

Requirements, Vision, and Potential Key Technologies for AI Edge White Paper



AI Edge Alliance

2026.04

Content

Foreword	4
1. Background and Requirements.....	4
1.1 Industrial and Technological Background of DOICT Convergence	4
1.2. Global R&D Status.....	6
1.2.1 ITU-R's Vision on Integration of AI and Communications	6
1.2.2 Industrial R&D Status	7
1.2.3 Academic Research Status.....	9
1.3 Demand and Driving Forces.....	13
2. Technical Connotation of AI Edge.....	14
2.1 Definition and Key Features of AI Edge	14
2.1.1 Definition.....	14
2.1.2 Three Key Features of AI Edge	15
2.2 Technological Advantages of AI Edge.....	16
2.3 Innovations of AI Edge	17
3. Typical Application Scenarios and Potential Values of AI Edge.....	18
3.1 Industrial Robots and Intelligent Manufacturing.....	19
3.1.1 Scenario Description	19
3.1.2 Potential Value Analysis	20
3.2 Smart Energy and Grid Dispatch.....	21
3.2.1 Scenario Description	21
3.2.2 Potential Value Analysis	21
3.3 Smart Agriculture and Unmanned Agricultural Machinery	22
3.3.1 Scenario Description	22
3.3.2 Potential Value Analysis	23
3.4 Communications and Regulation of Low-Altitude UAV	24
3.4.1 Scenario Description	24
3.4.2 Potential Value Analysis	24
3.5 Embodied Robot Training Field.....	26
3.5.1 Scenario Description	26
3.5.2 Potential Value Analysis	26
3.6 Edge-Enhanced Immersive XR	27
3.6.1 Scenario Description	27
3.6.2 Potential Value Analysis	29
3.7 Intelligent Driving and Vehicle-Road Collaboration	30
3.7.1 Scenario Description	30

3.7.2 Potential Value Analysis	30
3.8 Emergency Communications and Support	31
3.8.1 Scenario Description	31
3.8.2 Potential Value Analysis	32
3.9 Smart Sports	33
3.9.1 Scenario Description	33
3.9.2 Potential Value Analysis	34
3.10 Robotic Guide Dog.....	35
3.10.1 Scenario Description	35
3.10.2 Potential Value Analysis	36
4. Technical Directions and Main Challenges of AI Edge	37
4.1 System Architecture	37
4.2 AI for Edge Technology	40
4.2.1 The Rise of AI-for-Edge.....	40
4.2.2 Core Values of AI-for-Edge	40
4.2.3 Core AI Requirements of Edge Networks	42
4.2.4 Key Paths to AI-for-Edge Performance Improvement	44
4.2.5 Testing and Verification of AI-for-Edge Models and Algorithms	46
4.3 AI over Edge Technology	47
4.3.1 Multi-modal Sensing and Fusion.....	48
4.3.2 Model Lightweighting and Low-Latency Inference Technologies	48
4.3.3 Cloud-Edge-Device Collaboration for Large, Medium, and Small Models...	50
4.3.4 AI Agent Technology	53
4.3.5 End-to-End Information Service Technologies for Embodied AI.....	55
4.3.6 Data Security and Privacy	57
4.4 Chip and Computing Power Foundation	60
4.4.1 Chip Architecture Innovation with Integration of Communication, Sensing, Intelligence, Computing, and Control	60
4.4.2 Intelligent Scheduling Engine for Global Heterogeneous Computing Power	63
4.4.3 Open Ecosystem for Intelligent Computing Power	66
4.5 AI Edge System, Platform, and Testing	69
4.5.1 AI Edge System and Platform	69
4.5.2 AI Edge Testing.....	72
5. Conclusions	74
References	74
List of abbreviations.....	78
List of contributors to the white paper	80

Foreword

Mobile communications networks are transcending traditional single-connectivity services, accelerating their evolution toward integrated information services that integrate communications, sensing, intelligence, computing, and control. The functions of communications networks will be further decentralized, forming super-converged edge network nodes. Through the platformization and opening of network connectivity resources, computing power resources, and storage resources, these nodes will provide users with low-latency, intelligent, and customized services to meet the development demands of multi-scenario services in 5G and future 6G. In this context, AI Edge has emerged. AI Edge is a comprehensive mobile information service infrastructure. Based on a network-embedded heterogeneous, open, programmable, and shared computing power base, it achieves deep convergence of data technology, operational technology, information technology, and communications technology (DOICT) by understanding channel environments and user needs. It also enables on-demand orchestration of functions such as mobile edge information services, network function virtualization (NFV), and network-native AI and self-management. This white paper systematically elaborates on the background, driving forces, domestic and international research status, basic technical connotations, and core features of AI Edge, as well as its typical application scenarios and potential values. It provides in-depth discussions on technical directions such as the AI Edge system architecture, AI for Edge technology, AI over Edge technology, chip and computing power foundations, and AI Edge systems, platforms, and testing. Furthermore, it points out potential research directions and technical challenges for AI Edge.

1. Background and Requirements

1.1 Industrial and Technological Background of DOICT Convergence

Driven by the sweeping wave of digitalization, data technology (DT), operation technology (OT), information technology (IT), and communication technology (CT) are converging at an unprecedented pace. This trend of convergence is an inevitable outcome propelled by the dual engines of industrial transformation and technological innovation.

From an industrial perspective, industries worldwide are accelerating their digital transformation. The manufacturing industry aims to achieve intelligent manufacturing upgrades through technological convergence, enhancing production efficiency, reducing costs, and improving product quality and customization levels. For instance, automotive manufacturers leverage DT to mine and analyze vast amounts of data on the production line, utilize OT for precise control of the manufacturing process, rely on IT to build intelligent management systems, and employ CT to enable efficient communication among equipment and between the factory and external entities, thereby establishing a highly automated and intelligent manufacturing system. The energy industry follows a similar path. In smart grid construction, DT is used to analyze power supply and users' demand, OT is used to ensure the stable operation of power systems, IT is used to enable energy management informatization, and CT

is used to support real-time transmission of data towards remote control center, enhancing energy utilization efficiency and power supply reliability. **From a technological perspective**, each technology faces its own developmental bottlenecks, urgently requiring convergence for breakthroughs. In the CT domain, while 5G has brought significant performance improvements, it still falls short in meeting the stringent requirements for low delay, high reliability, and massive connectivity in scenarios such as the Industrial Internet and autonomous driving, necessitating collaboration with other technologies. In the IT domain, the development of cloud computing has driven the centralization of computing resources, but constraints such as data transmission delay and privacy/security issues limit its application in certain scenarios, requiring optimization through methods such as edge computing. OT focuses on specific industrial scenarios, and its closed nature and limitations in data processing capabilities have become prominent during digital transformation, urgently requiring the introduction of DT and IT technologies for open interconnection and intelligent upgrades. DT, in turn, relies on CT, OT, and IT to acquire multi-source data and leverage them to realize the value of data.

➤ **Technological trend 1: Intelligent collaborative evolution**

AI, as the core of DT, will be deeply embedded in OT, IT, and CT. In the OT domain, AI empowers industrial robots to perform more precise and flexible operations, making intelligent decisions based on real-time sensing data. In the IT domain, cloud computing platforms utilize AI for intelligent resource scheduling, improving service quality. In the CT domain, communication networks leverage AI to optimize network planning, fault diagnosis, and traffic management, achieving network self-optimization and self-healing. The intelligence fostered by multi-technology convergence is evolving towards end-to-end intelligent collaboration, forming a unified intelligent decision-making system from the equipment layer, network layer, to the application layer. An example is the collaborative closed loop in smart factories, involving real-time sensing of equipment status, automatic optimization of network transmission, and dynamic adjustment of production plans.

➤ **Technology trend 2: Enhanced edge convergence**

Edge computing has become a key hub for DOICT convergence. At the network edge, CT provides network connectivity, OT equipment generates data, IT provides computing and storage resources, and DT performs data analysis and processing. Through edge convergence, data can be quickly processed locally without entirely uploading to the cloud, which reduces delay, alleviates network burden, and enhances data security. For instance, in intelligent transportation, roadside edge nodes process traffic flow data collected by cameras (DT) in real time, integrate it with traffic signal control (OT), provide feedback to vehicles and traffic management centers via wireless communication (CT), and simultaneously utilize edge computing (IT) capabilities for real-time decision-making to optimize traffic signal timing.

➤ **Technology trend 3: Building unified standards and an open source ecosystem**

Establishing unified standards and an open source ecosystem is crucial. Currently, the lack of standardized technical standards across industries hinders the large-scale adoption of DOICT convergence. For example, the multitude of industrial communication protocols makes it difficult for equipment from different manufacturers to interconnect. To address this, the industry is actively promoting standardization efforts, such as developing unified data interfaces, communication protocols, and security specifications. Open source projects are also

emerging, which fosters technology sharing and innovation, reduces the threshold and R&D cost for enterprises, and attracts more participants to build a converged technology ecosystem.

➤ **Industry value 1: Enhancing industrial efficiency and innovation capability**

DOICT convergence breaks down information silos across industries, enabling the free flow and deep mining of data, and catalyzing new business models and applications. Taking the healthcare industry as an example, by leveraging CT for remote communication of medical equipment, IT for building medical information systems, DT for analyzing patients' data such as medical records and images, and OT for ensuring precise operation of medical equipment, innovative services such as telemedicine and intelligent diagnosis can be developed. This enhances medical efficiency and quality, providing patients with more convenient and precise medical services. In the manufacturing industry, convergence enables supply chain collaborative optimization, real-time production process monitoring, and intelligent production scheduling, significantly improving production efficiency and resource utilization.

➤ **Industry value 2: Optimizing user experience and service quality**

In the consumer domain, converged technologies deliver more intelligent and convenient experiences to users. In smart homes empowered by DOICT, users can remotely control appliances via smartphones (CT communication). Devices automatically sense the environment and user habits (OT sensing and DT analysis) and intelligently adjust their operating status (IT control), creating a comfortable and energy-efficient living environment. In smart mobility, based on CT communication and DT data analysis, the navigation system plans the optimal route in real time, while the vehicle's advanced driver assistance system (ADAS) (combining OT and IT) ensures driving safety, enhancing travel efficiency and comfort.

➤ **Industry value 3: Promoting industrial upgrading and economic growth**

DOICT convergence drives the digital and intelligent transformation of traditional industries, fosters emerging industries, and becomes a new engine for economic growth. Traditional agriculture, leveraging converged technologies, develops smart agriculture, achieving precision planting and breeding, improving agricultural production efficiency, and promoting agricultural modernization. Simultaneously, it catalyzes emerging industries such as the Industrial Internet, intelligent logistics, and digital finance, creating new jobs and new profitable source, enhancing a country's overall industrial competitiveness, and promoting sustainable economic development.

1.2. Global R&D Status

1.2.1 ITU-R's Vision on Integration of AI and Communications

In the current era of rapid information technology development, the convergence of communication and AI has become a key trend in the development of 6G. In June 2023, the International Telecommunication Union -Radio Communication Sector (ITU-R) released the *Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond*, explicitly identifying "the convergence of AI and communications" as one of the six major application scenarios for 6G, thereby setting the direction for the global development of 6G.

Judging from the outputs of relevant ITU-R working groups, in future 6G networks, AI will no longer be a communication aid but will be deeply embedded in all aspects of the communication system. On the one hand, communication systems provide ubiquitous

connectivity for AI, enabling AI services to reach various devices and users, thus achieving universal intelligence. For example, in distributed AI model training, 6G networks, with their high-capacity, low-latency, and high-reliability communication capabilities, can facilitate efficient data and model exchange among intelligent terminals, enhancing training efficiency while protecting user privacy.

On the other hand, AI empowers communication systems, enabling intelligent O&M and performance optimization. With the help of AI algorithms, 6G networks can analyze massive amounts of data in real time, intelligently allocate communication resources, and quickly respond to issues such as network congestion and signal interference, thereby enhancing network flexibility and adaptability. Taking the application of collaborative robots in 6G scenarios as an example, the real-time model inference capabilities provided by network-native intelligence can meet their stringent requirements for low-latency, high-accuracy AI services.

Furthermore, among the emerging technology trends identified by ITU-R for IMT-2030 (6G), technologies for AI-native communications (AI-native air interface design and AI-native radio network) are significant. This implies that AI and communication will deeply collaborate from foundational aspects such as air interface design and network architecture construction, establishing a new communication paradigm. 6G will not only need to serve as a connectivity-oriented infrastructure but also natively introduce support for AI within its architecture. The AI applications carried by 6G will feature characteristics such as fragmented AI demands, three-dimensional coverage, diversified interaction, open and customized AI services, and integrated capabilities. Through deep integration and optimization of resources across multiple dimensions such as connectivity, computing power, data, and algorithms, 6G can effectively guarantee the quality of AI services, which will also become a key driving force for transformation at the service level.

1.2.2 Industrial R&D Status

➤ AI-RAN Alliance

The AI-RAN Alliance is an international collaborative organization focused on promoting the deep convergence of AI with radio access networks (RAN). It was officially established and announced on February 26, 2024, at the GSMA Mobile World Congress (MWC 2024) in Barcelona, Spain. The establishment of the AI-RAN Alliance aims to enhance the performance, efficiency, and flexibility of RAN by integrating AI technologies into cellular communication networks, thereby driving the development of 5G and the upcoming 6G networks. Its core missions include: improving mobile network efficiency, reducing power consumption, transforming existing infrastructure, and creating new business opportunities for telecom companies leveraging AI in the 5G and 6G era. The AI-RAN Alliance focuses its research and innovation efforts on the following three main areas: 1) AI for RAN: Utilizing AI to improve the spectrum efficiency, energy efficiency, and cost-effectiveness of RAN; 2) AI and RAN: Deeply converging AI with RAN processes to achieve efficient resource utilization and innovate AI-driven revenue models; 3) AI on RAN: Deploying AI services at the RAN edge to enhance operational efficiency and provide new intelligent services to end users. At MWC Barcelona 2025, the AI-RAN Alliance showcased multiple demonstrations covering AI-for-RAN, AI-and-RAN, and AI-on-RAN, involving key areas such as air interface technology, energy-saving measures, spectrum sensing, and network orchestration.

➤ Next G Alliance

The Next G Alliance defines AI as a "capability multiplier" for 6G networks. In its released *Roadmap to 6G*, it proposes an "AI-driven network autonomy" system. This system comprises a three-layer intelligent architecture: the edge layer focuses on real-time decision-making (e.g., millisecond-level interference suppression), the regional layer is responsible for collaborative optimization (e.g., cross-base station resource scheduling), and the core layer handles global strategies (e.g., service load prediction). The Next G Alliance has particularly validated the application of AI in dynamic spectrum sharing, achieving adaptive allocation between licensed and unlicensed spectrum through reinforcement learning algorithms, improving spectrum utilization by over 40%. Its perspective holds that the convergence of communication and AI requires building a synergistic ecosystem of "hardware-algorithms-data." It is currently promoting the standardization of chip-level AI acceleration units and collaborating with universities on security research for federated learning in network optimization.

➤ **EU HeXa-X project**

HeXa-X, as the EU's flagship 6G project, proposes the concept of an "AI-native network architecture," deeply embedding AI into the full-stack design from the air interface to the core network. The released phase II technical report outlines three main directions: First, AI-driven integration of communication, sensing, and computing, achieving joint scheduling of communication, sensing, and computing resources through multi-task learning models. Second, the network intelligent orchestration, ensuring end-to-end service quality based on digital twins and reinforcement learning, with service availability reaching 99.999% in industrial scenario verifications. Third, a trustworthy AI framework, addressing data security and model robustness issues through federated learning and differential privacy technologies. The project emphasizes that 6G requires the establishment of an "AI-as-a-Service (AIaaS)" platform. Currently, it has already completed the specification of cross-vendor AI model interfaces, providing a technical benchmark for industry chain collaboration.

➤ **6GANA (6G Network AI Alliance)**

6GANA members include 31 organizations such as operators, equipment manufacturers, internet service providers, and universities. Its mission and goal are to actively promote the AIaaS transformation of 6G networks from both technology and ecosystem perspectives. Specifically, 6GANA aims to build consensus on network AI through joint research across the entire ecosystem, covering information and communication technology (ICT) equipment manufacturers (such as chipmakers, network infrastructure providers, and mobile network operators), vertical industries, AI service providers, AI solution providers, AI academia, and other stakeholders. It seeks to drive AI as a new capability and service for 6G networks, accelerating the arrival of the era of ubiquitous intelligence. Centering on the core proposition of "mutual empowerment between networks and AI," 6GANA has constructed a "three-horizontal, three-vertical" technical system: horizontally covering the three major fields of network architecture, data governance, and security and trustworthiness; vertically spanning the three stages of requirement definition, technology R&D, and industrial implementation. Its released *6G Endogenous Intelligence White Paper* systematically proposes a collaboration paradigm of AI4NET (AI enhancing networks) and NET4AI (networks supporting AI): In AI4NET, it verified a network fault self-healing solution based on graph neural networks, improving recovery speed by 50%; in NET4AI, it designed an edge-cloud collaborative model training framework, tripling distributed training efficiency. 6GANA promotes technology

implementation through cross-industry working groups and has currently developed 12 technical specifications for scenarios such as Vehicle to Everything and industrial control, driving the convergence of communication and AI from concept to industrialization.

➤ **IMT-2030 (6G) Promotion Group**

The IMT-2030 Promotion Group positions "Intelligent Connectivity of Everything" as the core vision for 6G. In its *White Paper on 6G Vision and Candidate Technologies*, it clarifies a "double helix" development path for the convergence of communication and AI. On the one hand, it promotes the reconstruction of communication systems via AI, introducing deep learning-assisted waveform optimization in air interface design, increasing communication rates by 25% in complex environments. On the other hand, it aims to build a communication infrastructure supporting AI services, proposing the concept of an "AI-agent Communication Network (ACN)" to ensure real-time edge AI inference through ultra-low-latency communication. The promotion group, jointly with industry, academia, and research institutions, has established a 6G AI testbed, completed the verification of key technologies such as reconfigurable intelligent surface and AI collaborative transmission, and formulated the *6G AI Technology White Paper*, providing a Chinese solution for the formulation of global 6G standards. It emphasizes that integrated development needs to balance technological innovation and industrial maturity and is currently promoting the collaborative optimization of AI model lightweighting and communication protocol simplification.

1.2.3 Academic Research Status

The convergence of communication, sensing, intelligence, and computing is not only a consensus in the industry but also a cutting-edge research direction of concern in academia. In the field of communication and AI convergence, a large amount of research conducted in the early stages has achieved fruitful results in areas such as physical layer enhancement, spectrum efficiency improvement, network fault diagnosis, and energy efficiency optimization based on the AI-empowered communication system performance optimization. In the past year or two, research has primarily concentrated on the application of generative AI (GenAI) technologies in communication networks. In the field of communication and sensing convergence, the research focus has gradually shifted from the sharing and coexistence of communication and sensing over shared frequency spectrum, radio frequency (RF) hardware and software resources, to new stages like multi-station collaborative wireless communication and sensing as well as multi-modal network communication and sensing convergence. The application scenarios have also expanded from target sensing in Vehicle to Everything and low-altitude unmanned aerial vehicle (UAV) to intelligent transportation and security surveillance. In the field of communication and computing convergence, academia primarily conducts research along four paths: external computing power (e.g., mobile edge computing (MEC)), networked computing power (embedding computing power information into routing protocols), native computing power (such as C-RAN and O-RAN), and intelligent computing power (e.g., AI-RAN). The following is a brief introduction to the latest progress in several key research directions.

➤ **Telecom Foundation Models**

Large Language Models (LLMs) have the potential to revolutionize the Sixth Generation (6G) communication networks. However, current mainstream LLMs generally lack the specialized knowledge in the telecom domain. In this context, a research team from Abu Dhabi's Technology Innovation Institution (TII) and Khalifa University first proposed a design

framework to adapt general-purpose LLMs to the telecom field^[1]. Evaluation results show that the fine-tuned LLM, i.e., TelecomGPT, significantly outperformed state-of-the-art LLMs, including GPT-4, Llama-3, and Mistral in telecom mathematical modeling baseline tests, and excelled in various evaluation baselines such as TeleQnA, 3GPP technical document classification, telecom code summarization, generation, and completion.

➤ **AI-assisted physical layer design**

In terms of physical layer communication, the first typical use case is beamforming. AI models, pre-trained on extensive beamforming scenario datasets, can predict optimal beams that maximize signal strength and minimize interference. This can be achieved by leveraging multi-modal data to provide additional information about blockage probability, user state, and activity. Second, AI models can be used for channel state information (CSI) estimation purposes for the uplink and downlink transmissions. Through self-attention mechanisms and the generative capabilities of AI models, it is envisioned that models will capture the inherent relationship between uplink and downlink transmissions and utilize 3D multi-modal environmental data (including cameras, radar, LiDAR, and GPS) to select the best uplink-downlink beam pair, perfectly aligning the angle of arrival (AoA) and angle of departure (AoD) at specific user locations. Third, regarding millimeter wave (mmWave) beam prediction, a team from Southeast University transformed the mmWave beam prediction problem into a time series forecasting task. They aggregated historical observation data through a cross-variable attention mechanism, used a trainable tokenizer to convert it into text-based representations, employed Prompt-as-Prefix (PaP) technology for context enhancement, and leveraged the powerful capabilities of LLMs to predict future optimal beams^[2]. Fourth, in the field of joint source channel coding (JSCC), research has integrated channel and source coding into semantic-aware JSCC. GenAI models contribute to achieving efficient JSCC schemes to improve wireless communication performance^[3].

Furthermore, GenAI is also being used to enhance receiver performance. Diffusion models (DMs), which learn to remove noise progressively, have been widely applied in artificial intelligence-generated content (AIGC) in recent years. To verify whether DMs can be applied to wireless communication to help receivers eliminate channel noise, a research team from Shanghai Jiao Tong University proposed the Channel Denoising Diffusion Model (CDDM) for wireless communication at GLOBECOM 2023^[4]. CDDM can act as a new physical layer module after channel equalization to learn the distribution of channel input signals and then use the learned knowledge to remove channel noise. Experimental results show that CDDM further reduces the mean squared error (MSE), demonstrating better performance.

To achieve efficient characterization of wireless channels, a team from Peng Cheng Laboratory proposed a multi-task wireless foundation model, WirelessGPT. This model adopts a BERT-like Transformer architecture and adapts to wireless applications through technologies such as three-axis attention mechanisms and multi-scale data encoding. It can empower various communication and sensing tasks such as channel estimation, channel prediction, localization, and activity recognition^[5]. A team from Beijing University of Posts and Telecommunications constructed wireless channel foundation models, ChannelGPT and ChannelDS, based on the idea of fine-tuning LLMs, achieving excellent performance in downstream tasks like channel prediction^[6]. A team from Peking University proposed the concept of "Synesthesia of Machines (SoM)," based on the task-driven AI-native philosophy, to achieve intelligent

integration of communication and multi-modal sensing. Their developed LLM4CP (for channel prediction), LLM4WM (for wireless physical layer multi-tasking), and WiFo (the wireless foundation model) provide new ideas for the integration of communication and sensing in 6G networks^[7].

➤ **AI-empowered wireless sensing**

Deep learning (DL) models have fundamentally advanced the development of wireless sensing solutions, where RF data can be acquired and mapped to two-dimensional images for sensing applications, including localization, remote sensing, and resource allocation. GenAI models can enable efficient multi-modal localization solutions. The generalizability and self-attention nature of these AI models can be key to detecting the context and situational information of network users and nodes, capturing mutual displacement between multiple images, and correlating these images and their changes with the corresponding electromagnetic behavior of wireless signals.

RF signal generation technology is of great significance for wireless sensing systems. To address the current lack of high-quality time-series RF signal generation models, reference^[8] established a novel Time-Frequency Diffusion Theory and proposed the first generative diffusion model for RF signals, RF-Diffusion. It enables diverse, large-scale, high-precision automatic generation of time-series RF signals and has been successfully applied to a series of key tasks such as Wi-Fi sensing data augmentation and FDD channel estimation.

➤ **GenAI-assisted semantic communication**

Semantic communication (SemCom) faces many challenges in constructing background knowledge bases for training semantic coding models. The recently emerged generative artificial intelligence (GAI) technology holds promise for assisting background knowledge construction in SemCom and enhancing the interface capability of semantic coding models. In this context, reference^[9] proposed a GAI-assisted SemCom framework that uses GAI to assist in generating samples for training semantic coding models based on user context information. Compared to traditional SemCom, Gen-SC achieves higher semantic precision when the original training samples are insufficient. Reference^[10] proposed a semantic communication framework dedicated to image data (LAM-SC) based on AI LLMs. It designed a SAM-based Knowledge Base (SKB) and proposed an attention-based semantic integration (ASI) method and an adaptive semantic compression (ASC) coding method. Reference^[11] proposed a general generative semantic communication framework based on pre-trained foundation AI models. The framework comprises three core modules: multi-modal semantic decomposition and synthesis, semantic-aware multi-stream transmission, and low-latency semantic power allocation.

➤ **LLM-based multi-agent**

The rapid development of LLMs brings enormous opportunities to 6G communication, such as allowing users to input task requirements to LLMs via natural language for network optimization and management. However, directly applying native LLMs to 6G will encounter various challenges, such as a lack of professional communication data and knowledge, and limited model capabilities in logical inference, evaluation, and optimization. To address these challenges, reference^[12] designed an LLM-enhanced multi-agent system for 6G communication. This system constructs a professional knowledge base and tools for 6G

communication and possesses capabilities beyond the original LLM in planning, memory, tool utilization, and reflection.

➤ **AIGC services based on cloud-edge-device collaboration**

To provide low-latency and customized AIGC services, it is essential to adopt a collaborative cloud-edge-device AIGC framework. Some high-performance terminal devices can directly run AIGC models to provide services for themselves, further improving real-time response performance and security compared to edge computing. Simultaneously, mobile terminal devices can offload AIGC tasks to edge or cloud servers, enabling flexible service configuration. Considering the lightweight nature of terminal devices, models that run on these devices typically need to undergo compression and quantization to reduce computing and storage resource overhead. Reference^[13] proposed an edge adapter model that achieves a trade-off among inference accuracy, delay, and resource consumption. In terms of architecture design, reference^[14] proposed a bottom-up BAIM architecture to maximize the utilization of user data and knowledge extracted by edge expert models. This architecture effectively combines Pathways and Mixture of Experts (MoE) models, enhancing the efficiency and user experience of GenAI services through the collaboration of large cloud models and small edge models.

➤ **Wireless network architecture design for AI tasks**

Current wireless networks designed as "data pipelines" are not suitable for accommodating and utilizing the capabilities of GenAI. To this end, reference^[15] proposed a network architecture that integrates GenAI capabilities to manage network protocols and applications. By constructing a semantic-based GenAINet, semantic concepts are extracted from multi-modal raw data, a knowledge base representing their semantic relationships is built, and then GenAI models are used for planning and inference. In this mode, agents can quickly learn from the experiences of other agents to make better decisions and communicate more efficiently. Reference^[16] proposed the endogenous intelligent network architecture NetGPT, leveraging resource imbalances in cloud-edge computing to achieve efficient collaboration between LLMs of different scales at the cloud and edge. In contrast to AI exogenous networks with decoupled communication and computing resources, NetGPT can utilize converged communication and computing to deploy smaller LLMs at the edge and larger LLMs in the cloud, and intentionally implement cloud-edge collaborative computing to provide personalized content generation services.

