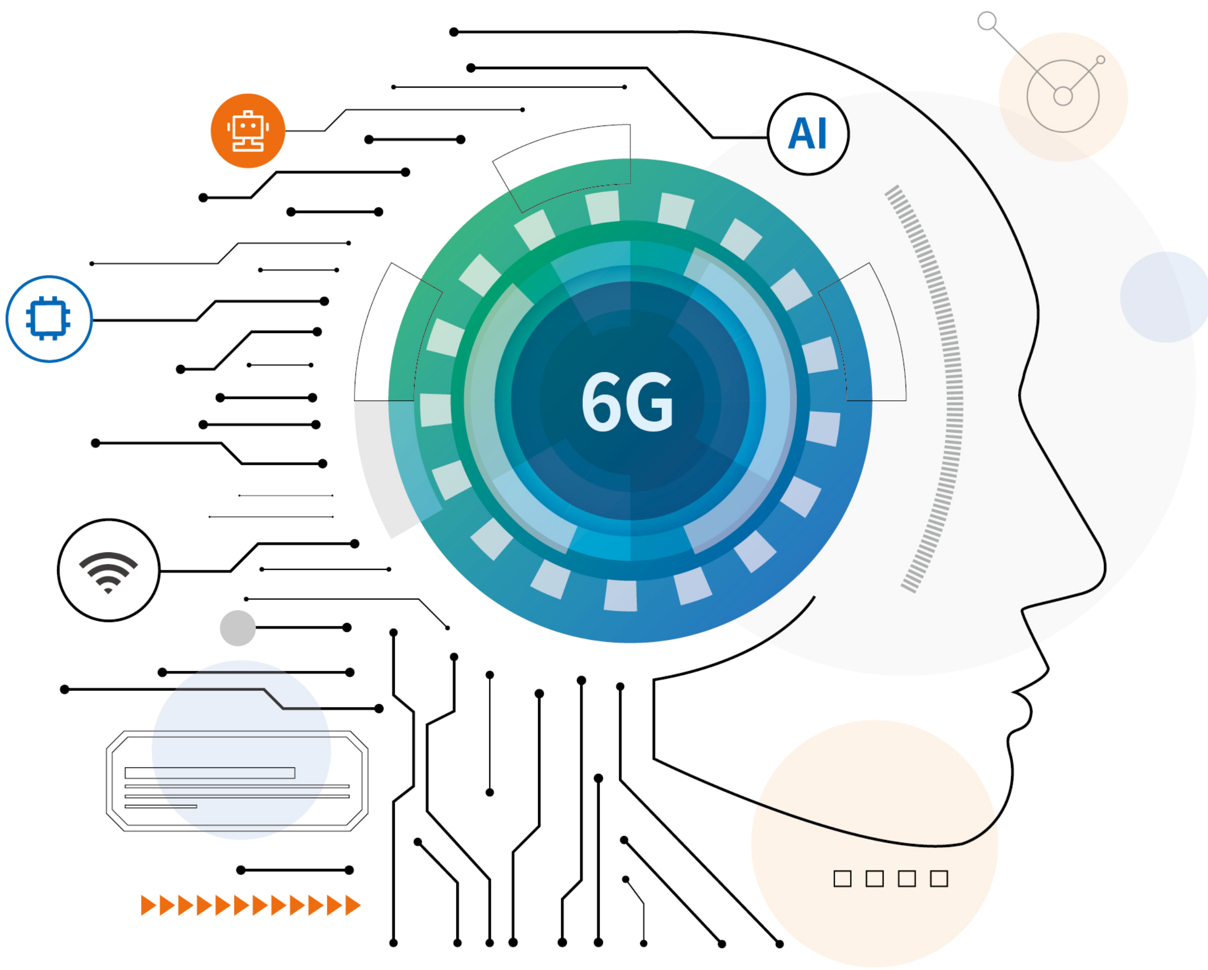
6G NETWORKS INTEGRATED WITH AI

ENDOGENOUS INTELLIGENCE·
CAPABILITY EXPOSURE·DIGITAL TWIN



DT Mobile Communications Equipment Co.,Ltd.
CICT Mobile Communication Technology Co.,Ltd.
ZGC Institute of Ubiquitous-X Innovation and Applications
State Key Laboratory of Wireless Mobile Communications (CICT)

Abstract

This White Paper provides a systematic exposition of the transformative trends and technological pathways for 6G networks integrated with artificial intelligence (AI), and identifies endogenous intelligence as the foundational core of the 6G mobile communication system. In response to the exponential growth of emerging service demands for an ultimate experience, 6G networks are evolving from the conventional role as a "data pipeline" into an "intelligent service engine", thereby driving a profound and all-round transformation in the communications domain.

As detailed in this White Paper, the integration of AI and 6G networks will bring about four core transformations: First is the revolutionary evolution of terminal forms, which enables intelligent agent terminals to supplant traditional terminals; Second is the intelligent interconnection upgrade of application scenarios, which realizes the leap forward from the "Interconnection of Everything (IoE)" to the "Intelligent Interconnection of Everything (IIoE)"; the intelligent restructuring of network architecture, Third is transitioning AI from an "external auxiliary tool" to an "endogenous integration"; the integrated development of service systems, which moves from a "communication-centric" model toward the "in-depth integration of communication, sensing, computing, and intelligence".

Based upon the three key characteristics of endogenous intelligence, capability exposure, and digital twin, this White Paper constructs a hierarchical development pathway for AI-integrated 6G networks: the network element level focuses on high-performance and efficient deployment, the upper layer emphasizes multi-task and multi-network element collaboration, and the top layer centers on advanced intelligence empowered by large models. Core technological innovation focuses on the collaboration framework of "large models + small models + network intelligent agents". With three key technological enablers—the channel foundation model, the network operation large model and the network intelligent agents—there achieves accurate alignment between the propagation characteristics of physical signals and the content of human social interactions.

This White Paper also conducts an in-depth analysis of the operational mechanisms and key technologies underpinning the exposure of 6G intelligent capabilities to empower emerging AI applications, as well as how intelligent digital twin

pioneers a new paradigm of network intelligence. Through the dual-loop collaborative mechanism of “internal closed-loop verification and external closed-loop feedback”, the intelligent digital twin system provides a secure verification environment for validating AI strategies and accelerates the advancement of network intelligence. Leveraging the characteristics of Endogenous Intelligence, Capability Exposure, and Digital Twin, the collaborative evolution of small models, large models, and intelligent agents enables an effective breakthrough of technological bottlenecks, facilitating the realization from the vision of “interconnection of everything” to “intelligent interconnection of everything”. This will lay a robust technological foundation for the digital economy and the intelligent society. This evolution will not only redefine the paradigm of the communications industry, but also propel the digital transformation across various industries into a new stage, ultimately materializing the vision of an intelligent ecosystem characterized by “Networks as a Platform, Capabilities as a Service”.

6G NETWORKS INTEGRATED WITH AI

Contents

Introduction 01

1 / Transformation and Trends in AI-6G Integration 03

1.1 Core Transformations in AI-6G Integration 04

1.2 Trends in AI-6G Integration 05

2 / The 6G Intelligent Network Architecture 08

2.1 Design principles for 6G intelligent network 09

2.2 6G intelligent network architecture 11

3 / 6G Network Evolution Driven by Endogenous Intelligence 15

3.1 6G Endogenous Intelligence Operating Mechanism 16

3.2 Key Technologies for 6G Endogenous Intelligence 17

3.3 Development Trends of 6G Endogenous Intelligence 23

4 / Enabling Emerging Applications via 6G Intelligent Capability Exposure 24

4.1 Operation Mechanism of Enabling Emerging Applications via 6G Intelligent Capability Exposure 25

4.2 Key Technologies for Enabling Emerging Applications via 6G Intelligent Capability Exposure 26

4.3 Developing Trends of Enabling Emerging Applications via 6G Intelligent Capability Exposure 29

5 / 6G iNDT Leads the New Paradigm 31

5.1 6G iNDT Operational Mechanism 32

5.2 Key Technologies of the 6G iNDT 34

5.3 Development Trends of the 6G iNDT 36

Summary and Prospect 38

Introduction

Amid the full commercial deployment of 5G networks and the rapid advancement of artificial intelligence (AI) technologies, the communications industry is undergoing a strategic transformation from the Interconnection of Everything (IoE) to the Intelligent Interconnection of Everything (IIoE). Against this background, the evolution toward the sixth-generation mobile communication system (6G) represents not merely a technological upgrade, but a fundamental paradigm shift for the communications industry.

At present, the communications industry is confronted with unprecedented challenges and opportunities. Emerging services such as Extended Reality (XR)/holographic communication, distributed swarm intelligence and next-generation intelligent terminals have not only raised demands for ultra-high performance and extreme connectivity of networks, but also created development opportunities for new intelligent services. Existing communication network architectures are unable to meet the processing requirements of these novel applications, necessitating an urgent upgrade toward intelligence. AI technologies need to evolve from playing tools in the 5G era to becoming endogenous elements of 6G, deeply integrated into network architectures, and enable the core capabilities including self-perception, self-decision-making, self-optimization, self-execution and self-evolution. In this context, 6G networks with deep AI integration have emerged as an inevitable trend for the development of the communications industry.

Rooted in this era background, this White Paper systematically elaborates on the transformative trends and technological pathways of 6G networks deeply integrated with AI. We propose that the evolution of 6G networks is anchored in three core characteristics: endogenous intelligence, capability exposure, and digital twin, with the goal of building a new generation of intelligent information and communication infrastructure. Endogenous intelligence signifies that AI will become an intrinsic endogenous element of network architectures, fully enabling self-perception, self-decision-making, self-optimization, self-execution and self-evolution. Capability exposure refers to providing AI computing power, data and service capabilities through an architecture with exposure services, thereby constructing a service-oriented system characterized by Networks as an AI Platform. Digital Twin entails leveraging digital twin platforms to accelerate AI model iteration and optimization through the dual-loop collaborative mechanism—comprising internal closed-loop verification and external closed-loop feedback—which is a core mechanism defined in this paper.

At the technical architecture level, this White Paper proposes a hierarchical intelligent architecture of cloud-network-edge-end, where AI capabilities are endogenously embedded within every layer. The end layer focuses on lightweight inference, while base stations and edge nodes deliver real-time intelligent computing, the network layer enables regional collaborative intelligence, and the central intelligent brain achieves global optimal scheduling. Meanwhile, by adhering to design principles—including the servitization of AI elements, collaborative control of AI resources in multi-dimensional heterogeneous networks, and cross-layer/end-to-end joint intelligent collaboration, we construct an efficient, flexible and sustainable intelligent network system capable of adapting to the evolution of intelligent services.

