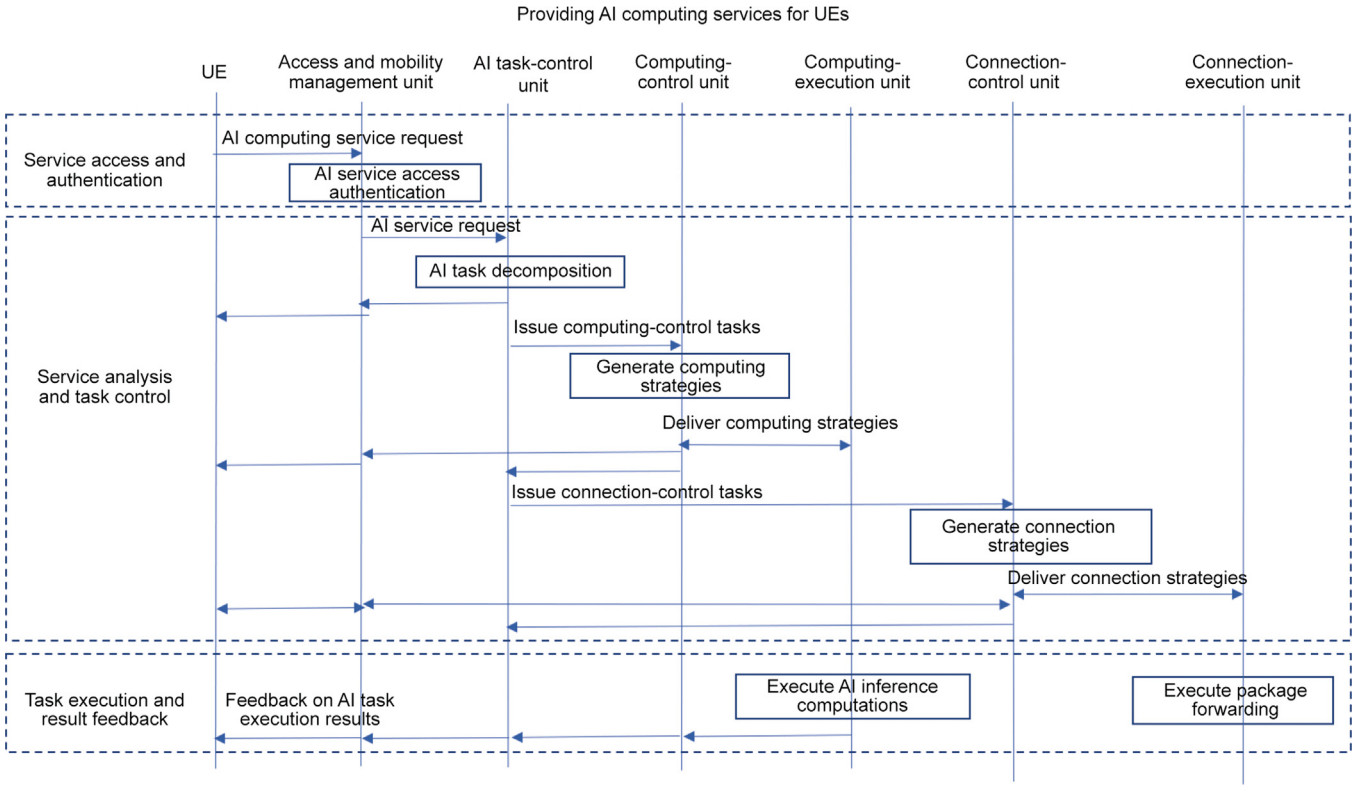

Fig. 6. Typical task process design for network-enabled AI.

element functionalities. The network analyzes AI service requirements and orchestrates multi-element capabilities to match the services precisely. This process includes planning and task orchestration, which are themselves challenging and critical for AI applications in the network. For example, AI agent-based intelligent applications are becoming increasingly prevalent, exhibiting heterogeneous business logic and multifaceted service requirements. To address this complexity, application logic analysis is decoupled from control processes, focusing specifically on analyzing and processing AI-driven services. This ensures that the variability of services does not disrupt fundamental network control functions, maintaining scalability.

The decoupled element function and functionality design also introduce the need for integrated control, which interfaces with application logic analysis at the top level and controls the functions of multiple element functionalities at the bottom level, serving as a bridge for multi-element capabilities to support diverse businesses. The output of the element function and functionality definition, illustrated in Fig. 7, reflects these considerations. The AI service-orchestration function is responsible for AI service requirement analysis and AI workflow design. The task-control function operates as an integrated control layer that processes the application logic output from the AI service-orchestration function and coordinates cross-element control across three foundational elemental functions: the connectivity function, the data function, and the computing function. These functions are further decoupled into control and execution functionalities to separate control from execution. More specifically, the connectivity function is split into enhanced connection control and enhanced user plane forwarding; the data function is divided into data control and data execution; while the computing function is separated into computing control and computing execution. Guided by the principle of simplicity, external element functionality invocation is designed to be straightforward, while internal functionalities are consolidated to ensure efficiency and stability.

3.2.3. Hierarchy definition

Building on the task and element functionality definitions, the next step involves defining the hierarchical layout of element functionalities across the UE, RAN, and CN. This step dynamically positions and distributes element functionalities to ensure an optimal match between resource supply and task demand [29]. Achieving this objective requires careful consideration of several factors, including the diverse and evolving demands of AI tasks, the hierarchical resource availability in 6G network topologies (spanning UE, RAN, CN, edge clouds, and central clouds), and the dynamic and complex characteristics of the air interface environment.

From a demand perspective, for network-enabled AI scenarios, the demand exhibits both diverse and multidimensional characteristics. Here, diversity refers to AI's pervasive integration into various aspects of society, necessitating 6G networks to support a wide range of intelligent applications. Multidimensionality reflects the differentiated requirements of AI-enabled applications, including varying levels of "real-timeness," mobility, and privacy. In terms of AI-empowered network scenarios, it involves leveraging AI to increase the efficiency and functionality of every subsystem of the 6G network. The real-time requirements of AI computing vary across network subsystems and levels; for example, AI inference in the CN may tolerate delays of several seconds, while higher layers of the RAN protocol stack require inference within hundreds of milliseconds. At the physical and medium access control (MAC) layers, AI inference demands are even stricter, with processing times needing to fall between 1 and 10 ms.

From a supply perspective, the 6G network system features a multi-level, distributed architecture encompassing UE, RAN, CN, and cloud platforms. This architecture results in significant differences in the distribution of multidimensional resources such as