➤ **LLM-based wireless resource management and optimization**

Due to the expanding range of user demands, optimizing various wireless user tasks poses significant challenges to network systems. Despite the progress in deep reinforcement learning (DRL), the need to customize optimization tasks for individual users complicates the development and application of a vast number of DRL models, leading to substantial computing resource and energy consumption, and potentially inconsistent results. To address this issue, reference^[17] proposed a novel approach that utilizes a MoE framework, assisted by an LLM, to effectively analyze user goals and constraints, select specialized DRL experts, and weigh every decision made by participating experts. The approach proposed in the paper reduces the need to train new DRL models for each unique optimization problem, thereby lowering energy consumption and the implementation cost of AI models.

➤ **Service quality guarantee for personalized requirements**

Personalized service should become one of the key capabilities of future 6G networks. Reference^[18] proposed the concept of multi-dimensional indicator convergence to quantify and meet highly differentiated user demands. This work proposed the concept of a "service requirement zone (SRZ)" to characterize and visualize the comprehensive service requirements of individual tasks on the user side. SRZ is defined by an eight-dimensional radar chart covering eight key performance indicators: delay, energy consumption, storage, rate, security and privacy, reliability, knowledge, and cost, setting clear boundaries for the user's personalized quality of experience (QoE). Building on this, the work further introduced "user satisfaction ratio (USR)" as an evaluation indicator on the system side to measure the overall service capability of the system in meeting tasks with different SRZs. These concepts provide a theoretical foundation and evaluation framework for achieving customized services centered on each person, offering important reference value for AI to be deployed to the edge to achieve refined and personalized resource scheduling and service guarantee.

To advance research and global cooperation in the field of integration of communication and AI LLMs, the IEEE Communications Society established the Large Generative AI Models in Telecom Emerging Technology Initiative (GenAINet ETI) in early 2024. This ETI is an academic organization involving global scholars and industry experts. Its main goal is to create an open research platform for discussing LLM technologies in communication networks. In May last year, GenAINet ETI released the first academic research paper collection on LLM technologies in communication networks, summarizing the latest research results in related fields^[19].

1.3 Demand and Driving Forces

Cross-industry integration and the continuous expansion of application boundaries are significant trends in the development of mobile communications. The demand for multi-scenario applications of 5G technology has driven the advancement of network functions virtualization and slicing, as well as the deep integration of Information and Communication Technology (ICT). Future 6G vertical industry applications, particularly in the Industrial Internet, will further promote the hyper-convergence of DT, OT, IT, and CT. Based on a network-embedded heterogeneous, open, programmable, and shared computing power foundation, it will become possible to provide vertical industry users with intelligent, deterministic, customized, and low-latency integrated mobile information services encompassing communication, sensing, intelligence, computing, and control. As a crucial carrier for hyper-converged DOICT technology, the edge network demonstrates unique value by leveraging its proximity to users to offer low-latency advantages.

On the one hand, the rapid development of AI technologies has generated a strong demand for computing power. Currently, computing power resources are no longer confined to the centralized, large-scale processing capabilities provided by cloud computing centers, but are more widely distributed at the network edge and across various terminal devices. In this context, a critical challenge urgently requiring resolution for AI development is how to efficiently integrate and utilize these distributed computing power resources, enabling "on-demand access" to computing power, extending its reach to provide low-latency, highly reliable computing services for intelligent applications in vertical industries, and realizing the

ubiquitous intelligence vision of "compute anywhere." The distributed computing power integrated within the edge network can serve as a novel type of network resource. By aggregating and utilizing fragmented computing power, it can provide "last-mile" computing power delivery services, constructing an indispensable networked extension of computing power resources.

On the other hand, emerging service scenarios such as autonomous driving, digital low altitude, robotics, and smart factories impose extremely stringent requirements on end-to-end delay, jitter, and reliability. To achieve high-quality service capabilities closer to users, network functions are continuously migrating towards the edge, establishing a new paradigm of distributed, hierarchical intelligent services. Through the flexible and efficient sharing of mobile edge network resources, mobile network functions and intelligent information services can be deployed in proximity at the network edge. Leveraging the connectivity and sensing capabilities of the edge network, it can fully interact with the physical world and enable a rapid closed-loop of sensing, inference, and execution, empowering new AI agent terminals such as low-altitude UAVs, robots, and autonomous vehicles for various low-latency embodied applications.

2. Technical Connotation of AI Edge

2.1 Definition and Key Features of AI Edge

2.1.1 Definition

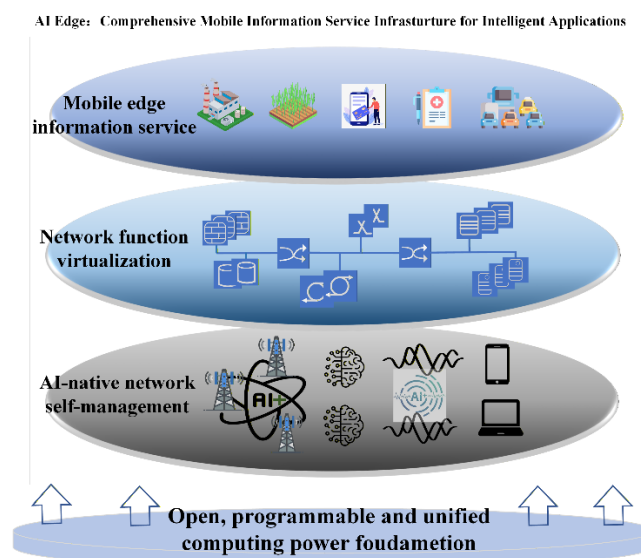


Figure 1 Definition of AI Edge

AI Edge is a comprehensive mobile information service infrastructure for intelligent applications. Based on an open, programmable, unified computing power architecture, it simultaneously realizes three major functions:

- Mobile edge information service
- Network function virtualization

- AI-native network self-management

2.1.2 Three Key Features of AI Edge

AI Edge possesses key features such as sharing, scalability, and hierarchy, elaborated as follows.

Sharing: AI Edge is highly compatible with heterogeneous computing power, such as CPU, GPU, NPU, FPGA, and SoC. Through software programmability, it achieves functional integration and capability sharing of communication, AI, sensing, network control, and computing services on a unified hardware foundation. Specifically, various heterogeneous computing hardware like CPUs, GPUs, NPUs, FPGAs, and SoCs form a unified computing power foundation. Through virtualization and pooling technologies, the computing power resources of different hardware are integrated into a unified virtual computing power resource pool. This resource pool, via designed standardized unified software interfaces, enables centralized invocation and flexible scheduling of heterogeneous computing power, thereby promoting the sharing and on-demand use of computing power resources. Functions such as communication signal processing, AI services, edge sensing, and network control do not need to concern themselves with underlying hardware differences, and they can directly obtain the required computing resources from the virtual computing power pool. With the help of a unified scheduler and collaborative management mechanism, the system can dynamically respond to service demands, achieving elastic allocation and efficient management of resources, thus improving overall computing power utilization efficiency and service deployment flexibility.

Scalability: AI Edge not only horizontally integrates computing power resources across adjacent base stations to build an elastic and expandable edge computing power network but also vertically enables cross-layer distributed intelligence through efficient cloud-edge-device collaboration, supporting the scalability of mobile information services across the entire network. Specifically, on the one hand, AI Edge constructs an edge computing power network supporting high-speed interconnection among multiple nodes. Relying on mechanisms such as computing power sensing, computing power routing, and computing power scheduling, it horizontally expands AI Edge's collaborative computing capability by cross-domain integrating computing power resources from multiple edge nodes through methods such as computing task decomposition and migration, cross-domain collaborative model inference, and data parallel training. On the other hand, AI Edge also vertically interconnects heterogeneous computing power resources across the hierarchical cloud-edge-device network. Through the adaptive deployment of large and small models on the cloud, edge, and device sides, it achieves dynamic configuration and efficient utilization of multi-scale computing power. Furthermore, through mechanisms such as in-network collaborative inference, bidirectional updates, and co-evolution of large and small models, it effectively promotes deep collaboration between centralized large-scale computing power and distributed fragmented computing power, comprehensively improving multi-dimensional service quality, including inference delay and device energy consumption.

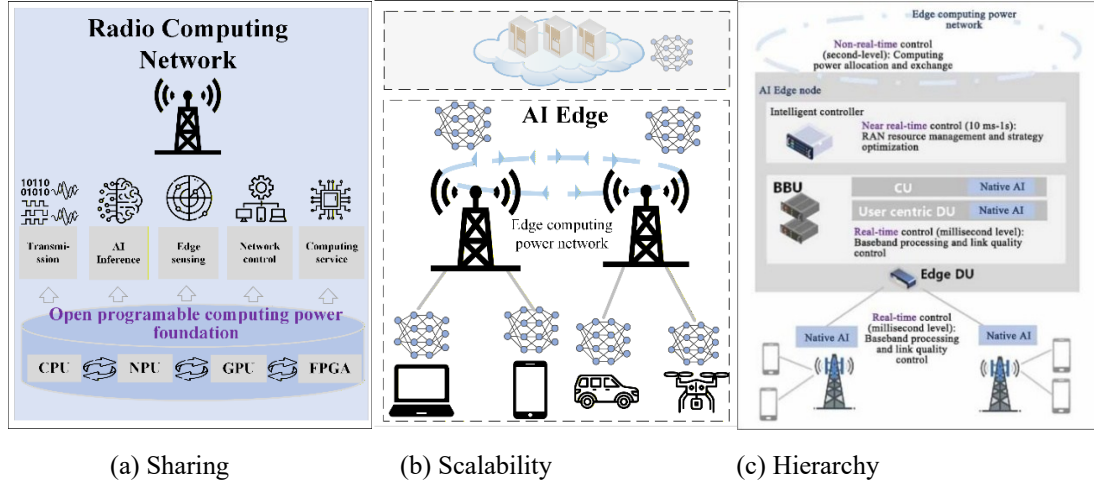


Figure 2 Key Features of AI Edge

Hierarchy: AI Edge fully incorporates the latest concepts and technologies from the AI field. It characterizes complex wireless environments using channel foundation models (like WirelessGPT), effectively utilizes the shared foundation, and directly maps user demands into fundamental capabilities of the communication networks based on the Agentic AI technology to invoke corresponding functions to achieve real-time, near-real-time, and non-real-time hierarchical network autonomous management and control from millisecond-level to second-level. Specifically, based on channel foundation models, AI Edge characterizes the multi-dimensional relationships of wireless channels and signals across the time, space, and frequency domains. Building upon this, various downstream task models embedded within wireless access units and baseband processing units complete communication sub-tasks such as channel estimation, channel prediction, beam management, interference suppression, and decoding optimization, achieving millisecond-level real-time optimization processing for RF and baseband signals. Simultaneously, leveraging intelligent optimizers deployed at edge nodes, AI Edge perceives and parses user demands in real time, dynamically generating near-real-time intelligent network control strategies, thereby significantly enhancing user experience. Additionally, through collaborative computing within the edge computing power network, AI Edge generates a non-real-time, high-performance global management solution for cross-node communication, sensing, and computing resources, achieving efficient coordination and unified orchestration of multi-edge node resources.

2.2 Technological Advantages of AI Edge

Based on the above features, AI Edge demonstrates significant advantages and value, as detailed below.

Firstly, AI Edge, based on a shared foundation, enables interoperability of DOICT capabilities and the sharing of computing power resources. It natively supports new network functions such as sensing, control, forwarding, routing, and data management, thereby significantly improving the utilization efficiency of network resources. Through the deep integration and efficient collaboration of communication, sensing, computing, control, and intelligence, it achieves closed-loop support from environmental sensing, data transmission,

intelligent analysis, and precise control, providing end-to-end integrated service support for intelligent applications.

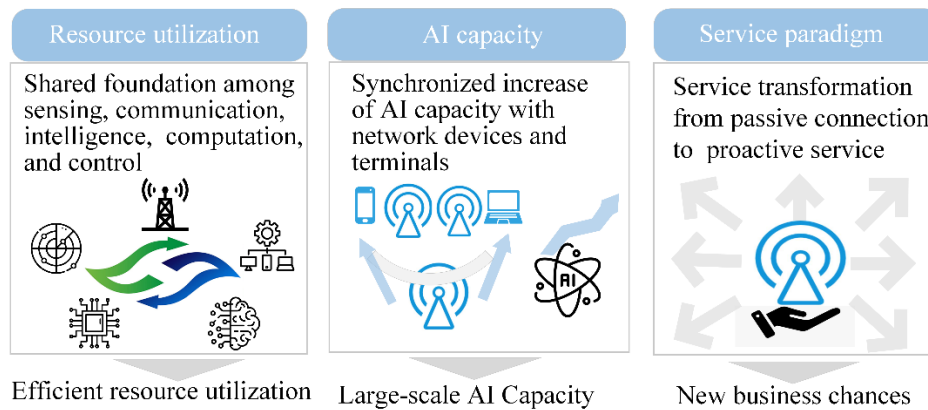


Figure 3 Key Advantages of AI Edge

Secondly, by aggregating massive distributed computing power, including from the cloud, base stations, and devices, it not only facilitates "open-source" sharing paradigm of computing power but also transforms numerous devices from mere consumers of computing power into suppliers, thereby enabling available computing power and intelligent service capabilities to increase in tandem with network scale expansion, empowering large-scale AI applications. In AI Edge, AI capabilities are no longer confined to centralized processing by a distant "cloud brain." Instead, by deeply embedding into billions of devices and network equipment, it constructs a ubiquitous neural network tightly coupled with the physical world with continuously sensing and computing, establishing a new paradigm of pervasive, seamless general intelligent services everywhere.

Thirdly, through the construction of native AI capabilities, it proactively senses complex wireless environments and differentiated user service demands, promoting the transformation of the communication network's service paradigm from "passive response" to "active prediction," enabling the leap from "best-effort" to "on-demand guarantee" and promoting new value growth for the traditional "pipeline". Communication networks are no longer limited to traditional data transmission and connectivity functions. By embedding sensing, cognition, and decision-making capabilities, based on advanced prediction and intelligent orchestration of business intent, user behavior, and environmental state, they autonomously allocate computing power, storage, and communication resources, dynamically optimize service paths and quality, providing users with precise, adaptive, and personalized intelligent service experiences.

2.3 Innovations of AI Edge

The innovations of AI Edge can be reflected in four aspects, as detailed below.

First, revolutionizing network technology: AI Edge breaks the traditional 5G system model of separate implementation for the RAN, user plane function (UPF), network control, and MEC. It supports the deep integration of DOICT technologies encompassing RAN, AI inference, edge sensing, and control, driving the evolution of RAN into the radio computing network (RCN), thereby endowing mobile communication networks with new connotations. Furthermore, AI Edge is expected to achieve precise characterization of the physical

environment, business characteristics, network load, and user habits through cognition of the global environment, enabling network self-learning, self-optimization, and self-evolution.

Second, expanding the concept of Edge: In traditional mobile networks, Edge is a single edge network functional entity, while AI Edge will construct a distributed, cloud-edge-device collaborative shared computing power foundation, enabling the distributed deployment and cross-layer, cross-domain extension of AI service capabilities. In AI Edge, terminal devices are no longer viewed as isolated data collection points or mere service consumers, but are inherently integrated as part of the Edge, enhancing the sensing, computing, and response capabilities of the edge layer, and building a collaborative, intelligent, globally interconnected network of ubiquitous edge nodes.

Third, enhancing AI capabilities: The AI applications supported by AI Edge are not merely information retrieval, content generation, or task planning. Instead, by leveraging the network's connectivity and sensing capabilities to fully interact with the physical world, they achieve a rapid closed-loop of "sensing, inference, and execution", promoting the practical implementation of network endogenous AI. By feeding real-world data back to AI models, it drives the evolution of AI toward a new paradigm capable of real-time sensing and inference. AI models are no longer fixed sets of neural network parameters, but become a new generation of intelligent agents with proactive sensing, dynamic adaptation, continuous updates, and autonomous evolution, laying a key technical foundation for building next-generation sustainable and adaptive AI.

Fourth, reconstructing the interaction paradigm: AI Edge, based on Agent technology, achieves user intent understanding as well as automated orchestration and scheduling, thereby breaking the traditional interaction mode based on fixed protocols and processes, and realizing intent-based intelligent interaction. Specifically, by analyzing and understanding user voice or text input, it identifies personalized requirements of users for network service quality, autonomously generates dynamic network configuration strategies, and drives underlying physical functional units for efficient configuration, thus achieving a closed-loop intelligent service from "user expression" to "system response".

3. Typical Application Scenarios and Potential Values of AI Edge

AI Edge, through its integrated capabilities of "communication, sensing, intelligence, computing, and control", is reshaping the underlying logic of industrial digitalization. From a technical perspective, it breaks through the "connectivity-centric" limitations of traditional communication networks, deeply coupling communication (ubiquitous connectivity), sensing (environmental insight), intelligence (real-time decision-making), computing (edge computing power), and control (precise execution) to form a closed-loop collaborative technical system. This convergence not only brings performance leaps, such as reducing device response delay from seconds to milliseconds in industrial scenarios and improving airspace control precision to the meter level in low-altitude scenarios, but also reconstructs the industrial value distribution model.

In the business dimension, AI Edge has given rise to a new paradigm of "capability as a service": hardware vendors shift from selling devices to providing long-term services of "terminal + edge node", algorithm companies achieve recurring revenue through model subscriptions, and industry customers pay based on actual value delivered (for example, settlement for every 1% reduction in failure rate). This model drives the transformation of the industry chain from "one-time transactions" to "symbiotic value addition". It is estimated that by 2030, the market size related to AI Edge will exceed USD 500 billion in just the three major fields of intelligent manufacturing, intelligent transportation, and the low-altitude economy, becoming a core engine for digital economic growth.

3.1 Industrial Robots and Intelligent Manufacturing

3.1.1 Scenario Description

In industrial production environments, the "communication, sensing, intelligence, computing, and control" technology system of AI Edge creates a "precise sensing - intelligent decision-making - efficient execution" closed-loop workflow for industrial robots. Industrial robots are equipped with a variety of sensors, such as vision cameras, force sensors, and LiDAR. Through communication technologies such as 5G or Industrial Ethernet, they interact with edge computing nodes in real time, enabling comprehensive sensing (communication and sensing) of information such as material position, shape, assembly precision requirements in the surrounding environment, as well as their own joint status and operation trajectory.

Edge nodes within the workshop collect real-time data on equipment vibration, temperature, and energy consumption (communication and sensing) via industrial wireless (e.g., 5G-Advanced) and IoT sensors. Edge AI models (such as motor fault diagnosis algorithms based on federated learning), supported by local computing power (computing), achieve millisecond-level identification and root cause analysis of equipment anomalies (intelligence). Subsequently, they link with PLC control systems to automatically adjust machine tool parameters, trigger shutdown warnings, or dispatch maintenance robots (control), building an unmanned production system.

Edge AI algorithms utilize localized computing power resources for rapid analysis and processing of sensing data. For example, deep learning-based visual recognition algorithms can accurately identify components of different models on the assembly line and determine their gripping position and posture. By real-time parsing of force sensor data, the robot's gripping force can be adjusted to avoid damaging precision components. Simultaneously, the edge system, combined with production tasks and real-time working conditions, employs intelligent scheduling algorithms to generate optimal action strategies for robots. For example, in multi-robot collaborative operation scenarios, it plans the operation sequence and paths for each robot to avoid collisions (intelligence and computing). Finally, control instructions are rapidly transmitted to the robot's execution units, such as motor drives and joint controllers, achieving precise control of the robot's movements to complete complex production tasks such as material handling, part assembly, and product inspection (control). This scenario is widely applied in industries such as automotive manufacturing, electronics processing, and logistics warehousing. For example, in automotive assembly workshops, robots leverage AI Edge technology to

efficiently complete tasks such as door installation and component welding; on 3C product production lines, the high-precision chip mounting and inspection can be achieved.

3.1.2 Potential Value Analysis

Through the deep collaboration of "communication, sensing, intelligence, computing, and control", AI Edge endows industrial robots with enhanced environmental adaptability, task execution capabilities, and intelligent decision-making levels. Its core value lies not only in improving production efficiency and product quality, but also in generating significant economic and social benefits for the manufacturing industry through innovative business models, driving the industry towards a new era of intelligent and flexible production.

➤ Technological value

Overcoming the limitations of "limited sensing and slow response" in traditional industrial robots: The AI Edge technology system elevates the sensing accuracy of industrial robots to the sub-millimeter level. Compared with traditional solutions, the recognition accuracy rate for small components increases from 85% to 98%. The real-time decision-making capability of edge AI reduces the robot's response delay from 200-300 ms, typical of traditional cloud processing, to within 50 ms, meeting the stringent requirements of "instant response and precise operation" on high-speed production lines, thereby significantly reducing production errors and defect rates.

Achieving a technological leap in "complex condition adaptation and flexible production": Through multi-modal sensing data fusion and edge intelligence algorithms, industrial robots can operate stably in harsh environments with complex lighting, vibration, and other factors, and can quickly switch production tasks to adapt to flexible production demands for small-batch, multi-variety manufacturing. For example, in electronics manufacturing, the operating procedures can be reprogrammed and adjusted within a short timeframe to produce different models of electronic products, reducing production line changeover time from several hours to within half an hour.

➤ Business value

Equipment upgrade and service subscription model: Manufacturing enterprises can purchase new industrial robots equipped with AI Edge technology or upgrade existing equipment, with equipment suppliers charging a one-time upgrade fee. Simultaneously, enterprises can subscribe to edge AI algorithm services on demand, such as optimization algorithms for specific processes, billed based on usage duration or the number of invocations, creating a continuous revenue stream for suppliers while reducing enterprises' technology R&D costs.

Production efficiency improvement and cost savings: AI Edge-empowered industrial robots can increase enterprise production efficiency by 30%-50%, reduce labor costs by 20%-40%, and simultaneously decrease raw material waste and equipment wear. For example, by adopting the relevant technologies, an automotive manufacturer can shorten the production time of a single vehicle by 2-3 hours, achieving annual cost savings of tens of millions of yuan, thereby enhancing its market competitiveness.

➤ Social value

Promoting the high-end transformation of manufacturing: It facilitates the upgrade of traditional manufacturing to intelligent manufacturing, improves the overall level of the

national manufacturing industry, strengthens international competitiveness, attracts the return of high-end manufacturing industries, and promotes industrial structure optimization.

Alleviating labor shortages and the skills gap: Against the backdrop of rising labor costs and a shortage of skilled technical workers, the intelligent upgrade of industrial robots can reduce enterprises' reliance on a large workforce for repetitive tasks. It can also reduce the training cycle and difficulty for new employees from 3-6 months to 1-2 months, thereby promoting the sustainable development of the manufacturing industry.

3.2 Smart Energy and Grid Dispatch

3.2.1 Scenario Description

In smart energy and grid dispatch scenarios, the "communication, sensing, intelligence, computing, and control" technology system of AI Edge constructs a closed loop of energy management consisting of "global sensing - intelligent decision-making - dynamic regulation". Grid edge nodes integrate IoT sensors (such as smart meters, photovoltaic inverters, and energy storage battery monitoring modules) and millimeter-wave communication modules to collect multi-dimensional data in real time, including distributed energy output (photovoltaic power and wind power), user load fluctuations, and transmission line status (temperature and current) (communication and sensing). The edge AI engine employs spatio-temporal sequence prediction algorithms (like the LSTM model) to forecast power generation of new energy sources and electricity load, and utilizes edge computing power to rapidly perform supply-demand balance calculations and optimal power flow analysis (intelligence and computing). Subsequently, it issues instructions to energy storage systems, adjustable loads (such as charging piles and industrial motors), and substation control systems to dynamically adjust charging/discharging strategies, load priorities, and power allocation across transmission lines (control), thereby achieving collaborative optimization of source, grid, load, and storage.

This scenario covers diverse energy contexts: In distributed photovoltaic power stations, the edge system adjusts inverter output in real time to smooth out power fluctuations; in urban distribution networks, AI Edge enables off-peak charging scheduling of charging piles to prevent transformer overload; within industrial parks, edge nodes link microgrids and the main grid to optimize the ratio of self-owned power plants to purchased electricity, reducing energy costs.

3.2.2 Potential Value Analysis

Through the deep collaboration of "communication, sensing, intelligence, computing, and control", AI Edge upgrades the power grid from "passive dispatch" to an "actively sensing, intelligently responding" smart energy network. Its core value lies not only in the leap in technical performance, but also in balancing new energy consumption, grid security, and energy costs through commercialization model innovation. It provides key technological support for building a new power system, promoting the transformation of the energy industry towards efficiency, cleanliness, and sustainability.

➤ Technological value

Overcoming the bottlenecks of "slow response and difficulty in new energy consumption" in traditional grid: The millisecond-level sensing and decision-making capabilities of AI Edge compress the grid's response time to load fluctuations from seconds to within 50 ms, improve

the prediction accuracy of new energy power generation to 90% (compared to ~70% with traditional methods), and reduce the curtailment rate by 15 percentage points. Through distributed control of edge nodes, the line loss rate of distribution networks decreases from 8% to below 5%, solving the problems of "insufficient computing power and high latency" in centralized dispatching.

Achieving "complex scenario adaptability - safety redundancy": Through multi-source data fusion (such as meteorological data and historical load), AI Edge can predict grid risks up to 12 hours in advance under extreme weather conditions (such as typhoons and cold waves) and automatically initiate load transfer plans. This elevates power supply reliability to 99.99%, reducing power outage duration by 50% compared to traditional models.

➤ **Business value**

"Platform + value-added services" model: Power grid enterprises can establish an AI Edge energy management platform, charging new energy power stations for data access and dispatch services (billed as a proportion of power generation). They can also offer "load optimization packages" to industrial users, with revenue sharing based on energy saving (e.g., taking a 10%-20% share of the electricity cost savings achieved through off-peak power consumption).

Equipment manufacturer ecosystem model: Energy storage equipment manufacturers can embed edge AI control modules, charging via "equipment sales + algorithm subscription" (e.g., providing dynamic charge/discharge strategies, with annual service fees per system exceeding 10,000 yuan). New energy vehicle enterprises, leveraging vehicle-to-grid (V2G) technology and partnering with the edge platform, provide vehicle owners with "off-peak charging discounts + grid ancillary service revenue", forming a win-win closed loop for users and the grid.

➤ **Social value**

Promoting the green transformation of the energy structure: AI Edge technology can support high-proportion integration of new energies (such as increasing the proportion of wind power and photovoltaic to over 40%), reducing annual carbon emissions by over 100 million tons, and aiding the achievement of "Dual Carbon" goals. A 20% increase in the distributed energy consumption rate will be made, which is equivalent to adding 10 million kW of clean power supply capacity.

