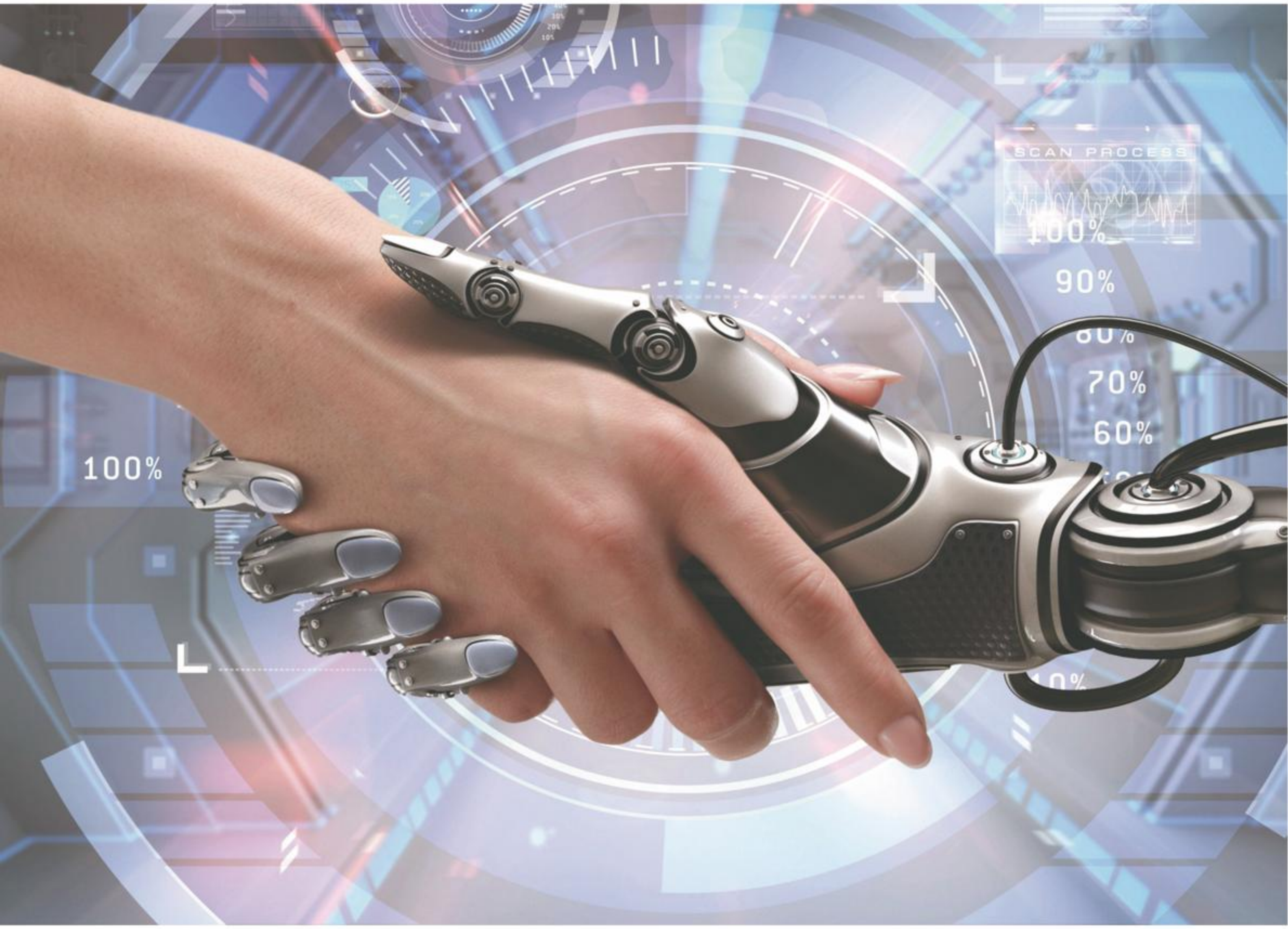
ZTE



ZTE AIR Core Technology System White Paper

TABLE OF CONTENTS

FIGURES

TABLES

1 Preface

In today's era of digitalization and intelligence transformation, significant changes are taking place in both the mobile network and Artificial Intelligence (AI) fields. In the mobile network domain, 5G has already commercially used, while he 6G system starts to formulate standards. The development of 5G applications is at a critical juncture where the focus is shifting from technology-driven to value-oriented. AI has become a key driver for service innovation and efficiency enhancement. According to the timelines set by the ITU and 3GPP, 6G is expected to be commercially deployed by 2030. As a new scenario first mentioned in the IMT-2030 Vision Proposal, the integration of communication and AI has prompted global mobile network researchers to consider what new characteristics networks should possess as future communication paradigms evolve from "Internet of Everything" to "Intelligent Internet of Everything".