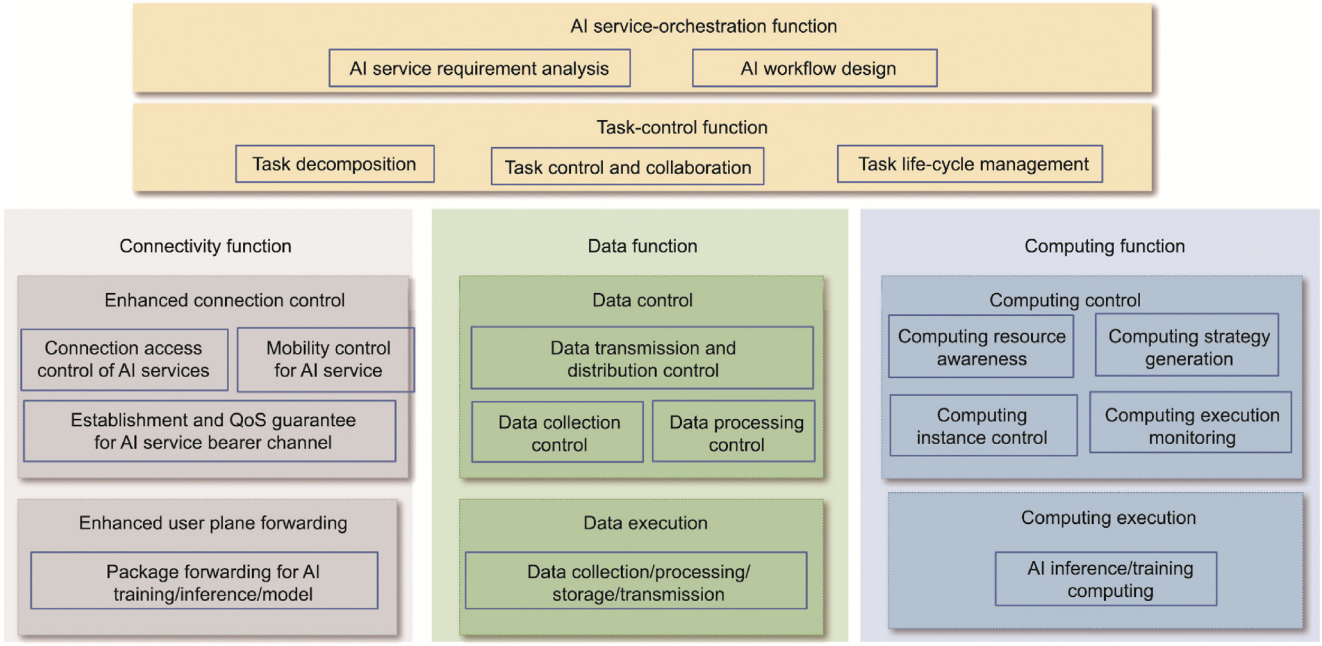

Fig. 7. Element functionalities for AI tasks.

computing, data, and connectivity across these levels. Computing resources exhibit a gradient of capability: Computing power and pooling efficiency increase when moving closer to the higher and more centralized level of the network. However, the diverse nature of UE introduces variability in computing capacity and power consumption; for example, the computing capabilities of wearable devices, AI smartphones, and smart vehicles are gradually increasing.

Data resources demonstrate complementary characteristics: The closer to the CN, the greater the distance from the data sources, which increases transmission delay; however, the geographical scope and breadth of accessible data expand. For example, at the CN level, it is easier to integrate multi-domain data, such as network state information alongside user behavior and application context, which is critical for enabling 6G native AI.

Connectivity resources also follow a distinct pattern. Centralized processing close to the CN causes latency increases due to extended transmission distances. However, the air interface between the UE and the RAN is a dynamic and inherently unpredictable transmission channel, making it the most variable link in the entire network, and an extremely challenging factor in network optimization to ensure a seamless user experience. Fig. 8 illustrates the interplay between diverse AI task requirements and the multi-layered resource-supply characteristics, including computing, data, and connectivity, across the 6G system's hierarchical architecture.

The key challenge in this step lies in how to determine the optimal placement and distribution of element functionalities within a complex and dynamic environment, such as varying channels, user mobility, and topologies. The objective is to balance capability, efficiency, and quality, ensuring optimal alignment between demands and resource supply. To address this challenge, we propose a unified structural framework combining distributed AI execution with hierarchical centralized collaborative control. Fig. 9 illustrates this unified framework.

Within this framework, the distributed AI execution focuses on addressing diverse latency requirements, privacy needs, and transmission costs associated with AI services. The distribution of functional elements related to AI execution follows three key principles: ① The on-demand principle, in which element functionalities such as data collection and AI inference should be distributed at the level and nodes where AI service happens; ② the proximity principle, in which the element functionalities should be distributed as close to service demands as possible, to minimize end-to-end latency, ensure data privacy, and reduce transmission overhead; and ③ the collaboration principle, in which the element functionalities should be distributed across UE and network nodes that share collaborative relationships.

The hierarchical centralized collaborative control primarily addresses latency, reliability, network usage efficiency and collaborative control efficiency. The distribution of computing-control functions is guided by two primary considerations: ① CN centralized control, in which the CN is responsible for centralized AI service functions, including service requests, authentication, and orchestration; and ② RAN regional real-time control, in which the RAN focuses on real-time control at the regional level, enabling the real-time scheduling of communication, computing, data, and AI resources; centralized coordination among multiple UEs and base stations (BSs); and adaptation to dynamic changes in the air interface.

3.2.4. Connectivity definition

Building on the definition of elements, their functions, and hierarchy levels, the final step focuses on defining the connection relationships and interface protocols between element functions to enable seamless services across the architecture. Based on the principles of flexibility and on-demand matching, the interaction requirements between element functions are analyzed to design efficient, minimally necessary connection relationships and interface protocols.

The connectivity design for 6G inherits and extends the connection paradigms established in 5G while addressing new demands. Element control functions should adopt an end-to-end service-based design, emphasizing flexibility and dynamic orchestration. This approach increases service scheduling and orchestration capabilities through a unified service bus and standardized interfaces. Element control functions leverage hierarchical relationships and sequential invocations to coordinate tasks effectively [30].

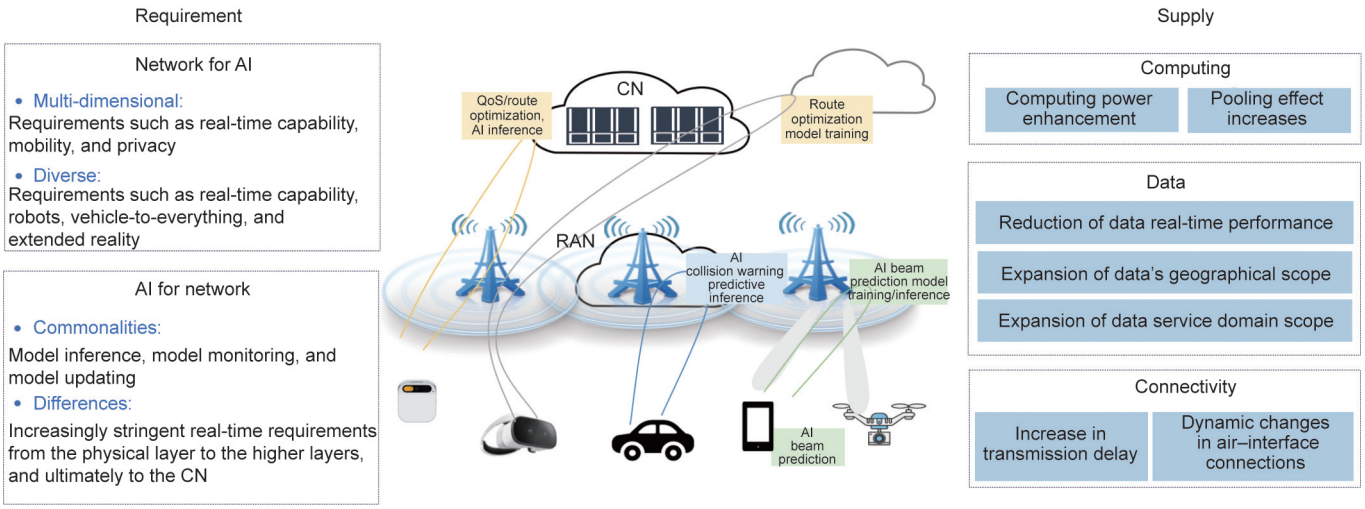

Fig. 8. Diverse AI task requirements and multi-layered resource supply.

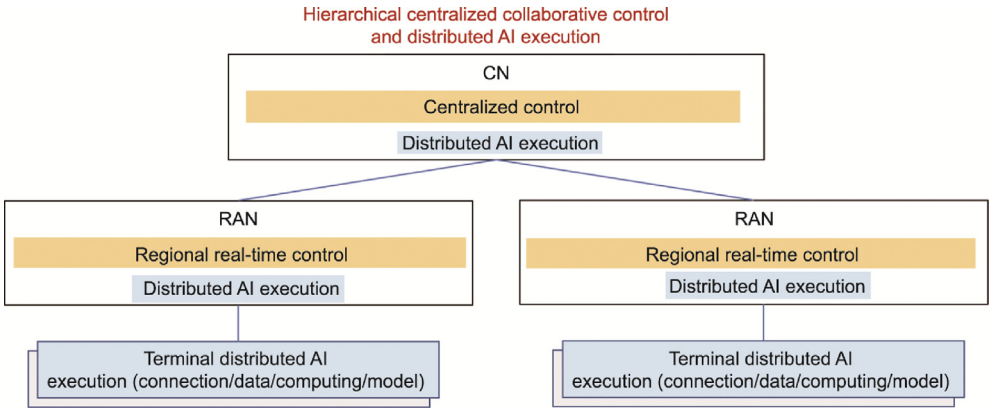

Fig. 9. A unified structural framework featuring distributed AI execution with hierarchical centralized collaborative control.