Reducing societal energy costs: Industrial users can lower their electricity expenses by 10%-15% through edge optimization, achieving annual cost savings exceeding 10 billion yuan. Residential users, via smart meters and edge dispatch, enjoy time-of-use pricing benefits, reducing average annual electricity bills by 8%-10%. It also enhances grid resilience during extreme weather conditions, minimizing socio-economic losses caused by disasters.

3.3 Smart Agriculture and Unmanned Agricultural Machinery

3.3.1 Scenario Description

In smart agriculture scenarios, the "communication, sensing, intelligence, computing, and control" technology system of AI Edge constructs a closed loop of agricultural production consisting of "full-domain monitoring - intelligent decision-making - precise execution". Edge nodes deployed in fields integrate soil sensors, UAV remote sensing devices, and LoRa/5G communication modules to collect real-time data on soil moisture, crop growth, pest/disease

signs, and meteorological conditions (communication and sensing). The edge AI engine utilizes algorithms such as image recognition (like leaf disease classification) and growth model prediction (like crop water requirement calculation), combined with local computing power, to rapidly generate precise management strategies for irrigation, fertilization, and plant protection (intelligence and computing). Subsequently, the edge system controls equipment such as fertigation machines, unmanned seeders, and UAV sprayers to implement variable-rate irrigation, targeted fertilization, and precise prevention and control of pests and diseases (control).

This scenario can cover diverse agricultural production needs: For example, in field crop cultivation, edge nodes dynamically adjust irrigation volume based on wheat growth stages; in facility agriculture (greenhouses), they automatically regulate shading nets and ventilation equipment by sensing temperature, humidity, and light intensity; in animal husbandry, by collecting animal health data via wearable devices, edge AI provides real-time disease risk warnings and triggering isolation instructions.

3.3.2 Potential Value Analysis

Through the deep collaboration of "communication, sensing, intelligence, computing, and control", AI Edge shifts agricultural production from experience-driven to data-driven. Its core value lies not only in yield and efficiency improvements, but also in bridging the urban-rural digital divide through technological accessibility, while enabling green and sustainable agricultural development. It provides a practical technological pathway for ensuring food security and promoting rural revitalization.

➤ Technological value

Overcoming the limitations of "extensive management and dependence on weather" in traditional agriculture: The distributed sensing capability of AI Edge increases the accuracy of soil, crop, and environmental data collection to over 90%, improving efficiency by 50 times compared to traditional manual inspections. Edge AI's intelligent decision-making reduces control errors for irrigation and fertilization volumes to 5%, solving the problem of "excessive input," while reducing pesticide usage through early pest/disease identification (with an accuracy rate of 95%).

Achieving technology accessibility for "the large-scale farming of smallholder": Lightweight edge equipment (like low-cost soil sensors) is suitable for smallholder production scenarios, and the deployment of edge computing power reduces per-plot management costs by 60%, breaking down the "high barrier" of smart agriculture and promoting technology penetration among individual farmers.

➤ Business value

"Hardware + SaaS" service model: Agritech enterprises provide edge sensing equipment (like smart sensors) and cloud management platforms. Farmers subscribe to AI planting solutions (like precision irrigation models for wheat) by mu. The annual service fee per mu is only tens of yuan, which lowers the threshold for use.

Industry chain collaboration model: Agricultural input enterprises (fertilizer and pesticide manufacturers) partner with AI Edge service providers to offer customized input packages based on crop data collected at the edge, billed according to actual usage. E-commerce platforms predict harvests using edge data, providing integrated "pre-sale + logistics" services, achieving precise linkage of production and sales.

➤ **Social value**

Enhancing agricultural production efficiency and sustainability: In grain cultivation, AI Edge technology can improve water resource utilization by 40%, reduce fertilizer and pesticide use by 30%, and increase yield per mu by 10%-15%. In facility agriculture, it can shorten vegetable harvest cycles by 20% through precise environmental control, increasing annual output by over 20 million tons.

Supporting rural revitalization and food safety: Smallholder farmers can increase their income through technology empowerment (for example, average annual income increase exceeding 3,000 yuan per farmer in a pilot village). Production data recorded by the edge system is traceable, providing a basis for agricultural product quality certification, promoting "from farm to table" safety control, and enhancing consumer trust.

3.4 Communications and Regulation of Low-Altitude UAV

3.4.1 Scenario Description

In UAV-based low-altitude application scenarios, the "communication, sensing, intelligence, computing, and control" technology system of AI Edge constructs a closed loop of low-altitude services consisting of "full-domain monitoring - intelligent collaboration - precise operation". Ground edge nodes collaborate with multi-modal sensors onboard UAVs (millimeter-wave radar, high-definition cameras, and BeiDou Navigation Satellite System) to sense the UAV position, flight status, airspace obstacles, and ground targets (such as logistics parcels and inspection equipment) in real time (connectivity and sensing). The edge AI engine employs airspace conflict prediction algorithms and path planning models (intelligence), combined with local computing power at edge nodes (computing), to dynamically generate UAV obstacle avoidance instructions and mission scheduling plans (e.g., multi-UAV collaborative delivery routes). Finally, it controls UAV takeoff/landing, trajectory correction, and task execution (such as precise delivery and equipment inspection) via integrated air-ground communication (such as 5G-Advanced and LTE-M) (control).

This scenario covers diverse low-altitude demands: In logistics, the edge system schedules UAV swarms to complete instant delivery within a "3 km radius in 15 minutes". In power line inspection, UAVs identify transmission line defects (e.g., damaged insulators) via edge AI, simultaneously generating repair coordinates. In emergency rescue, edge nodes quickly plan UAV search and rescue routes, locating trapped individuals using thermal imaging sensing.

3.4.2 Potential Value Analysis

Through the deep collaboration of "communication, sensing, intelligence, computing, and control", AI Edge provides a "safe, controllable, efficient, and economical" technological foundation for low-altitude applications. Its core value lies not only in overcoming airspace resource constraints, but also in propelling the low-altitude economy from "isolated pilots" to "large-scale operation" through commercialization model innovations. It is projected to drive related industries to create more than one trillion yuan in economic value by 2030, becoming a new growth pole of the digital economy.

➤ **Technological value**

Overcoming the bottleneck of "difficult supervision and low efficiency" in low-altitude applications: AI Edge achieves airspace sensing accuracy at the 0.5-meter level, enabling high-

density collaborative control of up to 500 UAVs per square kilometer, with conflict warning accuracy improved to 99% and efficiency increased 10-fold compared to traditional manual scheduling. Edge AI's real-time decision-making reduces UAV emergency response delay from 300 ms with cloud processing to 50 ms, meeting the "second-level response" requirements for disaster relief.

Achieving "complex environment adaptation and cost optimization": By offloading the data processing pressure of UAVs through edge computing power, onboard sensor costs per UAV are reduced by 40% (without the need for high-end onboard chips). In harsh weather conditions such as rain, snow, fog, and haze, multi-modal sensing fusion technology maintains a target recognition accuracy of over 85%, solving the traditional UAV problem of "limited visibility and poor identification".

➤ **Business value**

The "infrastructure + operations service" model: Local governments or enterprises build edge low-altitude control base stations and charge logistics, inspection, and other enterprises for UAV access fees (based on flight hours/sorties). Third-party service providers provide AI algorithm subscriptions (such as power defect identification models), and the average annual service revenue of a single industry can reach 100 million yuan.

Scenario-based solution model: For agricultural crop protection scenarios, a "UAV + edge AI" precision spraying solution is provided, with charges based on the treated area (the cost per mu is reduced by 60% compared with manual operations). In the field of urban security, UAVs and ground surveillance systems are integrated through an edge system to deliver an "aerial + ground" integrated security solution for property management companies and industrial parks, with annual subscription fees exceeding 10 million yuan.

The building of an "airspace services + computing power leasing" ecosystem: Logistics enterprises pay by flight mileage to use edge computing power and communication resources, reducing costs by 60% compared to traditional solutions based on satellite positioning. Local governments can conduct compliance management through low-altitude control systems, which can expand industries such as UAV logistics and emergency rescue to a scale exceeding one trillion yuan and create more than 500,000 jobs.

➤ **Social value**

Unleashing the potential of the low-altitude economy: AI Edge can support the large-scale deployment of UAV logistics, reducing the cost of intra-city on-demand delivery by 50% and benefiting livelihood sectors such as fresh produce and pharmaceuticals. Inspection efficiency in the power sector can be increased by 80%, with the line fault detection rate improving from 70% to 95%, thereby reducing the social losses caused by power outages.

Enhancing emergency response capabilities: In forest fires and flood disasters, UAV swarms enable full-area monitoring through edge-based coordination, increasing the efficiency of rescue force deployment by 3 times and reducing the time required to locate and rescue trapped individuals by 60%, thereby significantly minimizing casualties and property losses.

3.5 Embodied Robot Training Field

3.5.1 Scenario Description

The embodied robot training field relies on AI Edge's technology system encompassing "communication, sensing, intelligence, computing, and control", with a closed-loop training ecosystem featuring "environmental sensing - intelligent decision-making - action execution - feedback optimization". Within the training field, millimeter-wave radar, vision sensors, and 5G/6G edge nodes are deployed to collect real-time data on joint angles, motion trajectories, force feedback, and three-dimensional spatial information of the surrounding environment (sensing). The edge nodes integrate high-performance AI chips and leverage algorithms such as reinforcement learning and digital twins to construct virtual training scenarios, simulating real-world challenges, including extreme weather and complex terrain, while dynamically optimizing the robot's motion control strategies (intelligent computing). Benefiting from the low-latency characteristics of the edge system, control instructions can be transmitted to the robot's actuators within 10 milliseconds, enabling precise mapping between virtual training and physical actions. Meanwhile, multi-robot coordination algorithms are employed to schedule training resources and prevent equipment conflicts (control).

This scenario can support the training of multiple types of embodied robots. Industrial collaborative robots can learn precision assembly tasks through edge AI, household service robots can simulate furniture avoidance and human-robot interaction behaviors in virtual scenarios, and rescue robots can be trained to respond to emergencies in extreme conditions such as earthquakes and fires generated by the edge system.

3.5.2 Potential Value Analysis

The embodied robot training field leverages AI Edge technology to reshape the "learning-evolution" pathway of robots. Its core value lies not only in optimizing training efficiency and cost, but also in lowering the barrier to deploying robotic technologies through deep coordination of communication, sensing, intelligence, computing, and control. This enables embodied AI to move from the laboratory into diverse industries, becoming a key productivity tool in areas such as intelligent manufacturing and services related to people's lives.

➤ Technological value

Breaking through the "high-cost, high-risk" bottleneck of traditional robot training: edge digital twin technology reduces physical prototype wear by 70% and cuts the cost of extreme scenario training by 60%. The real-time decision-making capability of edge AI compresses robot action correction delay from seconds to milliseconds, increasing training efficiency by 3 times.

Achieving intelligent improvement in "data closed-loop and generalization capability": the edge system aggregates multi-robot training data and employs federated learning to train a universal motion model, increasing task success rates in unknown scenarios from 50% to 85% and addressing the challenge of "scenario transfer difficulty."

➤ Business value

B2B subscription model: Robot manufacturers pay to use the edge training platform based on training duration or scenario complexity, reducing the annual training cost per robot by USD 20,000. Third-party algorithm providers supply customized training models (e.g.,

reinforcement learning algorithms for precision tasks) and earn revenue shares based on the number of model invocations.

Capability output model: The training field operator packages edge training technology as "Training-as-a-Service (TaaS)" and provides API access to universities and research institutions, supporting robot algorithm validation and talent development, with annual service revenue potentially reaching millions of dollars.

➤ **Social value**

Accelerating the industrialization of embodied robots: In the industrial sector, collaborative robots can be deployed in just one month instead of six months. In the service industry, household robots achieve over 90% autonomous completion of daily tasks through edge training, helping to alleviate caregiving pressure in aging societies.

Building an inclusive ecosystem for robotics technology: Small and medium-sized manufacturers no longer need to build expensive training facilities. They can access advanced training capabilities through the edge platform, promoting balanced technological development across the industry. Driven by this, the global robotics market growth is expected to exceed USD 300 billion by 2030.

3.6 Edge-Enhanced Immersive XR

3.6.1 Scenario Description

Immersive XR includes virtual reality (VR), augmented reality (AR), mixed reality (MR), and others. Edge-enhanced immersive XR is positioned as one of the representative scenarios demonstrating the value of network intelligence and service collaboration. Figure 4 outlines the comprehensive capability boundaries required for an immersive experience through eight key dimensions in AR/VR gaming scenarios. These dimensions can be summarized as a set of interrelated indicators balancing delay and jitter, throughput and resolution, reliability and continuity, storage, energy efficiency and terminal thermal management, security and privacy, intelligent collaboration and personalization, and cost and scaling up. Taking VR as an example, end-to-end interaction requires a sensing-to-rendering closed loop within 20 milliseconds, stable high frame rates and field of view, precise head and hand tracking, and multi-user consistency. Additionally, energy consumption and terminal heat need to be controlled at the same time. This requires pushing capabilities such as render partitioning, gaze-aware coding, scenario semantic compression, resource prefetching, and asset caching to the edge, combined with link-layer rate adaptation and cross-access seamless switching, thereby dynamically scaling each "vertex" of experience octagon as needed.

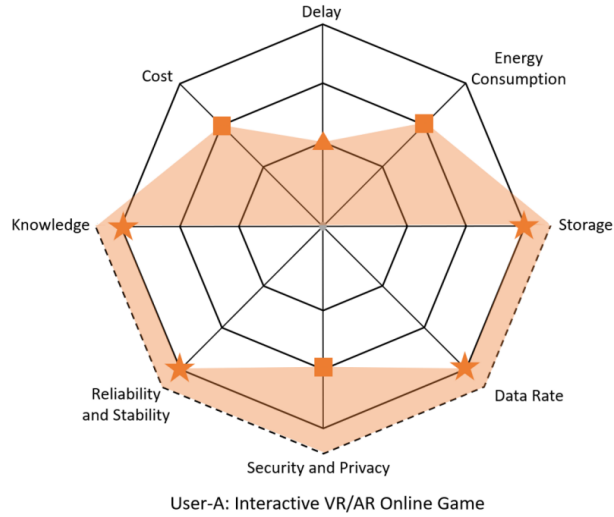


Figure 4. Schematic Diagram of AR/VR Game User Needs

Given that AR/VR experiences are constrained by terminal device computing power, battery life, and delay associated with cloud data return, AI Edge technology can extend natively integrated AI and computing capabilities on the edge side. By offloading complex rendering, computing, and AI processing tasks from terminal devices to nearby edge servers, AI Edge technology enables lightweight, high-fidelity, and highly interactive immersive experiences for users, as shown in Figure 5.

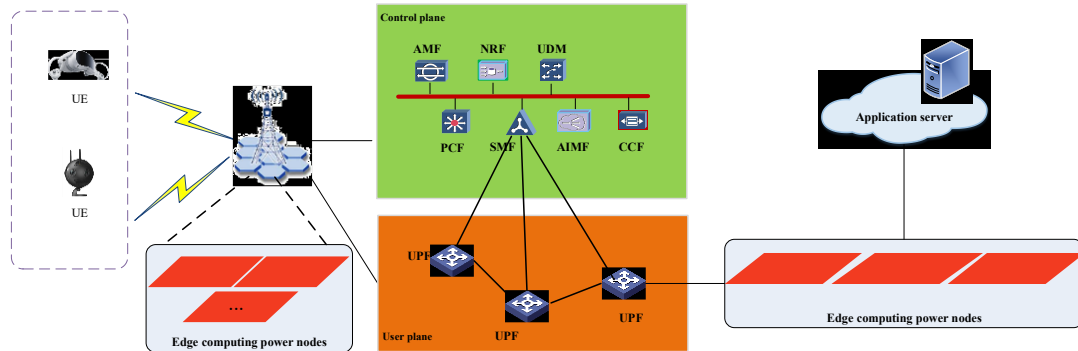


Figure 5 Edge-Enhanced AR/VR

On the one hand, computing power-intensive tasks such as image rendering, audio and video stream feature analysis, and video coding can be partially or fully offloaded from the cloud to edge computing power nodes. Through coordinated cloud-edge processing, transmission delay is reduced and bandwidth consumption is minimized. On the other hand, computing tasks performed by terminal devices (such as video decoding and tracking and positioning) can be offloaded to edge computing power nodes to enable coordinated device-edge processing. This addresses terminal devices' limitations in image rendering, mobility, and interactive experience, while reducing requirements for battery life, device size, and storage capacity, thereby lowering terminal device costs. Furthermore, in AR/VR business scenarios enhanced by AI Edge, AI applications can be used to analyze user behavior and preferences to adjust content recommendations based on user behavior in the virtual environments. AI technologies can also support XR content creation, as well as animation and virtual avatar generation. By offloading part of the AI-enhanced capabilities to edge computing power nodes,

efficient collaboration between device and edge computing power is achieved, delivering a superior immersive experience to users.

In specific scenarios, this technology can support a variety of applications, for example, industrial AR remote O&M (e.g., engineers use AR glasses to view internal equipment structures while edge AI simultaneously highlights fault points and provides repair instructions in real time), VR multi-user collaborative work (e.g., in a virtual conference room, edge nodes process motion-capture data from more than 10 users simultaneously, ensuring avatar interaction delay stays below 20 milliseconds), and immersive education (e.g., in VR anatomy courses, the edge system renders organ details in real time based on student gestures, while AI algorithms dynamically adjust the difficulty of the explanations).

3.6.2 Potential Value Analysis

AR/VR enhanced by edge leverages AI Edge technology to reshape the "device-edge-cloud" collaboration paradigm. This approach not only addresses core user experience challenges but also lowers industry entry barriers through technology inclusivity, accelerating AR/VR's transition from concept to large-scale commercial deployment and establishing AR/VR as a key support for metaverse infrastructure.

➤ Technological value

Breaking through the traditional AR/VR's "computing power bottleneck" and addressing "delay pain point": Edge nodes reduce terminal computing requirements by 60%, enabling lightweight AR glasses (weighing <100g) to deliver flagship-level experiences. Through dynamic bitrate adaptation and predictive rendering powered by edge AI, end-to-end delay is kept below 15 milliseconds, mitigating motion sickness and improving user experience satisfaction by 40%.

Achieving intelligent "environmental sensing-content adaptation": The edge system uses multi-modal sensing data (such as user gaze focus and ambient lighting) to automatically optimize rendering accuracy (e.g., 4K resolution in focal areas, 1080P in peripheral areas), reducing bandwidth usage by 30% while ensuring the clarity of key content.

Reducing costs and improving efficiency to foster industry ecosystem development: Through computing offloading, terminal devices can evolve toward being lightweight, long-lasting, and low-cost, significantly lowering user acquisition barriers and hardware upgrade costs. At the same time, processing massive raw data locally at the edge can greatly reduce backhaul bandwidth usage and central cloud computing costs, creating conditions for large-scale deployment and promoting healthy development of the industry ecosystem.

➤ Business value

B2B2C service model: Hardware manufacturers (e.g., AR glasses manufacturers) pre-install edge-adaptation modules and pay edge service providers based on device activations. Enterprise customers (e.g., manufacturers, educational institutions) subscribe to edge computing power and AI model services (e.g., industrial AR annotation algorithms) and are billed according to usage duration, with annual per-user payments reaching several hundred US dollars.

Content ecosystem revenue-sharing model: The edge platform aggregates AR/VR content creators and uses AI recommendation algorithms to increase content exposure. The platform shares revenue based on traffic (e.g., 10%–20%) while providing creators with low-code

development tools (such as edge AI-powered automatic modeling features) to lower the barrier to content production.

➤ **Social value**

Driving the transformation of AR/VR from "entertainment-oriented" to "production-oriented": In the industrial sector, equipment maintenance efficiency can be increased by 50%, and training costs can be reduced by 60%. In the telemedicine sector, surgeons are allowed to guide primary-level surgeries through AR, improving access to high-quality healthcare resources by 30%.

Expanding new scenarios in the digital economy: By 2027, edge-enhanced AR/VR is expected to drive market growth in areas such as virtual offices and digital twin cities to over USD 500 billion, creating more than one million new jobs (including roles like edge AR content designers and virtual space operators).

3.7 Intelligent Driving and Vehicle-Road Collaboration

3.7.1 Scenario Description

In intelligent driving scenarios, the AI Edge "communication-sensing-intelligence-computing-control" technology framework establishes a driving closed-loop featuring "environmental sensing, decision-making and planning, and precise control." With the integration of 5G/6G communication and edge computing, sensors such as cameras, millimeter-wave radar, and lidar on vehicles can collect real-time 360-degree environmental data around the vehicle, including the positions and speeds of other vehicles, pedestrian movements, traffic signal states, and road conditions (communication and sensing). Edge nodes on vehicles integrate AI chips to rapidly process massive sensing data using local computing power and AI algorithms such as target detection, semantic segmentation, and multi-target tracking. They can identify various objects, predict their motion trajectories, and generate optimal driving paths and decision instructions based on traffic rules and driving intentions (intelligence and computing). Finally, these instructions are transmitted to vehicles' power, steering, and braking systems within milliseconds, precisely controlling acceleration, deceleration, turning, and obstacle avoidance maneuvers (control).

This scenario can cover a variety of driving environments: On highways, edge AI assists vehicles in maintaining safe distances and performing automatic lane changes. In urban roads, it enables intelligent intersection navigation and handling of complex traffic conditions. In parking lots, it supports autonomous space searching and parking. Furthermore, in a vehicle-road collaborative mode, roadside edge nodes can interact with vehicles to further enhance driving safety and traffic efficiency.

3.7.2 Potential Value Analysis

Through the deep coordination of "communication, sensing, intelligence, computing, and control", AI Edge gradually transforms intelligent driving from a concept into practical applications. Its core value lies not only in the leap in technical performance but also in creating significant socio-economic benefits through innovative business models, driving profound transformation in the transportation and mobility sector.

➤ **Technological value**

Overcoming the traditional intelligent driving challenges of "limited sensing and decision delay": AI Edge's multi-sensor fusion sensing technology increases vehicle target recognition accuracy to over 98%, indicating a significant improvement over traditional single-sensor solutions. Edge AI's real-time decision-making capability reduces vehicle decision delay from 200–300 milliseconds with cloud processing to under 50 milliseconds, meeting the stringent requirements of "low delay and high reliability" in intelligent driving and effectively preventing collisions.

Achieving technical upgrades for "complex scenario adaptation and intelligent collaboration": By combining edge computing with AI, vehicles can maintain stable and reliable sensing and decision-making capabilities in adverse weather conditions such as rain, snow, and fog, as well as in complex environments like tunnels and urban canyons. In vehicle–road collaborative scenarios, the edge system enables real-time information exchange between vehicles, infrastructure, and other vehicles, supporting coordinated driving and improving overall road traffic efficiency by over 30%.

➤ **Business value**

Enterprise procurement and service subscription model: Automakers purchase AI Edge intelligent driving solutions and pay technology providers based on the number of vehicles provided with the solutions, enhancing vehicle intelligence and market competitiveness. Vehicle owners can subscribe to higher-level intelligent driving features on demand (such as fully autonomous driving in specific scenarios), creating additional after-sales revenue sources for automakers.

Data services and advertising model: Driving data collected through AI Edge (after anonymization) can be provided to insurance companies for precise risk assessment, enabling differentiated auto insurance pricing. At the same time, under compliance requirements, personalized advertising services can be pushed based on users' driving habits and preferences, creating new business value growth points.

➤ **Social value**

Enhancing traffic safety and mobility efficiency: Intelligent driving technology can reduce traffic accidents caused by human error, potentially lowering traffic-related fatalities by 80%. By optimizing traffic flow and alleviating urban congestion, daily commuting time is expected to decrease by 20%–30%.

Unlocking social productivity: The application of autonomous driving in logistics can significantly reduce truck drivers' labor intensity while improving transport efficiency and lowering logistics costs by 15%–20%. Additionally, it provides convenient mobility for special groups such as the elderly and people with disabilities, expanding their access to social activities.

3.8 Emergency Communications and Support

3.8.1 Scenario Description

In emergency communication scenarios, the AI Edge "communication-sensing-intelligence-computing-control" technology framework establishes an emergency support closed-loop featuring global sensing, intelligent networking, dynamic scheduling, and precise response. When disasters such as earthquakes or floods disable traditional communication

infrastructure, emergency equipment like UAV base stations and portable edge nodes can be rapidly deployed. With the integration of multi-band communications (e.g., satellite narrow band, 5G private networks) with environmental sensing (infrared imaging, vibration sensors), real-time information on personnel locations, building damage, and coverage blind spots in the disaster area can be obtained (communication and sensing). The edge AI engine uses this real-time data to generate optimal networking strategies (e.g., self-organizing network topologies) and rapidly performs channel resource allocation and load balancing calculation through edge computing power (intelligence and computing). Meanwhile, the system dynamically controls equipment transmission power and switches communication frequency bands, prioritizing the delivery of rescue instructions and life-detection signals to achieve low-latency communication among rescue personnel, command centers, and affected people (control).

This scenario can support a variety of emergency needs. For example, in earthquake rescue, edge nodes use AI to identify trapped individuals' mobile signal features, guiding rescue teams to precise locations. In flood-affected areas, UAV base stations collaborate with ground edge nodes to establish temporary communication coverage, ensuring real-time delivery of disaster relief supply dispatch instructions. In pandemic lockdown zones, edge computing distributes health code verification data to prevent network congestion from affecting emergency healthcare communications.

3.8.2 Potential Value Analysis

Through the deep coordination of "communication, sensing, intelligence, computing, and control," AI Edge upgrades emergency communications from a "passive repair" approach to a modern system of "proactive perception and intelligent response." Its core value lies not only in technical performance breakthroughs but also in saving lives and minimizing losses through efficient communication, while establishing a sustainable "government-led, market-participated" operational model that provides key technological support for the modernization of the national emergency management system.

➤ Technological value

Overcoming the traditional emergency communication challenges of "slow response and limited coverage": AI Edge's self-organizing network technology reduces emergency communication deployment time from several hours to 15 minutes and extends signal coverage radius to 5 km, while lowering communication interruption rates by 80% in complex terrains. Edge AI's intelligent traffic scheduling can increase the priority of rescue instructions to 99%, ensuring zero-delay transmission of key information.

Enabling intelligent "resource self-adaptation-risk anticipation": the edge system analyzes environmental sensing data (such as aftershock frequency and water level rise) to proactively adjust communication parameters (such as enhanced anti-interference coding), increasing communication reliability to 95% under extreme conditions and overcoming the limitations of traditional emergency communications that rely on "passive response."

➤ Business value

Government procurement + service outsourcing model: Emergency management authorities procure AI Edge emergency equipment and annual O&M services through public tendering, paying based on the number of deployments and communication duration. Telecommunications operators provide "dedicated emergency communication networks + edge

computing power" services, undertaking customized assurance needs from governments and enterprises (such as backup communications for large-scale activities).