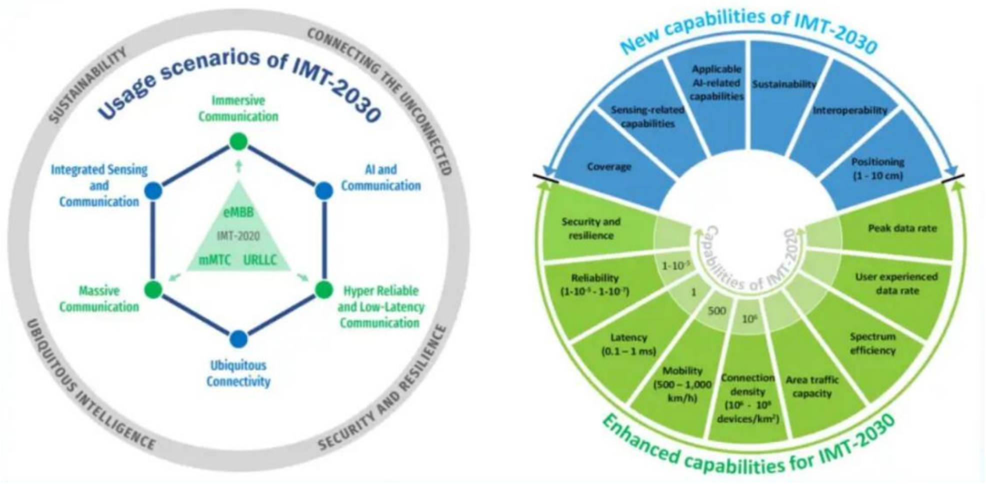

Usage scenarios of IMT-2030
SUSTAINABILITY
CONNECTING THE UNCONNECTED
UBIQUITOUS INTELLIGENCE
SECURITY AND RESILIENCE
Immersive Communication
Integrated Sensing and Communication
AI and Communication
eMBB
IMT-2020
mMTC
URLLC
Massive Communication
Hyper Reliable and Low-Latency Communication
Ubiquitous Connectivity
New capabilities of IMT-2030
Applicable AI-related capabilities
Sustainability
Sensing-related capabilities
Interoperability
Coverage
Positioning (1 - 10 cm)
Security and resilience
Peak data rate
Reliability (1-10⁻⁵ - 1-10⁻⁷)
User experienced data rate
Latency (0.1 – 1 ms)
Spectrum efficiency
Mobility (500 – 1,000 km/h)
Connection density (10⁶ - 10⁸ devices/km²)
Area traffic capacity
Capabilities of IMT-2020
1-10⁻⁵
1
500
10⁶
Enhanced capabilities for IMT-2030

Fig. 1-1 IMT-2030 application scenarios and key capabilities defined by ITU-R

Meanwhile, the AI field is undergoing profound transformations. Until now, AI has experienced three waves of enthusiasm and two troughs of disillusionment. In November 2022, OpenAI's release of ChatGPT, along with its adoption of Transformer algorithms and pre-trained large models for generative AI, pushed AI technology to unprecedented heights and marked a turning point for large model AI technologies, sparking a hype peak. Generative AI, with its ability to create new contents, mimic human creativity and innovation, plays a significant role across various sectors, driving prosperity and advancement in the AI domain. Scale creates miracles; larger models yield greater intelligence. As AI technology continues to evolve, industries across the board will be able to leverage AI more effectively to enhance operation efficiency and create business values, transitioning from "Digitalization" to "Digital Intelligence".

Therefore, the integration and mutual empowerment of mobile networks and AI will be the main theme of the industry in the coming years. On one hand, mobile networks can leverage AI to optimize networks, detect anomalies, improve energy efficiency and enhance security, driving the intelligence, innovation, and efficiency of mobile networks in multiple dimensions; on the other hand, an E2E efficient network can be built through mobile networks to support AI services, realizing full-scene training and inference among the cloud, edge, and terminals.

This White Paper analyzes the current challenges and opportunities in the development of core network, introduces AI technology to achieve an AI Core for service innovation and cost reduction and efficiency improvement, focuses on the role of new AI technologies in reshaping the core network architecture, and provides an outlook on the future evolution direction and challenges of core network.

2 Current Situation of AI Core

2.1 Embracing AI+, the Industry is Actively Promoting the Intelligence of Core Network

The current mobile network is facing multiple challenges. The first challenge is user churn, mainly manifested in the fact that OTT, with its flexible and rapid service innovation, accurate personalized services, and low costs, significantly diverts users from traditional mobile networks; the second challenge is the single and rigid nature of mobile network service applications, which cannot meet the needs of generalized and personalized application characteristics; the third challenge is the slowdown in mobile network traffic growth, with the revenue model based on traffic encountering bottlenecks, leading to slow or even declining growth in operator revenue; the fourth challenge is the insufficient autonomous decision-making capability of autonomous networks, mainly reflected in inadequate assessments of the effectiveness and impact of network optimization, the impact of changes, and the impact of major escape operations.

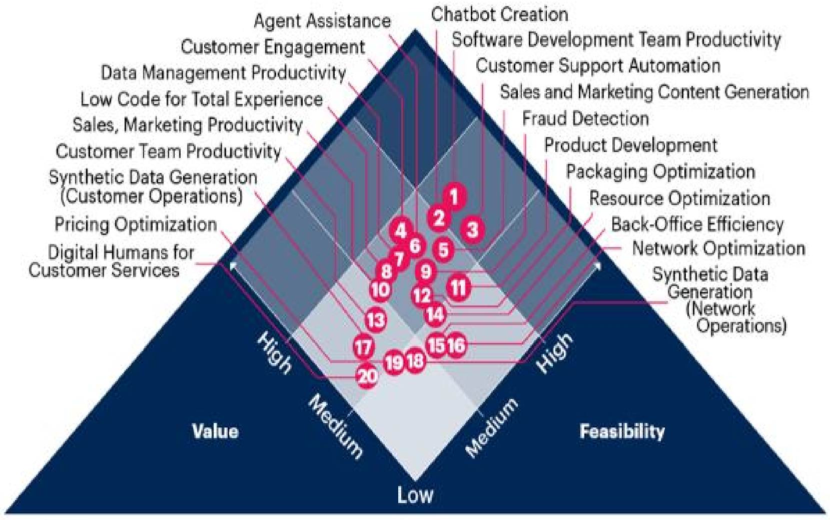
How to address the pain points of the current mobile network and accelerate the paradigm shift of the network by leveraging AI technology to build an AI Core is a topic that the industry is actively addressing, continuously promoting the development of intelligence in the core network.

Let's take a look at the macro policies. Currently, governments around the world are increasing their investment and policy support in the field of AI domain. For example, the European Commission has adopted the "Digital Europe Programme" work plan for 2024, which will provide funding support for digital solutions including AI and cybersecurity; China mentioned in the "Government Work Report" to deepen the R&D application of big data and AI, etc., carry out the "AI+" action, and create a digital industrial cluster with international competitiveness.

Let's take a closer look at the standards. Under the continuous advancement of 3GPP standards, core network intelligence is evolving from embedded intelligence to distributed and native intelligence. NWDAF, serving as an external embedded intelligence, has been continuously evolving from Release 15 to subsequent versions. In R15, NWDAF only supports network slice load analysis; in R16, a centralized architecture was defined, meeting basic data analysis requirements; R17 achieved a separation of training and inference architecture, defining the Analysis Logic Function (AnLF) and Model Training Logic Function (MTLF), and also established a hierarchical intelligent architecture supporting multi-NWDAF collaboration, introducing a data management framework to enhance data collection and analysis efficiency. In the future, NWDAF may further expand in terms of analytical capabilities and service scope to better adapt to growing service needs. For future 6G, AI will be integrated into the architecture, functional processes, and protocols from the design stage, providing native AI services, both internally and externally, as a basic service of 6G.