In the network-enabled AI scenario, external AI services access the network via control signaling. The network-access functions serve as the entry point for external AI services, linking terminals with the network. As top-level control functions, AI service-orchestration functions align AI service requirements with task-control functions. Task-control functions analyze these orchestration functions and connect them with multi-element functions. The element-control layer connects the task-control functions and each element's execution functions, sending action instructions to the underlying execution functions and resources based on task-control commands.

Execution-type functions differ from control-type functions by prioritizing optimal processing performance. While service-based mechanisms offer flexibility, scalability, and maintainability, they can introduce significant processing delays in scenarios with complex inter-service dependencies and numerous services, making them less favorable for stringent performance requirements. Therefore, computing-execution functions employ message interaction protocols or dedicated protocol stacks that ensure high transmission efficiency and low bandwidth overhead, such as the SRv6 protocol. For data execution, protocols that facilitate high-throughput and real-time concurrent data transmission and interaction are utilized, such as the quick user datagram protocol Internet connections (QUIC) protocol for data transmission and Kafka for data message processing and distribution. Adhering to the principle of flexibility, the connectivity design emphasizes platform-based structures and service-oriented mechanisms to support rapid service matching and dynamic adaptability.

3.3. Network iterative optimization

A preliminary design for the 6G AI-native network architecture is developed based on the decomposed use case tasks and the specified network element functions, including their deployment, connections, and interface protocols, as previously defined. It is essential to evaluate and iteratively optimize the architecture to assess the alignment between the design and user demands and to achieve global optimums in capability, quality and efficiency [31]. Consequently, we propose an evaluation method to jointly optimize these aspects of the network, which is detailed below.

Suppose that the network business domain includes N services, represented as $S = \{s_1, s_2, \ldots, s_N\}$, the network logical function domain provides M NFs, represented as $F = \{\mathrm{nf}_1, \mathrm{nf}_2, \ldots, \mathrm{nf}_M\}$, and the total amounts of computing and communication resources are limited to $\overline{\mathrm{CP}}$ and $\overline{\mathrm{CM}}$, respectively. Assume that users initiate requests for various types of services during a given time interval T, and each request is associated with specific parameters depending on the service type. Consider the jth request coming from a user for the service s_i ($i \in \{1, 2, \ldots, N\}$), with the quality requirement $Q_{i,j}$ regarding the specific capability performance $P_{i,j}$ (e.g., latency). Then, according to the quality requirement and network states, the network-scheduling policy generates a specific schedule

strategy, including the types and deployment locations of NFs to use and the amount of network resources to allocate. In this case, let $\mathrm{Comp}_{i,j}$ and $\mathrm{Comm}_{i,j}$ be the computing and the communication resources allocated to meet the quality requirement $Q_{i,j}$ imposed by the jth request of the service s_i.

The interactions between different NFs can be represented by a matrix $\boldsymbol{A}$, as shown in Eq. (1):

$$\boldsymbol{A}_{i,j}=\begin{bmatrix} a_{11}^{i,j} & \cdots & a_{1M}^{i,j} \\ \vdots & \ddots & \vdots \\ a_{M1}^{i,j} & \cdots & a_{MM}^{i,j} \end{bmatrix}_{M\times M} \tag{1}$$

where $a_{pq}^{i,j}\geq 0$ is the count of interactions required by the jth request for the service s_i between the two NFs (nf_p and nf_q) and is equal to 0 when there are no interactions between the two NFs. In particular, all entities in the pth column and row are equal to 0 ($a_{pq}^{i,j}=0$ and $a_{qp}^{i,j}=0$ with $\forall q\in\{1,\ 2,\ldots,\ M\}$) when the request does not use the nf_p.

The differences in the distances of the deployment locations of any two NFs can be represented by a matrix $\boldsymbol{D}$, as shown in Eq. (2):

$$\boldsymbol{D}_{i,j}=\begin{bmatrix} d_{11}^{i,j} & \cdots & d_{1M}^{i,j} \\ \vdots & \ddots & \vdots \\ d_{M1}^{i,j} & \cdots & d_{MM}^{i,j} \end{bmatrix}_{M\times M} \tag{2}$$

where $d_{pq}^{i,j}=d_{qp}^{i,j}\geq 0$ is the difference in the distance of the actual deployment locations between the NFs nf_p and nf_q.

Moreover, other parameters, such as the bandwidth used for transmission in a specific interaction, the amount of information contained in a specific interaction, and the energy consumed to process a unit of information, are all known as priors.

Let $f_{\mathrm{P}}(\cdot)$ and $f_{\mathrm{E}}(\cdot)$ be the capability and energy-consumption function of the jth request of the service s_i, respectively. Then the capability performance $P_{i,j}$ (e.g., considering latency as a capability, the sum of the transmission time and the processing time) and energy consumption $E_{i,j}$ (e.g., the total energy consumed by information transmission and processing) for the jth request of the service s_i are given in Eqs. (3) and (4).

$$P_{i,j}=f_{\mathrm{P}}(\boldsymbol{A}_{i,j},\ \boldsymbol{D}_{i,j},\ \mathrm{Comp}_{i,j},\ \mathrm{Comm}_{i,j}) \tag{3}$$

$$E_{i,j}=f_{\mathrm{E}}(\boldsymbol{A}_{i,j},\ \boldsymbol{D}_{i,j},\ \mathrm{Comp}_{i,j},\ \mathrm{Comm}_{i,j}) \tag{4}$$

In particular, it is important to note that the detailed calculations for Eqs. (3) and (4) are highly dependent on specific services. However, since this paper focuses on the architecture design rather than on scheduling algorithms, these detailed calculations are beyond its scope.

The performance of the network is usually measured by either consumed energy or delivered capability. In this case, optimizing one does not necessarily ensure the optimization of the other. For instance, optimizing energy consumption might compromise capability performance, and vice versa. However, our aim is to jointly optimize both energy consumption and capability performance. Thus, the ratio of $P_{i,j}$ and $E_{i,j}$ is considered, whose physical meaning represents the capability performance delivered relative to the amount of energy consumed. A higher value of this ratio (written as $U_{i,j}$) indicates that the network is able to better support the required capability while consuming lower energy on average. This ratio can be expressed as

$$U_{i,j}=\frac{P_{i,j}}{E_{i,j}} \tag{5}$$

Then, based on Eqs. (3)–(5), the overall network performance over a time interval T, taking into account the effects of capability, quality, and energy consumption, is represented by the following equation:

$$U=\frac{1}{T}\sum_{i=1}^{N}\alpha_i\omega_i\sum_{j=1}^{N_i}U_{i,j} \tag{6}$$

subject to the following resource limitation constraints:

$$\sum_{i=1}^{N}\sum_{j=1}^{N_i}\mathrm{Comp}_{i,j}\leq\overline{\mathrm{CP}} \tag{7}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N_i}\mathrm{Comm}_{i,j}\leq\overline{\mathrm{CM}} \tag{8}$$

where α_i is the successful rate of the achievement of the desired level of quality, and ω_i is the normalized weight of requests for the service s_i among all requests over the time interval T. Let N_i be the total number of requests for the service s_i over the time period, and let N_{T} be the total number of requests; then, the two factors α_i and ω_i are calculated as follows:

$$\alpha_i=\frac{\sum_{j=1}^{N_i}\delta(P_{i,j}\geq Q_{i,j})}{N_j} \tag{9}$$

$$\omega_i=\frac{N_i}{N_{\mathrm{T}}} \tag{10}$$

The function $\delta(\cdot)$ in Eq. (9) is given as

$$\delta(x\geq y)=\begin{cases}1, \text{if } x\geq y\\ 0, \text{if } x<y\end{cases} \tag{11}$$

where x and y are taken as any real values.