Equipment leasing + technology licensing model: Equipment manufacturers lease portable edge base stations to professional rescue teams such as fire brigades and the armed police, while licensing AI networking algorithms to industry customers. Fees are charged based on equipment model and scope of licensing, with the average annual revenue per system reaching several hundred thousand yuan.

➤ **Social value**

Enhancing disaster rescue efficiency and survival rates: In earthquake rescue scenarios, AI Edge technology can reduce the average time required to locate trapped individuals from six hours to one hour, increasing rescue success rates by 40%. In public health emergencies, it ensures smooth communication for medical supply dispatch, improving emergency response speed by 30%.

Reducing social losses and governance costs: Once critical communications can be rapidly restored, disaster-induced information silos can be mitigated, and economic losses can be indirectly reduced. (For example, during a provincial flood in 2023, emergency communication support helped cut agricultural losses by over 200 million yuan). At the same time, grassroots governments can be provided with an "integrated normal period–emergency" communication capability, which can be used for routine management tasks such as forest fire prevention and geological monitoring during normal periods, thereby improving resource utilization efficiency.

3.9 Smart Sports

3.9.1 Scenario Description

In recent years, emerging technologies such as smart wearable devices, 5G event live streaming, and AI-powered exercise prescriptions have continuously promoted the deep integration of sports and technology. "Smart sports" is gradually becoming a key engine for nationwide fitness programs, and the development of competitive sports and the sports industry. Taking basketball as an example, intelligent agents (such as athletes, referees, and coaches) within the venue must engage in real-time interaction and collaborative decision-making in a highly dynamic, high-speed environment. This places extremely stringent demands on low-latency network transmission and rapid edge AI inference.

In smart sports scenarios, traditional cloud AI architectures, constrained by long transmission links and high end-to-end delay, struggle to meet the requirements for real-time performance in such scenarios. For example, professional athletes typically need to keep their average reaction time within 150 milliseconds, while that of ordinary individuals is around 250 milliseconds. Under a cloud AI architecture, excessive delay in referees' decisions, coaches' tactical adjustments, or the transmission of athletes' movement data would directly affect event fairness and tactical effectiveness.

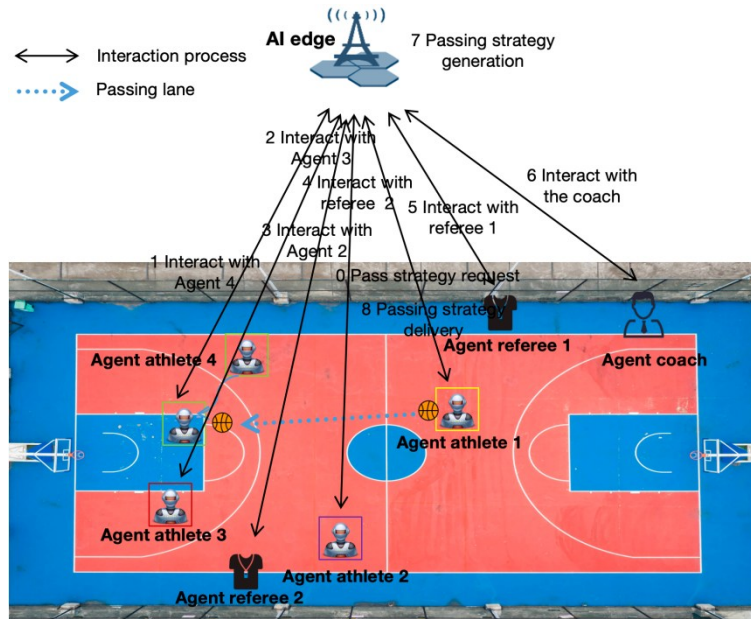


Figure 6 An Example of AI Edge Empowering Smart Sports

AI Edge brings AI inference and sensing capabilities down to the RAN side, enabling deep integration of communication and computation. It represents a promising technology for dealing with such scenarios. The RAN not only handles data forwarding but can also directly access real-time sensing information from the venue environment and engage in high-frequency interactions with multiple terminals/agents, functioning as a "brain" for cluster-level decision-making. As shown in Figure 6, passing and movement information among athletes can be rapidly shared, referees' decisions can be instantly conveyed to athletes and coaches via AI Edge, and coaches' tactical adjustments can be delivered to multiple on-court athletes through AI Edge with millisecond-level delay. In this process, a single interaction can involve data volumes of several hundred kilobytes, while the round-trip delay between the RAN and MEC or the cloud can range from tens to over a hundred milliseconds. By enabling multi-agent interactions directly at the RAN side through AI Edge, additional delay caused by multi-hop information forwarding can be avoided.

3.9.2 Potential Value Analysis

By deploying inference and control capabilities at the RAN side, AI Edge significantly shortens communication paths, reduces end-to-end delay, and achieves faster response speed than traditional MEC and cloud AI. With AI models and computing resources closer to the terminal, communication links are shorter, and the delay is lower. At the same time, integrating communication, AI inference, and control logic effectively reduces the overall system delay. For example, the adoption of an edge computing architecture can significantly reduce transmission delay, allowing experiential delay in real-world scenarios to be kept within tens of milliseconds, far better than the hundreds of milliseconds typical of cloud or remote MEC. Obviously, an edge computing architecture can meet the strict real-time requirements of smart sports and similar scenarios. Athletes, coaches, and referees can share venue views and state information with ultra-low delay, enabling rapid collaborative decision-making and precise control. For example, during high-speed movements, AI Edge can identify and predict motion

trajectories in real time, enabling referees' officiating and coaches' instructions to be more timely and accurate.

➤ **Technological value**

The technological value of AI Edge is mainly reflected in the following aspects: 1) Significantly reduced end-to-end delay: Cloud round-trip delay is approximately 80 milliseconds, while RAN-side AI Edge can reduce the delay to around 20 milliseconds, saving about 60 milliseconds and meeting the 150-millisecond reaction threshold required by athletes. 2) Optimized bandwidth usage: Assuming 16 agents, 10 interactions per second, and 200 KB per interaction, the traffic reaches about 31.25 MB/s (approximately 250 Mbps), totaling around 88 GB per match. With local processing at the RAN side, backhaul traffic can be reduced by 10–100 times. 3) Improved robustness: By reducing reliance on remote links, the reliability of critical officiating and control functions is enhanced in weak network environments. 4) Real-time collaboration capability: RAN-side AI Edge can rapidly integrate multi-agent data to enable millisecond-level pass recommendations, hazard alerts, and referee-assisted decisions.

➤ **Business value**

In smart sports scenarios, AI Edge can generate business value in the following aspects: 1) Value-added event services: Spectators can pay to access low-latency tactical viewing perspectives. For example, if 1,000 users pay USD 2 per match, an additional USD 2,000 can be generated per game. 2) Cost optimization: AI Edge significantly reduces the expenses of bandwidth and cloud computing. 3) B2B subscription services: Providing low-latency decision systems to coaches, referees, and clubs. 4) Revenue generated by advertising and AR broadcasting: Low-latency processing enables more diverse visual content and embedded advertising. 5) System sales and O&M services: Deployment of RAN/Edge hardware and software, along with long-term service level agreements (SLAs), can generate stable revenue.

➤ **Social value**

With AI Edge, low-latency referee-assistance systems can reduce 40%–80% of critical misjudgments, thereby enhancing fairness in events. Real-time monitoring of heart rate and movement trajectories helps lower the risk of collisions and injuries, significantly improving athlete safety. The deployment of smart sports solutions in public venues enhances user experience and participation rates, contributing to the broader promotion of nationwide fitness programs. In addition, localized processing and differential privacy mechanisms effectively safeguard data security.

3.10 Robotic Guide Dog

3.10.1 Scenario Description

The robotic guide dog application leverages the convergence of communication, sensing, intelligence, computing and control" to build a closed loop of environmental sensing, intelligent planning, action, execution, feedback, optimization. It focuses on the key requirements of visually impaired people for daily mobility and cross-scenario navigation.

There are approximately 17 million visually impaired individuals in China, yet there are only about 400 guide dogs, making independent mobility for the visually impaired one of the crucial social challenges. Although devices like robotic guide dogs have shown potential in research, they face challenges such as high costs, limited battery life, and safety concerns.

Thanks to AI Edge's multi-dimensional environmental sensing, scalable edge computing capability in network, and efficient cloud-edge-device collaboration, robotic guide dog devices can be made lighter and more cost-effective with extended terminal battery life and improved safety, making large-scale adoption feasible. The robotic guide dog terminal dynamically offloads computing tasks to the AI Edge, continuously analyzing data and assessing risks. For example, it can automatically distinguish between stationary guardrails and moving bicycles, and recognize green light countdowns. The analysis results from the edge can be quickly transmitted back to the robotic guide dog's actuators and voice module to achieve precise steering, emergency braking, and real-time voice broadcasts, such as "There is a puddle 3 meters ahead". With these data and tasks offloaded to AI Edge, the local computing load on the terminal is significantly reduced, lowering terminal costs and making reliable mobility assistance more affordable for a larger number of people with visual impairments. In addition, the device can be made lighter and more energy-efficient, enhance battery life in guide devices.

AI Edge can support robotic guide dogs across multiple mobility scenarios: basic scenarios include walking on urban sidewalks, traffic light recognition, and street-crossing guidance; advanced scenarios include locating shelves in supermarkets and guiding users to accessible elevators in metro stations.

3.10.2 Potential Value Analysis

AI Edge empowers robotic guide dogs to redefine the smart mobility service system for visually impaired people. Its core value lies not only in enhancing the intelligence and reliability of guide devices but also in leveraging network-empowered terminals to overcome conventional assistance tools' limitations in accessibility and battery life, promoting the wide usage of intelligent services.

➤ Technological value

Overcoming the computing power bottlenecks and high power consumption of conventional guide devices: with complex computing tasks dynamically offloaded to AI-Edge, the local computing load of the robotic guide dog terminal is reduced, eliminating the need for expensive onboard computing chips and potentially lowering the cost of the robotic guide dog.

Battery life is extended, increasing the cruising distance and extending the activity radius. This meets the needs of visually impaired individuals for longer trips, such as shopping or traveling from home to the metro station and then to the workplace.

Leveraging the communication-sensing-computing collaboration capability of AI Edge, the robotic guide dog can analyze the environment in real time and automatically integrate AI models updated at the AI-Edge, such as newly added "shared bike lane avoidance" or "construction zone detour" models. Users do not need to manually upgrade the terminal device's hardware or software, enabling continuous access to more comprehensive safety features and allowing the device's intelligence to evolve with the network.

➤ Business value

The practices of the terminal business model inspire the development of the following two commercial models: 1) The operator provides users with the robotic guide dog along with communication and computing packages, similar to a contract mobile phone model. 2) Users buy the robotic guide dog directly from the manufacturer, while the operator offers device operation support and sells the corresponding communication and computing services to the user. Diversified business models can meet the needs of different users, while lowering the

usage barriers for end users, further enhancing the scalability of the solution. For operators, AI model operations, computing power leasing, and value-added services can help create a long-term, stable revenue system. For terminal manufacturers, leveraging open network computing capabilities helps to reduce terminal costs and overcome device computing capability limitations. For public welfare organizations, collaboration with operators and device manufacturers enables rapid technology iteration and reliable identity authentication, and promotes the implementation of charitable assistance projects. Ultimately, a collaborative and mutually beneficial industry landscape among operators, terminal manufacturers, and public welfare organizations can be shaped.

Expanding a diversified business ecosystem: based on an open architecture, third-party service providers can develop value-added features, such as real-time location sharing for family members, which can be integrated into the robotic guide dog service via the operator's platform, creating a "basic service + value-added service" commercial ecosystem.

➤ **Social value**

Tapping the potential for visually impaired people to achieve independent mobility: AI Edge capabilities can be leveraged to develop low-cost, long-endurance, and lightweight robotic guide dogs, thereby enabling large-scale adoption. AI Edge empowerment of robotic guide dogs is expected to substantially increase the rate of independent mobility among visually impaired people. By enabling them to autonomously participate in daily activities such as shopping, medical visits, and commuting, it can reduce reliance on family members and volunteers, while significantly enhancing social participation and personal confidence.

4. Technical Directions and Main Challenges of AI Edge

4.1 System Architecture

The system architecture of the AI Edge network is divided into four components: distributed nodes, super edge nodes, core nodes, and edge intelligent computing power orchestration and management. The architecture follows the principles of "edge autonomy, hierarchical deployment, and global collaboration," ensuring an efficient, flexible, resilient, and elastically scalable system in terms of computing power distribution, intelligent scheduling, and service provisioning, thereby meeting the diverse demands of future edge-intelligent applications for end users. Meanwhile, the system also supports access to specific edge network capabilities and the sharing of wireless computing power resources, in order to meet the requirements of vertical industry users for diverse intelligent applications.

- **Distributed units** are positioned at the forefront of the radio access network and act as the key enablers for bringing AI capabilities down to the "last hop." These units are typically deeply integrated with distributed units (DUs), small cells, and remote radio frequency units (RRUs/AAUs), featuring near-end data collection, signal pre-processing, and lightweight AI inference capabilities. By offloading certain AI functions to the front end of the access network, the wireless system can achieve

millisecond-level response time, supporting ultra-low-latency application scenarios such as wearable devices and Vehicle-to-Everything networks, thereby effectively reducing link overhead and enhancing user experience.

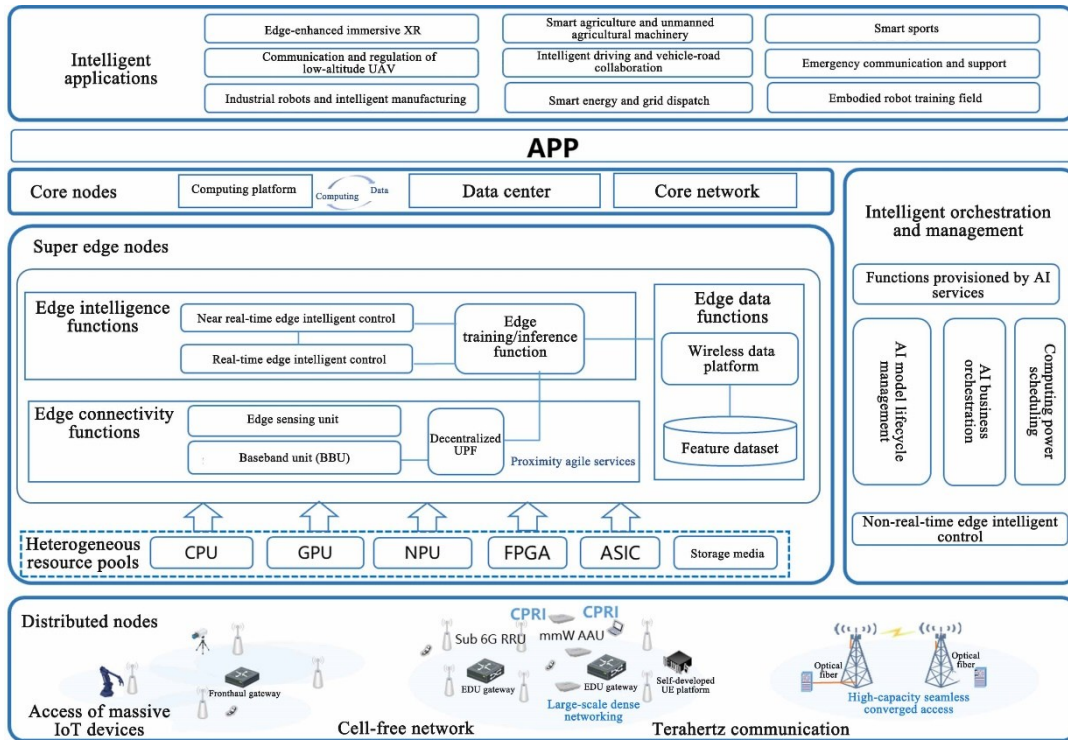


Figure 7 AI Edge System Architecture for DOICT Convergence

- **Super edge nodes** cover the traditional radio access network as well as some sinked core network functions, serving as a critical computing platform for processing data from terminals and distributed units. These nodes can leverage heterogeneous computing resources such as CPUs, GPUs, NPUs, and FPGAs to enable edge connectivity, edge intelligence, and edge data functions, allowing real-time processing and intelligent inference of multi-modal data locally. In terms of operation, super edge nodes possess both local edge autonomy and cross-domain collaboration capabilities. They can adaptively schedule computing power and model resources locally to provide proximity agile services without data leaving the domain, while also coordinating and interacting with other nodes through cross-domain collaboration mechanisms to form a hierarchical intelligent service network. This layer ensures service continuity and stability in complex environments while meeting the stringent real-time and reliability requirements of large-scale intelligent applications.
- **Core nodes** are typically deployed in large-scale data centers or cloud platforms, providing more powerful computing and storage capabilities. They are mainly responsible for tasks such as large AI model training and optimization, cross-regional inference, and global data aggregation and analysis. Core nodes can periodically perform model iteration and optimization based on service feedback from edge nodes, and distribute the optimized models to super edge nodes and distributed

units through unified interfaces, thereby establishing an edge-core model iteration closed loop. At the same time, core nodes are responsible for coordinating execution strategies for various AI tasks within the network and integrating model data, as well as providing unified authentication, management, and billing for various AI services requested by end users and industry users.

- **Intelligent orchestration and management** are required throughout the entire AI Edge architecture, providing the network with intelligent control and scheduling capabilities. Intelligent orchestration and management mainly involve functions such as non-real-time intelligent control, model management, service orchestration, computing power scheduling, and intelligent service provision. At the computing power layer, the orchestration system can dynamically schedule network and computing power resources across distributed units, super edge nodes, and core nodes. At the service layer, the orchestration mechanism supports rapid deployment of AI models, online optimization, and cross-scenario migration, enabling different services to flexibly leverage network AI capabilities. At the data layer, the orchestration system promotes continuous evolution and iteration of AI services through a closed-loop mechanism featuring edge data collection and preprocessing combined with core node-level optimization and feedback. The management and orchestration functions not only significantly enhance the operational efficiency of the AI Edge network but also ensure systematic coordination across multiple intelligence layers, collectively empowering 6G-based AI applications.

The architecture features the following characteristics and advantages:

- (1) Edge autonomy:** Unlike traditional solutions' passive scheduling that relies on central clouds, AI Edge enables nearby sensing, decision-making, and execution at edge nodes, significantly reducing delay and enhancing the quality of services. Based on edge nodes with self-sensing, self-diagnosis, and self-optimization capabilities, intelligent tasks can be processed locally without data leaving the domain, enabling proximity agile services. This edge-autonomous approach not only enhances network resilience and fault tolerance but also provides low delay, high reliability, and strong data security guarantees for critical industry applications, such as industrial control, intelligent wearable devices, and low-altitude UAVs.
- (2) Cross-domain collaboration:** AI Edge leverages a unified intelligent management and orchestration system to enable the coordinated scheduling of heterogeneous network and computing resources, simultaneously supporting traditional communication services and emerging AI applications, such as industrial internet of things, embodied AI, and low-altitude economy intelligent networks. Its core lies in the dynamic sensing of the entire environment and data-driven intelligent optimization. AI Edge can comprehensively consider the radio access network status, service characteristics, network load, and user behavior to manage and utilize data throughout its entire lifecycle. It also dynamically matches computing power resources and AI models across distributed nodes, super edge nodes, and core nodes. As a result, global resource optimization can be achieved, with network computing power resource utilization enhanced, empowering diverse intelligent applications through device–network–cloud collaboration.
- (3) Heterogeneous computing power integration:** At the computing power layer, the AI Edge architecture is highly open and inclusive, seamlessly integrating diverse heterogeneous computing resources. Through software programmability and virtualization technologies, AI

Edge unifies communication, AI, sensing, network control, and computing services on a single hardware base, enabling deep sharing of computing power and network capabilities. This heterogeneous computing power integration not only improves resource utilization efficiency but also drives the evolution of the radio access network (RAN) toward a radio computing network (RCN), progressively transforming the network from a purely communication infrastructure into an integrated platform that combines computing and intelligent services. **(4) Intelligent programmability:** Through open interfaces and an AI-driven functional framework, AI Edge can support rapid customization and deployment of diverse services. Network functions and applications can be flexibly combined and dynamically optimized through hardware–software decoupling, modular design, and reconfigurable mechanisms.

This intelligent programmability provides strong support for the incubation and deployment of innovative applications, such as customized control in To Business smart factory scenarios, real-time diagnostic services in smart healthcare, and even immersive interactions and embodied AI applications in the future 6G era.

4.2 AI for Edge Technology

4.2.1 The Rise of AI-for-Edge

With the rapid evolution of 5G and 6G communications technologies and the explosive growth of Internet of Things (IoT) devices, data volume at the edge is increasing exponentially, and the demand for real-time services becomes more significant, which drives the deep integration of edge computing and AI. Thus, AI for Edge has emerged as a key path to achieving intelligent edge^[20]. AI for Edge refers to a technical paradigm in which AI technologies are deeply integrated into edge environments, enabling intelligent sensing, real-time modeling, decision optimization, and resource scheduling at edge nodes, thereby delivering low-latency, high-reliability, and highly efficient intelligent services. It aims to address challenges such as the explosive growth of edge data, increasing real-time service requirements, as well as complex and dynamic network environments. AI-enabled edge computing enables intelligent collaboration and optimization of multi-dimensional resources related to communication, computing, and storage^[21]. The following section systematically elaborates on the technical challenges and implementation approaches of AI for Edge from three dimensions: technological motivations, core requirements, and key pathways for performance enhancement.

4.2.2 Core Values of AI-for-Edge

➤ Real-Time modeling capabilities in dynamic and complex environments

In edge intelligence scenarios, the dynamic and complex nature of the environment is a common challenge. Wireless channels are jointly affected by multipath propagation, user mobility, blockages, and external interference, resulting in highly time-varying and unpredictable channel characteristics. At the same time, traffic at the service layer also exhibits significant fluctuations. For example, data burst from applications such as short-form videos, online gaming, virtual reality (VR), and augmented reality (AR) may cause rapid and drastic changes in system load. Traditional approaches based on mathematical models or static assumptions often struggle to remain effective in such rapidly evolving environments. In contrast, AI models, with their strong generalization and online learning capabilities, can track

changes in real time and capture patterns in complex environments, enabling efficient modeling and adaptive optimization.

➤ **Nonlinear device compensation and signal fidelity capabilities**

The performance of the wireless air interface directly determines the throughput, latency, and stability of the edge network. In high-frequency broadband, multi-antenna, and complex electromagnetic environments, signal transmission is not only affected by multipath fading and interference but also constrained by the nonlinearity of radio frequency components (such as power amplifiers, ADCs/DACs, and antenna arrays), resulting in degraded system performance. Traditional compensation methods based on analytical models struggle to cope with variations in hardware characteristics and dynamic environmental conditions. AI technologies provide a new approach for air interface performance optimization through data-driven modeling and predictive capabilities. For example, AI models can learn the nonlinear characteristics of radio frequency components from measured signals. It enables functions such as digital predistortion (DPD), IQ imbalance compensation, and quantization noise suppression, while also as adapting to temperature drift or device aging. In addition, AI can leverage time-series modeling and reinforcement learning to jointly predict channel state information (CSI) and hardware performance, enabling early identification of link degradation trends. It can then dynamically adjust modulation and coding, power control, and beamforming strategies to achieve proactive optimization over the air interface. AI-enabled nonlinear compensation and air interface optimization can maintain link stability and efficiency under complex hardware and channel conditions, driving wireless systems toward self-sensing, self-optimization, and self-healing capabilities.

➤ **Capability to solve complex, high-dimensional optimization problems**

In edge intelligent system, the system optimization objectives are typically multi-dimensional. Specifically, service requirements such as low latency and high reliability need to be satisfied, while multiple factors such as energy efficiency and spectrum utilization must also be taken into consideration. These problems are usually non-convex, highly combinatorial, and subject to complex constraints, which makes it difficult for classical analytical approaches or single heuristic methods to produce high-quality solutions within tight time budgets. Especially in scenarios involving a large number of users, heterogeneous service demands, and diverse resource types, the searching space grows rapidly with problem size, and conventional methods often struggle with the resulting dimensionality. In contrast, AI technologies, through reinforcement learning and neural network-based function approximation methods^[22], can efficiently explore high-dimensional spaces, rapidly generate near-optimal solutions, and improve solving performance through continuous learning.

➤ **Network state prediction and proactive O&M capabilities**

Network conditions often vary in bursts and show strong temporal dependency. For example, link quality may fluctuate due to interference or user mobility, and traffic demand may surge suddenly triggered by hotspot events or application behavior. Traditional network O&M relies on passive monitoring and reactive adjustments, often resulting in delayed responses and making it difficult to meet the requirements of high-reliability and low-latency services. In contrast, AI technologies, leveraging their strengths in time-series modeling, can accurately predict network states based on deep learning models such as Long Short-Term Memory (LSTM) networks, Gated Recurrent Unit (GRU) networks, and Transformers. At the

same time, by incorporating methods such as reinforcement learning^[23], AI can enable proactive O&M, resource reconfiguration, path switching, or load balancing before the predicted link degradation or traffic spikes.

➤ **Rapid strategy generation and deployment capabilities**

In edge networks, tasks such as resource scheduling, slice orchestration, task offloading, and interference management often require decision-making and execution within seconds or even milliseconds. As user mobility, bursty traffic, and interference conditions continue to change, traditional methods relying on static rules or offline optimization struggle to respond in time. It often leads to underutilized resources and degraded service quality. In addition, edge deployment places higher demands on cross-scenario adaptability. Traditional algorithms typically rely on scenario-specific channel models and parameter tuning, and their performance can degrade significantly once the environment or service types change. In contrast, AI technologies can utilize continuous learning in complex and dynamic environments, to capture key features and transfer experience rapidly to new scenarios. Such generalization and adaptability not only significantly reduce system iteration and maintenance costs but also provide support for the intelligent evolution of large-scale heterogeneous edge networks.

4.2.3 Core AI Requirements of Edge Networks

➤ **Capable of lightweight and efficient inference**

Given the constraints on computing power, memory capacity, and energy budgets, edge nodes require deployed AI models to be lightweight and capable of efficient inference. To meet the requirements, industry practices commonly use techniques such as model pruning, model quantization, knowledge distillation, and model sparsification. Among these, model pruning techniques can be implemented using structured pruning, in which entire components such as convolutional kernels or channels are removed. Alternatively, unstructured pruning removes individual weights, resulting in sparse weight matrices^[24]. Model quantization reduces storage footprint and improves computational efficiency by lowering the numerical precision of weights and activations. It is typically implemented through post-training quantization, quantization-aware training, and subsequent fine-tuning. Knowledge distillation leverages a powerful teacher model to train a smaller student model. The student learns from the teacher's soft predictions or intermediate features, improving accuracy while keeping the student model's lightweight nature. Model sparsification methods reduce unnecessary computations by introducing zero-valued weights or activations during training or inference. In addition, for distributed collaborative scenarios, split federated learning frameworks^[25] extend these lightweight techniques across heterogeneous devices. By partitioning models and employing priority scheduling mechanisms, such frameworks reduce local computational burden while enabling collaborative training among diverse edge nodes, thereby broadening the applicability of efficient AI in real-world edge deployments.