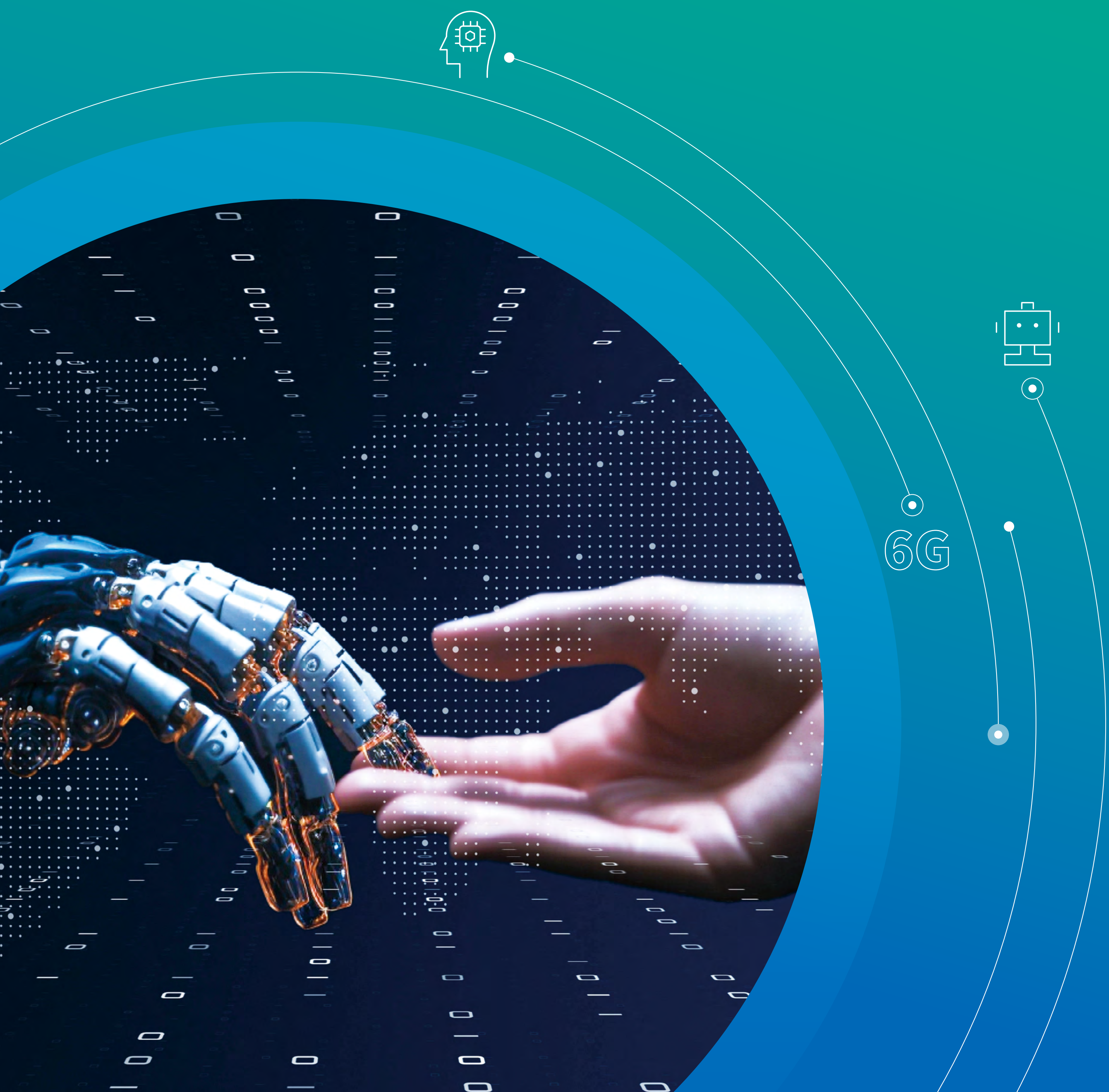
At the application level, this White Paper focuses on how the exposure of 6G intelligent capabilities empowers new applications. 6G networks will be no longer limited to providing basic connectivity services, instead, as a foundational intelligent infrastructure, they proactively exposure core capabilities such as edge computing resources, network data and AI models, and provide end-to-end full-link support for the development and deployment of new applications. Through three key enabling technologies—namely, AI model life cycle management (LCM), cloud-network-edge-end AI collaborative empowerment, and AI service Quality of Service (QoS) guarantee—we establish a complete empowerment chain. This chain integrates capability exposure, technological collaborative support, and AI value realization, thereby bridging the gap between network technical capabilities and industrial application demands.

Furthermore, this White Paper presents an in-depth analysis of the new paradigm for 6G networks brought about by intelligent digital twin—a core concept of 6G endogenous intelligence. Driven by the 6G endogenous intelligent architecture, digital twin and AI technologies are deeply integrated to evolve into an intelligent digital twin system with autonomous decision-making capabilities. Through a complete closed loop of data perception, knowledge generation and strategy implementation, the system facilitates real-time, dynamic and efficient collaboration between physical networks and digital twins, and supports the full LCM of network operations and service innovation.

This White Paper is compiled with the objectives of providing theoretical guidance and practical references for the deep integration of 6G networks and AI technologies. It seeks to drive the leapfrog development of the communications industry from IoE to IIoE, and to lay a solid technological foundation for the construction of the digital economy and the intelligent society. By systematically establishing a technological system characterized by endogenous intelligence, capability exposure, and digital twin, 6G networks with deep AI integration will achieve a qualitative transformation from being a traditional "communication pipeline" into a modern intelligent service hub. This transformation will provide a powerful new impetus for the digital transformation and intelligent upgrade of various industries worldwide.

01 Transformation and Trends in AI-6G Integration

The 6G network, deeply integrated with AI, addresses core demands arising from novel terminals and emerging applications, such as multi-dimensional data fusion and intelligent architecture, and drives comprehensive and profound transformations across the communications field.



1.1 Core Transformations in AI-6G Integration

The 6G network has evolved beyond a mere information transmission pipeline into a new form of digital infrastructure that integrates AI-powered communication, sensing, and computing capabilities. The pursuit of ultimate user experiences by emerging applications and business models, along with the deepening and expansion of high-value scenarios is driving the 6G network toward even higher level.

Business Model Innovation: AI facilitates the evolution of mobile communication services toward scenario-based and customized business models. By leveraging technologies such as user behavior modeling and intelligent traffic scheduling, 6G can precisely fulfill the differentiated requirements of emerging services like XR/holographic communication.

Endogenous Intelligence of Architecture: By leveraging technologies such as intelligent air interface, distributed intelligent agents, and AI-native protocol stacks, this approach overcomes the limitations of conventional architectures and enables full-domain coverage and ubiquitous intelligent connectivity.

Multi-dimensional Data Fusion: Through the integration of communication, sensing, computing, and intelligence, as well as the fusion analysis of multimodal data, an intent-driven intelligent network operation system is built to support scenario requirements in smart cities, industrial Internet, and other fields.

Driven by the aforementioned requirements, the integration of AI and 6G networks will bring about four major transformations:

Revolutionary Evolution of Terminal Forms: Terminals break through physical limitations and evolve into intelligent agent terminals, becoming smart network nodes equipped with environmental perception and collaboration capabilities.

Intelligent Upgrading of Application Scenarios: Driving the evolution from the IoE toward IIoE, and unleashing the full-domain empowerment potential through distributed intelligence and network capability exposure.

Intelligent Reconstruction of Network Architecture: This transformation shifts the role of AI from an external, AI-assisted paradigm to an embedded and AI-native integration within the network architecture. It thereby endows the network with predictive, cognitive, and self-evolving capabilities.

Integrated Development of Service System: It transforms from "communication-centric" to a tightly-coupled paradigm of "communication, sensing, intelligence, and computing". This expands service boundaries from singular connectivity to comprehensive intelligent enablement.

1.2 Trends in AI-6G Integration

AI is emerging as a central driver, fundamentally shaping the evolution of 6G network architecture. To realize the 6G vision of “ubiquitous intelligent connectivity”, the network architecture must be designed around a systematic technological framework characterized by three pivotal attributes, including endogenous intelligence, capability exposure, and digital twin. This section examines the advantages and limitations of key enabling technologies, proposes a hierarchical convergence development pathway, outlines the evolutionary timelines and deployment rationales for distinct technological components, and offers strategic guidance for the practical implementation of integrated systems.

Trends and Core Technological Characteristics

The explosive growth of future applications and services presents multidimensional technical challenges in areas such as air interface efficiency, network architecture, and service capability. Traditional rule-driven paradigms struggle to address these requirements, making AI technology a critical pathway for overcoming existing bottlenecks. An AI-integrated 6G network must be built around following three core characteristics:

First, an “endogenous intelligence” mode aims to achieve self-perception, self-decision-making, self-optimization, self-execution, and self-evolution. Enabled by a data-driven closed-loop mechanism and a hierarchical collaborative architecture, this approach achieves autonomous operation and delivers qualitative improvements in responsiveness, resource utilization, autonomous capability and adaptability.

Second, a “capability exposure” paradigm leverages an open architecture to unleash multidimensional data resources, AI models, computing resources, and service capabilities. This lowers the barrier to integrating AI components and enhances the efficiency and economic value of resource utilization.

Third, a “digital twin” capability relies on a digital twin platform to generate massive volumes of simulation data. Through a dual-loop collaborative mechanism of “internal closed-loop validation and external closed-loop feedback”, it accelerates the iterative optimization of AI models and provides a virtual-physical integrated environment for technology verification and exploration.

Core Technology Analysis

Centered on the vision of “Endogenous Intelligence, Capability Exposure, and Digital Twin”, technologies such as small AI models, large AI models, intelligent agents, and digital twins serve as key enablers. Each exhibits distinct advantages, application scenarios, and limitations, while also facing shared overarching challenges.

Small AI models demonstrate strengths in high efficiency, low complexity, and high accuracy, and have been deployed in fields such as network optimization and resource scheduling. However, they are constrained by limited generalization capability and reusability, requiring enhancements in efficient management and broader adaptability.

Large AI models and intelligent agents possess core capabilities, including knowledge extraction, multimodal data fusion, and cross-domain generalization, which can drive network evolution toward "intelligent and simplified architectures". Key challenges include ensuring multi-element coordination stability, balancing computational demands with performance, and addressing trade-offs between compatibility and security.

Digital Twins enable functionalities such as network element virtualization and environment simulation, alleviating data scarcity and shortening algorithm iteration cycles to support precise network optimization. Through tight coupling with large AI models and intelligent agents, digital twins can facilitate the progressive realization of advanced intelligent network capabilities.

Hierarchical Convergence Development Pathway

AI-integrated 6G networks will evolve in a layered and progressive manner. At the network element level, the focus is on high-performance and high-efficiency deployment. The higher layer emphasizes multi-task and multi-network-element collaborative optimization, while the top layer targets advanced intelligence enabled by large AI models. Small models will continue to serve specific niche use cases, while large models and intelligent agents will be progressively applied for system-level optimization, ultimately forming a landscape of "collaboration between small and large models, with intelligence empowering all domains".

The specific evolution pathway encompasses three key directions:

Integration of Large AI Models with the Network

This progression will occur in stages, following the sequence from "core network" to "radio access network" and from "operational intelligence to "runtime intelligence". Deployment will adopt two complementary modes: "network element integration" for real-time, performance-critical scenarios and "cloud-based deployment" or compute-intensive, latency-tolerant scenarios.

Integration of the Network and Intelligent Agents

This integration will first mature in operation and maintenance scenarios, subsequently expanding to the service architecture level. In the future, multi-agent systems will be deployed to handle complex collaborative tasks, with stable operation of single agents as a prerequisite.

Integration of Digital Twins

This capability will evolve progressively from the network element level to the network level and ultimately the full-domain level, enabling autonomous decision making and coordinated interaction between virtual and physical twin entities.

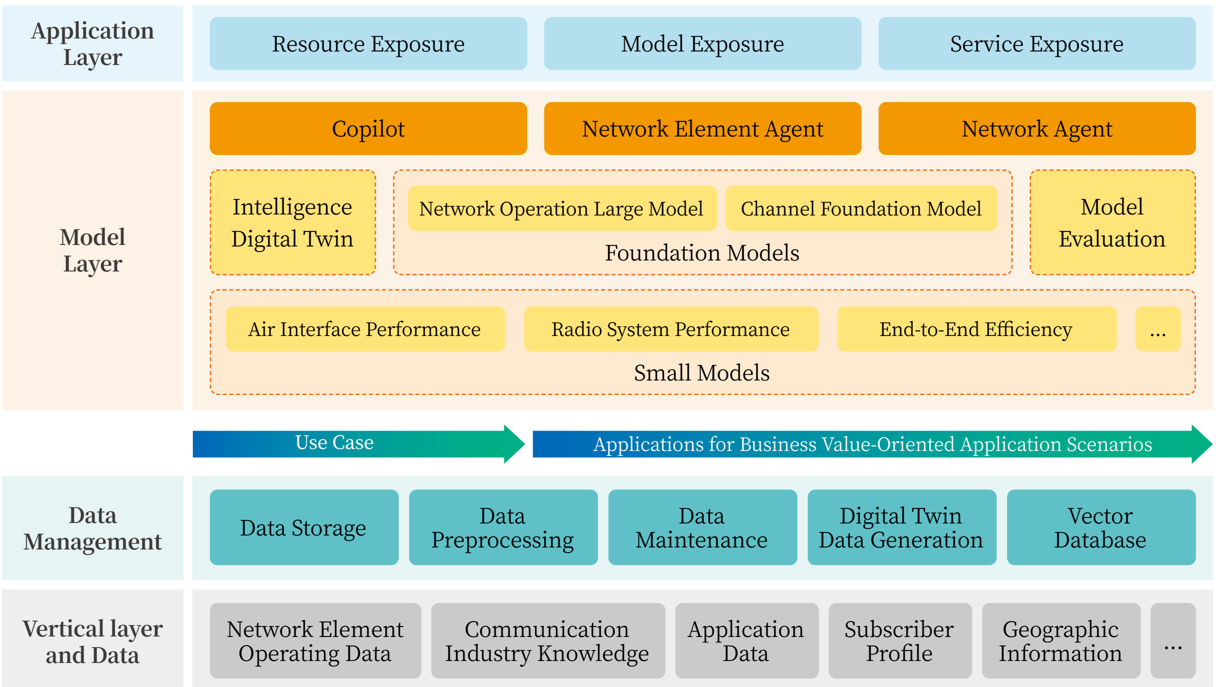


Figure 1-1 AI-network integration development path

The AI-integrated 6G network must ensure high coordination among network functions, data management, and computing resource management, while exhibiting sufficient elasticity and scalability. The architectural design should embody core principles such as AI element servitization, layered distributed collaboration and exposure of AI capabilities, thereby establishing an efficient, flexible and sustainable intelligent network system through comprehensive lifecycle management.