Looking at the operators, global operators are actively working on "AI+". Chinese operators are actively advancing AI+ strategies. Taking China Mobile as an example, it has realized a comprehensive AI strategy, moving from "5G+" to "AI+." It has already released 23 AI+ products and 20 AI+ DICT industry applications. The 5G New Calling has become

China Mobile"s AI+ strategic product. International operators are exploring the potential of AI through multi-party collaborations. In terms of industry cooperation, operators such as SK Telecom, Deutsche Telekom AG, Etisalat, and Singtel have signed a memorandum of understanding for AI cooperation, forming the Global Telco AI Alliance. The aim is to accelerate AI transformation in telecom services and develop new service growth points driven by AI models. In cooperation with cloud service providers, AT&T, in conjunction with Meta, Microsoft and OpenAI, is jointly advancing AI-based network optimization and fault location scenarios, and has completed commercial deployments. Telstra, in collaboration with Microsoft/OpenAI, is developing an AI-generated summary of telecom networks and a generative Q&A knowledge base based on Microsoft Azure Cloud and OpenAI, and has also completed commercial deployments. Based on operator practices with AI+, Gartner has summarized the following 20 major application scenarios, including experience-oriented ones such as metaverse service, avatars and efficiency-oriented ones like fault detection/network and resource optimization, as shown in the following figure.


Agent Assistance
Customer Engagement
Data Management Productivity
Low Code for Total Experience
Sales, Marketing Productivity
Customer Team Productivity
Synthetic Data Generation (Customer Operations)
Pricing Optimization
Digital Humans for Customer Services
Chatbot Creation
Software Development Team Productivity
Customer Support Automation
Sales and Marketing Content Generation
Fraud Detection
Product Development
Packaging Optimization
Resource Optimization
Back-Office Efficiency
Network Optimization
Synthetic Data Generation (Network Operations)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
High
Medium
High
Medium
Low
Value
Feasibility

Source: Gartner
797387_C

Fig. 2-1 Gartner "AI+" Telecom Network Application Scenarios

2.2 AI Core, Opportunities and Challenges Coexist

By introducing AI technology, unprecedented opportunities have been brought to the core network, such as helping the core network to generate new application scenarios and business models, enhancing network value, and also achieving self-awareness and self-optimization of the network, improving network reliability and user experience; however, it will also be accompanied by a series of challenges, including how to determine high-value application scenarios, how to obtain high-quality data, as well as data security and network security.

Opportunities:

- Service Upgrade: AI Core promotes the deep integration of AI and network technology, such as voice, messaging, etc. This integration can upgrade traditional basic SMS and voice call services to native intelligent, entry-point services with endogenous capabilities, for example, by introducing "AI+" New Calling to achieve video, data interaction, AIGC, intelligent agent AI entry, digital human and other new communication capabilities, making traditional voice services have the attributes of OTT social software and becoming an APP entry.
- Operational Transformation: AI Core can leverage technologies like AI large models and digital twins to establish a new paradigm of operations collaboration between Copilot and Agent, achieving intelligent O&M optimization for the network, enhancing the utilization of network resources, shortening the innovation cycle for new services, and boosting operational efficiency.
- Business Model Shift: By introducing AI technology, it not only elevates the network's intelligence and awareness capabilities, enabling differentiated 5G services, but also facilitates new service models, such as enhancing user experience through "AI+" Connectivity, and realizing a shift from traffic-based to experience-based service operations.

Challenges:

According to analysis from Gartner's report, introducing AI technology into telecommunication networks presents the following challenges, primarily manifested in the following areas:

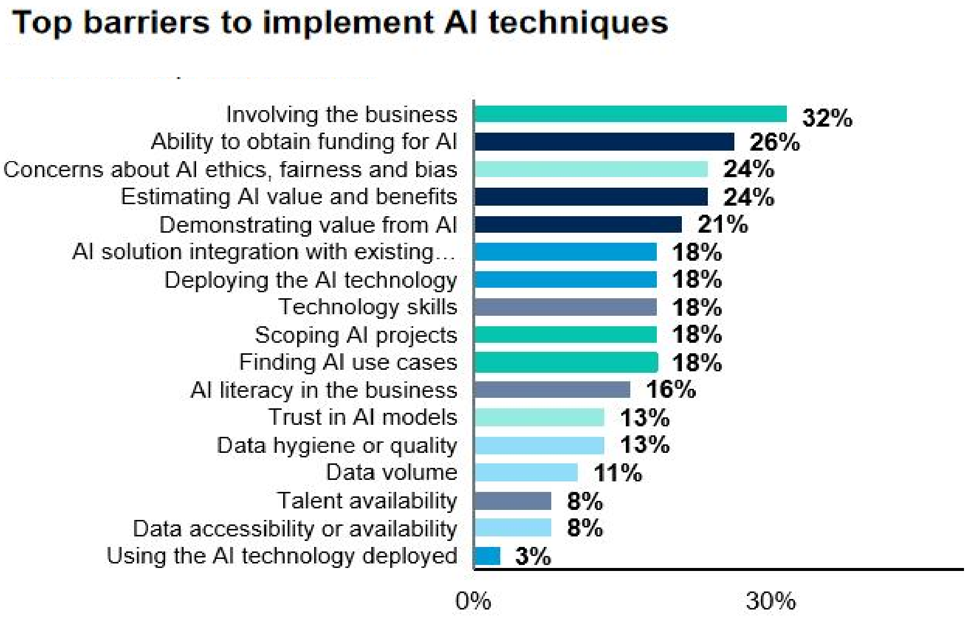


Fig. 2-2 Gartner's Major Challenges in Achieving "AI+" Telecommunication Networks

- Scenarios and Business Closed Loop: While there are numerous scenarios where AI technology can be applied, determining high-value scenarios and achieving business closure presents a significant challenge, for instance, the varying value delivered by differentiated scenes like network optimization and experience enhancement.

- Uncertainty of Large Models: AI large models lack explainability and controllability. How to ensure that the capabilities of AI large models introduced into the network are certain and controllable. Corresponding technical means must be in place to realize high-value scenarios that are implementable.