➤ **Supporting a collaborative paradigm for both foundation and lightweight models**

In edge intelligent networks, resources such as computing power, energy, and storage are limited, and different tasks have significantly varying model requirements. This makes it difficult for a single AI model to meet all demands simultaneously. A more practical approach is a collaborative setup: a general-purpose, communication-domain foundation model paired with task-specific lightweight models. This enables downstream models to be deployed and

updated flexibly across scenarios without retraining or heavily altering the foundation model. First, a domain-specific foundation model for communication is pre-trained with extensive communication data. Based on this, lightweight models for specific tasks can be quickly derived by using efficient parameter fine-tuning methods such as Low-Rank Adaptation (LoRA)^[26], adapter module addition, and prompt tuning. With containerization and virtualization, Model as a Service (MaaS) solutions can be employed within the edge intelligence ecosystem, enabling on-demand intelligence across diverse scenarios and accelerating large-scale deployment of edge AI.

➤ Capable of multi-modal sensing and cross-layer fusion

Edge networks need to simultaneously process multi-source heterogeneous data, including signal data from the physical layer, quality of service (QoS) indicators from the application layer, and external environmental sensing information (e.g., radar, cameras). Therefore, AI models must support multi-modal sensing and cross-layer fusion. Such models can perform unified modeling and joint analysis of multi-dimensional information about channel states, service loads, and environmental sensing, facilitating cross-layer data fusion and intelligent decision-making. For example, in integrated communication and sensing scenarios, fusing communication signals with sensing data can effectively improve positioning accuracy, link robustness, and resource scheduling efficiency, thereby enhancing overall network performance^[27]. In the AI-driven digital twin platform shown in Figure 8, for wireless scenario applications, combining 3D maps and radio frequency measurements can significantly reduce the overhead of channel probing in real environments, supporting efficient training and validation of AI models.

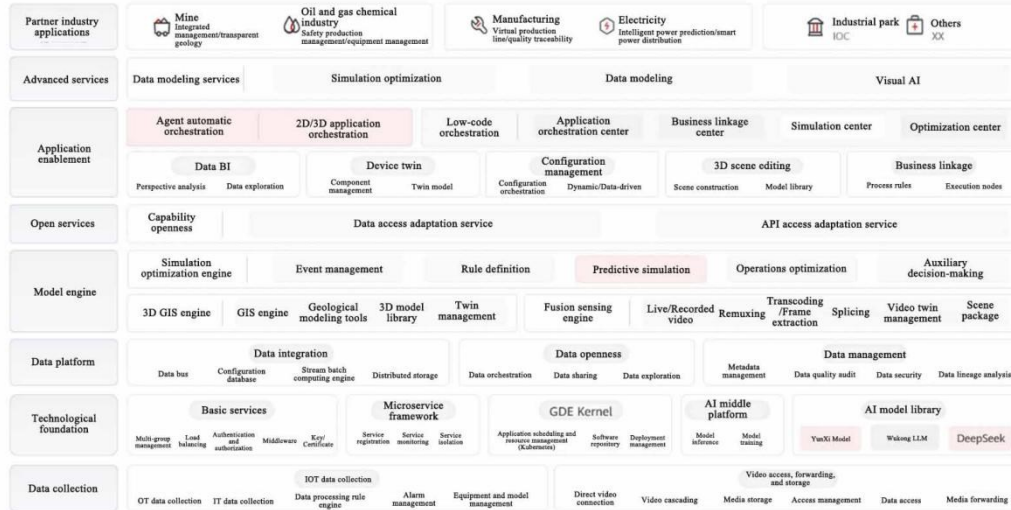


Figure 8 MetaWorks Digital Twin Platform

➤ Meeting the requirements for interpretability and testability

Edge networks must meet high reliability requirements, particularly in critical scenarios such as O&M and security. Therefore, AI models should be interpretable, capable of providing visualizations of decision logic and causal analysis, enabling O&M personnel to understand and trace model behavior. At the same time, AI models need to be testable, allowing systematic and quantitative evaluation of performance, robustness, and security. In laboratory environments, the stability and fault tolerance of AI models can be evaluated through channel

disturbances, bursty traffic, and attack simulations, ensuring their trustworthy and reliable deployment.

4.2.4 Key Paths to AI-for-Edge Performance Improvement

➤ High-efficiency characterization of wireless channels

To support low-latency embodied AI applications, edge networks must provide ultra-reliable communication, real-time channel prediction and link adaptation, high-precision positioning, as well as rapid environmental sensing and reconstruction. These capabilities heavily depend on acquiring timely and cost-effective channel acquisition and prediction. For traditional approaches, channel state information is obtained through measurements, and then interpolation or extrapolation is performed based on preset prior assumptions. However, as system complexity increases (e.g., with a growing number of antenna ports), the overhead of measurements becomes unsustainable. Meanwhile, fixed priorities often fail to adapt to diverse and dynamic scenarios. For data-driven approaches that have emerged in recent years, pre-collected transmit and receive signals are used to train deep neural network models. By learning the patterns of channel variations from data, these approaches can partially overcome the aforementioned challenges. However, most existing technical solutions rely on task-specific AI models, lacking multi-task and cross-scenario generalization capabilities. In addition, they do not fully integrate multi-modal information, which limits their effectiveness in supporting sensing applications. To address these challenges, in an AI Edge network, a multi-modal, multitask foundation AI model for wireless channels can be constructed. By pre-training the model on large-scale, multi-modal datasets (including channel measurement data and environmental sensing data, where environmental sensing data include laser point clouds, millimeter-wave radar data, electronic maps, base-station sensing data, images/videos, and GPS data), a unified and efficient feature representation of wireless channels can be learned. Subsequently, techniques such as fine-tuning can be used to adapt the model to downstream task models or algorithms, or to design independent task-specific adaptation heads for downstream tasks, thereby enabling cross-task and cross-scenario generalization^[28].

➤ Wireless air interface intelligent optimization

As the primary link between users and the network, the air interface directly impacts capacity, latency, and reliability, and is therefore central to RAN performance. In intelligent beam management and beamforming, directional signal transmission is achieved by dynamically adjusting the phase and amplitude of the antenna array. The system integrates real-time sensing of user location and channel environment to optimize beam pointing and beamwidth, supporting efficient coverage for single users as well as parallel scheduling for multiple users. This approach has been widely applied in scenarios such as mmWave communications and ultra-dense networking. In AI-enabled channel coding and signal detection, deep learning is used to adaptively generate coding schemes matched to the channel environment, thereby reducing bit error rates and simplifying complexity^[29]. In addition, by learning the characteristics of channel distortion, signal recovery capability can be improved. These approaches have been validated in dynamic scenarios such as vehicle networks and satellite communications. In adaptive modulation and coding strategy optimization, the modulation scheme and coding rate are dynamically adjusted based on real-time predictions of channel quality, achieving an optimized match between channel state and transmission efficiency. Compared with traditional strategies, the integration of channel prediction, QoS-

driven resource scheduling, and multi-user collaborative allocation mechanisms can enhance scheduling flexibility and system efficiency in multi-service scenarios. At the same time, it is necessary to explore goal-oriented CSI compression feedback, receiver design driven by both mechanisms and data, and pilot-free transmission schemes to meet the requirements of joint air-interface optimization in highly dynamic scenarios^{[30]-[31]}. Furthermore, end-to-end trained neural network models can be considered to replace traditional modules such as channel estimation, equalization, and demodulation, enabling the system to directly predict transmitted symbols from the original received signal and achieve intelligent operation of multiple receiver modules or even the entire receiver.

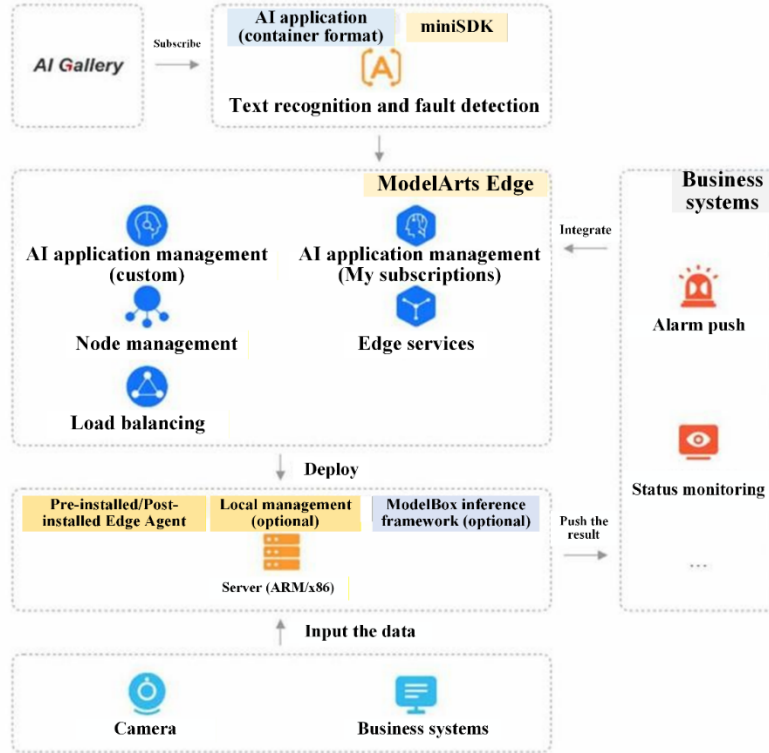


Figure 9 Edge-Cloud Collaboration Scenario Enabled by ModelArts Edge

➤ Joint resource scheduling

In the edge–cloud collaborative scenario shown in Figure 9, the integration of device, edge, and cloud computing brings challenges for multi-dimensional heterogeneous resource scheduling. It is necessary to build a joint resource scheduling mechanism that is prediction-driven, sensing-coordinated, and strategy-intelligent, enabling on-demand optimized allocation of spectrum, computing, storage, and communication resources. With AI-based prediction, user location, channel state, and load can be sensed in advance, allowing dynamic pre-allocation of spectrum resources to improve efficiency and communication continuity. By leveraging resource sensing and AI prediction, an integrated scheduling model for computing, storage, and communication can be built, supporting dynamic slicing and resource isolation to meet diverse service requirements. Intelligent load balancing can be achieved through reinforcement learning. By integrating the states of edge and central nodes, the system can adaptively schedule tasks and traffic, thereby enhancing service quality and system resilience.

➤ **Intelligent O&M and security**

As edge-intelligent networks become increasingly complex, traditional O&M models relying on manual operations and static rules become inadequate for dynamic scenarios and complex threats. The introduction of AI is driving the network O&M toward intelligent, adaptive, and closed-loop evolution. In terms of network security, AI can rapidly identify abnormal traffic through deep learning, improving both detection accuracy and response speed. AI-based methods are typically more adaptable to emerging attacks than rule-based systems. At the same time, robust learning techniques, such as adversarial training, can enhance the model's resistance to attacks, ensuring the stability of intelligent O&M systems^[32]. For energy management, AI can adjust device states based on traffic forecasts, balancing energy savings and performance. For example, intelligent sleep mechanisms can reduce energy consumption during low-load periods, enhancing green O&M. At the operational level, multi-source data modeling can enable automated fault detection and root cause analysis. Combined with adaptive recovery mechanisms, this supports self-healing, closed-loop O&M, improving network availability and robustness.

4.2.5 Testing and Verification of AI-for-Edge Models and Algorithms

To verify the effectiveness, reliability, and efficiency of AI for Edge technologies in practical deployments, it is necessary to establish a systematic testing system, with a focus on evaluating model lightweighting, dynamic robustness, and component-level performance gains.

➤ **Model lightweighting and performance benchmarking**

Lightweighting should be assessed on standardized hardware platforms, with core indicators including the accuracy–complexity trade-off after pruning/quantization (measured by FLOPs and model size) and the inference delay and energy efficiency ratio on edge chips, enabling informed model selection under resource constraints.

➤ **Dynamic environment robustness verification**

By introducing channel noise, data drift, and adversarial examples, the system evaluates the model's performance retention under disturbances. Together with resource perturbations (such as CPU/memory fluctuations), it also assesses stability under variable compute availability.

➤ **Component-level algorithm performance gain evaluation**

In a laboratory environment, high-fidelity wireless environments and end-to-end service links are constructed with air-interface simulators and network impairment testers to verify the performance of AI algorithms. Specifically, this includes: a) In high-mobility scenarios, air-interface simulators quantify the gains of intelligent beam management algorithms in terms of spectral efficiency and tracking accuracy; b) Network impairment testers inject delay, jitter, and packet loss, and traffic generators are used to simulate multiple services such as VR/AR and industrial Internet, so that the optimization effects of AI scheduling strategies on delay and resource utilization can be evaluated.

➤ **Online learning and collaborative ability test**

For continual learning, adaptation speed and resistance to catastrophic forgetting need to be verified. In collaborative scenarios such as federated learning, evaluation should cover multi-agent policy consistency and communication efficiency.

This model-level testing is guided by the principles of reproducibility, quantifiability, and traceability, providing an objective basis for algorithm selection. It serves as a prerequisite for integrating a selected algorithm into the AI Edge system platform, and its results directly support full-stack, system-level testing. (Detailed in Section 4.5.2).

4.3 AI over Edge Technology

AI-over-Edge technology aims to leverage the existing communications, computing, and storage resources of mobile communications networks to enable AI services to be deployed closer to the edge. Compared with AI services provided on the cloud, the uniqueness of Edge AI lies not only in its low latency and ability to effectively protect user data privacy, but also in the fact that the types of AI services are fundamentally different from those on the cloud. Specifically, the AI applications supported by AI Edge are not limited to information retrieval, content generation, or task planning. Instead, they leverage the network's connectivity and sensing capabilities to interact fully with the physical world, enabling a fast closed-loop of sensing, inference, and execution, and thereby natively supporting embodied AI applications. In addition, AI Edge can feed real-world data back into AI models, driving the evolution of AI paradigms toward real-time sensing and inference.

Current mobile communications networks do not yet natively support AI applications, with three major deficiencies. **First, lack of data:** Although the mobile communications networks have a large amount of data, the input modes are not diverse, which are mainly various RF measurement results, such as CQI/PMI/RI, CSI/DMRS, and mobility measurement. Multi-modal data is insufficient, such as movement trajectory, business characteristics, traffic distribution, high-precision maps, meteorological information, user profiles, and camera and sensor data. This seriously limits the representation, understanding, and generation capabilities of AI models deployed on Edge. **Second, lack of intelligence:** Mobile communications networks are rule-based. The basic logic of their operations (such as symbol decision, Modulation and Coding Scheme (MCS) adaptation, and channel estimation) is to make decisions based on measured values, preset rules, and static policies. However, AI applications are all based on intelligence, and their operating logic is to achieve understanding and generation based on perception. This means that intelligent processing modules must be added to existing mobile communications networks. **Third, lack of memory:** Mobile communications networks lack short-term memory (For example, actions of cell A that are completely independent of those of cell B; operations are completely independent for different RRC requests or even new session requests) and long-term memory (such as the tidal effect, the periodic changes brought about by commuting, and the steady-state trends brought about by events, concerts, live streaming, etc.). This makes it impossible to form a complete closed loop of perception, analysis, decision-making, and actions on the edge side, or to achieve reflection and iterative optimization through interaction with the environment.

Based on the above analysis, to realize the vision of AI over Edge, the problems of multi-modal sensing and fusion need to be solved first. On this basis, AI models are introduced at the edge, and single-point optimization technologies (such as model light weighting) and the cloud-edge-device collaboration architecture are fully leveraged to achieve efficient training and inference of edge-side AI. Furthermore, to form a closed loop from perception to decision-making and then execution, AI Agent technology is developed to support AI applications by providing edge networks with memory, learning, and collaboration capabilities. Finally, focusing on embodied AI applications, problems need to be solved, such as orchestration and exposure of information services, lifecycle management, cross-layer resource optimization, and data security and privacy, thus achieving end-to-end service quality assurance.

The following sections describe the key technologies involved in AI over Edge and the challenges it faces.

4.3.1 Multi-modal Sensing and Fusion

To support AI applications on the edge side, multi-modal sensing and processing capabilities need to be introduced. First of all, wireless sensing and communications should be integrated, and cross-device joint scheduling should be used to achieve higher-quality data collection. In addition, the capability to process multi-source sensing data locally is required. Analysis of scenarios such as autonomous driving shows that multiple cameras, millimeter-wave radars, and LiDARs assembled on a single vehicle can generate about 2.3 GB of data per second. Traditionally, the data is sent to the cloud for processing, which incurs a latency of more than tens of milliseconds. Edge computing can complete computing and decision-making locally. With the help of lightweight sensor fusion algorithms and embedded data preprocessing technology, high-quality features can be extracted when resources are limited. This improves the robustness and accuracy of sensing. Furthermore, filtering, noise reduction, compression, and other processing operations are performed at the beginning of data generation. Only valuable information is uploaded or processed. This greatly relieves the burden of subsequent computing and transmission and meets the delay requirement of 20 milliseconds or even lower. In terms of multi-modal fusion, the focus is on solving the alignment problem of multi-modal data features. In addition, it is also an important issue to ensure the robustness of multi-modal sensing and fusion so that the system can still operate safely and effectively when some capabilities do not work.

4.3.2 Model Lightweighting and Low-Latency Inference Technologies

In edge devices, the limited computing and storage resources make it difficult to directly deploy large-scale deep neural networks. Therefore, the model lightweighting technology becomes a critical path. Through model compression, pruning, quantization, and structural sparsification, the model size can be significantly reduced while ensuring that the precision is basically not affected. Knowledge distillation achieves a balance between computing efficiency and prediction precision by migrating the knowledge of large models to lightweight models. At the same time, the combination of hardware-friendly operator optimization with heterogeneous acceleration chips can further improve inference performance and ensure that the model can respond to user needs at the millisecond level. In addition, to meet the low delay requirements of different business scenarios, mechanisms such as dynamic batch processing, adaptive computing power allocation, and priority scheduling need to be introduced. This enables model

inference to handle sudden requests and stably support continuous tasks. The following section describes some of the key technologies involved in model lightweighting^[33].

➤ **Pruning**

Pruning reduces the model size by removing redundant neuron connections. Models after pruning have fewer parameters and require fewer computations. The inference speed can be improved under ideal conditions. However, it may be hard to efficiently accelerate inference on general-purpose hardware due to the irregular sparse weight matrix produced by unstructured pruning. Therefore, pruning may not always result in a reduction in actual delay. To gain the true advantage of low delay, it is often necessary to combine structured pruning (such as whole-layer or whole-filter pruning) with hardware optimization for the sparse matrix to ensure that computations of pruned models can be executed efficiently and in parallel. Nevertheless, pruning can significantly reduce the number of parameters without significantly sacrificing precision and is still one of the effective ways for edge deployment.

➤ **Quantization**

Quantization achieves model lightweighting by reducing the model weight and numerical precision of activation (for example, from 32-bit floating point to 8-bit integer). A low bit width not only compresses the storage space required for models but also uses efficient fixed-point operations in hardware to increase computing speed. For example, after a model is quantized from FP32 to INT8, the model size, memory usage, and inference delay can be significantly reduced, with only a slight impact on precision. Actual cases show that, compared with floating-point models, INT8 models after quantization witness a reduction in the model memory usage by about 4 times and have an inference delay less than one-third of the original one, while the precision loss is negligible. Therefore, quantization technology is applicable to edge devices with limited resources. It improves energy efficiency while meeting the requirements for real-time.

➤ **Efficient Model Architecture Design**

Efficient model architecture design is also a key means to reduce delay. Lightweight models (such as MobileNet series and EfficientNet) that are manually designed or obtained through Neural Architecture Search (NAS) have been specially optimized for mobile devices to achieve higher precision with less computing power and adapt to the real-time inference needs of edge hardware.

➤ **Distributed Knowledge Distillation**

Distributed Knowledge Distillation (DKD) technology is used to train powerful Teacher large models on the center side (such as cloud) with abundant computing power. With distillation policies, it helps migrate the knowledge contained in large models to lightweight Student models that can be efficiently deployed on various distributed edge nodes to achieve edge-device intelligent collaboration. Specifically, there are two mainstream DKD frameworks:

Cloud-side distillation: This involves training a Teacher large model on the cloud/center side, performing knowledge distillation using global or representative data, obtaining a global Student model, and distributing the global Student model to edge nodes for deployment and inference. In the model fine-tuning and updating stage, edge nodes upload locally collected data to the center side. Then, the Teacher large model is used to generate soft labels, and the global Student model is fine-tuned, updated, and sent to edge nodes. When the updated global Student model is distributed, incremental adaptation technologies such as low-rank adaptation (LoRA)

can be used to synchronize only incremental parameters rather than full parameters. This achieves efficient model synchronization and personalization and improves the flexibility and system efficiency of edge deployment. The advantage of this framework is that it can make full use of the center-side computing power and the LoRA technology for efficient transmission. The challenges it faces include data privacy protection and model personalization issues.

Edge-side distillation: This involves training a Teacher large model on the cloud/center side, distributing it to edge nodes, and distilling knowledge to obtain a Student model using local data on edge nodes based on the Teacher large model. In the model fine-tuning and updating stage, federated learning technology can be used to achieve multi-node collaborative distillation. The specific implementation is as follows: Edge nodes regularly upload knowledge summaries (such as the output distribution of the Student model, logits, and a small amount of "knowledge representation") to the center side. In this step, corresponding efficient access solutions, such as data compression and over-the-air computation (OAC), can be adopted according to the specific form of user-uploaded knowledge to greatly reduce the data transmission volume. Furthermore, the center side averages the uploaded knowledge summaries or integrates them by soft labels, generates "aggregated knowledge", and sends it to edge nodes. These nodes can use aggregated knowledge and local data for local training. The advantages of this framework are that it can protect user privacy, has a low data uplink transmission delay, and a high degree of model personalization. The challenges it faces include device heterogeneity, limited edge-side communication and computing resources, etc.

4.3.3 Cloud-Edge-Device Collaboration for Large, Medium, and Small Models

Scalability is one of the core features of AI Edge. AI Edge is not a single edge network function entity. It can integrate the computing power resources of adjacent base stations across areas in the horizontal dimension to build an elastic and scalable edge computing power network. It can also realize cross-layer distributed AI through efficient cloud-edge-device collaboration in the vertical dimension. In this way, it supports the scalability of mobile information services across the entire network. Depending on specific collaboration modes, the technology of cloud-edge-device collaboration for large, medium, and small models can be divided into the following categories:

➤ **Collaboration between large and small brains: Large Decision Model plus Small Execution Models**

As shown in Figure 10, in this mode, large AI models are responsible for central decision-making and user interaction, while a large number of edge lightweight models distributedly deployed in the wireless network are responsible for executing specific tasks. In addition, it is expected that large visual and language models will be embedded in large AI models in the future to enhance the understanding of users' and operators' needs. Similarly, wireless communication operators can send commands to large AI models based on predefined functions or customize instructions in real time using voice and other methods. The central large model can complete intent understanding and task decomposition based on user input and then orchestrate and schedule various subtask AI models on edge nodes. Each model focuses on a specific field, such as resource optimization in edge computing, content generation in SemCom, and intelligent scheduling in satellite communication. Therefore, large AI models in future wireless communication will essentially act as a collection of AI clusters and make decisions based on information obtained from various subtasks and user interaction information^{[34][35]}.

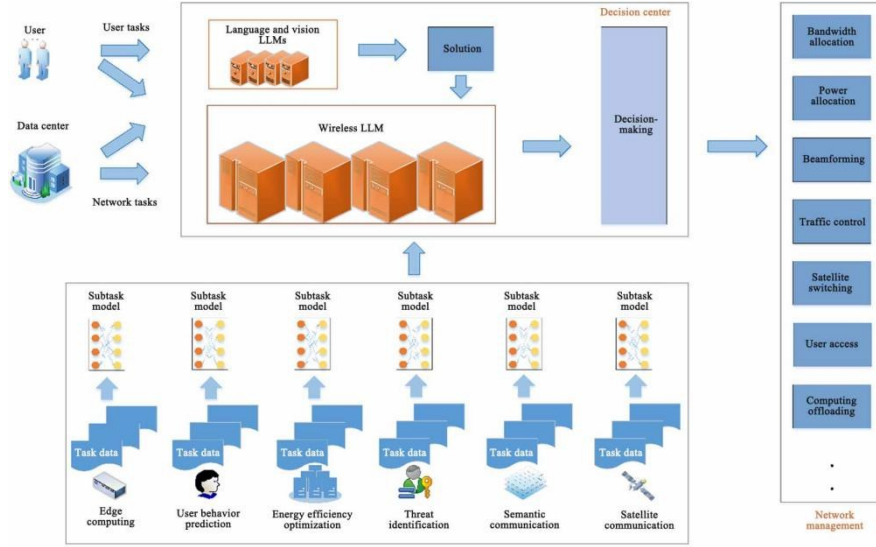


Figure 10 Diagram of Collaboration Between Decision-Making Large Model and Execution Small Model

➤ Split inference and split learning

For more complex inference tasks, they can be properly split between terminal devices and edge nodes. For example, use split inference or the early exiting mechanism of models to let terminal devices execute tasks at the first few layers of the network and directly give results when the output of the middle layer reaches the confidence level threshold, or send intermediate features to edge nodes to complete the remaining inference steps. This device-edge collaboration approach reduces the volume of raw data to be transmitted while ensuring precision, lowers end-to-end delay, and achieves faster response. It should be noted that split inference needs to balance communication overhead and computing savings and consider network conditions to select the optimal split point. To effectively implement split inference, dynamic offloading is a key technology. It can sense multi-dimensional factors in real time, such as network bandwidth, edge node load, and the task computing complexity. It uses graph-based optimization, reinforcement learning, and other methods to dynamically determine the allocation of tasks among different nodes and the optimal splitting method of the model, thereby reducing the computing load on individual nodes and improving the overall processing efficiency.

Split learning is proposed to reduce the storage usage and computing load on clients by splitting a model into multiple parts and training them separately on clients and servers. Split federated learning (SFL) is an important form of split learning. It improves the scalability of the split learning framework by computing learning tasks in parallel on multiple clients. PipeSFL is a fine-grained SFL framework. As shown in Figure 11, PipeSFL has two key mechanisms: 1) Priority scheduling mechanism on the cloud server side. It prioritizes the processing of split-layer activation values from clients with the worst performance to reduce idle resources on servers and other edge devices when poorly performing devices perform local backpropagation. 2) Hybrid training model. It allows asynchronous training within the same round and synchronous training between rounds. This avoids idle resources when servers receive split-layer activation values from all edge devices synchronously within a round^[25].