In summary, the AI-integrated 6G network represents an inevitable trend in the evolution of the communications industry, driving comprehensive transformation across network architecture, service systems, application scenarios, and terminal forms. Anchored in the core characteristics of "endogenous intelligence, capability exposure, and digital twin", and through the coordinated evolution of small models, large models, intelligent agents, and digital twins, it can effectively overcome performance bottlenecks and realize the 6G vision of advancing from the IoE to IIoE.

02 The 6G Intelligent Network Architecture

The AI-integrated 6G network is driving a comprehensive transformation from a communication infrastructure to a platform for all-domain intelligence. To meet the diverse and scenario-specific application demands of the future, the 6G network will dynamically schedule and integrate various AI technologies. Its focus extends beyond optimizing network performance to the core objective of empowering industries across the board, thereby providing efficient and precise intelligent services to all users. This objective requires that, from the top-level design phase, the 6G network architecture achieves tight coupling between connectivity capabilities and the three core AI elements: computing resource, algorithms and data. This integration establishes a natively intelligent, all-domain 6G intelligent network system.

Through distributed deployment and elastic coordination, the 6G network can provide AI as a Service (AIaaS) to network operators and external users on demand and efficiently. This approach not only ensures the quality and reliability of intelligent services but is also pivotal for enabling the network to achieve high-level autonomy and progress toward self-optimization. Furthermore, this exposed service model will promote deep integration and innovation across the intelligent ecosystem. In the co-evolution of AI technologies and the 6G network, the network architecture must also demonstrate strong compatibility and scalability to adapt to rapidly evolving AI paradigms such as large-scale models and intelligent agents. This will continuously enhance the service capability and adaptive evolution of the 6G intelligent network, solidifying the foundation for its technological implementation and industrial empowerment.

2.1 Design principles for 6G intelligent network

AI-as-a-Service: The 6G intelligent network architecture shall decouple core capabilities such as connectivity, computing resources, algorithms, and data into modular service units that can be independently invoked. Each AI element shall support elastic scaling based on actual demand, e.g. computing resources can be dynamically scheduled according to traffic fluctuations, and algorithm modules can be iteratively updated to suit different scenarios. Through flexible orchestration, various AI elements can be combined on-demand across different network layers. This breaks down resource silos, enhances the reuse efficiency and response speed of intelligent services, and enables efficient cross-domain resource sharing.

Multi-dimensional Heterogeneous AI Resource Coordination: The 6G network must achieve intelligent coordination of multi-dimensional heterogeneous resources, including spectrum, computing resources, and data. By leveraging AI to dynamically generate resource allocation strategies, the network will establish a complete closed-loop management cycle comprising demand identification, strategy formulation, and execution feedback. This approach moves beyond traditional static resource allocation. It enables precise resource-to-service matching and significantly improves spectrum utilization and computing energy efficiency.

Cross-layer/End-to-End Joint Intelligent Coordination: 6G shall break down the traditional boundaries among the physical, link, and application layers, establishing an AI-based, end-to-end global optimization framework. This framework will enable the deep integration of underlying transmission characteristics with upper-layer service requirements. It will eliminate information silos inherent in layered architectures and shift overall network performance from local optimization to a global optimum. thereby substantially enhancing end-to-end service quality.

Layered Distributed Intelligent Coordination: The 6G network will adopt a layered "cloud-network-edge-end" architecture, with differentiated AI capabilities deployed at each level, including terminals focusing on lightweight inference, base stations and edge nodes providing more real-time intelligent computing than cloud services, the network layer enabling regional coordination, and a central intelligent entity achieving global optimization. This supports distributed AI model training and inference, as well as multi-agent collaborative decision-making. It promotes hierarchical intelligent coordination across the entire chain.

Efficient Data Flow and AI Life Cycle Management (LCM): Efficient data flow is central to the 6G intelligent architecture. The architecture must incorporate robust capabilities for data collection, preprocessing, storage, and transmission. Simultaneously, it must embed native AI LCM to enable closed-loop control over the entire AI process, including data handling, model training, deployment, and updates.

Virtual-Physical Integrated Intelligent Decision-Making: The 6G intelligent architecture must be founded on the principle of virtual-physical integration. By creating a digital twin that mirrors the physical network in real time, and integrating AI engines with historical data and simulation exercises, the architecture supports predictive operations and strategy simulation. Through deep integration of digital twins and AI, the network system can achieve autonomous functions—including self-configuration, self-optimization, and self-healing. This drives the network's evolution from manual intervention toward intelligent autonomy.

Exposure and Customization of AI Capabilities: The 6G intelligent network shall expose its AI capabilities, including data, computing resources, and AI models to external parties through standardized interfaces. This supports the development of intelligent services across diverse industries according to their specific needs. Furthermore, it shall offer customized capabilities to meet the differentiated demands of various sectors, such as millisecond-latency AI inference services. This breaks down technological barriers, enables widespread social access to AI capabilities, and fosters an intelligent ecosystem based on "network as a platform, capability as a service." Ultimately, this facilitates digital transformation across all industries.

Based on the aforementioned design principles, the 6G intelligent architecture will embody three key characteristics: endogenous intelligence, capability exposure, and digital twin.

2.2 6G intelligent network architecture

Based on the 6G endogenous intelligence architecture design principles, AI technology evolutionary trends and AI service operational characteristics previously outlined, this white paper proposes a 6G intelligent network architecture. This architecture is characterized by three key technical features: endogenous intelligence, capability exposure, and digital twin.

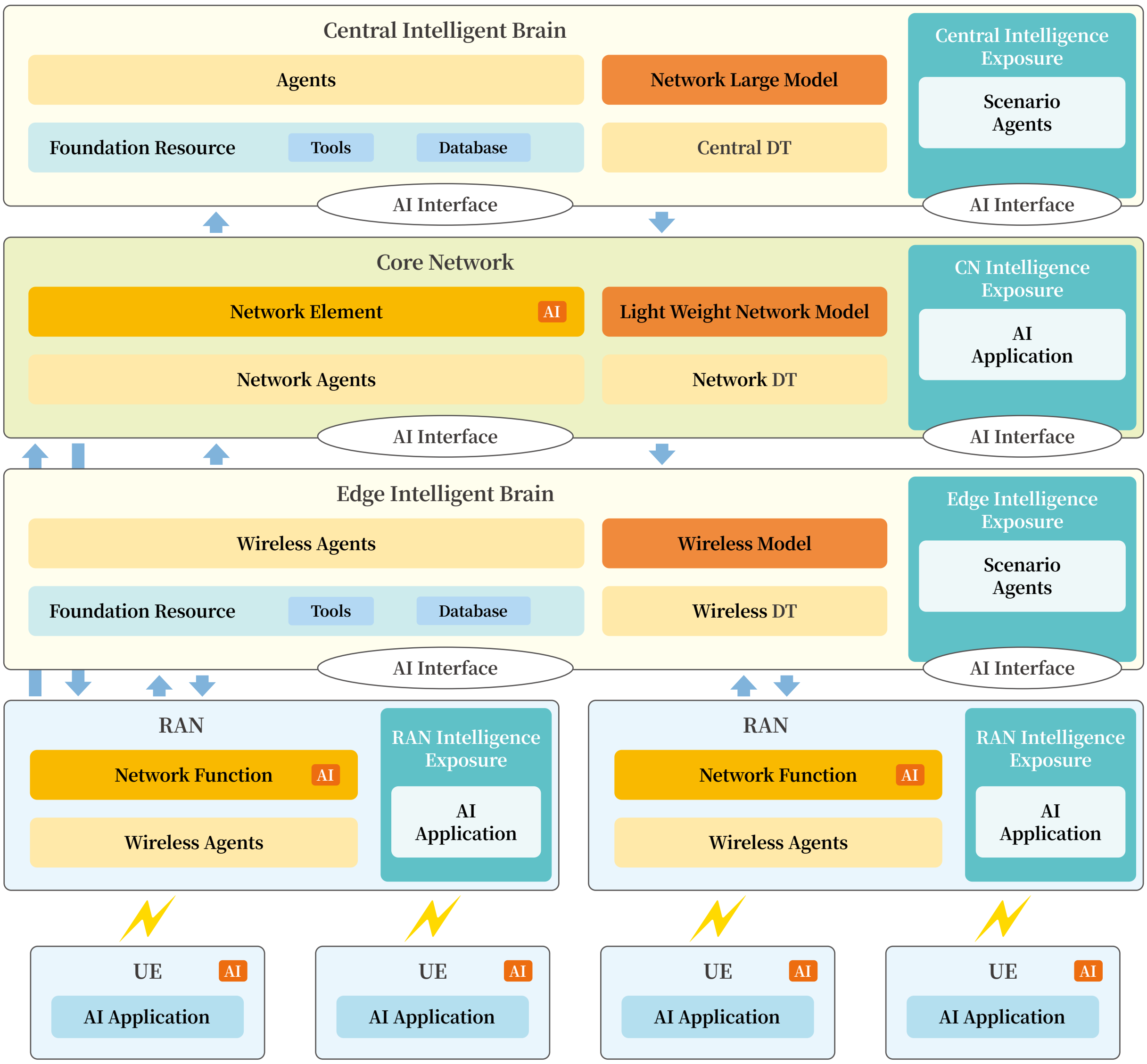


Figure 2-1 6G intelligent network architecture

The 6G intelligent architecture system proposed herein comprises five layers, including the Terminal Device, Radio Access Network (RAN), Edge Intelligence Brain, Core Network, and Central Intelligent Brain. Each layer is endowed with native AI capabilities, and they jointly establish a hierarchical, exposed AI computing system. Through standardized capability invocation and

AI data interaction interfaces that interconnect these layers, cross-layer collaboration and end-to-end efficient linkage are achieved. Concurrently, AI service capability exposure is supported. Within this multi-level intelligent system, the integration of the intelligent Network Digital Twin (iNDT) mechanism further enhances the intelligent autonomous capabilities of 6G networks.

The traditional RAN architecture is evolving toward endogenous intelligence, with AI capabilities continuously permeating downward. The physical layer introduces intelligent technologies, such as AI-based channel estimation and AI-enabled low-pilot transmission. Higher-layer protocols, such as Layer 2 (L2), implement intelligent decision-making mechanisms. These include dynamic resource scheduling and adaptive retransmission strategies. System-level intelligent management and control supports localized AI LCM on the wireless side through the integration of wireless tools and historical data. The core network supports the deployment of network-scale large models. It integrates data from both the access network and the core network to achieve end-to-end intelligent optimization, spanning from terminal device access to service offloading. Multiple types of intelligent agents can be deployed within the network. In conjunction with the network iNDT platform, these agents simulate the implementation effects of different AI strategies through virtual mirroring. This enables predictive assurance for the network.

The architecture of mobile communication networks has traditionally remained largely static throughout the lifecycle of each generational. must possess the capability to support and sustainably integrate emerging AI technologies. However, with the rapid advancement of AI technologies, the 6G network architecture must be capable of supporting and sustainably integrating emerging AI innovations. Built upon the deep convergence of fundamental elements such as connectivity, AI, data, and computing resources—the 6G functional architecture adopts a service-based design philosophy. To accommodate the deployment of large-scale network models, intelligent agents, and future novel AI technologies, a decoupled design approach is employed. This ensures that the 6G network not only incorporates deterministic native intelligent functionalities but also maintains the flexibility to adapt to the iterative evolution of AI technologies.

Endogenous Intelligence

In the architectural design of 6G networks, efficient AI data transmission and native network AI LCM constitute the critical foundation for endogenous intelligence. To overcome the limitations of past centralized AI deployment and management, the 6G architecture deeply integrates core elements—AI data, computing resources, and lifecycle management, into its network functions. This integration is achieved through principles such as "control plane and data plane separation" and service-oriented methodologies, thereby realizing the native capability of "network as an AI platform."