In essence, Eq. (6) serves dual purposes: It evaluates the network's operational performance and optimizes the network design concerning the selection and deployment of NFs as well as the resources allocated to them. Thus, an optimization problem is formulated to maximize Eq. (6) in order to obtain the optimal selection, deployment, and resource allocation of NFs.

$$\max\ U(\boldsymbol{A}_{i,j},\ \boldsymbol{D}_{i,j},\ \mathrm{Comp}_{i,j},\ \mathrm{Comm}_{i,j}) \tag{12}$$

$$\text{s.t.}\quad \sum_{i=1}^{N}\sum_{j=1}^{N_i}\mathrm{Comp}_{i,j}\leq\overline{\mathrm{CP}} \tag{13}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N_i}\mathrm{Comm}_{i,j}\leq\overline{\mathrm{CM}} \tag{14}$$

$$\alpha_i\geq\bar{\alpha}_i,\forall i\in\{1,\ 2,\ldots,\ N\} \tag{15}$$

where Eqs. (13) and (14) arise from the limitations of the total computing and communication resources, while Eq. (15) ensures basic service quality $\bar{\alpha}_i$ with $\forall i\in\{1,\ 2,\ldots,\ N\}$.

4. The 6G AI-native architecture

As the deployment of 5G networks accelerates and new application demands continue to evolve, 6G will need to address increasingly complex and diverse use cases. These include applications requiring ultra-high bandwidth, extremely low latency, and substantial computational capabilities, such as autonomous systems, immersive technologies, and large-scale AI-driven services. Traditional network architectures, however, struggle to meet these demands due to their lack of flexibility, adaptability, and intelligence. To address these limitations, this paper proposes a novel AI-native architecture for 6G, which integrates AI at the core of the network to enable intelligent resource orchestration, adaptive service management, and dynamic task execution.

The architecture is grounded in the principle of task-driven design. Instead of relying on static network resource allocation based on predefined policies, this approach tailors network behavior and resource management to the specific requirements of different tasks in real time. By aligning network operations directly with the demands of diverse applications, it ensures that resources are allocated efficiently and that the network remains highly adaptive to varying workloads and environmental conditions. This enables 6G networks to not only meet traditional communication requirements but also effectively support data-intensive and AI-powered applications, ensuring both high performance and intelligent adaptability.

The architecture is structured around three fundamental elements: connectivity, computing, and data, each of which plays a critical role in enabling intelligent, task-driven resource orchestration. These elements form the foundation of a system that can dynamically adjust to the specific needs of individual tasks. The architecture is divided into three core functional modules: service functions, control functions, and execution functions. Service functions focus on AI service orchestration and planning, ensuring optimal resource distribution based on the evolving demands of tasks. Control functions—including task collaboration, connection, computing, and data control—coordinate the network's behavior to meet the requirements of each specific task, ensuring that resources are managed dynamically. Execution functions are responsible for the physical execution of resources, ensuring that connectivity, computational power, and data (collection, processing, and storage) are deployed efficiently to support task execution. A schematic of the proposed AI-native architecture is shown in Fig. 10, which illustrates the interrelationships between the CN and RAN components, as well as the key functional modules.

This AI-native design goes beyond previous 6G proposals, which extended the 5G architecture by introducing dedicated computing and data planes. By integrating AI into the control, computing, and data planes, the architecture enables dynamic, intelligent decision-making across all levels of the network. This integration helps overcome the static nature of traditional network models, providing a more responsive, agile, and efficient system capable of handling the diverse and ever-changing demands of future applications [32].

In the following discussion, we will first explore the architecture's design and functionality in the CN, focusing on how the different modules collaborate to achieve intelligent resource management and task scheduling. We will then turn to the RAN, describing how it supports AI applications in conjunction with the CN. Finally, we will highlight the advantages of the AI-native architecture, particularly in addressing the four key challenges faced by 6G AI-native designs: capability generalization, quality harmonization, efficiency optimization, and the balancing of capability, quality, and efficiency.

4.1. Core network

The CN in the proposed AI-native 6G architecture serves as the central hub for the dynamic orchestration of resources, ensuring seamless task execution across multiple NFs. The architecture integrates AI across three core elements: connectivity, computing, and data. These elements enable intelligent resource management that adapts to the diverse requirements of both AI-driven applications (network for AI) and AI-enhanced network management (AI for network), ensuring optimal network performance and user service delivery.

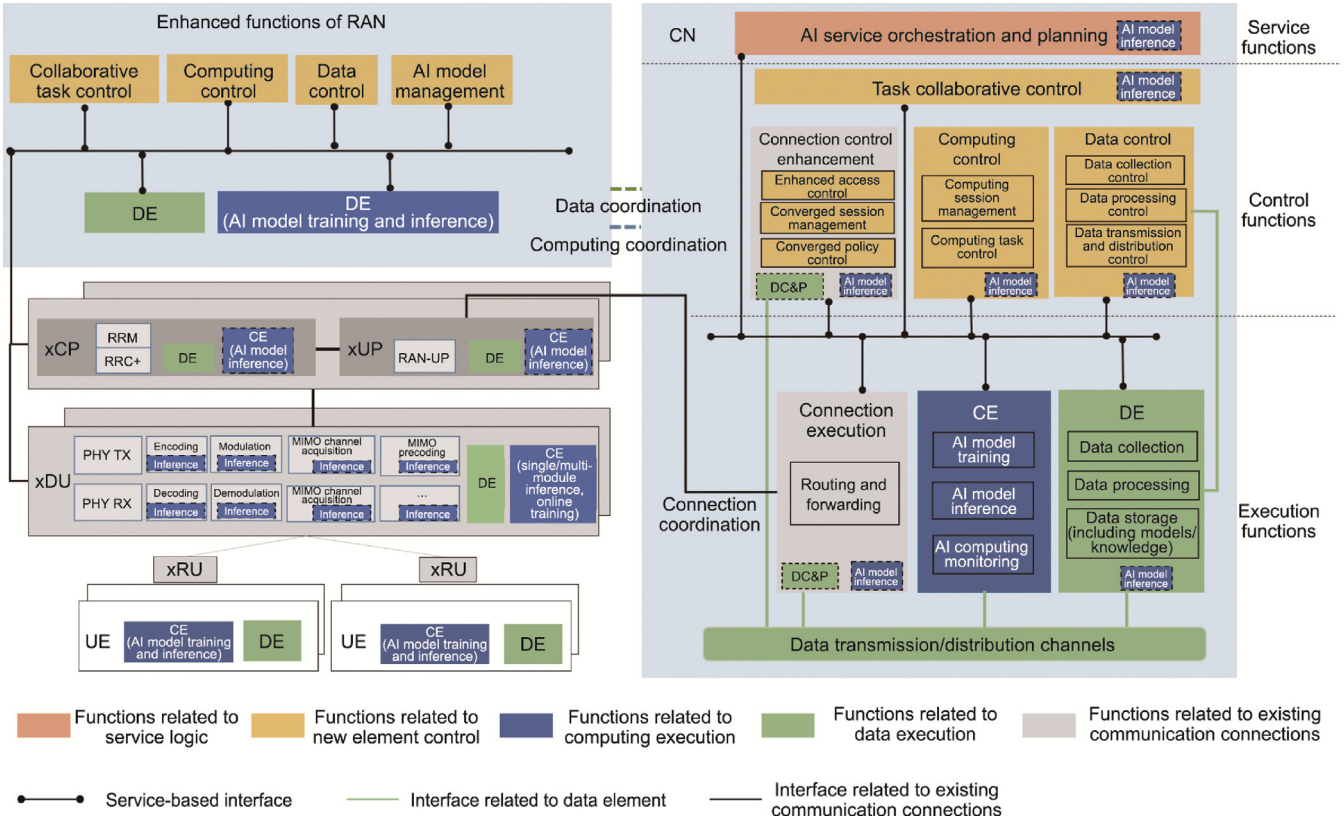

Fig. 10. The detailed design of 6G AI-native architecture. CE: computing execution; DE: data execution; DC&P: data collection and processing; xCP: x contol plane; xUP: x user plane; xRU: x radio unit; xDU: x distributed unit; PHY TX: physical layer transmitter; PHY RX: physical layer receiver; MIMO: multiple-input multiple-output.