- High-Quality Data: The enhancement of large models' capabilities depends on high-quality data, but often there is a lack of such data, primarily constrained by data privacy protection, high data acquisition costs, data scarcity, and the data's own quality and quantity not meeting requirements.

- Data Privacy and Security: AI Core requires processing large amounts of user data. Ensuring the security and privacy of this data has become a major challenge. Effective technical measures and management practices must be adopted to prevent data leakage and misuse.

- Talent Shortage and Skill Enhancement: The rapid development of AI Core technology poses higher demands on talent. However, the current talent pool in related fields is relatively insufficient, and skill levels are uneven. Therefore, efforts to strengthen talent cultivation and skill enhancement are needed to meet the demands of technological development.

In summary, while AI Core brings opportunities, it also comes with a series of challenges. To fully tap its potential and address these challenges, concerted efforts are needed from all sectors to strengthen technological innovation, improve regulations and policies, cultivate specialized talents, and promote the formulation and implementation of relevant standards and norms.

3 AI Core Technology System Architecture and Evolution Path

3.1 AI Core Technology System Architecture

Considering the development trends of AI technology and the analysis of opportunities and challenges in introducing AI technology into the core network, we know that traditional methods of scale expansion are gradually slowing down and becoming ineffective. The core network needs to be systematically restructured in four major areas - Services, Connections, O&M and Cloud Infrastructure - to incubate more service scenarios, enhance network value, improve network performance and reduce the difficulty and costs of O&M. To this end, ZTE has combined its own transformation strategy and front-line customer needs to propose the AI Core technology system architecture.

The AI Core technology system is composed of "three layers and one domain" (as shown in the figure below), which are the AI+ Cloud Infrastructure layer, the "AI+" Connectivity layer, the "AI+" Service application layer, and the "AI+" O&M domain. Among them, the "AI+" Cloud Infrastructure provides an innovative AI computing foundation for core network AI applications, including computing, storage, network and other hardware resources, as well as resource management and scheduling. It provides a variety of instances such as bare metal, virtual machines and containers, as well as the ability for fine-grained resource pooling. On top of this, a training and inference platform is built to provide AI empowerment capabilities for applications; "AI+" Connectivity provides intelligence for the 5GC control plane and user plane, stimulating traffic, enhancing experience, and mining data value; "AI+" Services provide intelligent capabilities for Messaging and New Calling, bringing interactive, intelligent and immersive new experiences to users, and offering unlimited communication possibilities. "AI+" O&M reshapes the operations paradigm, supporting the evolution of autonomous and intelligent networks toward unmanned operation.

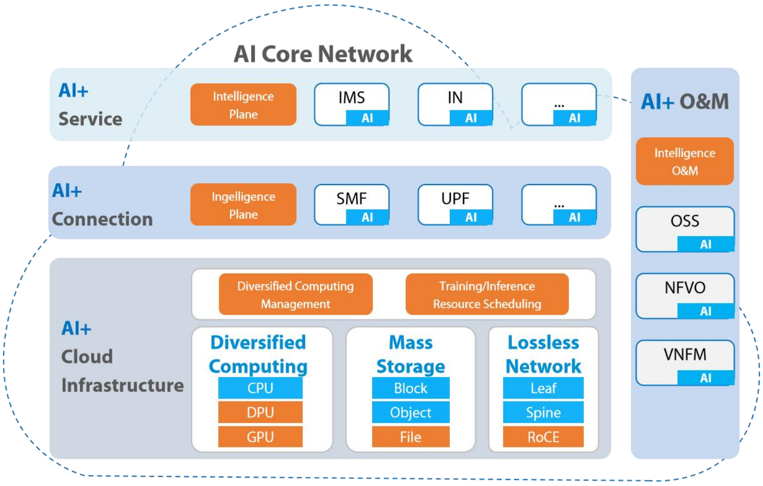
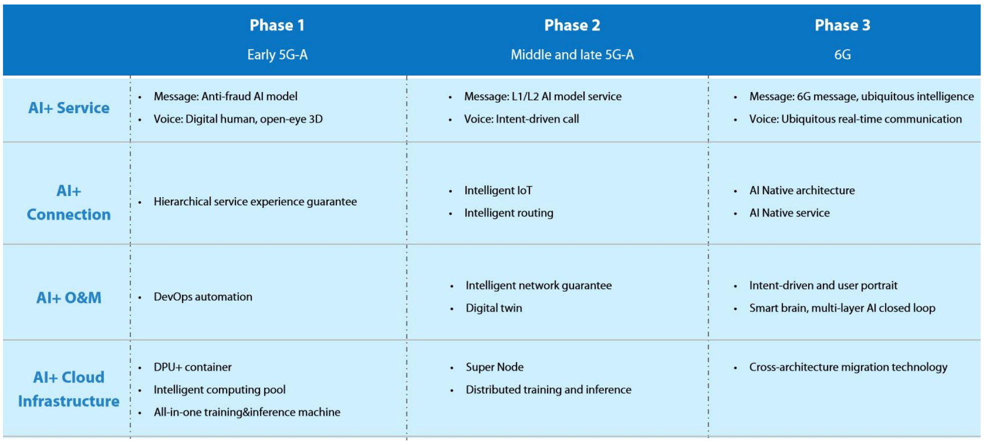


Fig. 3-1 “AI+” Core Network Technology System Architecture

3.2 Evolution Path of AI Core

AI technology is still evolving, mobile networks are undergoing continuous transformation, and network construction also takes a period. To unleash the full potential of AI technology while aligning with the pace of network development, we believe that the evolution of AI Core can be divided into three phases, as illustrated in the following figure:



	Phase 1 Early 5G-A	Phase 2 Middle and late 5G-A	Phase 3 6G
AI+ Service	• Message: Anti-fraud AI model • Voice: Digital human, open-eye 3D	• Message: L1/L2 AI model service • Voice: Intent-driven call	• Message: 6G message, ubiquitous intelligence • Voice: Ubiquitous real-time communication
AI+ Connection	• Hierarchical service experience guarantee	• Intelligent IoT • Intelligent routing	• AI Native architecture • AI Native service
AI+ O&M	• DevOps automation	• Intelligent network guarantee • Digital twin	• Intent-driven and user portrait • Smart brain, multi-layer AI closed loop
AI+ Cloud Infrastructure	• DPU+ container • Intelligent computing pool • All-in-one training&inference machine	• Super Node • Distributed training and inference	• Cross-architecture migration technology

Fig. 3-2 “AI+” Core Network Technology Roadmap

Phase One: “Enlightened Intelligence”. The most prominent feature of this era is that AI is transitioning from a "spark" to a "wildfire". Integrating AI large models deeply into networks, different operators and different service lines have varying needs. This involves numerous aspects of an operator's network planning, construction, maintenance, operation and IT. Each service scenario within these aspects has different requirements for model accuracy. Identifying suitable technical application scenarios becomes crucial, requiring joint efforts from equipment providers in the industry and pioneering customers willing to explore innovation.