The 6G architecture balances the determinism of core functions with scalability for future AI technologies. To ensure high reliability and deterministic performance, the foundational functional components, such as core modules for data, computing resources, and AI LCM, are designed with tight coupling and standardized service interfaces. For emerging large-scale network models and AI algorithms, a loosely coupled, plug-in architecture is adopted. This enables dynamic loading and flexible integration. An AI technology sandbox is established to permit third-party validation of new algorithms in an isolated environment. Once matured, these algorithms can be seamlessly integrated into the production network through a Resource Management (RM) registration mechanism. Core functions are decomposed into independently deployable units, with each module supporting version-independent evolution. This design facilitates smooth upgrades and compatible expansion during technological advancement without requiring modifications to the communication and AI LCM frameworks.

Capability Exposure

The 6G native endogenous network will establish a new generation of intelligent information and communication infrastructure, forming a service-oriented "network as an AI platform" system. This enables its evolution from a mere "connectivity provider" to an "intelligent service engine."

A coordinated cloud-network-edge-end hierarchical intelligent computing system forms the foundation basis for the exposure of 6G AI capabilities. Through deep integration of intelligent computing boards and edge inference units, 6G wireless base stations develop the novel capability of "base station as an intelligent node." The regional-level Edge Intelligent Brain node serves as an intelligent computing hub. It possesses low-latency AI service capabilities and enables expose operation of AI services by relying on an edge computing resource pool. The core network ensures the secure and controllable management of intelligent computing resources through security isolation mechanisms. Multi-tenant isolation technology achieves logical separation between communication computing resources and AI computing resources. Communication-core computing resources focus on real-time functions, such as signaling processing and session management, to ensure the high reliability of the 6G protocol stack. In contrast, AI-enhanced computing resources are exposed to third-party developers for intelligent services such as model training and inference.

The Central Intelligent Brain, serving as the "central intelligent entity" of the 6G network, functions as a global intelligent computing engine powered by large-scale intelligent computing cluster. It aggregates data resources from both the access network and the core network, trains cross-domain collaborative network optimization models, and deploys agents such as slice management Agents and QoS assurance Agents to enhance the agent ecosystem. Leveraging ubiquitous connectivity capabilities, it enables computing resources pooling between the Central Intelligent Brain and edge nodes. This facilitates the construction of a "Wireless Cloud" AI service system, which provides low-latency and diversified intelligent services for AI applications across various scenarios.

Network AI servitization empowers all industries through service-oriented encapsulation of network resources and exposed models such as Infrastructure as a Service (IaaS), Model as a Service (MaaS), and Agent as a Service (AaaS). This approach builds an "operators-developers-users" ecosystem and drives evolution from the "Internet of Everything" toward "ubiquitous intelligent connectivity".

Intelligent Network Digital Twin (iNDT)

Driven by the endogenous intelligent architecture of the 6G system, digital twin and AI technologies are deeply integrated and evolve into an iNDT with autonomous decision-making capabilities. This evolution breaks through the limitations of the traditional digital twin's "duplication + rule-driven" approach. By integrating cutting-edge technologies such as neural networks, reinforcement learning, large-scale models, and intelligent agents, it builds an iNDT system that covers "data perception—knowledge generation—strategy closed-loop."

Based on the digital twin mirror of device status, the iNDT, processes multi-dimensional fused data through edge nodes at the physical layer to achieve high-precision channel modeling. At higher layers, a distributed intelligent agent cluster is deployed. Combined with the network-scale large model of the central intelligent brain , it improves resource utilization and network stability through "twin simulation—intelligent decision-making."

It achieves efficient collaboration between AI strategies and the physical network through dual mechanisms: "internal closed-loop verification and external closed-loop feedback." In this process, physical network data updates the twin model in reverse to form an evolutionary closed loop. The iNDT is deeply coupled with 6G, building a "physical world—digital twin—intelligent decision-making" system. This help 6G achieve "self-adaptation, self-optimization, and self-growth", establishing an intelligent foundation base.

03 6G Network Evolution Driven by Endogenous Intelligence

Endogenous intelligence constitutes the core driving force behind 6G evolution, a progression dictated by its own inherent logic. To support complex scenarios, including cloud-network-edge-end collaboration and numerous novel AI applications, 6G necessitates urgent breakthroughs in intelligent technologies. From an evolutionary perspective, AI has transitioned from an auxiliary tool in the 5G era to an intrinsic element of 6G. It is deeply embedded within the network architecture to enable self-optimization and self-evolution capabilities. This integration not only addresses technical challenges, such as enhancing spectrum efficiency and reducing energy consumption, but also enables digital twin applications and empowers novel AI applications. Consequently, endogenous intelligence represents not only an inevitable choice for 6G to meet escalating demands and overcome technical bottlenecks, but also the foundational technological cornerstone for next-generation communications.

3.1 6G Endogenous Intelligence Operating Mechanism

6G endogenous intelligence networks are characterized by self-perception, self-decision making, self-optimization, self-execution, and self-evolution. They achieve autonomous operation through a data-driven closed-loop mechanism and a hierarchical collaborative architecture. The core advantage of endogenous intelligence, compared to plug-in intelligence, lies in its shift from 'add-on adaptation' to 'endogenous integration'. Relying on key technologies such as the Channel Foundation Model (CFM), Network Operation Large Model (NOLM), network agent, and large-small model collaboration, it achieves qualitative leaps in response efficiency, resource utilization, autonomous capability, and adaptability. This approach delivers inherent advantages in real-time performance, resource utilization, scenario adaptation, and cross-domain collaboration.

The operating mechanism of wireless communication with endogenous intelligence is depicted in Figure 3-1. Its collaborative paradigm follows the "large models + small models + network agents" architecture, which comprises the following components:

NOLM: A global-level large model that perceives network traffic demands and is responsible for global scheduling and decision-making across RANs.

CFM: A foundation model oriented toward the physical air interface. It perceives signal changes and provides unified capability support at the air interface.

Network Agents: An intelligent entity that represents demands, the air interface, and network elements (NEs). It is responsible for inter-layer interactions, interactions with other NEs, and the collaborative management of internal small models.

Through network agents, this mechanism achieves intelligent coordination across all levels and modules. It leverages the global decision-making capability of large models while enabling efficient local execution through small models.

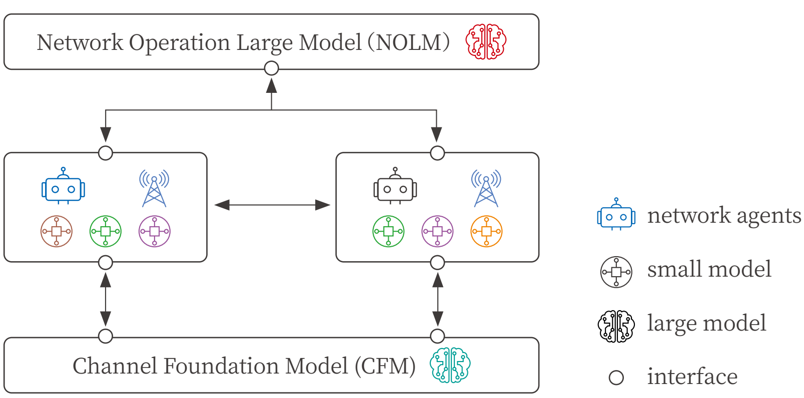


Figure 3-1 The operating mechanism of wireless communication with endogenous intelligence

3.2 Key Technologies for 6G Endogenous Intelligence

The core goal of 6G endogenous intelligence is to achieve precise mapping between the propagation characteristics of physical signals and user service requirements. Three key enabling technologies—the NOLM, the CFM, and the network agents—are pivotal to achieving this goal. Together, they will drive the 6G network' s paradigm shift from an "information connection tool" to an "intelligent service hub."

Wireless Profile representation: Channel Foundation Model

The CFM serves as the core physical sensing foundation for endogenous intelligence. It performs deep mining of multi-dimensional physical elements, including wireless channel characteristics, spectrum resource status, terminal access behavior, and network topology dynamics. Through end-to-end feature extraction and fusion modeling, the CFM constructs a high-precision, real-time digital fingerprint of the global wireless environment. This fingerprint provides high-reliability, fine-grained representation of wireless environment information for upper-layer decision-making tasks, such as resource scheduling and intelligent networking.

The CFM is pre-trained on large-scale, multimodal datasets, which enables cross-scenario adaptation and multi-task generalization. Its core advantage lies in overcoming the limitations inherent in traditional air interface AI technologies, which are often characterized by "single-scenario adaptation and single-task optimization," as illustrated in Figure 3-2. By integrating the physical laws of electromagnetic propagation with multimodal information, the CFM achieves deep modeling of core elements, including wireless channels and the physical environment. This provides a unified technical foundation for diverse tasks, such as beam selection, high-precision positioning, and resource allocation and scheduling for the NOLM.

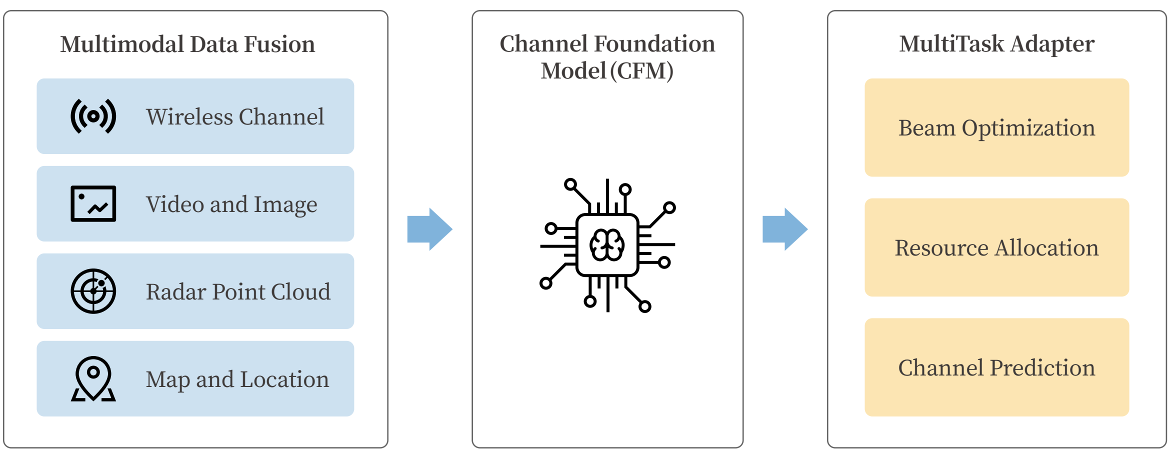


Figure 3-2 Technical logic of the channel foundation model

The value of the CFM is manifested in three key aspects. First, it addresses the challenge of multi-scenario adaptation. By learning generalizable wireless knowledge through pre-training, the model reduces data dependency and training overhead when deployed in new scenarios. Second, it enables coordinated optimization across multiple tasks by leveraging cross-task knowledge sharing. This approach avoids resource conflicts and functional fragmentation that can occur among isolated, single-task models. Third, it supports the endogenous intelligent network architecture by implementing a "perception-prediction-decision" closed-loop. This drives the air interface from passive adaptation toward active, intelligent evolution and provides crucial support for the operation of large models (e.g., the NOLM) within upper-layer networks and agents.

The key research areas for the CFM in 6G are outlined below:

Multimodal Data Fusion and Feature Extraction

Air interface scenarios involve multimodal inputs, such as wireless channel data, video and image data, point cloud data, map and location information. These diverse inputs necessitate that the foundation model possess robust cross-modal processing capabilities. To achieve efficient feature extraction and fusion alignment of multi-source data, the foundation model leverages core technologies, including specialized modality encoder design, cross-modal feature alignment mechanism, and lightweight modality extension.