The CN architecture is structured into three key functional categories: service functions, control functions, and execution functions. Each of these categories plays a critical role in the task-driven resource-allocation and management system, enabling the CN to effectively support both AI applications and advanced network management.

(1) **Service functions:** The service functions module is responsible for the high-level orchestration and planning of network services. Central to this module is the AI service-orchestration and planning function, which leverages AI to analyze service logic in order to dynamically generate adaptive network task strategies based on real-time network resource states, thereby fulfilling the service requirements. In network-for-AI scenarios, this module conducts in-network task analysis and planning for externally requested AI services, such as AI training and AI inference. In AI-for-network scenarios, it performs in-network task analysis and planning for the network's own intelligent operations, including service experience optimization and performance optimization.

(2) **Control functions:** The control functions module is responsible for managing the allocation and optimization of resources across different network tasks. This module integrates the control of tasks, connections, computing, and data to ensure efficient, real-time decision-making and resource allocation.

- Task collaborative control coordinates overlapping tasks that share network resources, such as connectivity, computing, and data. In network-for-AI scenarios, this function ensures that AI tasks such as model training and inference are scheduled to minimize interference. In AI-for-network scenarios, it optimizes the collaboration between AI-driven network-management tasks, such as traffic prediction and interference mitigation, and user-facing services.
- Connection control enhancement ensures robust and efficient network connectivity, adjusting network sessions and allocating bandwidth as required to meet dynamic task demand. This includes converged session management, which maintains continuity in data sessions across different network domains, and converged policy control, which enforces consistent policies for both AI tasks and user data transmission.
- Computing control is responsible for the intelligent allocation and management of computational resources. This function supports the allocation of compute power for both AI tasks in network-for-AI scenarios and resource-management tasks in AI-for-network scenarios. Key components include computing-session management, which oversees the life cycle of computing tasks, and computing task control, which optimizes the execution of these tasks, particularly AI model training and inference.
- Data control governs the flow and processing of data within the network. It encompasses data-collection control, which gathers relevant data for both AI and user-facing services, and data-processing control, which ensures that the data is processed and made available for subsequent analysis and task execution. The sub-function data transmission and distribution control ensures efficient data distribution, supporting real-time AI tasks and user traffic alike.

(3) **Execution functions:** The execution functions module is responsible for physically deploying and executing tasks based on the decisions made by the service and control functions. It ensures that the required computational, connectivity, and data resources are effectively utilized for the execution of AI-related and network-management tasks.

- Connectivity execution ensures the efficient routing and forwarding of data across the network. This function ensures that communication paths are optimized, minimizing latency and maximizing throughput for both AI-for-network and network-for-AI applications.
- Computing execution oversees the execution of computational tasks, such as AI model training and inference. In network-for-AI scenarios, this function is critical for processing the intensive computational tasks associated with AI model development and deployment [33]. This includes AI model training, which supports the development of AI models, and AI model inference, which applies trained models for real-time decision-making in the network. Additionally, AI computing monitoring ensures that computing resources are efficiently utilized, identifying and addressing performance bottlenecks.
- Data execution is responsible for managing data-related tasks, including data collection, data processing, and data storage. In the network-for-AI domain, this includes the collection and processing of AI-specific data (e.g., sensor data and training datasets) to support the training and inference of AI models. In AI-for-network scenarios, it focuses on the real-time handling of user data, ensuring the smooth execution of network-management tasks and service delivery. Moreover, data storage ensures that both AI models and user data are securely stored and readily accessible for future use, including updates and performance analysis.
- Data transmission and distribution channels ensure efficient data transfer across the network, enabling real-time communication for AI applications and network-optimization tasks.

The CN architecture of the AI-native 6G framework integrates AI-driven orchestration, dynamic resource control, and intelligent execution to support both network-for-AI and AI-for-network use cases. By integrating AI into the CN, the architecture facilitates the seamless deployment of AI-driven applications while increasing the efficiency of network management. This architecture allows 6G networks to be flexible, adaptive, and capable of handling the increasing complexity of AI applications, while simultaneously improving the efficiency and performance of the underlying network infrastructure.

4.2. Radio access network

The RAN, which provides third-party AI services to UEs in close proximity, also consumes AI services in AI-empowered network scenarios [34]. The RAN in the 6G AI-native architecture comprises two core aspects: connection-oriented functionalities and capabilities that go beyond connectivity. NFs such as the x radio unit (xRU), x distributed unit (xDU), and x centralized unit (xCU) primarily focus on traditional communication connections. However, integrating distributed AI execution capabilities into these NFs enables the RAN to achieve the optimal connectivity performance, exceptional user experience, and increased resource and energy efficiency required by 6G systems. The enhanced RAN functionalities are designed to support emerging capabilities such as intelligence and computing that extend beyond traditional connectivity. These enhancements also facilitate the intelligent evolution of communication connection functionalities, driving smarter and more efficient networks.

In the 6G AI-native architecture, the xDU's connectivity functions primarily handle the physical layer processing of the RAN, including channel coding, modulation, channel estimation/acquisition, and multiple-input multiple-output (MIMO) precoding. The x control plane (xCP) and x user plane (xUP) connectivity functions are responsible for the control plane of the radio protocol stack and the management of wireless resources, as well as user plane processing for wireless communications. The integration of data execution and AI inference will be intrinsic to the xDU, xCP, and xUP. For instance, channel coding, modulation, channel acquisition, and MIMO precoding will be further enhanced to support AI inference capabilities. Additionally, functions such as data collection, storage, and processing will play a crucial role in the xDU,

alongside AI inference and online training for multi-module joint optimization at the physical layer. On the UE side, AI inference, data collection, storage, and processing functions will also be inherently supported. As UE's computing capabilities improve, UEs will continue to strengthen AI training capabilities.

The enhanced functionalities of the RAN are designed to support new capabilities that go beyond traditional connectivity, focusing on intelligence and computing. These functionalities encompass both network-enabled-AI and AI-empowered-network scenarios' task control, computing management, data management, AI model management, data collection, storage, and processing, as well as AI model training and inference. The enhanced RAN will perform centralized collaborative real-time control at the regional level, connecting with multiple xCP, xUP, and xDU. By continuously monitoring the changes in multidimensional resources (i.e., communication, computing, data, and intelligence), it can dynamically orchestrate AI functionalities across multiple UEs and BSs. This includes integrated control and scheduling of communication, computing, data, and AI resources, efficiently adapting to variations in air interface and the status of computing resources in various network levels to meet pervasive AI service requirements.

The AI model training and inference capabilities of the enhanced RAN will facilitate the training and inference capability of AI models running on multiple xCP, xUP, and xDU, increasing the efficiency of network and computing resource utilization. The centralized AI model training and management functions will also improve AI models' generalization and accuracy performance running in the RAN through distributed AI model training, such as federated learning and transfer learning. In addition to improving AI-empowered-network scenarios, the computing execution within the enhanced RAN will support computational operations for AI applications, leveraging the network's connectivity assurances to provide low-latency and deterministic edge AI services. For example, in scenarios involving user mobility, the task- and computing-control functions can work together to implement a joint mobility-optimization scheme, taking into account user coverage status, network load, and computational capacity. This will allow for the optimal selection of connection and computing nodes, as well as UE–network collaboration mode adaptation, ensuring the continuity of service experience.

The enhanced RAN will interface with the CN in a service-oriented manner, facilitating the collaborative control and execution of functional elements with the CN. Additionally, the data and computing service provided by the RAN can be opened as services to the CN, enabling on-demand collaboration with the centralized AI service-orchestration and task-control functions of the CN.