Phase Two: "Enhanced Intelligence". As AI large models advance to a scale of over ten trillion parameters, their capabilities are further upgraded, potentially integrating into every aspect of industrial production. Meanwhile, after the exploratory phase of "Enlightened Intelligence", the core network has accumulated a comprehensive set of experiences and methods. Leveraging the capabilities of AI large models, it can uncover more AI scenarios for 5G-Advanced networks, further monetizing them and solidifying the foundation for 6G core networks with endogenous intelligence.

Phase Three: "Native Intelligence". This phase targets the endogenous intelligence of 6G, aiming to construct a core network with endogenous intelligence centered around AI large models and AI agents. During this phase, the AI Core will natively support integrated communication and AI services, achieving the capabilities of autonomous environmental perception, autonomous task generation, and autonomous task execution. This empowers rich and innovative services and supports the efficient and sustainable development of society.

We believe that the AI Core is currently in the first phase, with some domestic and international operators building the AI Core; looking to the medium and long term, we should focus on research and development of key technologies for "Enhanced Intelligence and Native Intelligence", quickly reaching industry consensus to accelerate the maturity of core technologies and the industry.

4 “AI+” Services: Unleashing Infinite Possibilities

4.1 “AI+” Messaging, Building Trusted Multi-Dimensional Intelligent Services

4.1.1 The Prominent Risk of Cyber Fraud

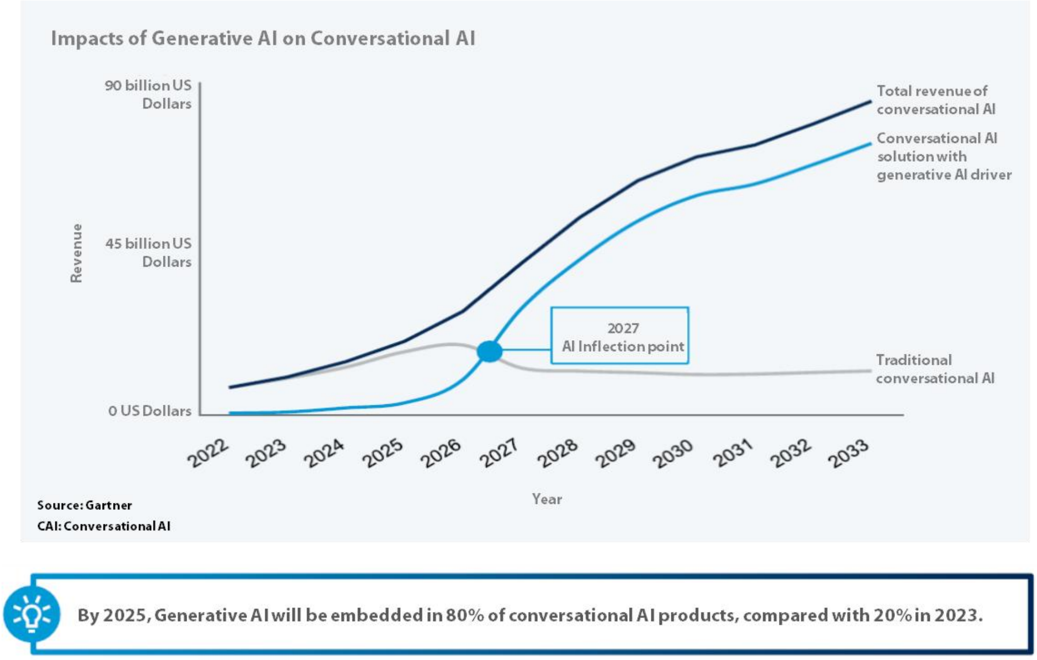
According to the Global Anti-Scam Alliance's (GASA) 2023 annual report, cyber fraud inflicted approximately $1.02 trillion in losses on global consumers, equivalent to 1.05% of global GDP. Secondly, based on the Global Economic Crime Survey (GECS), 51% of surveyed organizations experienced fraud in the past two years, the highest level in nearly 20 years. Telecommunications network services, including text messages and voice/video calls, are the most widely used communication methods globally, with approximately 9.3 billion users, making them prime targets for cyber fraud.

Fraudsters, through frequent changing in fraudulent content and various technical means, have made fraud increasingly complex and sophisticated, with new disguises emerging constantly. Telecommunication network fraud has become more rampant, causing not only huge economic losses to the public but also having a serious negative impact on society. Traditional governance systems, primarily based on static rules, are unable to dynamically adjust and identify new changes, resulting in low identification capabilities and efficiency. The cost of manual labor has significantly increased, and the service launch cycle is long, far from meeting the governance requirements of telecommunication networks.

4.1.2 AI+ Messaging Brings Technological and Value Upgrades

In response to the current situation and pain points of telecommunication network fraud governance, the application of the AI+ Messaging multimodal anti-fraud large model initiates a technological innovation and upgrade from traditional governance to AI governance. This ensures the trustworthiness of new communications, achieving efficiency, accuracy, and comprehensiveness.

According to market forecasts by Gartner, a global authoritative consulting firm, conversational AI, as an emerging technology, is facing broad prospects. Ideally, as shown in the following figure:


Impacts of Generative AI on Conversational AI

Source: Gartner
CAI: Conversational AI

By 2025, Generative AI will be embedded in 80% of conversational AI products, compared with 20% in 2023.