Multi-task Adaptation

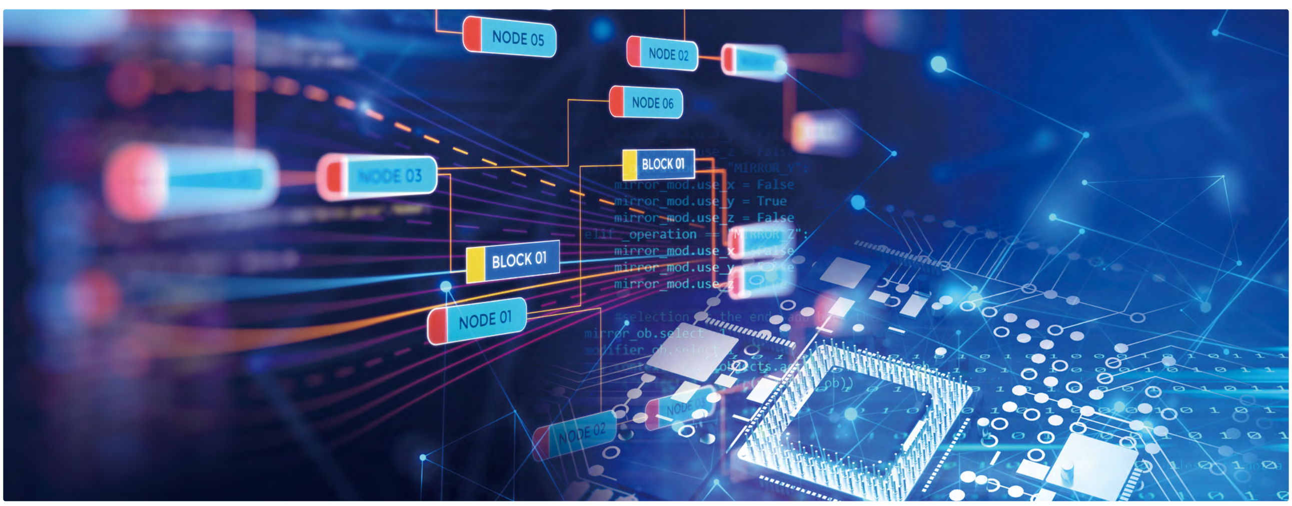
Leveraging the foundation model's general feature representation enables rapid fine-tuning for adaptation to multiple downstream subtasks. This approach significantly reduces model deployment costs compared to the alternative of training and deploying multiple specialized, single-task models. Furthermore, by utilizing cross-modal alignment and knowledge-sharing mechanisms, the model breaks down information barriers between tasks. This enables multi-task collaborative optimization and avoids the performance fragmentation and resource conflicts typical of isolated single-task models. Consequently, multi-task systems built upon this unified technical foundation, coupled with adaptive output design, can flexibly respond to the demands of task switching and parallel multi-task processing in dynamic 6G scenarios.

Autonomous Decision-Making

The "channel foundation model + downstream task-specific small model" paradigm integrates the learning capabilities of large models with established prior knowledge from wireless domains (e.g., channel fading models, power control theory). This integration empowers the subtasks with autonomous decision-making capabilities. For instance, in a dynamic interference scenario, a subtask model can autonomously adjust beam direction and transmission power to achieve real-time balance between interference suppression and communication quality. In addition, these decision-making outcomes provide critical feedback information to upper-layer intelligent agents and the NOLM.

Traffic Profile and Network Control Mapping: Network Operation Large Model

As intelligent communication networks evolve towards the 6G era, the NOLM is emerging as the core driver of network intelligence. It delivers value not only by solving complex network problems but also by adapting to the rapid iteration of network services. To realize this dual value effectively, breakthroughs in several core technologies are required. These breakthroughs are concentrated in three key directions: multi-source heterogeneous data fusion, cross-domain task processing, and reasoning capability enhancement. Collectively, these technologies form a complete technical framework for the NOLM, spanning from data ingestion to decision-making and from model training to deployment. This framework operates by establishing a high-quality cross-modal data foundation, enabling flexible and scalable task adaptation and knowledge transfer, and advancing models toward lightweight and efficient inference. More importantly, they provide a key enabling foundation for realizing the integrated network of communication, sensing, computation, and intelligence envisioned for the 6G era. The NOLM delivers two core values. First, the NOLM can



optimally combine complex network functions and configure multi-dimensional heterogeneous resources. This establishes an efficient collaborative working mechanism among network elements. Second, it better supports agile service iteration. The NOLM can dynamically calculate resource allocation weights and adjust network configurations in real time to comply with Service Level Agreement (SLA) requirements.

The key research areas for the NOLM encompass the following three aspects:

Multi-source Heterogeneous Data Fusion

Based on the 6G integrated network of communication, sensing, computation, and intelligence, a multimodal dataset is constructed. This dataset covers visual, linguistic, signal, and geographic environment data, enabling efficient multi-source data collection and collaboration. High-precision, dynamic reconstruction from physical to digital space is achieved through digital twin technology. Furthermore, multimodal data fusion is realized through strategy-chain-driven semantic and temporal alignment. This process provides a high-quality, consistent, cross-modal data foundation for large model training.

Cross-domain Task Processing

A hybrid expert foundation model is designed based on the 6G network architecture. This model supports flexible task routing and dynamic expert scheduling. Parameter-efficient fine-tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA), are utilized to enable rapid adaptation to tasks across different domains. A cross-layer and cross-domain coordination mechanism facilitates knowledge transfer and reuse, thereby enhancing the model' s generalization and collaborative reasoning capabilities for diverse and heterogeneous tasks.

Inference Capability Enhancement

For deployment in the 6G network environment, large models must be driven toward lightweight deployment and efficient inference. This involves strategies such as inference path optimization and dynamic early exit to reduce computational latency, combined with compression techniques like model pruning and knowledge distillation to reduce storage and transmission overhead. Additionally, the introduction of real-time dynamic retrieval augmentation and hierarchical quantization mechanisms can significantly improve model runtime efficiency and scalability at the terminal and edge sides while maintaining inference accuracy.

Intent-Driven Closed-Loop Real-Time Control: Network Agent

Acting as the core execution entity within the 6G endogenous intelligence architecture, the Network Agent drives real-time control of the full-chain "perception-decision-execution" intent loop. Its operation is guided by the overarching goal of achieving precise business intent implementation. The Network Agent, together with various other modules within the communication system, forms a functionally complementary and coordinated technological ecosystem.

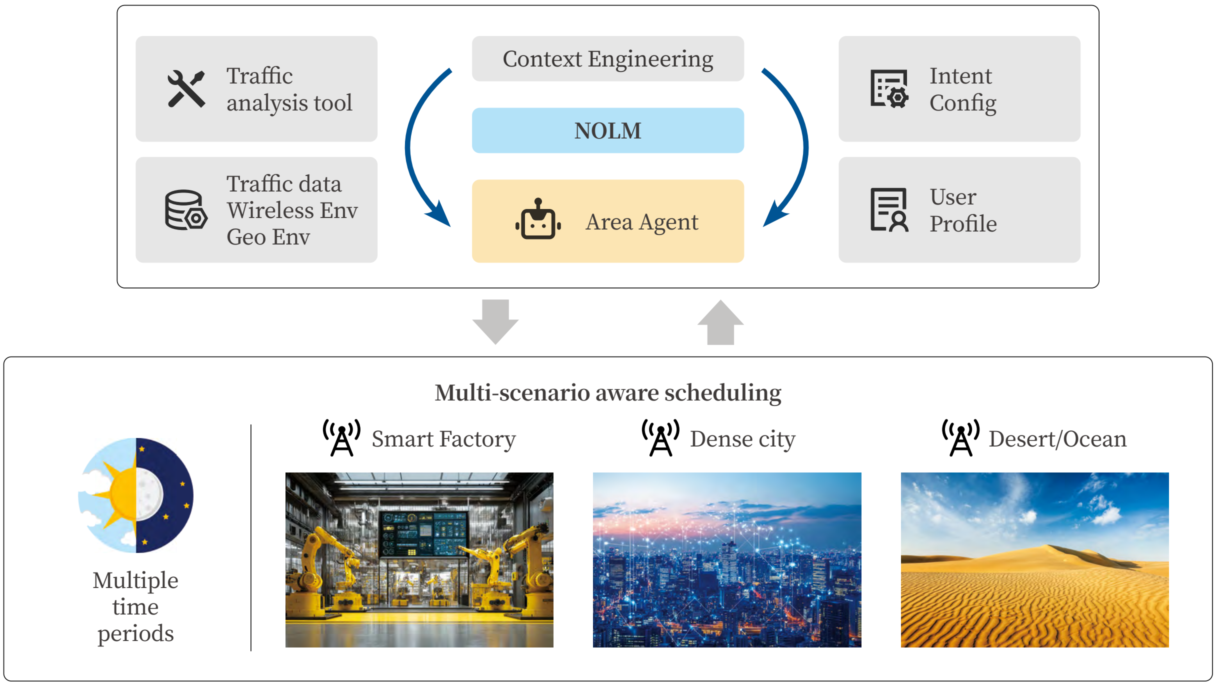


Figure 3-3 Technical logic of the network operation large model in operation

The core value of the Network Agent lies in its ability to efficiently coordinate the various endogenous models within the network. It achieves collaborative scheduling across different models by constructing fine-grained model capability profiles and implementing a dynamic task-matching mechanism. This process integrates real-time service scenario requirements with current network resource status. On one hand, the Network Agent coordinates with the NOLM for cross-domain resource orchestration. This collaboration promotes the generation of global strategies in optimization scenarios such as network-wide load balancing and long-term evolution planning. This integration avoids resource competition and functional redundancy, thereby ensuring both the scientific rigor of decisions and the efficiency of execution during network autonomous operations. Ultimately, the Network Agent serves as the entity that drives the network toward full-process autonomy, characterized by "self-perception, self-decision-making, self-optimization, self-execution, and self-evolution."

The key research areas for Network Agents encompass the following three aspects:

Intent Understanding

Intent Understanding is a core enabler for intelligent networks. It refers to the capability of a 6G network to automatically receive and interpret high-level, abstract business or operational intents from users or operators, and then accurately and dynamically translate them into specific, executable strategies, parameters, and action sequences at the lower-layer network level. Its essence lies in constructing a ‘translation’ and ‘decision-making’ hub. This hub aims to bridge the semantic gap between human natural expression and complex network technical implementation, thereby enabling a fundamental transformation in network management.

Tool Orchestration

Tools are defined as functional modules that can be invoked by other applications through API Interfaces. In a wireless network context, examples of such tools include a Traffic Prediction Tool, a Resource Allocation Tool, an Energy Saving Tool, and a Digital Twin Tool. Network Agents accomplish complex network tasks by orchestrating calls to appropriate tools. Leveraging network exposure capabilities, agents perform reasoning and planning based on the decomposition of increasingly complex tasks. They subsequently determine which tools to invoke and in what sequence to execute them. In response to the diverse service scenarios and differentiated performance requirements of 6G, Network Agents enable dynamic decision-making and task allocation. This capability is key to maintaining collaboration efficiency and service quality in complex, dynamic environments, thereby endowing the 6G network with more flexible and adaptive organizational intelligence.

Multi-Agent Collaboration

Collaboration among agents in the 6G network transcends mere functional module interconnection. Its core lies in establishing an orderly, efficient, and goal-aligned collective action mechanism among multiple autonomous agents, each endowed with perception, decision-making, and execution capabilities. This form of collaboration aims to overcome the inherent limitations of individual agents—such as constrained knowledge scope, limited environmental awareness, and resource constraints—by facilitating information sharing, task coordination, and strategy complementarity. These interactions trigger a qualitative leap in system-wide capability. True collaboration ultimately leads to the emergence of system-level intelligence where the whole is greater than the sum of its parts (“1+1>2”). This emergent intelligence endows the 6G network with significantly enhanced capabilities for solving complex problems and performing adaptive optimization.

3.3 Development Trends of 6G Endogenous Intelligence

As a cornerstone of the 6G network, endogenous intelligence is evolving through the deep integration of multiple technologies. Its development trajectory is characterized by multi-dimensional collaborative innovation and is progressively advancing toward large-scale commercial deployment.

In the dimension of collaborative closed loops, endogenous intelligence addresses limitations in channel foundation models and network models—such as insufficient prediction accuracy and constrained environmental interpretation. By integrating technologies like digital twin and real-time simulation, it establishes a dynamically mapped virtual network space. This enables precise perception and prediction of network status and service demands. On this foundation, an end-to-end intelligent closed loop encompassing perception, decision, control, and evaluation is formed. This drives the network's evolution from passive response to active optimization, ultimately achieving precise matching between resource scheduling and service demand.

Regarding efficient resource utilization, endogenous intelligence confronts the tension between soaring computational demands and constrained hardware resources. It leverages a dual mechanism of 'heterogeneous computing resource scheduling' coupled with 'large and small model collaboration' to overcome this challenge. Furthermore, by adopting a collaborative 'edge lightweight inference and cloud deep optimization' model, it effectively reduces computational costs, transmission latency, and energy consumption.