4.3. Advantages of the proposed architecture

The proposed AI-native architecture for 6G networks offers significant advantages in addressing key challenges related to capability generalization, quality assurance, efficiency optimization, and the balancing of capabilities, quality, and efficiency. By adopting a modular and task-driven design, the architecture ensures that the network can support a wide range of applications and services without requiring extensive reconfiguration, thereby enabling capability generalization. In terms of quality assurance, the integration of AI-driven resource management—particularly through functions such as connection-control enhancement and data control—ensures that both AI applications and traditional network services meet their respective QoS requirements, thus maintaining a high standard of service. Efficiency optimization is achieved through dynamic and intelligent allocation of resources across the computing, connectivity, and data planes, thereby minimizing waste and ensuring that both computational tasks and NFs are executed with maximum efficiency. Lastly, the architecture effectively balances the competing demands of AI-driven applications and conventional network operations by ensuring that resources are allocated according to task priority, preventing performance degradation in any area. Overall, this architecture provides a robust solution to the complex challenges of 6G networks, enabling flexible, high-performance, and resource-efficient operation across diverse applications.

When compared with the Hexa-X architecture for beyond 5G (B5G)/6G networks proposed in Hexa-X [10,11], the two architectures share fundamental commonalities: Both aim to achieve holistic convergence between AI/ML and network infrastructures while enabling service programmability to create agile network fabrics for future demands. However, they critically diverge in implementation. The Hexa-X architecture introduces an AIaaS framework comprising core modules (AI model repository, training, monitoring, and agent functions), which are managed via management and orchestration (M&O) systems to generate AIaaS services. These services, delivered through unified application programming interfaces (APIs) such as analytics, prediction, and classification to NFs, applications, and third parties, operate independently from other network frameworks, offering the advantage of minimal network modifications and backward compatibility. However, this decoupled design limits architectural-level integration between connectivity and AI functions, constraining the joint optimization of AI services. The proposed architecture addresses this issue by systematically decoupling AI functionalities into four core elements (computing, data, models, and connectivity) with intra-element decomposition across service, control, and execution logic. This approach enables unified orchestration and converged control mechanisms for multi-element joint optimization, achieving the architectural-level fusion of existing connectivity functions with AI capabilities. While the proposed architecture permits the on-demand generation of AI-connectivity co-optimized services for UEs, NFs, and third parties, it necessitates transformative changes to the network infrastructure, presenting more implementation challenges than the Hexa-X architecture.

4.4. Experimental verification

The proposed 6G AI-native architecture and design methodology were implemented on the Free5GC platform to deliver network-native AI computing services (as illustrated in Fig. 11). The experimental validation confirmed the feasibility of key architectural components and associated technology solutions. The workflow operates as follows: ① The UE initiates AI computing requests via non-access stratum (NAS) signaling to an enhanced access and mobility management function (AMF+) responsible for service admission control; ② a newly introduced service-orchestration and management function (SOMF) parses service requirements and generates orchestration policies; ③ the policies trigger the enhanced session-management function (SMF+), extending 5G's session management to govern converged connectivity-computing sessions through their life cycle, in order to coordinate with the computing-management function (CMF) for resource allocation; and ④ the computing-execution function (CEF) performs distributed computing task execution, and the enhanced user plane function (UPF+) is optimized for AI data transmission. Throughout the entire life cycle of the AI computing service, policy constraints are enforced by the enhanced policy-control function (PCF+), and the data repository function (DRF) provides centralized storage for AI datasets. This architecture embeds computational capabilities as intrinsic network elements, enabling a protocol-level convergence of connectivity and computing for native AI service delivery.

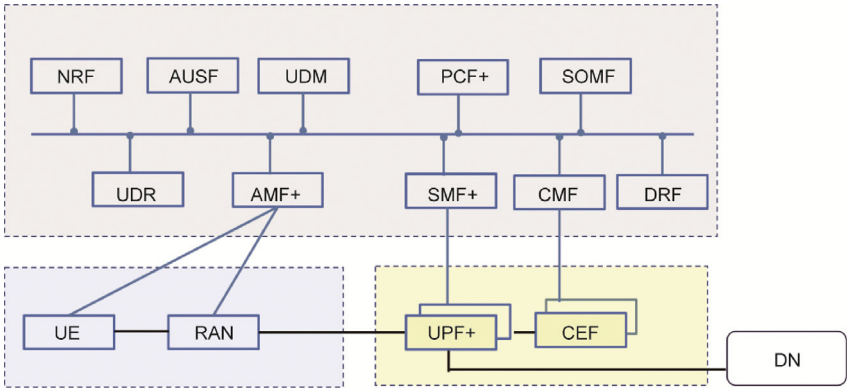

Fig. 11. Experimental system structure of the proposed 6G AI-native architecture. AUSF: authentication server function; UDR: unified data repository; AMF+: enhanced access and mobility management function; PCF+: enhanced policy-control function; SOMF: service-orchestration and management function; SMF+: enhanced session-management function; CMF: computing-management function; DRF: data repository function; UPF+: enhanced user plane function; CEF: computing-execution function.

Fig. 12(a) shows the experimental prototype platform. The left server contains containerized computing nodes, while the center server is responsible for strategy making, and the right server is responsible for the CN functions. Fig. 12(b) shows the experimental scenario of AI task orchestration. The UE initiates an AI-based video object detection task request to the CN, which performs AI task orchestration and scheduling based on real-time connectivity and computing-resource availability across the network infrastructure. When dynamic fluctuations in computing/network resources occur, the SOMF dynamically formulates new orchestration policies to redistribute the AI workloads among cloud, edge, or UE nodes. Figs. 12(c) and (d) show the process of an AI service making a request through NAS signaling, the SOMF orchestrating the tasks, and the SOMF re-orchestrating the tasks according to the status of the computing and connectivity resources. From this experiment, it can be seen that the architecture of converged connectivity-computing management and control shown in Fig. 11 meets the design requirements.

4.5. The relationship between the 6G AI-native architecture and the overall 6G framework

The 6G AI-native architecture represents a significant enhancement of the original 6G framework [25], refining its so-called "three bodies, four layers, five planes" structure by integrating AI-driven functionalities. This architecture strengthens key aspects

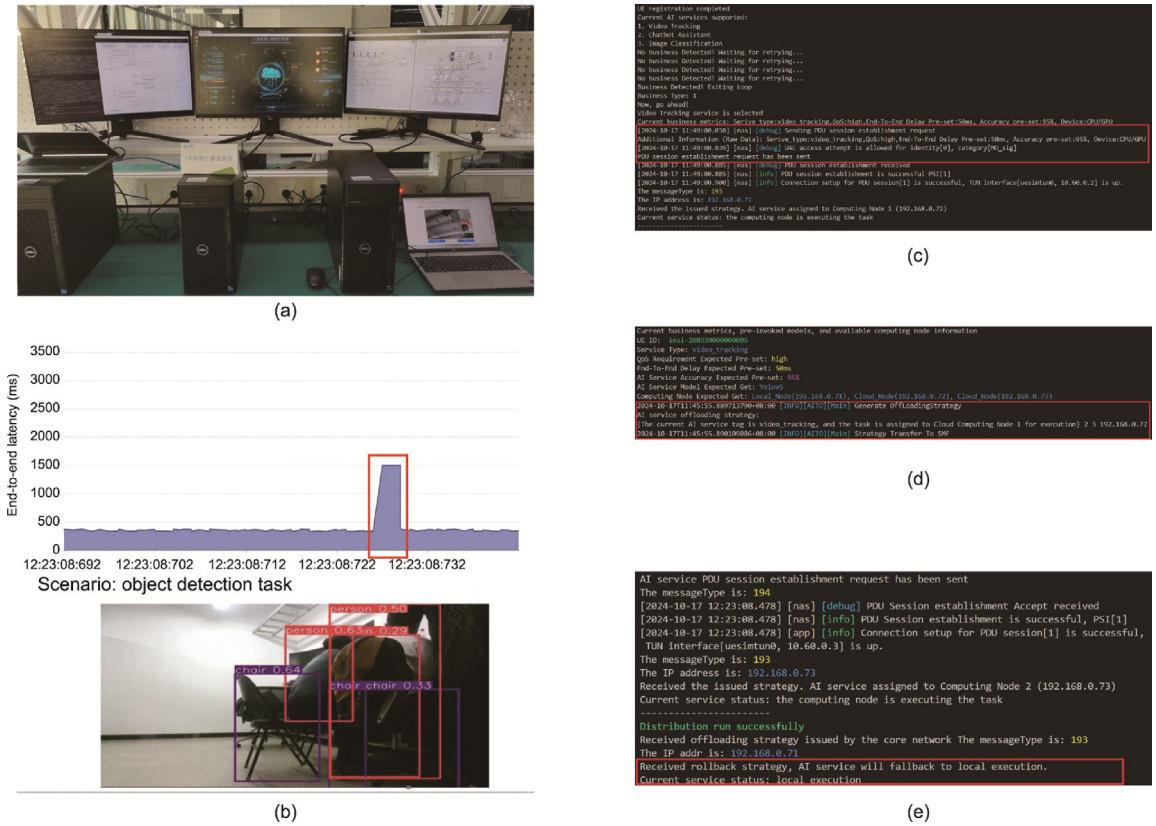

Fig. 12. Experimental scenario and results. (a) Experimental prototype; (b) experiment scenario for AI tasks (video object detection); (c) service-request process; (d) developing a strategy for cloud computing execution; (e) offloading the strategy optimization.