Fig. 4-1 Development Trend of Generative AI

- By 2033, revenue from conversational AI platforms will grow to $84.8 billion, driven by generative AI (pre-generated models), up from $6.9 billion in 2022, with a compound annual growth rate of 25.5%.
- By 2027, the revenue growth driven by generative AI in conversational interaction solutions will surpass that generated by traditional AI solutions. Technology vendors who do not take action before the tipping point arrives risk losing market share.
- By 2025, generative AI will be embedded in 80% of conversational AI products, compared to just 20% in 2023.

5G messaging is a crucial entry point for conversational AI. For ToC, the trend is to create high-frequency, essential entry points, where everyone has their own dedicated AI entity, with a focus on experience-first services that have significant long-term service potential. For ToB, Chatbots (service accounts) are the killer apps for 5G messaging, with a total of over 1.3 million in China, which will reach millions to tens of millions within three years. Chatbots across all industries have a strong demand for AI+ capability integration and service provision, with a high willingness to pay and quick monetization, making them economically valuable with great development potential. These factors present significant opportunities for AI+ messaging.

4.1.3 Three Layers and Two Planes: the "AI+" Messaging Creates Three Service Entries

By integrating the intelligence plane with the messaging plane, the network-side messaging platform seamlessly forms a three-layer, two-plane "AI+" messaging architecture, as illustrated in the following figure.

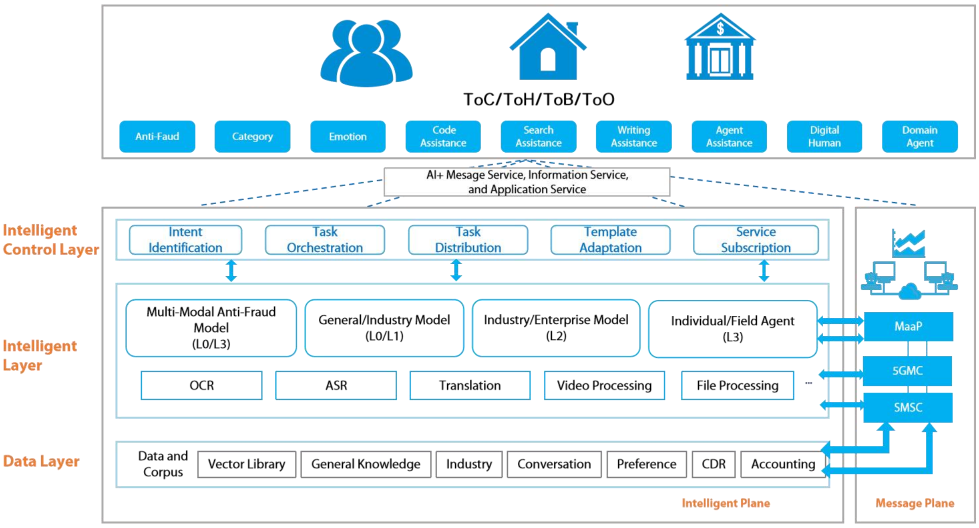


Fig. 4-2 Three Layers and Two Planes Architecture

The three-layer and two-plane architecture, with the intelligence control layer's intent recognition and capability orchestration as its core brain, aggregates and schedules capabilities across all layers. It builds an AI+ messaging service entry point, catering to ToC, ToB, ToH and others (ToO), to create three categories of comprehensive services: "AI+ messaging services, AI+ information services, AI+ application services".

- Three Layers

1. Data Layer: Based on daily operations of networks, platforms, or billing systems, it accumulates raw call records, data and linguistic resources. After data processing and cleaning, it provides core and long-term linguistic resources for the intelligence layer.
2. Intelligence Layer: Comprising large and small models and algorithms, it uses linguistic resources for training, refinement or continuous learning. It generates various AI large model applications that are tightly integrated with the messaging planes to complete service processes.
3. Intelligence Control Layer: Serving as the front-end smart brain, it is capable of recognizing target user intents, task orchestration, task distribution, template adaptation, and service ordering. It is the core control and direct supply layer for AI+ messaging comprehensive service capabilities.

- Two Planes

1. Message Plane: Provision of mobile communication messaging services and capabilities, offering services such as SMS, MMS, 5G messaging, and industry-specific messaging.
2. Intelligence Plane: As a newly added capability plane, it includes the intelligent control layer, intelligence layer and data layer. Integrated with the message plane, it collaborates to complete AI+ capability upgrades, providing end-users with AI+ messaging services, information services, and application services.

4.1.4 Multi-Modal Anti-Fraud Large Model, Safeguarding Credible Message Communication

As one of the most common channels for telecom fraud, scam messages constantly evolve and upgrade in content and form to penetrate the monitoring systems of telecom operators. They circumvent keyword rules through various means such as text mutation, escape characters, homophones, and visual similarities in content. In terms of number monitoring strategies, vast pools of numbers are used to avoid traffic and keyword thresholds. In sending methods, they often employ testing techniques to breach defenses at a single point, then flood the system with massive volumes. In media formats, more and more messages use images, audio, and video to replace text. Traditional anti-fraud management solutions suffer from poor accuracy, low efficiency, long response and handling cycles, difficulties in recognizing media content, and limitations in handling single languages.

AI+ multimodal large model for anti-fraud governance, as shown in the figure below, based on AI large language models, CV (Computer Vision) large models, and multimodal large models, capable of recognizing and processing media types including text, images, graphics, audio, video and documents, with powerful natural language semantic analysis, emotion recognition, logical reasoning, and greatly improved accuracy of conclusion ability. Multilingual capabilities can identify mainstream international languages and regional dialects, such as Chinese, English, French, Cantonese, Hokkien, Burmese, etc.