In the realm of security and trustworthiness, endogenous intelligence addresses emerging risks stemming from model uncertainty and data contamination by evolving toward an 'AI-endogenous security' framework. This is achieved by enhancing algorithmic capabilities to detect anomalous data and malicious behaviors, thereby establishing a multi-layered security protection system that covers data, models, and decision-making processes.

At the level of standards and ecosystem development, the maturation of agent-to-agent communication protocols is driving the standardization and encapsulation of related interfaces, data formats, and functional modules. This addresses compatibility challenges across different vendors and systems. Concurrently, by deepening the Industry-Academia-Research-Application collaborative innovation mechanism, key segments across the industrial chain—including chips, algorithms, platforms, and applications—are being integrated. This integration fosters joint advancement in testing and validation, pilot applications, and business model exploration. This collaborative process will effectively break down industrial barriers, build an open and collaborative integrated ecosystem, lay the foundation for the full commercialization of 6G, and ultimately drive the evolution of communication networks toward a highly autonomous and pervasive intelligent future.

The advancement of 6G endogenous intelligence relies not only on technological breakthroughs but also on systematic progress across four key dimensions, including closed-loop autonomy, efficient resource utilization, security and trustworthiness, and ecosystem collaboration. Collectively, these advancements will enable a next-generation intelligent network endowed with self-perception, self-decision making, self-optimization, self-execution, and self-evolution, thereby realizing the vision of true endogenous intelligence.

04 Enabling Emerging Applications via 6G Intelligent Capability Exposure

Under the trend of deep integration between the digital economy and intelligent technologies , AI is undergoing accelerated iterative upgrades. Emerging applications—including robot collaboration, intelligent agent interaction, and distributed large model training and inference—are continually emerging, profoundly reshaping production, operation, and development models across industries. The bidirectional collaborative evolution of 6G networks and AI has emerged as the core driving force behind industrial digital transformation. Meanwhile, 6G intelligent capability exposure, functioning as the core supporting infrastructure, provides a key pathway toward enabling the large-scale deployment of these emerging applications.

As a core manifestation of next-generation communication technology, 6G transcends the boundaries of traditional networks. Through the exposure of core capabilities and services—such as network computing resources, network data, and AI models, it establishes an intelligent infrastructure that enables the efficient operation of various emerging applications. This infrastructure offers full-process support in computing power supply, data provisioning, and scenario adaptation for AI technologies. Accordingly, this chapter systematically outlines the overall logic behind enabling emerging applications through 6G intelligent capability exposure. From three dimensions—operation mechanism, key technologies, and development trends—it comprehensively examines the core concepts underlying AI-integrated 6G network development. This analysis provides theoretical support and practical reference for building the 6G intelligent exposure ecosystem and fostering innovative development in emerging applications.

4.1 Operation Mechanism of Enabling Emerging Applications via 6G Intelligent Capability Exposure

As a new-generation information infrastructure that integrates communication, sensing, computing, and intelligence, 6G networks transcend the constraints of traditional architectures by establishing a service-oriented exposure system via global capability exposure. In resource exposure, 6G networks process multidimensional data generated from the convergence of communication, sensing, and computing, thereby providing high-quality data support for AI model training, as well as agent task planning and decision-making. Leveraging the "cloud-network-edge-end" collaborative hierarchical intelligent computing system, 6G networks intelligently orchestrate and offload computing tasks to optimal nodes, enabling efficient support for differentiated task requirements. In model exposure, 6G networks lower the barrier for developers to adopt AI technologies by exposing pre-deployed AI models. Additionally, they externalize the core capabilities of registered agents, thereby supporting flexible, on-demand invocation by applications. In service exposure, 6G establishes a quantifiable, differentiated QoS assurance system. This system enables accurate matching between AI capabilities and business requirements by providing ubiquitous, real-time, high-quality services along with standardized interfaces.

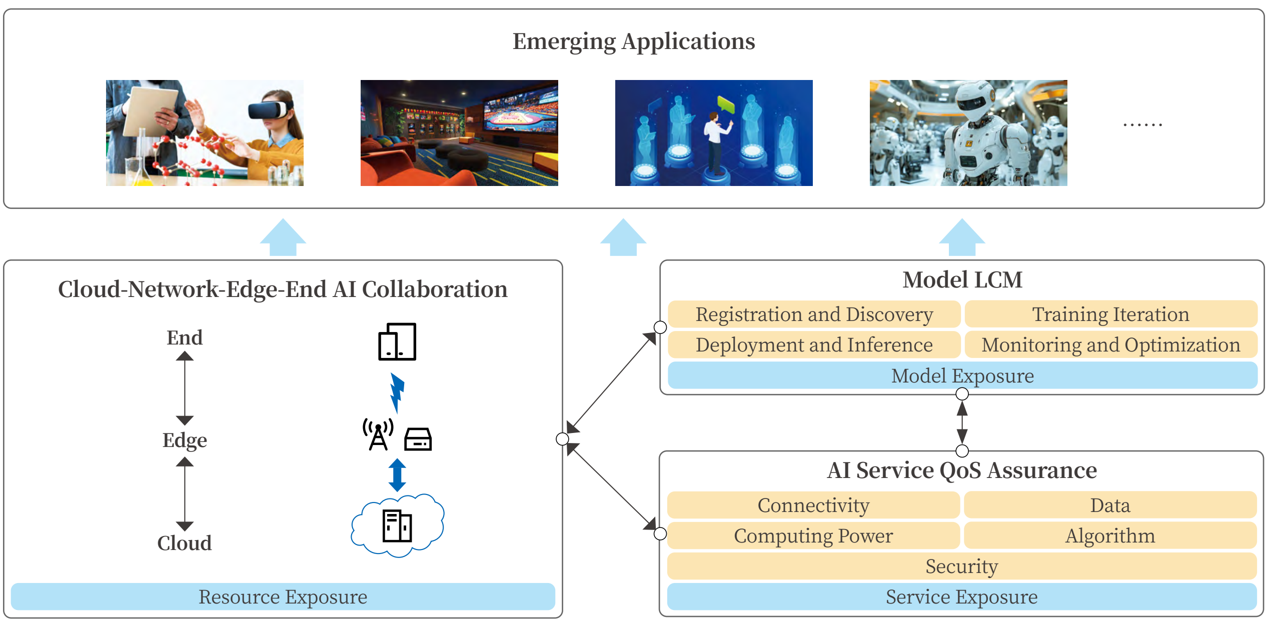


Figure 4-1 Operation mechanism of 6G intelligent capability exposure enabling the operation of emerging applications

The exposure of 6G intelligent capabilities is guided by the core principle of "full-dimensional capability release and systematic support." This approach effectively enables the closed-loop operation of emerging applications within the network environment and delivers stable, reliable underlying support. Figure 4-1 depicts the operation mechanism of emerging applications enabled by 6G intelligent capability exposure. Leveraging "cloud-network-edge-end" collaboration technology, the network delivers low-latency, high-bandwidth connectivity, along with global intelligent capability exposure and distributed collaborative scheduling. During "cloud-network-edge-end" task processing, the network possesses full AI LCM capabilities. These capabilities enable a full-process closed loop that encompasses model registration and discovery, training iteration, deployment and inference, as well as monitoring and optimization. Furthermore, throughout the lifecycle operation, the network dynamically orchestrates communication, computing, and data resources based on the real-time demands of models and agents. It also establishes a full-dimensional, differentiated QoS assurance system tailored to AI services. Ultimately, a full-process assurance system is established, spanning data supply, computing power support, and service adaptation. This system precisely fulfills the business implementation requirements of emerging applications.

4.2 Key Technologies for Enabling Emerging Applications via 6G Intelligent Capability Exposure

6G networks will evolve beyond mere communication pipelines to become intelligent infrastructures that proactively expose diverse core capabilities. These networks provide full-dimensional enablement for the development, deployment, collaborative operation, and value enhancement of emerging applications, particularly those represented by AI agents. This chapter elaborates on three key technologies that underpin this enablement paradigm: cloud-network-edge-end AI collaboration, AI model LCM, and AI service QoS assurance. Operating in synergy, these three technologies collectively construct a complete enablement chain: "network capability exposure-technical collaborative support-AI value realization."

Cloud-Network-Edge-End AI Collaboration

In the 6G era, the demand for real-time performance and intelligence from numerous emerging applications continues to escalate. Consequently, the traditional single centralized processing model can no longer satisfy the differentiated requirements of diverse application scenarios. By constructing a deeply collaborative and open "cloud-network-edge-end" intelligent architecture, the 6G network achieves full-domain exposure and distributed collaborative scheduling of core capabilities—including multidimensional data, heterogeneous computing power, AI models, and network connectivity. This architecture thereby provides core support for the large-scale deployment of emerging applications.

Regarding data processing, 6G networks accomplish full-domain convergence and fusion of multidimensional data. These networks aggregate sensing and business data collected by various terminals and AI applications toward edge nodes and cloud platforms, while also integrating supplementary data from other network nodes. This process results in a global dataset characterized by broader coverage and richer information dimensions, thereby establishing a high-quality data supply platform for AI applications.

With regard to computing power and models, the 6G network enables the hierarchical exposure of distributed, heterogeneous computing and storage resources. Terminals and edge nodes are dedicated to lightweight model training and local inference tasks to satisfy low-latency requirements, whereas the cloud supplies centralized, large-scale computing capacity along with massive storage resources for complex model training and global inference tasks. This establishes a coordinated framework characterized by "fast response at the edge and end" and "powerful computing capacity at the cloud."

Regarding network connectivity, the 6G network serves as the core collaborative infrastructure, delivering low-latency, high-reliability deterministic connectivity to facilitate full-domain data flow and computing power scheduling. This capability solidifies the foundational support required for the efficient operation of emerging applications.

AI Model Life Cycle Management

AI models serve as the core enabler of emerging applications. Therefore, supporting their full LCM constitutes the essential prerequisite and foundation for enabling the efficient operation of these applications via 6G intelligent capability exposure. Figure 4-2 illustrates the AI model LCM within 6G networks. By systematically exposing its endogenous intelligent capabilities, the 6G network achieves fine-grained and automated management of diverse AI models. This management encompasses registration and discovery, training iteration, deployment and inference, as well as monitoring and optimization.

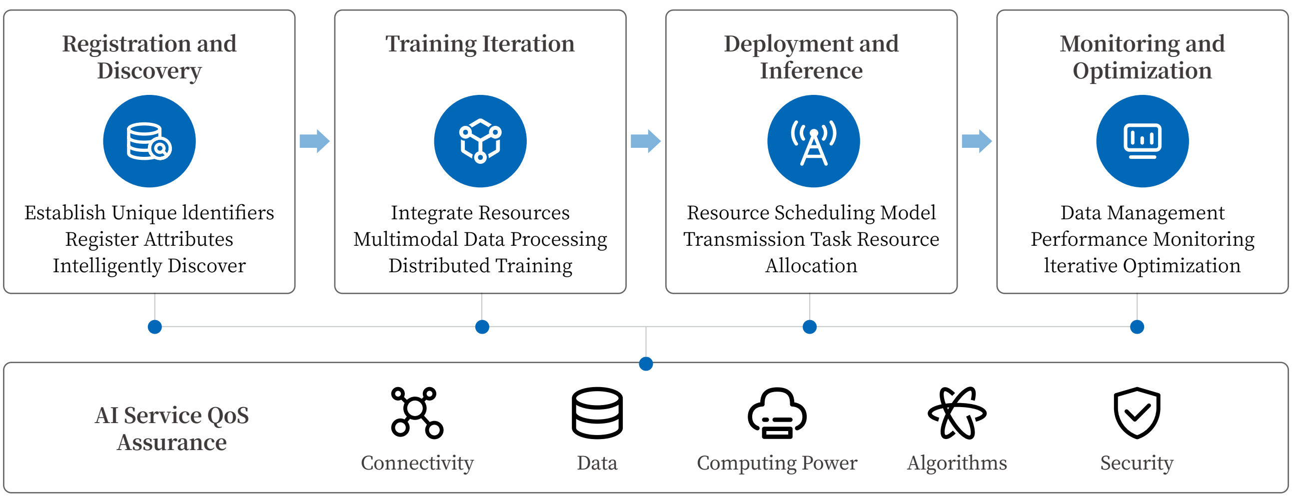


Figure 4-2 AI model life cycle management

The 6G network will assign a unique identifier to each AI model or intelligent agent and register and store its attributes, including functions and performance. This enables identity traceability and association throughout the entire lifecycle. Building upon this foundation, the 6G network exposes capabilities for model/intelligent agent discovery and collaborative scheduling. It can also intelligently match appropriate AI models based on task requirements, thereby laying the groundwork for collaborative sensing, data sharing, and other advanced functions in complex scenarios.