of the network, including the control plane, the user plane, the computing plane, and the data plane in the baseline 6G framework, providing a more intelligent, scalable, and efficient framework for supporting next-generation AI applications. Specifically, the introduction of AI-based control functions—such as task collaborative control, connection control, and computation control—expands the control plane, enabling real-time, dynamic resource management and optimized network operation. Additionally, the user plane enhancement and the newly established computing plane address the specialized requirements of AI services, facilitating low-latency, high-throughput data transmission, as well as supporting compute-intensive tasks such as AI model training and inference. The data plane is similarly augmented to handle large-scale data collection, processing, and storage, efficiently managing both traditional user data and AI-specific data streams. Furthermore, AI-driven orchestration enhances the management and orchestration body, improving network planning, automation, and overall performance. Likewise, the newly introduced digital twin body in the 6G overall framework [35,36] enables the construction of digital twins and simulation environments of physical networks. This supports the verification and closed-loop optimization of network AI strategies, thereby facilitating network autonomy. In conclusion, the AI-native architecture provides a crucial extension of the 6G framework, enabling seamless integration of AI capabilities while optimizing NFs across the computing, data, and control planes, and positioning the network for the future demands of AI-driven applications.

5. Standardization trends of network and AI integration and potential impacts for 6G

The standardization of network and AI integration has been gradually progressing since the advent of 5G. In terms of AI empowering the CN, TS 23.288—titled "Architecture enhancements for the 5G System (5GS) to support network data analytics services" [5]—is a 5G network intelligence specification that specifies the NWDAF responsible for network intelligence in the 3GPP Release 15 specification. Starting from that, the NWDAF evolved from a single NF into multiple logically distributed functions, including the analytics logic function (AnLF) for inference, the model training logical function (MTLF) for training, the data coordination and control function (DCCF), the multi-function analytics function (MFAF) for data bus, and the analytics data repository function (ADRF) for data storage. This evolution decouples data-related functionalities and separates AI execution functions into inference and training modules. Decoupling the multiple elements of AI functions and implementing distributed deployment within NWDAF functions improves efficiency and performance, thereby meeting the diverse needs of AI-empowered networks. However, the NWDAF and its functional modules remain independent functional units outside the existing NFs. For the RAN part, 3GPP specifies the data-collection enhancement to facilitate the RAN intelligence without introducing new AI/ML entities. AI/ML model training can be handled by the network-management system or the next-generation NodeB (gNB), while inference can be performed by both the gNB and UE. To facilitate agility and efficiency, the decoupling of AI element functions and the on-demand distributed deployment of AI functionalities are desired.

A series of standardization efforts in the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) have been initiated for future network AI and the integration of multiple elements into the architecture. The preliminary result have led to the publication of standard Y.3144, "Requirements and architecture for distributed core networks" [37], which proposes a distributed CN reference architecture for IMT-2020 and future networks. This architecture introduces capabilities for data storage and orchestration. The AI agent is designed as a functional entity that offers intelligent capabilities to the control plane, user plane, and data plane. Recommendation ITU-T Y.IMT2020-DIC, "Distributed intelligence collaboration for IMT-2020 and beyond" [38], is currently under development and aims to specify a combined, centralized, and distributed network intelligence coordination mechanism based on AI agents, thereby increasing the network's intelligence level. Recommendation Y.IMT2020-DDP, "Future networks including IMT-2020: Requirements and framework of distributed data plane" [39], is currently in progress, aiming to comprehensively design the specific functions, procedures, and interfaces of the data plane. Additionally, in Recommendation ITU-T Y.3401, "Coordination of networking and computing in IMT-2020 networks and beyond—capability framework" [40], a system framework for the integration of networking and computing is proposed, including a resource layer with network, computing, and storage capabilities; a control layer with resource-aware, scheduling, and control functions; a management and orchestration layer; and an operation and service layer. This framework provides integrated networking and computing services through the coordinated scheduling and control of network, computing, and storage resources.

The standardization progress outlined above indicates that the trend toward the integration of connection, data, computing, and AI has become a consensus within the industry. The functional design of new elements and the mechanisms for multi-element collaboration have emerged as key research directions for standardization. The advancement and research findings of these international standards serve as valuable references and contributions to the standardization of the 6G architecture.

The realization of 6G AI-native design will motivate the standardization of 6G from the perspectives of requirements and architecture. In September 2024, the first 6G requirements study item in 3GPP was officially initiated [2]. This study item considers the integration of mobile communications with computing and AI as one of the most important demand scenarios. Communication for intelligent entities and robot interconnection may surpass the existing person-to-person communication scenario, becoming a new direction for future network development.

The overall approach to standardizing 6G AI-native architecture is to build a service network platform and achieve an on-demand supply of AIaaS. This approach drives the architectural evolution by further decoupling or reconstructing AI elements and by building platform capabilities that involve data elements and computing elements. Through the encapsulation and on-demand combination of platform capabilities, computing services, and data services, comprehensive AI services are formed to support versatile services, including AI applications. Some AI services will be coordinated with NFs or will be built in an integrated manner to further empower the evolution of the network's own intelligent capabilities.

Specific standardization recommendations encompass the following key dimensions:

(1) A paradigm shift is required to transition from the 5G methodology of *post-hoc* AI integration to implementing native-AI design principles at 6G's architectural inception phase. This involves embedding two foundational frameworks within the network service function layer: a data service framework supporting end-to-end life-cycle management, encompassing data acquisition, processing, transmission, storage, and exposure to deliver communication-converged data services for AI applications; and a computing service framework enabling AI task offloading through protocol-compliant access mechanisms to provide communication-integrated computational support. Building upon these frameworks, the establishment of an AI service domain becomes imperative to deliver comprehensive AI capabilities for

pervasive AI agents and applications. Given the domain's operational requirement to dynamically interpret and adapt to complex, evolving service demands, the strategic incorporation of LLMs and AI agent architectures emerges as a technically viable implementation pathway.

(2) It is necessary to coordinate AI standardization efforts across the UE, RAN, CN, and network-management domains in order to establish AI/ML consistency alignment, thereby achieving unified consistency in identifiers, data formats, model interfaces, and AI service orchestration.

(3) An increased focus on 6G systems for AI [2] is called for, specifically targeting two critical dimensions: ① the construction of service architectures optimized for AI agent communication ecosystems, and ② the establishment of system coordination frameworks for AI inference that enable cross-domain synergy between terminal devices and network infrastructure, while facilitating collaborative interactions between foundation AI models and task-specific AI models.

6. Conclusions

This paper analyzed the requirements for the integration of 6G and AI and highlighted the gaps in current work—namely, that there is a lack of guiding principles and methodologies to design a 6G AI-native architecture. Next, stemming from system theory, innovative design principles and task-driven design approaches for 6G AI-native architecture were proposed. Furthermore, the paper explored the relationship between the proposed 6G AI-native architecture and the overall 6G framework. Finally, standardization practices around mobile networks and AI integration were shared, and potential standardization impacts of 6G AI-native design were investigated.

The industry is currently witnessing various 6G AI-native designs being proposed, but a detailed design of the functionality, interfaces, and procedures of a 6G AI-native architecture must be further converged for the unified standardization effort. The proposed design principles of practicality, simplicity, and flexibility, along with the task-driven "four-definitions" design methodology, establish a baseline for the design of a next-generation mobile network architecture.