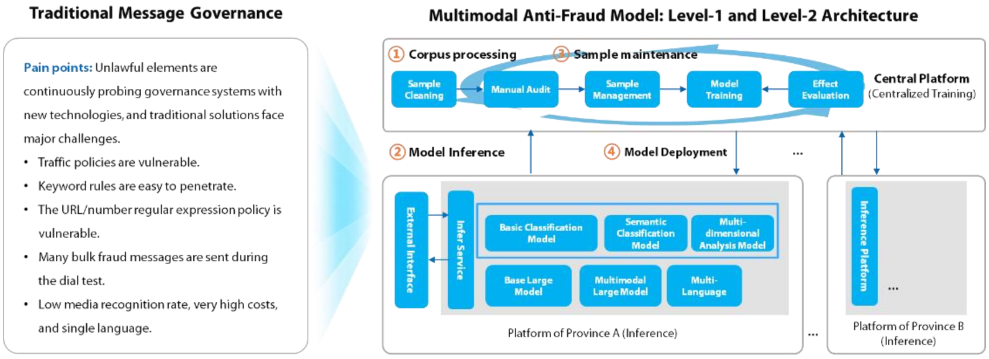

Traditional Message Governance
Pain points: Unlawful elements are continuously probing governance systems with new technologies, and traditional solutions face major challenges.
• Traffic policies are vulnerable.
• Keyword rules are easy to penetrate.
• The URL/number regular expression policy is vulnerable.
• Many bulk fraud messages are sent during the dial test.
• Low media recognition rate, very high costs, and single language.
Multimodal Anti-Fraud Model: Level-1 and Level-2 Architecture
① Corpus processing
③ Sample maintenance
Sample Cleaning
Manual Audit
Sample Management
Model Training
Effect Evaluation
Central Platform
(Centralized Training)
② Model Inference
④ Model Deployment
...
External Interface
Infer Service
Basic Classification Model
Semantic Classification Model
Multi-dimensional Analysis Model
Base Large Model
Multimodal Large Model
Multi-Language
Platform of Province A (Inference)
...
Inference Platform
...
Platform of Province B (Inference)

Fig. 4-3 Multimodal Large Model Architecture

The key technologies and features of multimodal large model anti-fraud are as follows:

1. Model Training: B+Z

B represents the base pre-trained large model, and Z stands for the anti-fraud message monitoring adaptation layer. The pre-trained large model B typically has a scale of tens to hundreds of billions of parameters, with strong semantic analysis, emotion recognition, logical reasoning, and conclusion abilities.

On this basis, the "anti-fraud-specific prompt words + anti-fraud sample refinement" model training method is superimposed, achieving high levels of fraud recognition accuracy and recall rate when dealing with fraud messages under hundreds of thousands of training samples.

2. Hallucination Elimination

Based on the "Prompt Role Guidance and Output Regulation" technology for identifying fraudulent messages, the common hallucination problems of large models are eliminated, enhancing the stability of anti-fraud identification by large models.

The prompt methodology framework is CRISPE, which consists of five parts: Capacity and Role, Insight, Statement, Personality, and Experiment. By constructing anti-fraud identification prompts according to the CRISPE methodology, they can accurately and reliably perform specific tasks, eliminating hallucination problems.

3. MLOps

Based on MLOps (Machine Learning Operations), a workflow and automated processes are established to improve the automation and operational capabilities of large models, achieving full tooling and task workflow automation. Sample data is automatically imported, automatically trained and fine-tuned, automatically deployed, and automatically inferred, realizing a closed-loop iteration of the service flow without human intervention, maintaining continuous updates of system capabilities, and significantly reducing maintenance costs and OPEX.

4. Inference Architecture

Deployment modes can be categorized into two architectures: training and inference integration and distributed inference.

(1) In the training and inference integration architecture, model deployment, training, refinement, and inference are all conducted within a single site. Under conditions of saving computational power and servers, training and inference can share computational resources.

(2) In the distributed inference architecture, it can be deployed as a two-level structure of centralized training and distributed inference. That is, there is no need to deploy training and refinement modules and conduct training and refinement at each distributed node. Instead, model training and refinement are concentrated at a central node, and the completed model is then distributed to each node for direct deployment and inference completion. This can significantly reduce the resources, time, and costs associated with repeated training and refinement.

AI combined with a multimodal large model for fraud prevention can achieve high-level accuracy in text and multimedia recognition, far exceeding traditional solutions. It significantly reduces the cost of manual review, decreases false positives to increase revenue, and reduces customer complaints, bringing comprehensive economic benefits.

In terms of social value, the application of multimodal large models significantly enhances the user's communication experience, creating a trustworthy and worry-free communication service. This maintains the communication brand's image and user satisfaction, encouraging more secure use of mobile communication services. It maximizes the reduction of property loss among the general public while making a significant contribution to social stability through the power of technology.

4.1.5 Endogenous Multi-Dimensional Agent, Unlocking Ultimate Smart Living

As a killer app for 5G messaging, Chatbots primarily interact through conversation. Currently, over 99% of Chatbots do not support AI conversation capabilities. Beyond pre-configured menus or fixed keywords, they are unable to recognize any user input or requests, leading to a severely inadequate experience.

If a Chatbot is to upgrade itself to support AI-intelligent conversation, there are two methods: Method one involves self-development, which has a high threshold, a long cycle, and very high costs. Method two requires calling third-party large model interfaces and utilizing their open capabilities. This necessitates custom development of API interfaces, joint debugging, and integration. API fees apply, and there are certain research and development costs as well as long-term usage costs. The QoS quality depends on the public network and cannot be guaranteed.

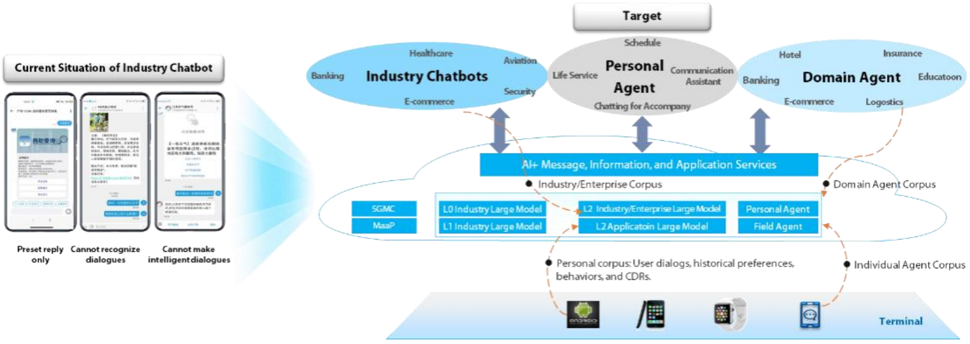


Fig. 4-4 Target Architecture for Multidimensional Intelligent Entities

Network-native multi-dimensional intelligent agents, as shown in the above figure, the Chatbot and domain-specific intelligent agent services for ToB have the following characteristics: seamless activation, no need for any development or modification; instant activation, can be activated in batches; QoS assurance, high security, data does not leave the network; lowest cost. For ToC, based on network, platform corpus, and data, long-term accumulation and learning are provided to offer the exclusive personal intelligent agent with most understanding of users.