The 6G network integrates distributed, heterogeneous storage and computing resources into a unified resource pool. During the model training phase, it delivers efficient resource support for processing massive multimodal training data and executing large-scale distributed training tasks. During the deployment and inference phase, it enhances model transmission efficiency and precisely allocates the required communication and computing resources for these tasks. Simultaneously, the network exposes secure data collection channels and a collaborative computing environment. It also implements unified management of interactive, status, and incremental data generated during model training and inference. Leveraging this data, the network conducts real-time monitoring and evaluation of model performance. Should performance degradation be detected, it can automatically trigger closed-loop remediation actions, such as parameter updates, iterative optimization, or model reselection and retraining.

The 6G network constructs AI model QoS indicators by considering the services involved in each lifecycle phase. It dynamically senses model performance status at every stage, establishes mapping correlations and enables collaborative control between network-level Key Performance Indicators (KPIs) and application-level QoS. Ultimately, this achieves end-to-end service quality assurance for AI applications.

AI Service QoS Assurance

The QoS assurance in traditional communication networks focuses primarily on connectivity-oriented performance metrics, such as latency and throughput. Beyond traditional communication resources, 6G networks integrate diverse resource elements—including distributed heterogeneous computing power, storage, data, AI algorithms, and tools—for the orchestration and scheduling of AI services. Consequently, constructing a multi-dimensional comprehensive evaluation system is essential to ensure the QoS enabled by endogenous intelligence.

Building upon the traditional QoS metric system, the 6G network expands to establish new QoS assurance metric elements covering the dimensions of connectivity, data, computing power, algorithms and security (as shown in Table 4-1). By incorporating the quantitative threshold requirements of specific scenarios, a differentiated QoS assurance system is established to ensure AI service quality compliance. Following the accurate identification of service requirements for emerging applications, 6G networks specify the corresponding network functional modules and perform quantitative definition along with dynamic configuration of QoS assurance indicators across all dimensions. This achieves precision QoS assurance tailored to user demands.

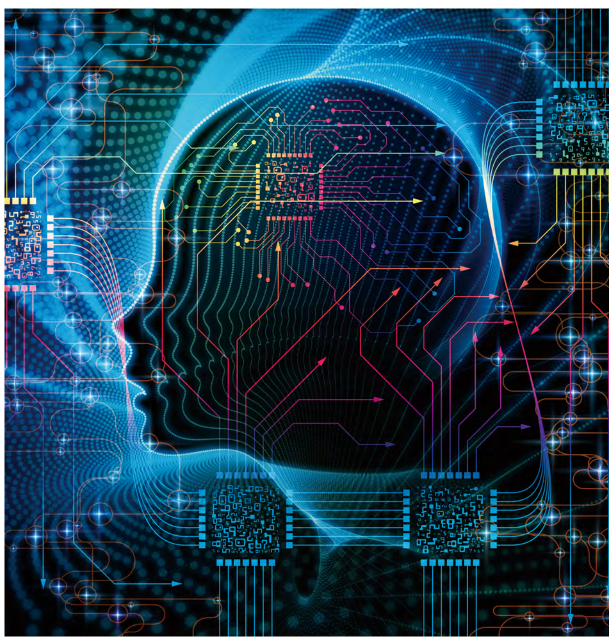
Table 4-1 Metric elements of AI Service QoS assurance

Elements	Introduction
Connectivity	Based on the connectivity requirements of massive emerging applications, differentiated communication metrics and QoS assurance mechanisms are designed on the basis of traditional QoS indicators
Data	Data quality assurance is provided for the full lifecycle management of data in multi-dimensional heterogeneous environments
Computing Power	By introducing network computing power-related metrics into the QoS assurance system, computing power resources are dynamically matched in combination with the characteristics of AI tasks to improve the utilization efficiency of computing power
Algorithm	Model availability, tool calling capability and other key aspects are provided for different application scenarios
Security	Multi-dimensional QoS assurance is implemented to cover the full-process security protection of data, algorithms, transmission and computing

4.3 Development Trends of Enabling Emerging Applications via 6G Intelligent Capability Exposure.

As 6G network technologies integrated with AI mature iteratively, 6G intelligent capability exposure will evolve from “technology enablement” to “value symbiosis,” becoming the core engine that drives the realization of “ubiquitous intelligent connectivity.” It will give rise to four core development trends that continuously redefine the innovation boundaries and industrial landscape of emerging applications.Technological collaboration is deepening toward global intelligent connectivity, thereby forging a highly efficient enabling foundation. Cloud-network-edge-end AI collaboration will transcend the existing hierarchical scheduling model, enabling dynamic linkage and intelligent scheduling of global computing power, data, and models. Integrated with distributed large models and multi-agent collaboration technologies, this collaboration will give rise to a global intelligent computing system characterized by “real-time edge-end response and global cloud-side optimization.” This system will substantially enhance the operational efficiency and adaptive capacity of emerging applications in complex scenarios. The LCM of AI models will advance toward automation and intelligence via the integration of digital twins and federated learning. This will enable secure collaborative training and rapid iteration of cross-domain models, thereby breaking down data silos and barriers to model reuse.

Ecological systems are expanding toward exposure symbiosis, thereby unleashing cross-boundary innovation vitality. 6G intelligent capability exposure will evolve from single-resource exposure to the full-chain exposure of "capabilities-services-ecology." By leveraging standardized interfaces and the AIaaS paradigm, a diverse ecosystem is established, encompassing developers, operators, and industry users. Cross-industry ecological integration is accelerating, leading to the deep embedding of network capabilities within various industries and their close coupling with industry demands. This drives the evolution of AI applications from single-point innovation to large-scale implementation, fostering a symbiotic dynamic in which networks enable industries and industries, in turn, feedback into networks. Service provision is evolving toward precise customization to meet scenario-specific demands. The AI service QoS assurance system will be further refined, evolving from multi-dimensional indicator management to scenario-specific dynamic adaptation. For differentiated scenarios—such as XR/holographic communication and industrial-grade collaborative control—customized solutions are provided for computing power allocation, latency assurance, and security protection. Meanwhile, service boundaries continue to expand, extending beyond the convergence of "communication-sensing-computing-intelligence" toward "full-factor intelligent enablement." This enables emerging applications to evolve from functional implementation to value enhancement.Security and trustworthiness have emerged as core pillars, consolidating the foundational framework for capability exposure development. As the scope of capability exposure expands, security and privacy protection will be embedded throughout the entire process. Through technologies such as endogenous security architecture design, data anonymization, and intelligent identity and access management, a tripartite assurance system is established, integrating "capability exposure-security management and control-privacy protection." Moving forward, 6G intelligent capability exposure will be anchored by technological collaboration, directed toward ecological symbiosis, guided by precise service provision, and bounded by security and trustworthiness. This integrated approach will continuously enable innovation in emerging applications.



05 6G iNDT Leads the New Paradigm

The endogenous intelligence of 6G systems drives the deep integration of digital twin and AI technologies, leading to the evolution of intelligent Network Digital Twin (iNDT). iNDT possesses autonomous decision-making capabilities. Through the construction of high-precision network mirrors, iNDT enables network twinning and environmental simulation. It alleviates data acquisition challenges and provides a testing ground for validating network strategies and fostering AI innovation. Gradually, iNDT fosters an industrial ecosystem encompassing technology development, solution implementation, and operational services. Furthermore, it empowers innovative applications across numerous vertical industries.

5.1 6G iNDT Operational Mechanism

As a core part of the 6G endogenous intelligence architecture, iNDT coordinates with all architecture layers. The multi-level deployment of the architecture provides a distributed and collaborative foundation for it. In turn, iNDT serves as a core enabler for implementing intelligent functionalities across the architecture, establishing a mutually reinforcing relationship. At the edge, the edge intelligence collaborates with the RAN to collect real-time multimodal fused data. By leveraging AI models within iNDT, the edge intelligence performs high-precision dynamic channel modeling. This provides accurate environmental sensing to support real-time intelligent decision-making at the edge. Through modeling network topology, service flows, and resource states, iNDT enables global situational awareness, policy simulation, and collaborative optimization.

iNDT operates based on a dual-loop collaborative mechanism, consisting of an inner closed-loop verification within the twin space and an outer closed-loop feedback from the physical network, as illustrated in Figure 5-1. This mechanism establishes a complete operational chain that includes virtual simulation, decision validation, network deployment, data feedback, and model iteration. This effectively enables real-time and highly efficient virtual-physical collaboration between AI strategies and the physical network.

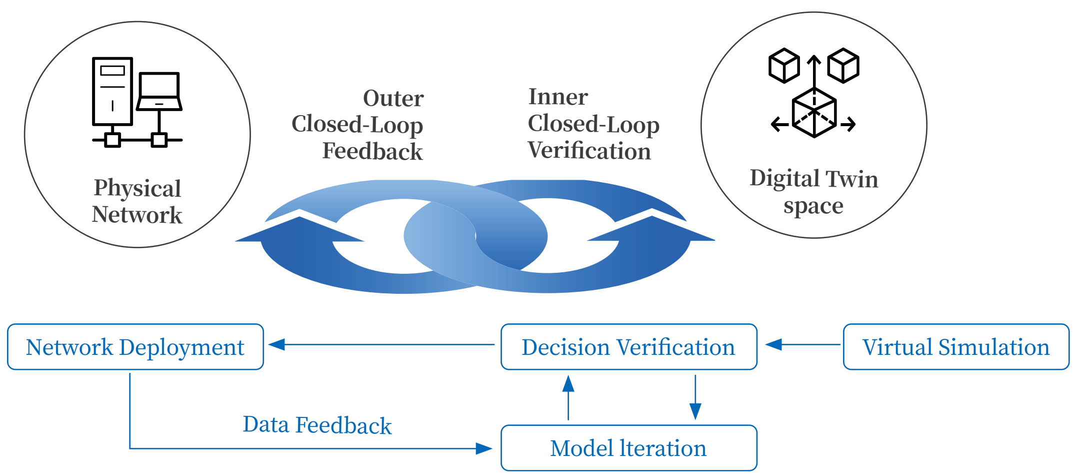


Figure 5-1 Dual-loop collaborative mechanism for virtual-physical interaction

Inner closed-loop: Within the digital twin space, the inner closed-loop conducts pre-evaluation and optimization of AI strategies. It replicates the network topology, wireless environments, and user behaviors to evaluate the performance and risks of resource scheduling, beamforming adjustments, and other schemes. By injecting disturbances—such as traffic surges and network failures—to test robustness, the system can automatically trace issues to their root causes in the data or algorithm layers for optimization. This process forms an iterative "verification-diagnosis-optimization" loop. This approach eliminates the risks inherent in live-network testing. Furthermore, it supports collaborative multi-model training, by simulating concurrent multi-task scenarios—for example, beam management, interference suppression, and traffic scheduling. This resolves strategy conflicts that arise from traditional, independently trained models.

Outer closed-loop: In the physical network, the outer loop employs a real-time and efficient virtual-physical mapping mechanism to facilitate the progressive deployment of well-validated AI models from the digital twin space. Real-time operational data from the physical network is continuously fed back to the digital twin space. This feedback dynamically refines both the training datasets and the decision rules of the intelligent agents. Through this bidirectional interaction, the accuracy of network state mapping is perpetually enhanced, ultimately establishing a virtuous evolution cycle. In this cycle, the physical network continuously informs and refines the digital twin.

As a key enabler within the 6G architecture, the iNDT delivers three core values:

Efficient Generator of Training Data

The iNDT generates large-scale, multi-dimensional, and intrinsically labeled synthetic data on demand, effectively addressing the challenges of scarce, high-quality training data for AI. Furthermore, it provides dedicated computing resources and exclusive training environments, which faciliates the development of highly generalizable and accurate AI models. Thereby, iNDT accelerates the optimization and iteration of 6G-related AI models, advancing network intelligence toward higher levels of autonomy.

Laboratory for 6G AI Technology Innovation

The iNDT enables real-time mapping and high-fidelity modeling of the physical environment. Leveraging this capability, it establishes a complete workflow from algorithm innovation to system-level verification. This provides a simulation environment for pioneering 6G AI technologies, significantly reducing the trial-and-error costs associated with AI strategies and accelerating model iteration. Moreover, it provides essential support for multi-agent collaborative training and decision-making, thereby continuously enhancing the decision capabilities of network intelligent agents.