For future work, there is value in further exploring key technical solutions for designing 6G AI-native architectures, while seeking trade-offs and iterative optimization in terms of capability, quality, and efficiency. Continuous efforts must be conducted by the industry to achieve consensus on adopting a 6G AI-native architecture in the upcoming 6G standardization work.

CRediT authorship contribution statement

Xiaoyun Wang: Project administration, Methodology, Conceptualization. **Lu Lu:** Writing – review & editing, Project administration. **Qin Li:** Writing – original draft. **Qi Sun:** Writing – original draft. **Nanxiang Shi:** Writing – review & editing. **Ziqi Chen:** Writing – original draft. **Tao Sun:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (2022YFB2902001).

References

[1] ITU-R M.2160-0: Framework and overall objectives of the future development of IMT for 2030 and beyond. ITU-R standard. Geneva: International Telecommunication Union (ITU); 2023.
[2] TR.22870: Study on 6G use cases and service requirements. 3GPP standard. Valbonne: 3rd Generation Partnership Project (3GPP); 2025.
[3] Lin X, Kundu L, Dick C, Velayutham S. Embracing AI in 5G-advanced toward 6G: a joint 3GPP and O-RAN perspective. IEEE Commun Stand Mag 2023;7(4):76–83.
[4] Sun Q, Li N, Chih-Lin I, Huang J, Xu X, Xie Y. Intelligent RAN automation for 5G and beyond. IEEE Wirel Commun 2024;31(1):94–102.
[5] TS. 23288: Architecture enhancements for 5G System (5GS) to support network data analytics services. 3GPP standard. Valbonne: 3rd Generation Partnership Project (3GPP); 2024.
[6] TS. 28104: Management and orchestration; management data analytics (MDA). 3GPP standard. Valbonne: 3rd Generation Partnership Project (3GPP); 2024.
[7] Huang Y, Li N, Sun Q, Li X, Huang J, Chen Z, et al. Communication and computing integrated RAN: a new paradigm shift for mobile network. IEEE Netw 2024;38(2):97–112.
[8] The outlook for IMT-2030 6G network architecture. Beijing: IMT-2030(6G) Promotion Group; 2023. Chinese.
[9] Merluzzi M, Borsos T, Rajatheva N, Benczúr AA, Farhadi H, Yassine T. The Hexa-X project vision on artificial intelligence and machine learning-driven communication and computation co-design for 6G. IEEE Access 2023;11:65620–48.
[10] Hexa-X architecture for B5G/6G networks. Report. Espoo: Hexa-X; 2023.
[11] Analysis of 6G architectural enablers' applicability and initial technological solutions. Report. Espoo: Hexa-X; 2022.
[12] AI-native wireless networks. Report. Washington, DC: Automatic Terminal Information Service (ATIS); 2023.
[13] Iovene M, Jonsson L, Roeland D, D'Anglo M, Hall G, Erol-Kantarci M, et al. Defining AI native: a key enabler for advanced intelligent telecom networks. Stockholm: Ericsson; 2023.
[14] ETSI GR ENI 051: Study on AI agents-based next-generation network slicing. European standard. Antibes: European Telecommunications Standards Institute (ETSI); 2025.
[15] Britto R, Murphy T, Iovene M, Jonsson L, Erol-Kantarci M, Kovács B, et al. Telecom AI native systems in the age of generative AI—an engineering perspective. 2023. arXiv:2310.11770.
[16] Wu W, Zhou C, Li M, Wu H, Zhou H, Zhang N, et al. AI-native network slicing for 6G networks. IEEE Wirel Commun 2022;29(1):96–103.
[17] Hossain AR, Liu W, Ansari N, Kiani A, Saboorian T. AI-native for 6G core network configuration. IEEE Netw Lett 2023;5(4):255–9.
[18] Qiu CX, Yang K, Wang J, Zhao SJ. AI empowered Net-RCA for 6G. IEEE Netw 2023;37(6):132–40.
[19] Jiang F, Peng Y, Dong L, Wang K, Yang K, Pan C, et al. Large language model enhanced multi-agent systems for 6G communications. IEEE Wirel Commun 2024;31(6):48–55.
[20] Jiang F, Peng Y, Dong L, Wang K, Yang K, Pan C, et al. Large AI model-based semantic communications. IEEE Wirel Commun 2024;31(3):68–75.
[21] Xu S, Thomas CK, Hashash O, Muralidhar N, Saad W, Ramakrishnan N. Large multi-modal models (LMMs) as universal foundation models for AI-native wireless systems. IEEE Netw 2024;38(5):10–20.
[22] Tarkoma S, Morabito R, Sauvola J. AI-native interconnect framework for integration of large language model technologies in 6G systems. 2023. arXiv:2311.05842.
[23] Von Bertalanffy L. The history and status of general systems theory. Acad Manag J 2017;15(4):407–26.
[24] Wang X, Duan X, Yao K, Sun T, Liu P, Yang H, et al. Computing-aware network (CAN): a systematic design of computing and network convergence. Front Inform Technol Electron Eng 2024;25:633–44.
[25] Wang X, Sun T, Duan X, Wang D, Li Y, Zhao M. Holistic service-based architecture for space–air–ground integrated network for 5G-advanced and beyond. China Commun 2022;19:14–28.
[26] Chen Z, Sun Q, Li N, Li X, Wang Y, Chih-Lin I. Enabling mobile AI agent in 6G era: architecture and key technologies. IEEE Netw 2024;38(5):66–75.
[27] Koch R. The 80/20 principle: the secret of achieving more with less: updated 20th anniversary edition of the productivity and business classic. London: Hachette UK; 2011.
[28] Domingo C, Tsukiji T, Watanabe O. Partial Occam's Razor and its applications. Inf Process Lett 1997;64(4):179–85.
[29] Wang C, You X, Gao X, Zhu X, Li Z, Zhang C, et al. On the road to 6G: visions, requirements, key technologies, and testbeds. IEEE Commun Surv Tutor 2023;25(2):905–74.
[30] Wang X, Sun T, Cui Y, Buyya R, Guo D, Huang Q, et al. Coordination of networking and computing: toward new information infrastructure and new services mode. Front Inform Technol Electron Eng 2024;25(5):629–32.
[31] Wang X, Liu C, He J, Shi N, Zhang T, Pan X. 6G Network architecture based on digital twin: modeling, evaluation, and optimization. IEEE Netw 2024;38(1):15–21.
[32] Wang X, Duan X, Sun T. Service-based network as a platform: research on a new information communication network architecture. Telecommun Sci 2023;39:20–9.
[33] Pan J, Cai L, Yan S, Shen X. Network for AI and AI for network: challenges and opportunities for learning-oriented networks. IEEE Netw 2021;35(6):270–7.
[34] Khan N, Schmid S. AI-RAN in 6G networks: state-of-the-art and challenges. IEEE Open J Commun Soc 2024;5:294–311.

[35] Duan X, Wang X, Lu L, Shi NX, Liu C, Zhang T, et al. 6G architecture design: from an overall, logical and networking perspective. IEEE Commun Mag 2023;61(7):158–64.
[36] Jiang HG, Sun T, Zhou C, Duan X, Lu L, Chen D, et al. Digital twin network (DTN): concepts architecture and key technologies. Acta Autom Sin 2021;47 (3):569–82. Chinese.
[37] Y.3144 (09/24): Future networks including IMT-2020: requirements and functional architecture of distributed core network. ITU-T standard. Geneva: International Telecommunication Union (ITU); 2024.
[38] Y.IMT2020-DIC: Distributed intelligence collaboration in IMT-2020 and beyond. ITU-T standard. Geneva: International Telecommunication Union (ITU); 2025.
[39] Y.IMT2020-DDP: Future networks including IMT-2020: requirements and framework of distributed data plane. ITU-T standard. Geneva: International Telecommunication Union (ITU); 2025.
[40] Y.3401: Coordination of networking and computing in IMT-2020 networks and beyond-capability framework. ITU-T standard. Geneva: International Telecommunication Union (ITU); 2024.