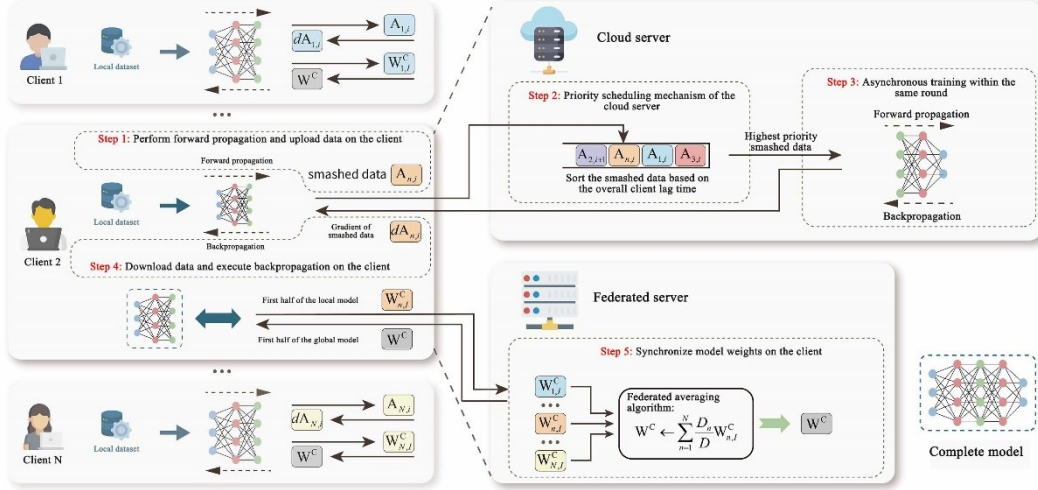


Figure 11 PipeSFL Framework Workflow

➤ Collaborative inference

Collaborative inference achieves fast response by running lightweight models on the device side. In case of complex or high-precision tasks, large models on the edge side take control. This forms a complementary mechanism featuring "fast response + high precision". By introducing a resource-aware adaptive scheduling policy, the system can remain stability and high efficiency under dynamic changes in device computing power, network bandwidth, and task complexity. This significantly improves the robustness and scalability of edge intelligence.

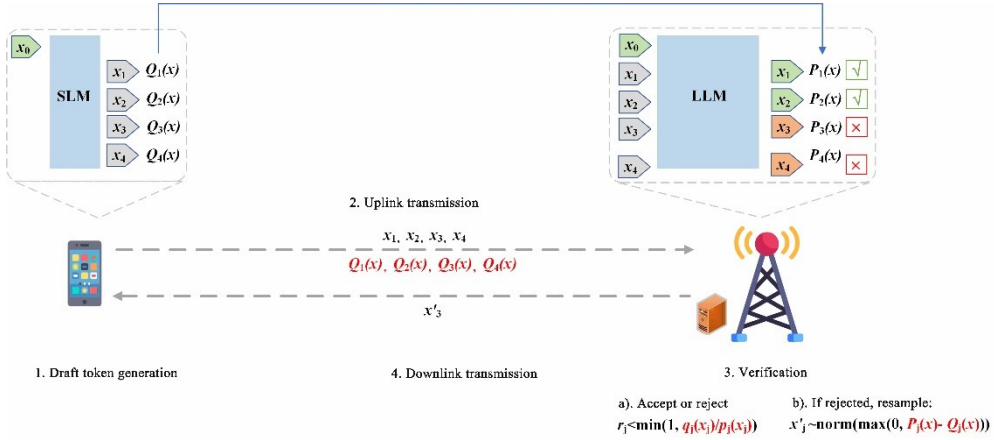


Figure 12 Distributed Speculative Decoding

As an important implementation of collaborative inference, speculative decoding has received increasing attention in recent years. Its core idea is as follows: First, a lightweight "draft small model" is used to quickly generate candidate sequences, and then a "target large model" with more parameters and enhanced capabilities is used to verify and resample. This significantly improves the inference efficiency and reduces the response delay while ensuring the generation quality^{[36][37]}. Combined with the distributed computing feature of AI Edge, the framework of distributed speculative decoding is proposed as follows^[38]: The draft small models are distributed to user devices, and the target large model is deployed on edge nodes. As shown in Figure 12, devices quickly generate multiple candidate sequences locally and send them to edge nodes, where the target large model performs final verification and resampling.

This architecture not only reduces the redundant computing overhead of edge nodes in the decoding stage but also greatly improves the inference throughput, enabling it to better meet the needs of efficiently using computing power and optimizing user experience in future wireless networks.

However, such a distributed deployment is still constrained by communications overhead: Each time a draft token is generated, its corresponding vocabulary probability distribution needs to be uploaded to a base station or an edge server for verification. Therefore, the data transmission volume increases linearly with the vocabulary size. For example, for a vocabulary with a size of 32k, the transmission volume per token is about 500 kbits in the case of representation in FP16. The above problems can be solved using the on-demand collaboration mechanism based on confidence level, separation of verification and resampling, sparse transmission, and other methods^[39].

➤ Collaborative training and model update

The various modes discussed above are mainly aimed at model inference. In addition to inference, cloud-edge-device collaboration also helps achieve more efficient model training and continuous updates. For example, in the training stage, the powerful computing power cluster on the cloud side is responsible for training and iterating complex large models. Optimized lightweight models are then distributed to edge nodes for nearby deployment. In addition, local data can be used on the edge side or device side to update some model parameters and aggregate parameters with those on the cloud side through federated learning or distributed optimization methods. This improves the generalization capability of models and effectively protects user privacy.

To endow edge AI with continuous adaptation and evolution capabilities, incremental learning and continuous learning technologies enable models deployed at the edge to use a small amount of new data generated locally for fine-tuning and iterative updates without leaving the production environment. In this way, models can quickly adapt to data distribution changes and emerging scenarios, and frequent dependencies on cloud-side retraining are greatly reduced, thus ultimately building a distributed intelligent network that is both efficient and agile and has self-evolutionary capabilities.

4.3.4 AI Agent Technology

In traditional mobile communication networks, system decisions often rely on predefined rules and static policies, lacking the capability to sense, memorize, and dynamically respond to environmental changes. An AI Agent is a system that can sense and understand objectives and take autonomous actions to execute tasks in specific environments^[40]. By introducing AI Agents on the edge side, wireless systems are expected to address the challenges of "lack of memory" and "dynamic decision-making", empowering edge networks with continuous learning, reflection, and collaboration capabilities.

Figure 13 illustrates the core components of a wireless AI Agent and its working mode on the edge side. We see that an AI Agent consists of four modules: Perception, Memory, Action, and Planning. Its core functions include^[41]:

- 1) Environmental perception and memorizing mechanism: AI Agents can continuously collect and store network status information (such as channel quality, user mobility, and business load) and form short-term and long-term memories to support context-aware intelligent decision-making.

2) Autonomous decision and dynamic planning: Based on reinforcement learning and online learning algorithms, AI Agents can adjust network parameters (such as power control and resource allocation) in real time without human intervention, optimizing network performance and energy efficiency autonomously.

3) Multi-agent collaboration: Through distributed negotiation and collaborative learning, multiple AI Agents can achieve joint optimization between edge nodes to avoid conflicts and improve overall system efficiency. This feature is particularly applicable to ultra-dense networks and mobility management scenarios. Moreover, the orchestrate capabilities in the form of "workflows" (such as sensing-planning-action, combined with memory) on the edge side enable a stable and security operation in base station. Multiple parallel processes can be executed with local data in user environments.

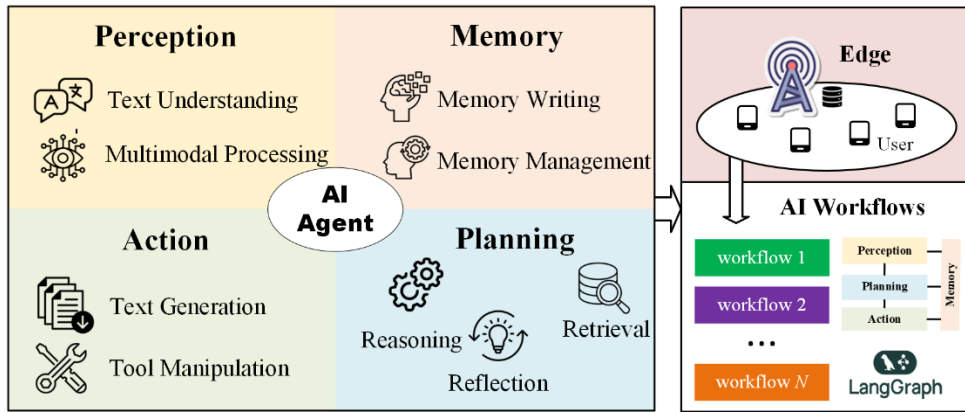


Figure 13 The Core Components of AI Agent and Its Working Mode on the Edge Side

Key technologies for AI Agent design and deployment on the edge side:

1) Memory and reflection mechanism: The AI Agent architecture with historical data analysis and experience playback capabilities is built to support policy iteration and long-term network behavior modeling (such as tidal effect and event-driven traffic prediction).

2) Multi-agent collaboration algorithm and protocol: The AI Agent orchestration framework is built to manage task allocation, communication, and collaboration among multiple edge AI agents. The multi-agent collaboration framework based on game theory, federated learning, and consensus mechanism is studied to ensure efficient, secure, and scalable collaboration in decentralized environments. For example, the collaborative training mode based on federated learning can be used to allow multiple AI Agents to jointly optimize models without sharing local data, protecting privacy and improving collective intelligence.

3) Lightweight agent design and deployment: To alleviate the severe constraints of computing power, storage, and power consumption on the edge side, it is necessary to comprehensively use technologies such as model architecture design, model compression and acceleration, continuous learning, and adaptive capabilities to strike a balance between efficiency and performance. Moreover, agent inference and training technologies with a low overhead and high concurrency degree are developed to achieve "small but flexible" agent deployment.

4) Decision-making and planning with edge optimization: The core objective of AI Agents is to make reliable and efficient autonomous decisions for wireless networks under the constraints of limited computing power and energy consumption. Therefore, AI Agents typically employ a hierarchical and lightweight decision-making architecture. At the upper

layer, AI Agents split complex tasks into executable subtask sequences based on lightweight reinforcement learning methods. At the lower layer, local optimization algorithms are adapted to achieve fast response and action execution for subtasks. Meanwhile, to adapt to the dynamic and uncertain nature of edge environments, AI Agents need to have online learning and adaptation capabilities, be able to continuously adjust policies based on feedback, and avoid performance degradation caused by data distribution changes. This enables AI Agents to perform local rolling optimization under multiple constraints and achieve real-time response and autonomous decision-making in dynamic environments.

5) Efficient resource management and collaborative computing: Efficient resource management and collaborative computing of AI Agents are the key requirements for their autonomous intelligence. To alleviate the challenges of limited edge node computing power, AI Agents apply strict energy consumption constraints, and changing network conditions, resource management mechanisms to effectively achieve dynamic adaptive scheduling policies. In addition, AI agents monitor the status of local computing resources, storage resources, energy consumption, and network resources in real time. They also perform intelligent decision-making based on task priority, delay sensitivity, and energy consumption indicators: i) For lightweight real-time tasks, inference is completed locally preferentially to ensure low delay; ii) For computing-intensive tasks, computing tasks are split and distributed to edge servers or the cloud through computing offloading technology, forming a cloud-edge-device three-layer collaborative computing paradigm.

To sum up, AI Agents on the edge side need to achieve efficient, secure, and scalable deployment through memory and reflection, multi-agent collaboration, and lightweight design to adapt to limited computing power^[42]. The application scenarios of AI Agents include:

- 1) Intelligent wireless resource management: AI Agents can dynamically adjust spectrum, power, and antenna parameters to improve network capacity and coverage.
- 2) Service orchestration and lifecycle management: In edge computing environments, AI Agents can realize on-demand deployment, migration, and termination of services to ensure Quality of AI Service (QoAIS).
- 3) Intelligent service of integrated sensing and communication: Combined with multi-modal sensing data (such as RF measurement, vision, and positioning), AI Agents can provide end-to-end situational awareness services, such as intelligent transportation dispatching and industrial IoT monitoring.

In Summary, AI Agent technology is one of the core enabling technologies to realize the vision of "AI over Edge". Empowering edge networks with capabilities to memorize, learn, and collaborate, it eliminates the limitations of "mechanical execution" in traditional communication systems and lays a solid foundation for building next-generation intelligent edge networks with continuous evolution capabilities.

4.3.5 End-to-End Information Service Technologies for Embodied AI

With the deep integration of AI and robot technology, embodied AI is becoming an important carrier for the next-generation information services and one of the important application scenarios of AI over Edge. It provides users with an unprecedented immersive and active service experience through sensing, interaction, and action by intelligent agents (such as robots, autonomous driving vehicles, and XR devices) in the physical world.

Embodied AI information services refer to a technical system in which intelligent agents with a physical "body" understand the environment through multi-modal sensing and perform embodied actions, thereby providing users with intelligent and contextualized services. This requires that the underlying information service technology must achieve an end-to-end closed loop covering sensing, computing, and execution and can dynamically adapt to complex and changing physical environments. The key technologies involved include:

➤ **End-to-end intelligent services deeply coupled with communication and sensing**

The deep integration of communication and sensing is the foundation of embodied AI services. End-to-end intelligence aims to reconstruct the traditional linear serial architecture of "sensing-communication-computing-execution" into a collaboratively optimized whole from raw sensor data to final service actions. Communication is no longer only a pipeline for transmitting data but has become part of the sensing and control system. Communication and sensing resources are dynamically allocated and adjusted according to the priority of the current task. Technologies such as deep learning are used to jointly optimize sensing modules, control policies, etc. This helps make the best trade-off when resources are limited and implement end-to-end intelligent services.

➤ **Intelligent orchestration and openness of services**

The capabilities of a single embodied AI agent are limited. In the future, an ecosystem with multiple-agent collaboration and cloud-edge-device collaboration will definitely be developed. Intelligent orchestration and openness of services are key to achieving this vision. Various embodied AI services (such as navigation and identification) are atomized and modularized, the capabilities of agents are encapsulated into network services that can be uniformly discovered and called, and a unified dynamic orchestration engine for orchestration management is built. The orchestration engine senses the environment status in real time by understanding the task logic and dependencies between services. It automatically discovers, combines, orchestrates, schedules, and executes a series of atomic services according to users' upper-layer instructions. At the same time, it supports dynamically adjusting service chains and task allocation policies by combining AI technologies, such as knowledge graph and reinforcement learning. By building an open service framework, service-oriented encapsulation and secure and trusted calling can be realized. An open platform can be created to allow third-party developers to register and publish new intelligent services, forming a rich service ecosystem.

➤ **On-demand flexible provisioning and lifecycle management of edge computing services**

Embodied applications are extremely sensitive to delay, and the computing load fluctuates greatly. Centralized cloud computing cannot meet the needs. Therefore, edge computing is required. Distributed computing resource pools can be built at the network edge to dynamically create, migrate, scale in/out, and release computing services according to the real-time needs of embodied AI agents. Computing power-aware routing and network protocols can be used to achieve intelligent scheduling and distribution of computing tasks through wireless computing power networks. A wireless computing power network management framework can be designed to perform full-lifecycle management on edge service instances and achieve seamless migration of services during the movement of agents to ensure task execution continuity.

➤ **QoAIS guarantee technology**

QoAIS is a core indicator for measuring the quality of intelligent information services. It is a multi-dimensional measurement system that goes beyond traditional network QoS (such as bandwidth and delay). It involves multiple layers, including sensing, communication, computing, and control, and covers multiple dimensions such as task success rate, task completion time, and energy consumption efficiency. By monitoring multi-dimensional indicators in real time, it breaks down barriers between layers and conducts cross-layer joint resource scheduling and optimization.

4.3.6 Data Security and Privacy

The deployment environment for edge AI is changing from highly controlled cloud data centers to the open, discrete real world. This fundamental change poses unprecedented challenges to the security paradigm. Therefore, the focus on security is not a supplement but the primary prerequisite for the successful implementation of edge intelligence. Its uniqueness stems from multiple factors: (1) The physical security boundary becomes blurred, and edge devices may be deployed in unattended public places, being prone to physical contact and tampering. (2) Network connections are unreliable and changeable. Unstable networks not only affect performance but also greatly increase the risk of communication monitoring or man-in-the-middle attacks. (3) Devices have highly limited resources and are hard to support complex traditional security software, making them a vulnerable spot in the security chain. (4) Heterogeneous environments are extremely complex, with coexisting hardware and software stacks from different manufacturers and under different architectures. This greatly expands the attack surface. In view of these challenges, it is urgent to build a corresponding threat model whose attack surface covers every layer of the system: (1) At the data layer, attackers can contaminate training sets through data poisoning or steal privacy data under transmission. (2) At the model layer, there are threats of model stealing, reverse engineering, and misleading inference results through adversarial example attacks. (3) At the infrastructure layer, edge nodes or terminal devices may be hijacked and become members of a botnet. (4) In the communication layer, man-in-the-middle attacks can tamper with or interrupt key instructions and data flows. Systematic identification of these threats is a prerequisite for building an effective security protection system.

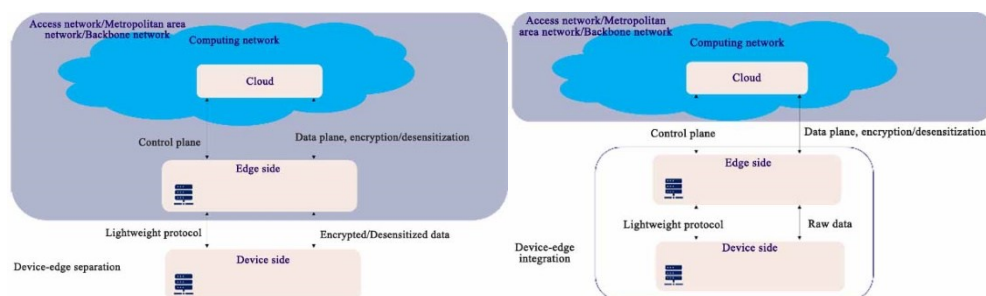


Figure 14 Edge-Device Separation and Edge-Device Integration Modes

In the AI over Edge era, regardless of device-edge integration or device-edge separation (as shown in Figure 14), security and privacy are the bottom-line focuses that must be faced, as long as data is transmitted to places other than the local area. In the edge-device separation mode, devices have "no computing power" or "little computing power", and all AI inference tasks are completed at the edge. Typical examples include security Internet protocol cameras

(IPCs) and industrial sensor nodes. In the device-edge integration mode, devices have "strong computing power" and can locally complete most AI tasks. Incremental capabilities are requested from the edge/cloud only for special needs. Typical examples include intelligent robots and autonomous driving vehicles.

Data security and privacy protection technologies related to AI Edge include but are not limited to:

➤ **Distributed trust**

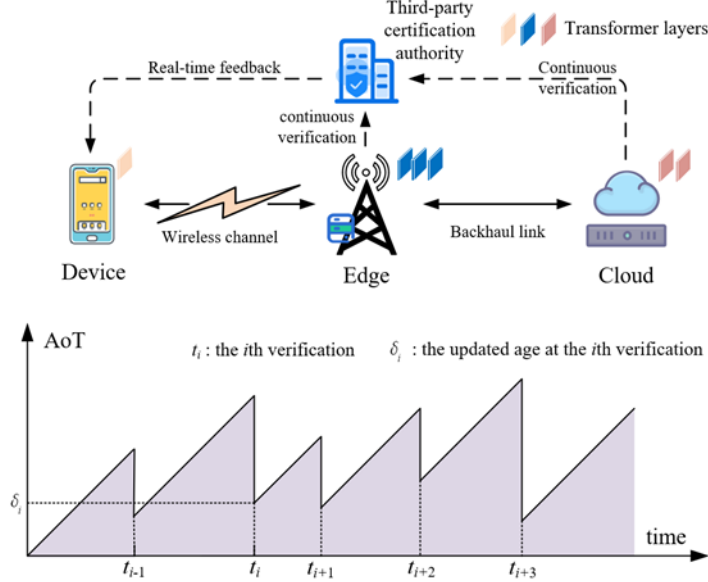


Figure 15 Continuous Authentication of Cloud-Side and Edge-Side Identities

In open and decentralized edge computing environments with blurred boundaries, users, devices, and edge workloads must undergo strict identity-based authentication and authorization before accessing any resources. Micro-isolation technology is used to achieve refined access control and ensure that attackers cannot perform lateral movement easily, even if a single node is compromised. Furthermore, authentication is designed as a continuous process. The dynamic access control mechanism continuously evaluates the security status of access subjects, such as the device fingerprint, model version, geographical location, and abnormal behavior. It can combine the concept of Age of Trust (AoT)^{[43][44]} to calculate the risk level in real time and dynamically adjust access permissions, thus forming adaptive, risk-based security policies.

➤ **Data masking**

Sensitive data is deformed, blurred, or replaced to minimize the sensitivity of data without affecting the purpose of data analysis or business logic, thereby protecting personal privacy and commercial secrets. Its goal is to make the data "usable but invisible" (the sensitive part).

➤ **Homomorphic encryption (HE)**

Edge devices can send sensitive data that has undergone HE (such as personal health data and video clips) to the edge server or the cloud. The server can directly perform AI model inference or training/ computing without knowing the content of the data (encrypted) and return the encrypted result. Only edge devices with the key can decrypt the final result.

➤ **Differential privacy (DP)**

An appropriate amount of random noise is injected into the data before the data leaves edge devices or when it is uploaded to a model for updates. The noise is large enough to mask the data contribution of any individual, preventing information about specific individuals from being inferred based on the output. The noise is also small enough so that it does not have a statistically significant impact on the analysis result of the entire dataset.

➤ **Secure multi-party computation (MPC)**

The input data of each party is split and obfuscated through a cryptographic protocol. As a result, during the computing process, neither party can view the raw data of other parties, but can obtain the correct computing result at the end. The main protocols for implementing MPC are those based on Garbled Circuit (GC), Secret Sharing (SS), and HE.

➤ **Trusted execution environment (TEE)**

A hardware-isolated "secure enclave" is created on edge devices (such as chips in mobile phones and tablets). Sensitive AI models and data can be processed and stored inside the TEE. They are completely isolated from the operating system and other applications running on devices. Even attackers with root privileges cannot access the content of the TEE. The current mainstream trusted hardware technology routes in the industry include ARM Trust Zone, Intel SGX, AMD SEV, RISC-V keystone, etc.

➤ **Blockchain technology**

In distributed and decentralized edge AI deployment environments, establishing cross-subject trust and ensuring transparent and auditable operation processes have important reference value for large-scale applications. Blockchain technology, with inherent immutability and traceability of its distributed ledgers, provides vital foundational capabilities for building trusted edge AI. Firstly, in terms of model trustworthiness, blockchain is used to record the traceability information of the entire model lifecycle. The source and hash of training data, the hyperparameters and environment of the training process, the version update and iteration records, and the logs of final distribution and deployment on edge nodes are all permanently recorded on the chain in a tamper-proof manner. This approach provides verifiable evidence for model trustworthiness and effectively prevents the spread of malicious models or models with contaminated data. This is especially important in multi-party collaboration scenarios. Secondly, in terms of the audit and accountability of AI inference, logs of key inference decisions, including the hash of input data, model version, inference result, and timestamp, can be stored on the chain in real time. This not only provides a data basis for subsequent responsibility definition and dispute arbitration but also provides a high-quality and reliable feedback data closed loop for model optimization and iteration. Finally, the distributed consensus mechanism of blockchain itself establishes a trusted network among edge nodes that does not require a central authority. Any attempt to tamper with logs or data will be rejected by the network, effectively preventing data forgery or malicious activities when a single or multiple edge nodes are compromised due to malicious attacks.

➤ **Model security technology**

In terms of model security, the adversarial example defense technology actively identifies and resists carefully constructed malicious inputs by introducing adversarial examples during training or deploying detection filters during inference. This ensures that the judgment of edge AI models is not misled and maintains decision reliability. At the same time, to protect the AI model assets with a huge resource investment during R&D, model watermarks are embedded

in them. The unique identification information is hidden in model parameters or structures. This provides strong evidence for model ownership confirmation and leakage source tracing to effectively curb illegal copying and abuse of models.

4.4 Chip and Computing Power Foundation

As the core infrastructure of all-domain intelligence, the AI Edge chip and computing power foundation builds a heterogeneous computing foundation with deep integration of "communication, sensing, intelligence, computing, and control" through multi-level technological breakthroughs. Its core adopts a multi-dimensional heterogeneous computing architecture to achieve hardware-level integration of general computing units, domain-specific accelerators, and multi-modal interfaces in the space of a single chip. Relying on three-dimensional stacking and storage-computing integration technologies, it breaks through the memory wall limitations and significantly reduces data transmission energy consumption and delay. The dynamically configurable architecture supports the flexible adaptation of communication and AI computing power through a scalar-vector-matrix three-dimensional fusion mechanism, thus meeting the composite needs of millisecond-level real-time control and distributed cognitive collaboration. The global computing power scheduling engine builds a cross-framework collaborative hub. Combined with green scheduling policies with trusted isolation mechanisms, it drives the global optimization and secure flow of heterogeneous resources. The endogenous security system integrates the hardware root of trust and a multi-level protection architecture to ensure function reliability and protect data privacy in physically exposed environments. Particularly, the architecture deeply integrates the core advantages of modularity, scalability, openness, and collaboration brought by Reduced Instruction Set Computer V (RISC-V). It also gives full play to its potential in customized computing unit design, multi-precision computing power support, and software-hardware ecology unification. In this way, it promotes the realization of full-stack support capabilities from the chip physical layer to the distributed operating system layer and provides a flexible and scalable intelligent foundation for various industries.

4.4.1 Chip Architecture Innovation with Integration of Communication, Sensing, Intelligence, Computing, and Control

The core of AI Edge chip design for all-domain intelligence is to build a dynamic capability matrix that supports the future development of edge intelligence. The future development of edge intelligence is highly dependent on whether the chip architecture can achieve deep collaboration between physical-layer resources and digital-layer capabilities. This requires AI Edge chips to break through traditional rigid constraints and establish a computing power foundation with multi-dimensional integration of "communication, sensing, intelligence, computing, and control" that adapts to the elastic evolution of businesses. Through the multi-dimensional integration and dynamic scheduling of heterogeneous computing resources, the chips can meet the composite needs of millisecond-level real-time control, distributed cognitive collaboration, and cross-domain autonomous decision-making. In this process, it is particularly important to introduce the RISC-V ecosystem. Its advantages of open source, scalability, and modularity provide underlying technical support for the chip architecture, allowing

communication, sensing, computing, and control functions to achieve efficient collaboration and dynamic adaptation under a unified instruction set framework. To achieve this goal and sustain continuous innovation, it is urgent to establish an open and collaborative technology foundation. This involves adopting modular design standards to enable flexible composition of functional units, implementing cross-protocol hardware abstraction layers to ensure seamless interoperability across multi-domain devices, and deeply integrating memory-computing integration with three-dimensional integration processes. Ultimately, a full-stack collaboration system covering device physics, circuit architecture, and a distributed operating system can be built, transforming AI Edge chips into the invisible engine that powers the intelligent transformation of various industries.