Accelerator of Network Precision Optimization

The iNDT establishes a virtual-physical interaction between the physical network and the twin space. Combined with the dual-loop collaborative mechanism, it minimizes the impact of new strategies on the live network. Verified solutions are deployed in the physical network, and their execution results are fed back to the twin space for iterative optimization. This mechanism ensures continuous improvement of user experience without user awareness and guarantees precise optimization of network strategies. iNDT establishes a closed virtual-physical interaction loop between the physical network and the digital twin space. Guided by the dual-loop collaborative mechanism, this interaction minimizes the operational risks and potential impacts of deploying new strategies on the live network. Verified solutions are deployed to the physical network, and their execution results are fed back to the digital twin space for iterative optimization. This closed-loop process ensures the continuous improvement of user experience transparently and guarantees the precision optimization of network strategies.

5.2 Key Technologies of the 6G iNDT

As a critical support for 6G endogenous intelligence, the iNDT relies on foundational technologies, including multidimensional data acquisition and sensing, as well as high-fidelity dynamic modeling on both the network and wireless environments. Collectively, these technologies establish a digital capability foundation spanning the entire network lifecycle. As shown in Figure 5-2, multidimensional data acquisition and sensing technology enables real-time collection and fusion of multidimensional data from the physical network. This process provides the essential data foundation for subsequent high-fidelity dynamic modeling of the network and wireless environments. For wireless modeling, environmental reconstruction information derived from 3D environment modeling serves as input for AI-based wireless channel modeling, which subsequently generates high-fidelity wireless channel models. Together, these technologies transform the complex physical network and its wireless environment into a computable, simulatable, and predictable digital system. Driven by technological paradigm innovation, they provide a new pathway for upgrading all aspects of network operations.

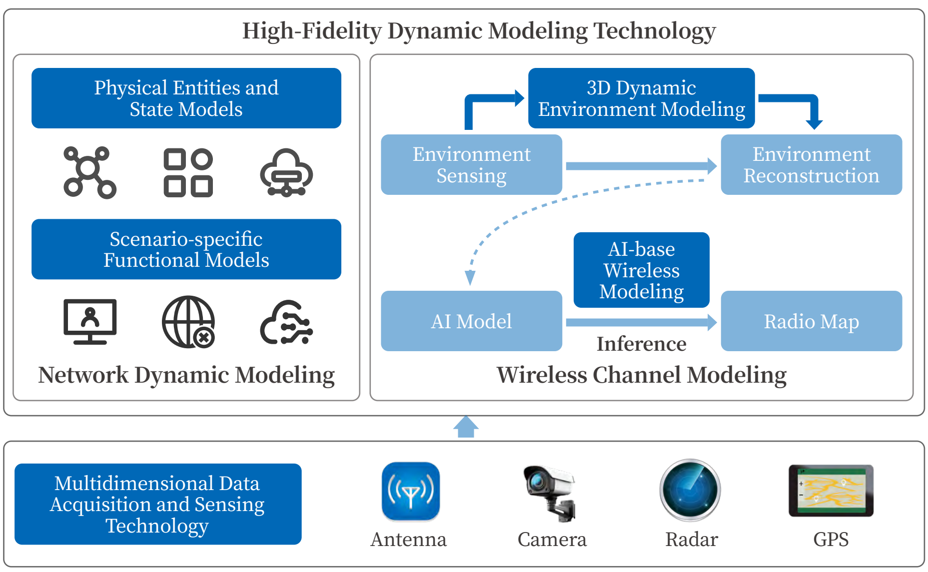


Figure 5-2 Key technologies of the iNDT

Multidimensional Data Acquisition and Sensing Technology

Multidimensional data acquisition and sensing constitute the foundational first step in building a digital twin. Its objective is to achieve real-time, accurate acquisition and fusion of data from the physical network. Given the complexity and diversity of network scenarios, a hierarchical and categorized data acquisition mechanism is essential. This mechanism must adapt to the distinct data characteristics of different network layers (e.g., RAN, core network) while balancing multiple requirements, such as transmission efficiency and quality control. Large volumes of data require compression or semantic feature extraction to alleviate bandwidth pressure on transmission interfaces. AI-enhanced data governance techniques are employed to ensure data quality and consistency. Standardized interface specifications are implemented to overcome data silos resulting from multi-vendor equipment deployments. To address challenges including user privacy protection and sample scarcity in special scenarios, intelligent technologies such as generative AI and federated learning are leveraged. These technologies enable privacy-preserving anonymization and sample augmentation while ensuring data security and regulatory compliance.

Network Dynamic Modeling Technology

This technology establishes a mirror of network states and includes both fundamental models and scenario-specific functional models. The foundational models focus on the physical entities and their states, providing precise digital representations of core elements including network topology, network element characteristics, and resource distribution. The functional models are designed for the deep integration of functional modeling for the multi-scenario, multi-task environments of future networks. This establishes models for functional and scenario models for network operational status assessment, trend prediction, and decision optimization. By further integrating physics-based models with AI technology, it performs dynamic simulation and trend forecasting of user distribution, service flows, and resource states. Ultimately, this process culminates in a globally accurate and bidirectionally interactive iNDT, providing a scientific and precise digital foundation for network operational decision-making and technology iteration.

Real-Time Dynamic Modeling Technology for Wireless Environments

This technology aims to build a high-fidelity iNDT that is dynamically coupled with the physical wireless propagation environment. It enables accurate digital representation of multi-dimensional wireless environments and electromagnetic wave propagation prediction, across diverse scenarios such as urban, indoor, and industrial settings. The modeling relies on two main dimensions:

3D Dynamic Environment Modeling

This involves the collaborative operation of multimodal sensing devices—including cameras, LiDAR, Millimeter-Wave (mmWave) radar, and GPS—integrated with communication terminals. This synergy enables the online, joint collection and dynamic updating of environmental geometric information, dynamic characteristics, and the electromagnetic parameters of physical materials. Among these, the automated acquisition and assignment of electromagnetic parameters for complex media are critical for transforming geometric visualization into an electromagnetic computable model. By combining lightweight, efficient 3D reconstruction algorithms with edge computing resource scheduling technologies, this approach achieves real-time replication dynamically changing scenes, while balancing modeling accuracy and computational efficiency. This overcomes the limitations of traditional modeling, which focuses only on static geometry and lacks dynamic adaptability and physical realism, thereby enabling high-fidelity dynamic environment reconstruction.

Knowledge and AI-Driven Wireless Channel Modeling

This approach addresses the limitations of traditional methods: the insufficient accuracy of traditional empirical models, the high computational complexity of deterministic ray tracing, and the high cost of measurement-based methods. Leveraging advanced generative models (e.g., diffusion models), it integrates multiple sources of prior knowledge, including urban geographic information, ray tracing simulation data, and historical measurement data. This enables rapid generation of wireless channel models for wide-area scenarios, significantly reducing modeling time and cost. However, since the training data for AI-based wireless channel modeling is often derived from simulation tools (e.g., ray tracing), the resulting models cannot perfectly replicate the complex radio wave propagation effects found in the real world. Consequently, hybrid modeling techniques that fuse knowledge-driven and data-driven methods are becoming a trend. By actively or passively collecting a small amount of real measurement data from the network, the system dynamically calibrates prediction errors of the twin model. This process continuously improves the accuracy and reliability of the wireless channel modeling.

5.3 Development Trends of the 6G iNDT

The integration of digital twin and AI technologies is evolving towards greater intelligence and autonomy, offering broad industrial application prospects. This fusion is not only reshaping operational paradigms within the information and communication technology sector but also providing a novel pathway for digital transformation across diverse industries. It has demonstrated significant application value in fields such as urban governance and transportation operations, driving leapfrog improvements in operational efficiency and intelligence levels. The 6G iNDT faces multiple challenges, which in turn drive its evolution across four key dimensions: technological approaches, operational modes, ecosystem development, and security management.

At the modeling level, a core challenge lies in balancing high fidelity with high efficiency. 6G networks are expected to support increasingly complex and diverse scenarios. Difficulties such as accounting for dynamic obstacles and acquiring material electromagnetic parameters present significant challenges in terms of computational load and latency for high-precision modeling. Consequently, iNDT is evolving towards a hybrid modeling approach that fuses knowledge and AI. This approach integrates the self-learning capabilities of AI with physical models, aiming to achieve real-time decision-making and rapid response while maintaining high model fidelity.

The emergence of iNDT is driving a shift in its operational mode from traditional offline simulation toward real-time interaction. Conventional offline mechanisms no longer meet the demands of twin-based services. Accordingly, the dual-loop virtual-physical collaborative mechanism of iNDT is accelerating toward real-time synchronization. Leveraging edge computing technologies, it enables millisecond-level sensing and modeling of the wireless environment. This breaks the spatial and temporal limitations of traditional offline simulation.

Regarding ecosystem development, iNDT is evolving from single-scenario applications towards a cross-domain collaborative twin ecosystem. By integrating data from multiple sectors-such as government , transportation, energy, and industry-it aims to construct a pan-domain intelligent twin system. This promotes deep integration between communication technologies and vertical industry applications, thereby reshaping industry operational paradigms.

The trends towards real-time operation and cross-domain collaboration impose stringent requirements on network performance. For instance, twin service scenarios such as Unmanned Aerial Vehicle (UAV) swarms and autonomous driving demand millisecond-level latency guarantees while industrial quality inspection requires large bandwidth to support stable transmission of high-definition images and 3D data. Meanwhile, the massive data streams generated by continuous operation of twin systems place stringent demands on large-scale network connectivity and data processing capabilities. Future efforts should focus on optimizing the entire process, including low-latency data acquisition, high-bandwidth transmission, and efficient processing. This is critical for ensuring precise virtual-physical synchronization.

The entire lifecycle of iNDT will generate substantial data flow, making the standardization of data security and compliance management an imperative trend. There is a growing need to ensure privacy and data security throughout acquisition, transmission, and sharing, especially in cross-domain collaborative scenarios involving multi-party data interactions with inherent risks. It is necessary to continuously improve data compliance management mechanisms and strengthen security controls over cross-domain data exchanges. Then it could establish a robust security foundation for the development of 6G iNDT.

Summary and Prospect

Looking ahead, the deep integration of AI and 6G networks will usher the communications industry into an intelligent era and drive the comprehensive transformation of networks from "connectivity providers" to "intelligent service engines." Based upon the technological system characterized by endogenous intelligence, capability exposure and digital twin, 6G networks are expected to exhibit the following key development trends:

Endogenous intelligence will evolve from "single-point optimization" toward "global optimization." With the in-depth integration of the network operation large models, channel foundation models and network intelligent agents, 6G networks will realize a fully closed-loop intelligent operation. This operational paradigm will advance from the basic perception-decision-execution cycle to the sophisticated capabilities of self-perception, self-decision-making, self-optimization, self-execution and self-evolution. Networks will gain the ability to proactively understand service demands, predict network states, dynamically optimize resource allocation and achieve autonomous evolution of the network itself.

The exposure ecosystem will shift from "technological exposure" to "value co-creation." 6G networks will foster a more open AI service ecosystem, and support developers in rapidly accessing AI capabilities through standardized API interfaces, thereby materializing the vision of "Networks as a Platform, Capabilities as a Service". Operators will transform from traditional communication service providers into builders of the intelligent service ecosystem, and collaborate with industry partners to develop innovative applications, unlocking the commercial value of the AI-6G integration.

Intelligent digital twin will expand from "single-scenario application" to "full-domain collaboration." As the in-depth integration of multi-dimensional data within 6G networks deepens, intelligent digital twin systems will break through the limitations of single network scenarios and evolve into a cross-industry, cross-regional, full-domain intelligent digital twin system. Within such twin system, the mechanism of internal closed-loop verification and external closed-loop feedback will become more sophisticated, enabling real-time optimization and sustained evolution of network strategies.

In summary, as the integration of AI and 6G continues to deepen, communication networks will evolve from an infrastructure for IoE into an intelligent hub for "the IIoE, providing core support for the digital economy and the intelligent society. In this process, the synergistic development of technology, ecosystem and security will serve as a pivotal driving force, propelling 6G networks from "technological leadership" to "value leadership." This will truly realize the vision of IIoE and inject new and robust momentum into the digital transformation of human society.

CICT CICT Mobile ZGC-XNET

CICT Mobile Communication Technology Co., Ltd.
www.cictmobile.com

ZGC Institute of Ubiquitous-X Innovation and Applications
www.zgc-xnet.com