➤ **Multi-dimensional heterogeneous computing engine**

AI Edge chips need to complete the four-dimensional integration of general computing units, domain-specific accelerators, multi-modal interfaces, and control engines in the space of a single chip. Through three-dimensional stacking and through-silicon vias (TSV) vertical interconnection technologies, the micrometer-level near-memory integration of computing cores, high-bandwidth memories, and communication baseband units is realized to eliminate the delay and energy efficiency loss caused by data transmission. The spatiotemporal multiplexing pipeline mechanism is adopted at the architecture level to enable heterogeneous tasks, such as industrial protocol parsing, tensor-accelerated computing, and RF signal processing, to be executed concurrently with a unified hardware resource pool. Leveraging the high flexibility and scalability provided by the RISC-V open instruction set, various processing units can achieve instruction-level dynamic optimization and functional reconfiguration according to task requirements, further enhancing the efficiency and adaptability of computing power integration. This deep integration not only reduces end-to-end processing delay to the nanosecond level, but also enables the chip to achieve an exponential increase in computing power density under sudden load scenarios compared to traditional architectures, providing underlying support for high-concurrency AI Edge intelligent services.

➤ **Cognition-driven dynamically configurable architecture**

The core of dealing with the mixed workload of communication and AI lies in building a modular and tailorable heterogeneous converged architecture. The AI Edge chip adopts a scalar-vector-matrix three-dimensional computing fusion mechanism: through a domain-specific architecture (DSA) composed of "scalar processor core + one-dimensional computing array + two-dimensional computing array", it realizes multi-dimensional collaboration of computing resources and efficient linkage between storage and computing. This architecture fully integrates the modularity and scalability of the RISC-V instruction set to achieve seamless integration from general computing to domain-specific acceleration and deep fusion at the instruction set level. Hierarchical storage and high-speed interconnection networks support different computing units to undertake differentiated tasks, ensuring professional efficiency in communication vector operations and AI matrix operations, and enabling resource complementarity through bidirectional expansion of computing array capabilities when the load fluctuates, thereby significantly improving the overall computing power utilization rate. This architecture breaks down the barriers between traditional communication and AI computing power, forming a scalable pool of hardware resources.

To address the diverse requirements in edge computing scenarios, chips must possess the dual capabilities of dynamic hardware resource sizing and flexible software algorithm adaptation. For scenarios with high AI requirements (such as autonomous driving and smart cities), parallel computing capabilities are enhanced and low-latency data transmission is ensured by activating full-two-dimensional computing arrays and optimizing inference algorithms. For scenarios with high communication requirements (such as drone swarms and emergency communications), priority is given to configuring one-dimensional computing arrays and high-speed communication modules, and optimizing signal processing and multi-node collaborative algorithms. For scenarios with low power consumption requirements (such as smart agriculture monitoring), data collection and lightweight inference processes are optimized by reducing the computing scale and enabling the low power consumption mode. Leveraging the unified and flexible instruction set support provided by the RISC-V ecosystem, this architecture effectively avoids ecosystem fragmentation caused by the traditional "one chip, one scenario" approach, significantly improving R&D efficiency and cross-scenario compatibility, and providing a flexible and efficient hardware foundation for the large-scale implementation of AI Edge.

➤ **Adaptive system with multiple energy efficiency optimizations**

To address the extreme energy efficiency constraints of AI Edge scenarios, chips need to build a cross-layer optimization system from the transistor level to the system level. Voltage and frequency island technology divides the chip into multiple independent power supply domains, supporting dynamic voltage and frequency adjustment with microsecond-level precision, reducing the chip's power consumption by more than 90% during switches between quiescent state and peak load. The burst load predictor preloads computing resources by pre-analyzing instruction flow characteristics, eliminating the energy downtime during the traditional power-on and power-off process. More importantly, future chips can incorporate optoelectronic interconnects, which will increase data transmission throughput while further reducing data transmission power consumption between systems. Combining the streamlined and efficient features of the RISC-V instruction set with the energy efficiency control mechanism continuously optimized by the open-source community, the chip can further reduce power consumption at the instruction execution level. This system with multiple energy efficiency optimizations ensures that the chip can seamlessly switch between microwatt-level standby and full-load operation state, enabling the AI Edge chip to run efficiently and reliably for extended periods, and truly promoting the large-scale implementation of AI in the physical world.

➤ **Endogenous security and trusted execution architecture**

With the increasing prevalence of edge intelligence devices and the growing value of the data they carry, chips are facing more severe threats of physical exposure. This makes it crucial to enhance protection capabilities at the hardware level and requires chips to embed hardware-level roots of trust and dynamic defense mechanisms. A four-layer protection system will be implemented throughout the entire chip design stack: Physically unclonable functional units generate an uncopyable encrypted identity for each computing task; homomorphic acceleration cores ensure that sensitive data remains encrypted throughout the entire computing cycle; formal verification hard cores perform mathematical proof-level verification of industrial control instruction flows; radiation-resistant fault-tolerant design maintains functional

continuity in space and industrial environments with strong interference through dynamic redundancy switching. Leveraging the auditability and transparency of the RISC-V open-source instruction set and combining its modular security extension mechanism, this architecture can effectively achieve a deep integration of hardware Trojan detection, bypass channel protection, and fault self-healing capabilities. It natively unifies hardware-level security and functional safety, providing a functional safety and information security integrated protection system for key application areas such as autonomous driving, telemedicine, and industrial control.

➤ **Distributed intelligent collaborative acceleration unit**

To support the cloud-edge-device collaborative architecture of AI Edge, the chip needs to have a built-in cross-layer collaborative acceleration unit. The federated learning hardware engine supports local parameter aggregation and encrypted gradient exchange, enabling distributed nodes to complete model co-evolution under the premise of privacy protection; the knowledge distillation accelerator compresses the global model into a lightweight expression that can be carried at the edge, reducing the collaborative communication overhead by more than 90%; the spatiotemporal alignment interface performs hardware-level timestamp calibration and spatial coordinate mapping on multi-source sensing data, providing microsecond-level precision environmental situation consensus for scenarios such as vehicle-road collaboration and drone swarms. Leveraging the hardware and software compatibility provided by the RISC-V open instruction set standard, more efficient instruction set-level collaboration and task scheduling can be achieved between different devices and platforms. These mechanisms enable the collaborative inference throughput to increase nearly threefold for every doubling of the edge node scale, achieving superlinear growth in intelligent capabilities.

➤ **Unified instruction set framework for heterogeneous computing power**

The instruction set for AI computing power adaptation in communication requires building a technical system with the "open-source architecture as the basis, standard collaboration as the outline, and intelligent scenario matching as the key". With the modular extensibility of the open-source instruction set architecture as its core, it creates a flexible instruction framework that combines generalizability and customization. Through scalable instruction clusters, it supports core edge needs such as multi-precision computing and real-time response, and enables seamless switching from general computing to domain-specific acceleration. The focus is on promoting the construction of a cross-scenario standard system, driving the coordinated unification of instruction set specifications with edge device interfaces and security protocols, solving the interoperability problem with heterogeneous hardware, and forming an open and compatible technology ecosystem. The dynamic adaptation capability of scenario sensing is enhanced, intelligent scheduling and energy efficiency balance of computing power resources are realized with the aid of the software and hardware collaborative optimization mechanism, and hardware-level security and distributed collaborative instruction design are integrated, providing underlying support for trusted interconnection and collaborative decision-making of edge nodes, and helping to build an elastic and scalable edge computing power base.

4.4.2 Intelligent Scheduling Engine for Global Heterogeneous Computing Power

The instruction set for AI computing power adaptation in communication requires building a technical system with the "open-source architecture as the basis, standard collaboration as the

outline, and intelligent scenario matching as the key". With the modular extensibility of the RISC-V open-source instruction set architecture as the core, a flexible instruction framework that combines generalizability and customization is created. Through a scalable instruction set, it supports core edge needs such as multi-precision computing and real-time response, and enables seamless switching from general computing to domain-specific acceleration. Due to its open, simple, and modular design, the RISC-V instruction set has significant scalability and customization capabilities and enables instruction-level optimization and functional expansion for integrated AI and communication computing scenarios, greatly improving energy efficiency and real-time performance. The focus is on promoting the construction of a cross-scenario standard system, driving the coordinated unification of instruction set specifications with edge device interfaces and security protocols, solving the interoperability problem with heterogeneous hardware, and forming an open and compatible technology ecosystem. The dynamic adaptation capability of scenario sensing is enhanced, intelligent scheduling and energy efficiency balance of computing power resources are realized with the aid of the software and hardware collaborative optimization mechanism, and hardware-level security and distributed collaborative instruction design are integrated, providing underlying support for trusted interconnection and collaborative decision-making of edge nodes, and helping to build an elastic and scalable edge computing power base.

➤ **Computing power sharing mechanism driven by the converged architecture**

On the AI Edge computing power platform, computing power scheduling and sharing are key links to achieve efficient resource utilization and ensure low-latency task execution. Backed by an integrated computing power medium and a full-stack collaborative architecture, full-link computing power scheduling and sharing are carried out with inputs at the algorithm, configuration, and hardware levels. At the algorithm level, the mechanism supports multi-format task input. AI tasks are carried by .ONNX format models, and communication tasks are written in a custom C-like language, providing diverse task sources for computing power scheduling. At the configuration level, it integrates the hardware resource status (load and available resources of scalar units, one-dimensional computing arrays, and two-dimensional computing arrays) and software parameters (task priority, delay requirements, computing power requirements, etc.) to provide constraints for scheduling decisions. Furthermore, the computing power scheduling and sharing mechanism can dynamically adjust resources, breaking through the limitations of traditional architectures in terms of the static allocation of hardware resources. When a certain type of computing unit is fully loaded (for example, when a one-dimensional computing array is fully loaded during communication peak hours), load can be offloaded across units to realize resource complementarity. The computing load of some communication tasks can be offloaded to idle two-dimensional computing arrays. Alternatively, when the concurrency of AI inference tasks increases, idle one-dimensional computing arrays can be allocated to undertake lightweight AI tasks, ensuring that the overall computing power resources are always efficiently utilized and avoiding resource waste caused by "uneven load distribution". Meanwhile, with the help of a unified intermediate representation and task feature recognition, the communication-AI fusion application of AI Edge can support pipeline-type scheduling of cross-type tasks, allowing different computing arrays to alternately process tasks at different stages and thereby improving the overall task execution efficiency.

➤ **Green computing power scheduling policy**

Closely aligned with the sustainable development needs of edge infrastructure, this policy, with "coordinated optimization of computing power and energy" as its core orientation, deeply binds the computing power requirements of edge nodes with the supply characteristics of distributed energy (such as photovoltaics and energy storage), forming a closed-loop system of "energy fluctuation sensing - dynamic migration of computing power - energy efficiency priority management". In practical applications, it can both shift non-real-time tasks (such as video data backtracking analysis and lightweight model training) to periods with abundant energy through off-peak scheduling, and prioritize the computing power supply for critical services (such as emergency monitoring and public services), while reducing the overall energy consumption and carbon footprint of edge clusters. This not only alleviates the energy supply constraints of edge nodes, but also promotes the deep integration of digital infrastructure and green energy systems, providing a low-carbon and low-cost path for edge intelligence deployment across various industries.

➤ **Trusted computing power isolation and dynamic authentication mechanism**

To address the security challenges in high physical exposure of edge devices and prominent security risks in cross-domain collaboration, this mechanism, with "endogenous security + dynamic defense" as its core, builds a trusted computing power system from hardware root trust to full-link protection. On the one hand, by using computing power isolation technology, independent operating spaces are allocated for different scenario tasks (such as industrial control instruction execution and user privacy data processing) to avoid data leakage or task interference. On the other hand, a dynamic authentication mechanism is established to verify the identity of edge nodes, data integrity, and task legality in real time, so as to ensure the security of computing power calls even when the device is physically touched or the network environment is complex. This mechanism builds a strong protective barrier for scenarios with high security requirements such as autonomous driving, telemedicine, and the Industrial Internet, solving the trust problem in the large-scale application of edge intelligence. It supports security collaboration across nodes and levels.

➤ **Cross-framework computing power collaboration hub**

To address the resource island problem caused by "framework fragmentation and hardware heterogeneity" in the edge computing power ecosystem, a unified computing power collaboration hub is created, which is the core of this strategy—by building standardized interfaces and adaptive adaptation engines, it aims to break down the collaboration barriers between different AI frameworks such as TensorFlow Lite and PyTorch Edge and between heterogeneous hardware platforms such as ARM, RISC-V, and x86. It can automatically convert different framework models and optimize their deployment, allowing the same edge node to flexibly undertake inference tasks for training multiple frameworks; at the same time, it promotes the pooling of "cloud-edge-device" heterogeneous computing power resources, allowing elastic sharing of computing power across devices and levels. This not only lowers the technical threshold for deploying AI models across scenarios and avoids enterprises from repeatedly investing computing resources due to differences in frameworks, but also releases the aggregated value of fragmented edge computing power, accelerates the transformation of algorithm innovation into industrial applications, and supports the flexible and differentiated implementation of edge intelligence in fields such as retail, agriculture, and logistics.

4.4.3 Open Ecosystem for Intelligent Computing Power

The intelligent computing power open ecosystem, with developer empowerment as its core, builds a technical support chain covering the entire process. The communication-AI converged computing architecture achieves a hardware-insensitive description of communication and AI tasks through a unified programming language. Combined with the compiler's intelligent mapping mechanism, it dynamically allocates task instructions to heterogeneous units such as scalar processors and one-dimensional/two-dimensional computing arrays, breaking through the hardware adaptation barriers faced by traditional architectures. The full-stack development support system provides a full lifecycle toolchain from functional simulation to prototype implementation, with a built-in communication-AI fusion verification environment and modular operator interfaces, significantly reducing the development threshold for complex edge applications. Drawing upon the core advantages of the CUDA ecosystem, the unified programming paradigm platform establishes a development framework that seamlessly integrates communication and AI, supports multi-format model migration and cross-precision computing needs, and provides elastic and scalable deployment capabilities for differentiated scenarios such as retail, agriculture, and logistics.

➤ **Communications-AI converged computing architecture (AI-Unified Radio Architecture, abbreviated as AURA)**

The AURA computing architecture consists of the Venus programming language, the Zoozve compiler, and a basic operator library. It addresses the pain points of complex hardware adaptation and fragmented task description in traditional architectures, and builds an efficient bridge between the hardware layer and the development platform layer. The Venus programming language supports a unified description of communication and AI tasks. Developers do not need to distinguish the type of task (communication or AI) and the corresponding hardware requirements (such as baseband or NPU). They only need to use Venus to write task logic to achieve a unified expression across hardware, which greatly simplifies the complex task description. As the core of code compilation and hardware mapping, the Zoozve compiler has two key capabilities: First, it compiles the unified task code written by Venus into machine instructions that can be executed by the AI Edge chip; second, through an intelligent mapping mechanism, it automatically allocates the compiled instructions to the corresponding computing units within the chip (such as mapping communication instructions to a one-dimensional computing array and AI instructions to a two-dimensional computing array), thus completing the hardware adaptation of communication and AI operators. Especially for long vector tasks, Zoozve breaks through the limitations of RISC-V Vector Extension (RVV) in terms of the number of static registers and power-of-2 grouping. Through strip-mining-free design and data adaptive register allocation policies, it supports arbitrary length vectors and register grouping configurations, which can reduce the number of dynamic instructions for communication tasks such as fast Fourier transformation (FFT) by at least 10 times, while increasing the chip area by only 5.2%. It improves compilation efficiency while avoiding the performance loss and hardware resource waste caused by traditional compilation. The basic operator library focuses on "hardware-level operator encapsulation," covering core basic operators in the fields of communication and AI. In the communication field, including FFT, modulation and demodulation, and channel coding and decoding, it can directly match the hardware characteristics of one-dimensional computing arrays. In the AI field covering

Conv2D/3D, all-inclusive connectivity, and activation functions (GELU/SiLU), it is adapted to the parallel capabilities of two-dimensional computing arrays. It also supports operator hardware mapping optimization, ensuring that each type of operator can match the architectural advantages of the corresponding computing unit, and providing high-performance underlying operator support for the Echo platform.

➤ **Full-stack development support system**

With the goal of "lowering the development threshold and accelerating application implementation", the Echo platform encapsulates the underlying capabilities of the AURA architecture in a "developer-friendly way", builds a support system covering the entire "development-verification-deployment" process, and provides one-stop services from functional simulation to prototype implementation. During the development and verification phase, the built-in application function and performance simulation tool of the platform enable developers to perform full-process verification of communication-AI fusion tasks (such as "5G signal processing + AI channel estimation") before physical hardware development. This verifies the functional correctness of the task logic (such as communication protocol compliance and the accuracy of AI inference results) and quantifies performance metrics (such as end-to-end delay, multi-dimensional computing unit utilization, and computing power resource consumption). It also helps identify potential issues such as hardware compatibility conflicts and computing power bottlenecks in advance, avoiding rework in later development. At the operator call level, the platform performs secondary encapsulation of the AURA basic operator library and provides commonly used communication-AI computing modules (such as 5G physical layer processing modules and lightweight AI inference modules), which are made available to developers through standardized interfaces. Developers do not need to put efforts on the underlying hardware implementation of the operators; they can directly call modules to complete the development of basic functions or extend custom modules based on interfaces (such as 6G semantic communication encoding modules), balancing ease of use and scalability. Regarding toolchain support, the platform provides a full-process toolkit from code writing, compilation, and debugging to deployment. The code writing link is compatible with mainstream development habits, using C-like and Python-like programming syntax. The compilation link integrates the Zoozve compiler to achieve automatic mapping from Venus code to hardware instructions. The debugging link supports fine-grained monitoring at the computing unit level (such as one-dimensional/two-dimensional computing array load and data flow path). The deployment link allows for one-click adaptation of applications to single-chip or multi-chip clusters, ensuring a smooth and efficient development link. At the prototype implementation level, the platform has built-in application prototype demos for typical scenarios such as 5G/LTE and AI channel estimation. Developers can quickly modify parameters (such as communication frequency bands and AI model precision) or extend functions (such as adding multi-terminal collaborative logic) based on these demos, significantly shortening the cycle from solution design to actual application implementation and providing convenient support for the large-scale promotion of AI Edge.

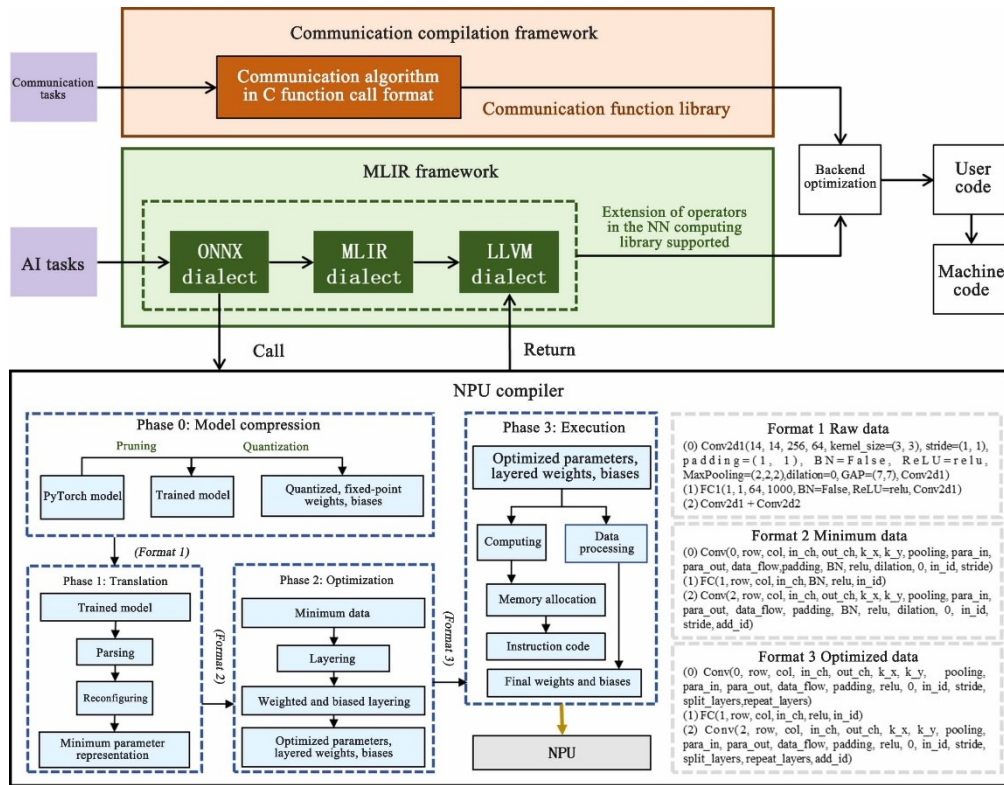


Figure16 Programming Framework for Transforming Communication and AI Fusion Tasks into a Unified Intermediate Representation

➤ Unified programming paradigm platform for communications and AI

In the traditional discrete architecture, communication and AI tasks require adapting to different dedicated development tools due to the fixed functional boundaries of hardware units. This leads to a fragmented programming model, which greatly increases the development threshold and learning cost of AI Edge applications, as well as inefficient task collaboration caused by the fragmented development link. The Echo platform focuses on the goal of "one toolchain covering all development scenarios of AI Edge". Drawing upon the core advantages of the CUDA ecosystem's unified programming paradigm and rich operator support, it builds a unified programming ecosystem for communication and AI. In the three aspects of model migration, task development, and operator support, it completely solves the pain points of traditional AI Edge computing power bases in terms of "heterogeneous programming languages of multiple computing units and fragmented development links", as shown in Figure16. The Echo platform provides a unified programming model for communication and AI tasks, allowing developers to develop a unified programming paradigm without distinguishing between task types (communication or AI) and corresponding hardware requirements. Whether in signal modulation and demodulation and FFT operations in the communication field, or in convolution and attention mechanism inference in the AI field, task logic can be described within the same programming framework.

➤ High-performance operator library system

Currently, AI Edge scenarios face the bottleneck of difficulty in coordinating low communication delay and high AI computing power in core applications such as industrial control and vehicle-road collaboration. Traditional computing architectures are confronted with

serious resource waste due to task fragmentation. The system constructs a "basic operator layer - fusion operator layer - optimization technology layer" architecture: The basic operator layer adapts communication algorithms (FFT/LDPC encoding/decoding) and AI computing (convolution/all-inclusive connectivity) to the characteristics of heterogeneous computing units through hardware-aware encapsulation, thus solving the pain point of task fragmentation; the fusion operator layer develops innovative communication-AI hybrid operators (such as joint computing of AI channel estimation and precoding), breaking through the delay bottleneck by reducing cross-module data transfer; the optimization technology layer adopts mixed-precision computing, operator fusion, and the zero-copy mechanism to improve energy efficiency while maintaining communication precision. This system provides a high-efficiency computing power base for real-time collaborative computing on the edge side, promotes the evolution of integrated sensing and computing, and supports the intelligent upgrade of scenarios such as autonomous driving and Industrial Internet.

➤ **Open operator ecosystem that empowers developers**

Developers face three major bottlenecks: complex multi-platform hardware adaptation, low debugging efficiency, and a fragmented ecosystem, which result in a long development cycle for the innovative applications of AI Edge and difficulty in large-scale implementation. This ecosystem builds cross-platform capabilities through a "multi-framework compatible interface layer": The CUDA compatibility layer enables zero-code migration of PyTorch/TensorFlow models, the OAI interface encapsulates the 3GPP and O-RAN protocol stacks to seamlessly integrate traditional base station software, and the domain-specific language (DSL) extension improves the development efficiency of complex communication-AI fusion tasks with Python-like syntax and predefined templates; combined with a "full-cycle development toolchain", a visual performance profiler automatically locates operator bottlenecks, an automatic optimization engine intelligently recommends policies such as mixed precision switching, and a one-click deployment tool automates the entire process from development to deployment through containerization technology. Developers can encapsulate custom algorithms to achieve "one-time development for reuse in multiple scenarios", significantly lowering the threshold for implementing complex algorithms on the edge side, accelerating the implementation of AI applications in fields such as communications, industry, and transportation, and promoting the development of AI Edge technology from single-point innovation to large-scale collaborative development across industries.

4.5 AI Edge System, Platform, and Testing

4.5.1 AI Edge System and Platform

The AI Edge system, as a type of comprehensive information infrastructure for intelligent applications, comprises four core layers: hardware infrastructure layer, operating system layer, software function layer, and application layer.

The hardware infrastructure layer forms the physical foundation of the system, encompassing core elements such as sensing and intelligent terminals, heterogeneous computing power and storage resources, and access and transmission network infrastructure. This layer integrates computing, storage, and network resources to provide heterogeneous

computing power, high-speed interconnection communication, multi-modal access capabilities, and resource pooling services, providing unified hardware abstraction and resource support for upper-layer operating systems and software functions.

The operating system layer, as the core of system software directly deployed on the hardware, is responsible for abstracting and uniformly managing the underlying heterogeneous physical resources. This layer deeply integrates cloud-native technology paradigms. Backed by a lightweight kernel architecture, a deterministic real-time scheduling mechanism, and an endogenous security framework, it achieves unified scheduling and highly reliable operation of heterogeneous computing power resources such as CPUs, GPUs, and NPUs, providing a stable, secure, and deterministic execution environment for upper-layer edge applications.

The software function layer is built on top of the infrastructure and operating system layers. By constructing a distributed data aggregation, federated learning collaboration, edge inference service, and heterogeneous computing power awareness and scheduling system, it provides full-link support for scenario-based applications in data aggregation, collaborative training, and inference deployment, realizing the global collaboration of cloud, edge, and terminal computing power and the efficient and scalable operation of AI tasks. First, backed by underlying abstract hardware resources and real-time operation management, a capability middle platform with edge computing and data functions as its core is formed, providing computing and data services for basic functions such as the wireless network, core network, intelligence, and sensing. Second, the basic function modules provide the application layer with capabilities such as connectivity, AI, sensing, computing, and data services. Third, the management orchestration module consists of business orchestration, network orchestration, cross-domain management, and resource orchestration, with "endogenous intelligence" as its core mechanism. It relies on the integrated scheduling of network, computing, and other resources, realizes hierarchical autonomous network control on demand in real-time, near-real-time, and non-real-time modes, and supports mobile information edge services.

The application layer uses Agentic AI technology to directly map user needs into the basic capabilities of the communication network. It then uses the software function layer to orchestrate on demand, call the corresponding functions, and finally feed back the service status to the application layer, forming a closed loop of "sensing-decision-execution". The implementation of this process relies first on the app's open interfaces and flexible architecture, enabling upper-layer applications to seamlessly call the underlying communication and sensing capabilities. On this basis, fine-grained customization of differentiated user needs is achieved with the help of AI Edge's DOICT fusion technology, thereby supporting the rapid deployment and large-scale promotion of AI applications in vertical industries. Ultimately, an AI-as-a-Service (AIaaS) "application store" ecosystem with AI Edge as the core is built, providing various industries with intelligent service modules that can be subscribed to and combined, and promoting the deep integration of networks and intelligence in terms of capabilities and value.

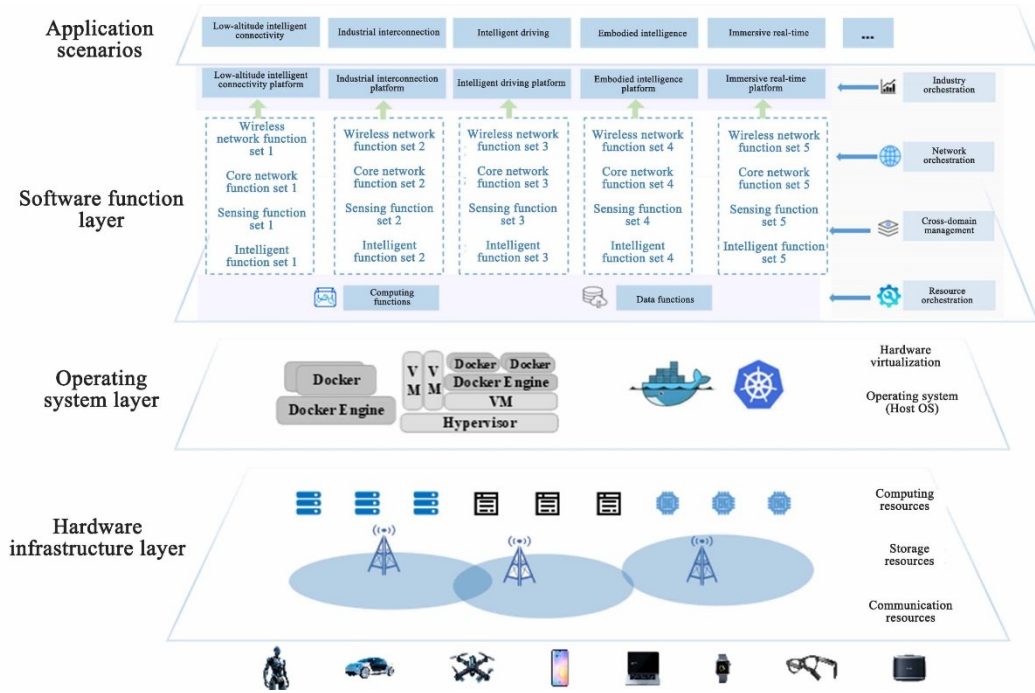


Figure 17 Core Layers of the AI Edge Platform

The core technology direction of the AI Edge system platform revolves around "integration" and "synergy," namely, achieving deep integration of computing power, networks, and intelligence through hardware pooling, OS abstraction, and software intelligence. The main technological challenges at present are as follows.

The system faces significant challenges due to hardware heterogeneity, manifested in diverse computing architectures (such as CPUs, GPUs, NPUs, and domain-specific chips), significant differences in computing power forms, and a variety of communication protocols and access technologies for terminal devices. This deep heterogeneity leads to complex software and hardware coupling, extremely high requirements for driver adaptation and system compatibility, and consequently causes serious ecosystem fragmentation problems, posing substantial obstacles to the unified management of system resources, the large-scale deployment of application services, and cross-platform portability.

In scenarios with high performance requirements, the first ultimate performance challenge is to provide nanosecond-level response precision and 99.9999% reliability for critical tasks in a shared heterogeneous resource environment through hardware virtualization, real-time kernel scheduling, and deterministic network transmission, so as to ensure that the stringent performance boundaries of industrial applications can still be met under extreme loads. The second is to dynamically determine the execution location of tasks on terminals, edge nodes, or the cloud based on the real-time network status, multi-level resource availability, and business service quality requirements, so as to achieve an efficient task offloading policy.

In mobile edge scenarios (such as intelligent connected vehicles and drone swarms), the high-speed mobility and dynamic topology changes of terminals place stringent requirements on business continuity. It is necessary to ensure zero interruption of session state and seamless business migration during cross-base station switching and cross-domain routing. This imposes extremely high requirements on the functional decentralization of the access network and core network, the coordinated scheduling of wireless resources, and the distributed session

management mechanism. It is also necessary to achieve near-field optimization of the control plane and user plane and millisecond-level seamless switching capabilities.

Building a secure and reliable system with full coverage is the core requirement of edge intelligence systems. Lightweight cryptographic protocols and hardware-accelerated encryption mechanisms need to be adopted to ensure end-to-end confidentiality and integrity of transmitted data while maintaining low delay and high throughput. At the same time, with privacy computing technologies (such as federated learning and secure multi-party computation), secure collaboration and value mining among distributed nodes can be achieved without the need for centralized collection of raw data, ultimately achieving the compliance goal of "moving value but not moving data".

4.5.2 AI Edge Testing

As AI models are being deployed on the edge side, the complexity and heterogeneity of AI Edge systems have increased significantly. Verifying their performance, security, and stability across multiple layers and scenarios has become a key prerequisite for large-scale deployment. The testing should not only cover the correctness of AI Edge's functions, but also verify its ability to provide continuous and stable services in dynamic network environments, cross-domain computing power scheduling, and complex application scenarios. Therefore, it is necessary to establish a full-stack testing system that matches the system architecture, and realize end-to-end verification from the hardware infrastructure layer to the application layer, so as to provide methodological and tool support for the implementation of highly reliable AI Edge systems.

➤ Layered testing framework

Verification of the hardware infrastructure layer: The focus is on verifying the performance of heterogeneous computing power and communication resources, such as terminals, base stations, and servers, under AI-driven conditions. Using tools such as terminal simulators and channel simulators, complex channel characteristics such as multipath fading, neighboring cell interference, and frequency offset distortion are reproduced, and the effectiveness of AI in RAN function optimization, such as channel estimation, signal equalization, and nonlinear compensation, is evaluated. The goal of the testing is to confirm whether AI can achieve robustness and gains at the physical layer in highly dynamic environments.

Verification of the operating system layer: The focus is on heterogeneous computing power abstraction, real-time scheduling, and endogenous security frameworks. A computing power load generator and a scheduling monitoring platform are used to simulate multi-task concurrency and sudden computing demands to test the efficiency and fairness of AI in scheduling heterogeneous resources such as CPUs/GPUs/NPUs. Meanwhile, encrypted transmission and virtualization isolation tests are introduced to verify the security protection and low delay capabilities of the operating system layer.

Verification of the software function layer: The test objects include network orchestration, cross-domain management, resource scheduling, and multi-modal sensing. A business traffic generator and a multi-user business simulator are used to construct mixed business load scenarios, such as voice, video, and IoT, to evaluate the policy generalization and dynamic adaptation capabilities of AI in traffic scheduling, congestion control, and QoS/QoE assurance. For edge-cloud collaborative capabilities, edge inference nodes and cloud training

platforms can be established in an experimental environment to verify the delay and consistency of task offloading, parameter synchronization, and distributed scheduling.

Verification of the application layer: The application layer carries typical industry platforms, including low-altitude intelligent connectivity, Industrial interconnection, intelligent driving, embodied AI, and immersive real-time. The testing focuses on scenario-based end-to-end verification:

- In low-altitude intelligent connectivity, the simulation of drone swarms switching across base stations and link fluctuations is used to test the stability of AI in flight path planning and resource scheduling.
- In Industrial interconnection, high-concurrency control messages and burst loads are injected to verify the ability of AI to ensure millisecond-level delay and production line stability.
- In intelligent driving, high-speed movement and the Doppler effect are reproduced to test the real-time performance of AI in V2X collaboration and task offloading.
- In embodied AI, the multi-modal sensing and motion control flow of robots is simulated to evaluate the response and robustness of AI in the closed-loop sensing-decision-execution.
- In immersive real-time scenarios, high-bandwidth AR/VR business streams are generated, with delay jitter and bandwidth fluctuations superimposed to test AI's dynamic optimization capabilities for QoS/QoE.

➤ **Key test requirements**

AI Edge testing needs to meet the following four core requirements:

Verification of cross-layer consistency: Achieve end-to-end overall evaluation from physical layer air interfaces and network layer business to application layer platforms, ensuring that the AI optimization effect is consistent throughout the entire stack.

Multi-dimensional scenario reproduction capability: Support accurate simulation of complex environments such as channel time-varying characteristics, multi-user interference, cross-domain switching, and high-speed movement.

Verification of dynamic adaptability and closed-loop optimization: Evaluate the initial performance of AI and test its online learning and self-optimization capabilities, forming a closed-loop iteration of "testing-evaluation-optimization-re-verification".

Assessment of robustness and security: Inject high noise, abnormal traffic, and adversarial examples into the network impairment tester and attack simulation platform to test the fault tolerance and security boundaries of AI under extreme conditions.

➤ **Tools and platform support**

To implement the above testing scenarios, a standardized tool system and integration platform need to be built:

- Laboratory simulation environment: Create a controllable and reproducible verification environment using terminal simulators, channel simulators, network impairment testers, and business traffic generators.
- Cross-layer monitoring and analysis system: Achieve synchronous collection and joint analysis of indicators from the physical layer, business layer, and application layer.

- Closed-loop verification mechanism: Drive the online adjustment of AI models through real-time feedback, forming a dynamic optimization iterative process.
- Visualized testing platform: Provide an intuitive display of multi-dimensional indicators and open API interfaces to connect with third-party tools and automated O&M systems.

In summary, AI Edge testing is an important component of the system and platform. It not only covers the computing power and security verification of the underlying hardware and operating system, but also extends to end-to-end scenario-based verification of software functions and application layers. The building of a consistent, reproducible, and quantifiable testing system across layers allows systematically evaluating the empowerment value of AI in edge networks, laying a solid foundation for the large-scale deployment and trusted application of AI Edge systems.

5. Conclusions

Leveraging the unique advantages of edge network nodes in low latency due to closeness to users, AI Edge will enable the open and efficient utilization of the endogenous computing power of RAN, empowering low-latency and high-reliability integrated mobile information services that cover communications, sensing, intelligence, computing, and control. If the computing power network is the "main artery" of the information age, then AI Edge is like countless "capillaries", providing on-demand access to distributed computing power at the edge. Leveraging the shared foundation of distributed computing power at the edge side, AI Edge not only builds a comprehensive mobile information service infrastructure that supports the hyper-convergence of new DOICT technologies, but also creates an open and inclusive ecosystem of mobile information network vertical industry applications, accelerating the implementation of intelligent applications across various industries. This white paper provides a detailed discussion of the industry and technological background, technological drivers, core technological features and advantages, potential scenarios and needs, and possible technological directions of AI Edge. It aims to catalyze further discussion in both the academia and the industries. In the future, AI Edge will address a series of challenges, including building an open computing platform compatible with various heterogeneous computing powers, designing new network architectures and defining network functions, developing efficient edge AI technologies, operator libraries and toolchains, edge-side apps, computing power trading and billing mechanisms, and data security and privacy protection mechanisms.

References

- [1] H. Zou, Q. Zhao, Y. Tian, L. Bariah, F. Bader, T. Lestable, and M. Debbah, "TelecomGPT: A framework to build telecom-Specific large language models", <https://arxiv.org/abs/2407.09424>, Jul. 2024.
- [2] Y. Sheng, K. Huang, L. Liang, P. Liu, S. Jin, and Y. Li, "Beam prediction based on large language models," *IEEE Wireless Communications Letters*, vol. 14, no. 5, pp. 1406-1410, May

2025.

- [3] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, 'Large generative AI models for telecom: The next big thing?', *IEEE Communications Magazine*, vol. 62, no. 11, pp. 84-90, Nov. 2024.
- [4] T. Wu, Z. Chen, D. He, L. Qian, L. Xu, M. Tao, W. Zhang, "CDDM: Channel denoising diffusion models for wireless communications", in *Proc. IEEE GLOBECOM 2023*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 7429-7434.
- [5] T. Yang, P. Zhang, M. Zheng, Y. Shi, L. Jing, and J. Huang, "WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication," *IEEE Network*, early access, Jun. 2025.
- [6] L. Yu, L. Shi, J. Zhang, J. Wang, Z. Zhang, Y. Zhang, and G. Liu, "ChannelGPT: A large model to generate digital twin channel for 6G environment intelligence," <https://arxiv.org/abs/2410.13379>, Oct. 2024.
- [7] B. Liu, S. Gao, X. Liu, X. Cheng, and L. Yang, "WiFo: Wireless foundation model for channel prediction," *Science China Information Science*, vol. 68, no. 6, Jun. 2025.
- [8] G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. Han, "RF-Diffusion: Radio signal generation via time-frequency diffusion", in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom'24)*, Washington D.C., USA, Nov. 2024, pp. 77-92.
- [9] F. Zhao, Y. Sun, L. Feng, L. Zhang, and D. Zhao, "Enhancing reasoning ability in semantic communication through generative AI-assisted knowledge construction", *IEEE Communications Letters*, vol. 28, no. 4, pp. 832-836, 2024.
- [10] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, X. You, "Large AI model-based semantic communications", <https://arxiv.org/abs/2307.03492>, Jul. 2023.
- [11] L. Qiao, M. B. Mashhadi, Z. Gao, C. H. Foh, P. Xiao, and M. Bennis, "Latency-aware generative semantic communications with pre-trained diffusion models", *IEEE Wireless Communications Letters*, vol. 13, no. 10, pp. 2652-2656, Oct. 2024.
- [12] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, D. Niyato, O. A. Dobre, "Large language model enhanced multi-agent systems for 6G communications", *IEEE Wireless Communications*, vol. 31, no. 6, pp. 48-55, Dec. 2024.
- [13] M. Xu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, A. Jamalipour, "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 1127-1170, Second Quarter, 2024.
- [14] Y. Tian, Z. Zhang, Y. Yang, Z. Chen, Z. Yang, R. Jin, T. Q. S. Quek, and K. K. Wong, "An edge-cloud collaboration framework for generative AI service provision with synergetic big cloud model and small edge models", *IEEE Network*, vol. 38, no. 5, pp. 37-46, Sep. 2024.
- [15] H. Zou, Q. Zhao, L. Bariah, Y. Tian, M. Bennis, S. Lasaulce, M. Debbah, "GenAINet: Enabling wireless collective intelligence via knowledge transfer and reasoning,"

<https://arxiv.org/abs/2402.16631>, Feb. 2024.

- [16] Y. Chen, R. Li, Z. Zhao, C. Peng, J. Wu, E. Hossain, H. Zhang, "NetGPT: An AI-native network architecture for provisioning beyond personalized generative services", *IEEE Network*, vol. 38, no. 6, pp. 404-413, Nov. 2024.
- [17] H. Du, G. Liu, Y. Lin, D. Niyato, J. Kang, Z. Xiong, D. Kim, "Mixture of experts for network optimization: A large language model-enabled approach", <https://arxiv.org/abs/2402.09756>, Feb. 2024.
- [18] Y. Yang, et al., "6G network AI architecture for everyone-centric customized services," *IEEE Network*, vol. 37, no. 5, pp. 71-80, Sept. 2023.
- [19] IEEE GenAINet ETI, website: <https://genainet.committees.comsoc.org/home-2/>.
- [20] IMT-2030(6G) Promotion Group. Research on 6G AI-as-a-Service (AIaaS) Requirements. 2023.
- [21] R. Singh, S. Gill, "Edge AI: A survey", *Internet of Things and Cyber-Physical Systems*, vol.3, pp. 71-79, 2023.
- [22] Q. Hu, Y. Cai, Q. Shi, et al., "Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems", *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1394-1410, Feb. 2021.
- [23] Q. Luo, J. Zhang, S. Hu, and et al., "Joint task migration and resource allocation in vehicular edge computing: A deep reinforcement learning-based approach," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 6, pp. 9476-9490, Jun. 2025.
- [24] S. Liu, G. Yu, R. Yin, and et al., "Joint model pruning and device selection for communication-efficient federated edge learning," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 231-244, Jan. 2022.
- [25] Y. Gao, B. Hu, M. B. Mashhadi, and et al., "PipeSFL: A fine-grained parallelization framework for split federated learning on heterogeneous clients," *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 1774-1791, Mar. 2025.
- [26] E. J. Hu, Y. Shen, P. Wallis, and et al., "LoRA: Low-rank adaptation of large language models," <https://arxiv.org/abs/2106.09685>, Jun. 2021.
- [27] F. Liu, Y. Cui, C. Masouros, and et al., "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728-1767, Jun. 2022.
- [28] T. Jiao et al., "Addressing the Curse of Scenario and Task Generalization in AI-6G: A Multi-Modal Paradigm," *IEEE Transactions on Wireless Communications*, vol. 24, no. 9, pp. 7377-7391, Sept. 2025.
- [29] G. Zhang, H. Li, Y. Cai, and et al., "Progressive learned image transmission for semantic communication using hierarchical VAE," *IEEE Transactions on Cognitive Communications and Networking*, early access, Feb. 2025.

- [30] L. Sun, Y. Wang, A. Swindlehurst, and X. Tang, "Generative-adversarial-network enabled signal detection for communication systems with unknown channel models," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 47-60, Jan. 2021.
- [31] Y. Wang, L. Sun, and A. Swindlehurst, "Knowledge-driven signal detector for uplink transmission in IoT networks with unknown channel models," *IEEE Internet of Things Journal*, vol. 11, no. 15, pp. 25839-25852, Aug. 2024.
- [32] B. D. Son, N. T. Hoa, T. V. Chien, and et al., "Adversarial attacks and defenses in 6G network-assisted IoT systems," *IEEE Internet of Things Journal*, vol. 11, no. 11, pp. 19168-19187, Nov. 2024.
- [33] J. Xiong, M. Wang, D. Zhou, and et al., "Edge intelligence: A review of deep neural network inference in resource-limited environments," *Electronics*, vol. 14, no. 12, p. 2495, 2025.
- [34] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 19-25, Jan. 2020.
- [35] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84-90, Aug. 2019.
- [36] W. Zhao, W. Jing, Z. Lu, and X. Wen, "Edge and terminal cooperation enabled LLM deployment optimization in wireless network," In *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, Hangzhou, China, Aug. 2024, pp. 220-225.
- [37] S. Oh, J. Kim, J. Park, S. Ko, T. Q. S. Quek, and S. Kim, "Uncertainty-aware hybrid inference with on-device small and remote large language models," <https://arxiv.org/abs/2412.12687>, Dec. 2024.
- [38] J. Ning, C. Zheng, and T. Yang, "DSSD: Efficient edge-device deployment and collaborative inference via distributed split speculative decoding," in *Proceedings of the ICML Workshop on Machine Learning for Wireless Communication and Networks (ML4Wireless)*, Vancouver, Canada, Jul. 2025.
- [39] C. Zheng and T. Yang, "Communication-efficient collaborative LLM inference via distributed speculative decoding," <https://arxiv.org/abs/2509.04576>, Sep. 2025.
- [40] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges," <https://arxiv.org/abs/2505.10468>, May 2025.
- [41] J. Tong, W. Guo, J. Shao, and et al., "WirelessAgent: Large language model agents for intelligent wireless networks," <https://arxiv.org/abs/2505.01074>, May 2025.
- [42] Y. Shen, J. Shao, X. Zhang, and et al., "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Communications Magazine*, vol. 62, no. 10, pp. 140-146, 2024.
- [43] Y. Xiao, Q. Du, W. Cheng, P. D. Diamantoulakis, and G. K. Karagiannidis, "Age of trust (AoT): A continuous verification framework for wireless networks,"

<https://arxiv.org/abs/2406.02190>, Jun. 2024.

- [44] Y. Xiao, Q. Du, W. Cheng, G. K. Karagiannidis, A. Nallanathan, and M. Guizani, "Redefining information freshness: AoGI for generative AI in 6G networks," <https://arxiv.org/abs/2504.04414>, Apr. 2025.

List of abbreviations

3GPP: 3rd Generation Partnership Project

6GANA: 6G Alliance of Network AI

AAU: Active Antenna Unit

AI: Artificial Intelligence

AlaaS: AI as a Service

AoT: Age of Trust

API: Application Programming Interface

AR: Augmented Reality

ASC: Adaptive Semantic Compression

ASI: Attention-based Semantic Integration

B2B2C: Business to Business to Consumer

BAIM: Big Artificial Intelligence Model

BERT: Bidirectional Encoder Representations from Transformers

CDDM: Conditional Denoising Diffusion Model

CPU: Central Processing Unit

CQI: Channel Quality Indicator

CSI: Channel State Information

DM: Diffusion Models

DMRS: Demodulation Reference Signal

DOICT: Data Technology, Operation Technology, Information Technology, Communication Technology

DP: Differential Privacy

DRL: Deep Reinforcement Learning

DSA: Domain Specific Architecture

FDD: Frequency Division Duplex

FP32: Floating-Point 32

FPGA: Field-Programmable Gate Array

GAI: Generative Artificial Intelligence

GenAI: Generative Artificial Intelligence

GenAINet ETI: Large Generative AI Models in Telecom Emerging Technology Initiative

Gen-SC: Generative Artificial Intelligence based Semantic Communication

GLOBECOM: Global Communications Conference
GPU: Graphics Processing Unit
GRU: Gated Recurrent Unit
HE: Homomorphic Encryption
INT8: Integer 8
IoT: Internet of Things
ITU-R: International Telecommunication Union-Radiocommunication Sector
JSCC: Joint Source-Channel Coding
LAM-SC: Large AI Model-Based Semantic Communications
LLM4CP: Large Language Model-empowered Channel Prediction
LLM4WM: Large Language Model for Wireless Multi-Tasking
LoRa: Long Range Radio
LoRA: Low-Rank Adaptation
LSTM: Long Short-Term Memory
MaaS: Model as a Service
MEC: Mobile Edge Computing
MoE: Mixture of Experts
MPC: Secure Multi-party Computation
MR: Mixed Reality
MSE: Mean-Square Error
MWC: Mobile World Congress
NAS: Neural Architecture Search
NPU: Neural Processing Unit
PaP: Prompt-as-Prefix
PLC: Programmable Logic Controller
PMI: Precoding Matrix Indicator
QoAIS: Quality of AI Service
QoE: Quality of Experience
QoS: Quality of Service
RAN: Radio Access Network
RCN: Radio Computing Network
RF: Radio Frequency
RI: Rank Indicator
RISC-V: Reduced Instruction Set Computer - V
RRU: Remote Radio Unit
SaaS: Software as a Service
SAM: Segment Anything Model
SemCom: Semantic Communication
SKB: Segment Anything Model based Knowledge Base

SKT: SK telecom
 SoC: System on a Chip
 SRZ: Service Requirement Zone
 TaaS: Training as a Service
 TEE: Trusted Execution Environment
 UPF: User Plane Function
 USR: User Satisfaction Ratio
 V2G: Vehicle-to-Grid
 VR: Virtual Reality
 WiFo: Wireless Foundation Model

List of contributors to the white paper

This white paper was prepared under the leadership and organization of Peng Cheng Laboratory and jointly completed by more than 50 experts and scholars from 20 domestic and foreign enterprises, universities, and research institutions. Peng Cheng Laboratory is responsible for the summary and finalization of materials. Below is the list of contributing organizations and individuals for each chapter (in no particular order).

Chapter		Contributing organization	Contributing individual
1. Background and Requirements		China Mobile	Yingping Cui, Xinyao Wang, and Tianjiao Chen
		Asmote Technology Co.,Ltd.	Lei Chen and Jindong Peng
		Peng Cheng Laboratory	Li Sun and Ning Huang
		Tieto China Co., Ltd.	Qing Ji
		Hong Kong University of Science and Technology	Yang Yang and Mulei Ma
2. Technical Connotation of AI Edge		Peng Cheng Laboratory	Tingting Yang, Ning Huang, and Li Sun
3. Typical Application Scenarios and Potential Value of AI Edge		Asmote Technology Co.,Ltd.	Lei Chen and Jindong Peng
		CICT Mobile Communications Technology Co., Ltd.	Yapeng Wang
		Hong Kong University of Science and Technology	Yang Yang and Mulei Ma
		China Mobile	Yingping Cui, Xinyao Wang, and Tianjiao Chen
		China Telecom	Chao Wu and Qingtian Wang
4. Technical	4.1 System Architecture	China Telecom	Yue Wang, Qingtian Wang, Zexu Li, and Jingyi Wang

Directions and Main Challenges of AI Edge		ZTE	Li Yang, Wenwen Sun, and Feng Xie
		Hong Kong University of Science and Technology	Yang Yang and Mulei Ma
	4.2 AI for Edge Technology	Huawei Technologies Co., Ltd.	Jian Wang and Huiguo Gao
		Shanghai University	Ting Zhou and Shengli Liu
		Institute of Computing Technology, Chinese Academy of Sciences	Congcong Wang, Yanli Qi, and Hanxiao Yu
		Xi'an Jiaotong University	Yichen Wang
		The Chinese University of Hong Kong, Shenzhen	Shuguang Cui, Jinke Ren, Zezhong Zhang, and Jie Xu
		University of Surrey	Chong Huang and Pei Xiao
		Peng Cheng Laboratory	Tingting Yang and Li Sun
		Beijing Dotouch Information Technology Co., Ltd.	Hang Wang and Jun Mao
		Asmote Technology Co.,Ltd.	Zhenbing Zhang and Lei Chen
		Nokia Communications (Shanghai) Co., Ltd.	Tao Tao and Liyu Cai
	4.3 AI over Edge Technology	Peng Cheng Laboratory	Li Sun and Ce Zheng
		CICT Mobile Communications Technology Co., Ltd.	Yapeng Wang
		Hong Kong University of Science and Technology	Zhang Jun
		The Chinese University of Hong Kong, Shenzhen	Shuguang Cui, Jinke Ren, Zezhong Zhang, and Jie Xu
		Xi'an Jiaotong University	Qinghe Du, Yuquan Xiao, and Chaoyang Zhang
		Shanghai University	Ting Zhou and Shengli Liu
		University of Surrey	Chong Huang and Pei Xiao
		Tieto China Co., Ltd.	Qing Ji
	4.4 Chip and Computing Power Foundation	SmarCo	Zixin Yang
		Shanghai University	Ting Zhou, Zhiyuan Jiang, and Shengli Liu
		Biren Technology	Liucheng Duan and Ze Liu
	4.5 AI Edge System, Platform, and Testing	China Unicom	Fuchang Li, Tao Zhang, and Yanjun Ma
		Purple Mountain Laboratories	Yongming Huang, Zening Liu, and Jianjie You

		Beijing Dotouch Information Technology Co., Ltd.	Hang Wang and Jun Mao
		Asmote Technology Co.,Ltd.	Liang Gu, Jindong Peng, and Wenting Guo
Foreword, Conclusion, and Others		Peng Cheng Laboratory	Li Sun and Ning Huang

During the planning and writing of the white paper, Prof. Xiaohu You, Academician of the Chinese Academy of Sciences and Chairman of the Expert Committee of AI Edge Alliance, and all members of the Expert Committee, including Tingting Yang, Jianjun Wu, Guangyi Liu, Yue Wang, Erni Zhu, Fuchang Li, Xiangyang Duan, Yan Li, Shuguang Cui, Ge Li, Zhaoyang Zhang, Yang Yang, Yiqing Zhou, Ting Zhou, Chengjun Sun, Shaohui Sun, Xiaotao Chang, Pei Xiao, Lin Cai, Sumei Sun, and Mérouane Debbah, provided many valuable opinions and suggestions on the content of the white paper. We would like to express our highest respect and heartfelt gratitude to the experts mentioned above for their hard work and careful guidance!