Based on AI+ messaging network-native multi-dimensional intelligent agents, there are the following key technologies and features:

1. AI conversation for various types of messages

Supports multi-dimensional intelligent agent capabilities for conversation with SMS, MMS, and 5G messages.

2. Bidirectional multimodal interaction

A multimodal large model combines the understanding and generation capabilities of large language models with other modalities (such as images, audio, video, etc.) to integrate various types of inputs and outputs such as text, images, sounds, and videos, providing a richer and more natural interaction experience.

3. Internet search generation

In agent conversations, questions often arise regarding data with real-time characteristics, such as weather or stock prices. In such scenarios, the agent needs to be capable of connecting to the internet to search for and obtain the most recent, updated results or the latest officially refreshed information. The search results are then fed as input into a multimodal large model, supplemented with prompt words, to generate the final outcome.

On the other hand, the deep integration of generative AI with search is driving search engines from a "retrieval" to a "retrieval + generation" upgrade. Generative search achieves enhancements in three aspects: intelligent integration and organization of information, content creation, and personalized content experiences.

4. Document Summarization and Synthesis

With multimodal large model technology, document content can be automatically analyzed and distilled to generate concise and clear summaries or synthesis reports. This technology can significantly improve the efficiency and quality of document processing, enabling users to quickly grasp the core content of a document and save time and effort. Examples include key information extraction from documents, abstract generation, trend analysis, and data visualization.

5. Intent Recognition

The goal of intent recognition is to categorize users' natural language inputs into specific tasks or actions, providing the system with accurate contextual understanding. This significantly enhances the efficiency of problem-solving, reduces ineffective communication between users and systems, and requires support for understanding multiple intents from users as well as managing multi-turn conversations. Intent recognition serves as the entry point and foundation for an agent's capabilities; precise intent recognition is essential for effectively executing tasks for single and multi-agent systems.

The reconstruction and transformation of AI+ messaging accelerates the construction of a comprehensive intelligent agent service entry based on native device access points, direct dialing numbers, and secure trustworthiness. In terms of business models, AI+ messaging achieves a business model upgrade, moving from the traditional model where charges are primarily based on "per message" and package fees to an enhanced model that includes differentiated AI service fees, creating new growth points for next-generation messaging communications.

4.1.6 The Future of AI+ Messaging, Intelligently Driving the World

The mobile communication messaging service has evolved from the SMS era of past 2G/3G/4G networks, through the current 5G messaging era, and is heading towards the AI+ messaging era in future 6G networks. The goal and direction of AI+ messaging is to drive the world with intelligent messages, with the core development philosophy and ideas of "Message communicating everything, Messaging as an Intelligent Service for everything, and Messaging as AI driving everything". See the following figure:

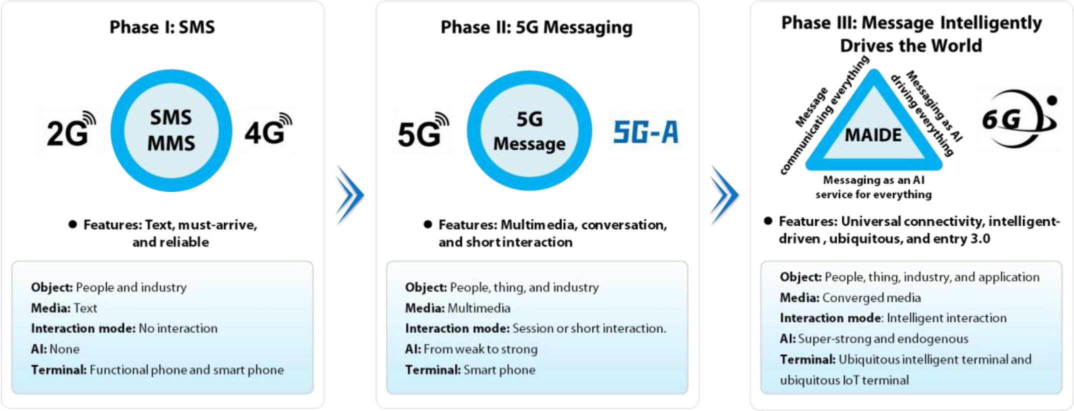

Phase I: SMS
2G
SMS
MMS
4G
● Features: Text, must-arrive, and reliable
Object: People and industry
Media: Text
Interaction mode: No interaction
AI: None
Terminal: Functional phone and smart phone
Phase II: 5G Messaging
5G
5G Message
5G-A
● Features: Multimedia, conversation, and short interaction
Object: People, thing, and industry
Media: Multimedia
Interaction mode: Session or short interaction.
AI: From weak to strong
Terminal: Smart phone
Phase III: Message Intelligently Drives the World
Message communicating everything
Messaging as AI driving everything
MAIDE
Messaging as an AI service for everything
6G
● Features: Universal connectivity, intelligent-driven , ubiquitous, and entry 3.0
Object: People, thing, industry, and application
Media: Converged media
Interaction mode: Intelligent interaction
AI: Super-strong and endogenous
Terminal: Ubiquitous intelligent terminal and ubiquitous IoT terminal

Fig. 4-5 Future Development of AI+ Messaging

Intent-driven and intelligent connection of all things are important directions for the future development of networks and AI+. The essence of intent-driven is intelligent driving based on message flow, including intent recognition, decomposition, execution and feedback. Therefore, the intelligent driving of the world and all things by messages is an inevitable trend in the future.

1. Message Communicating Everything: Future message communication will be directed towards the delivery of messages to "people, all things, applications, and industries", ensuring that messages can be communicated, understood, and executed. Messaging as a Service facilitates communication between people and all things.

2. Messaging as an AI Service for Everything: Messages will be fully integrated with AI, with AI+ messaging applications everywhere, Messaging as an AI, and Messaging as an AI Service.

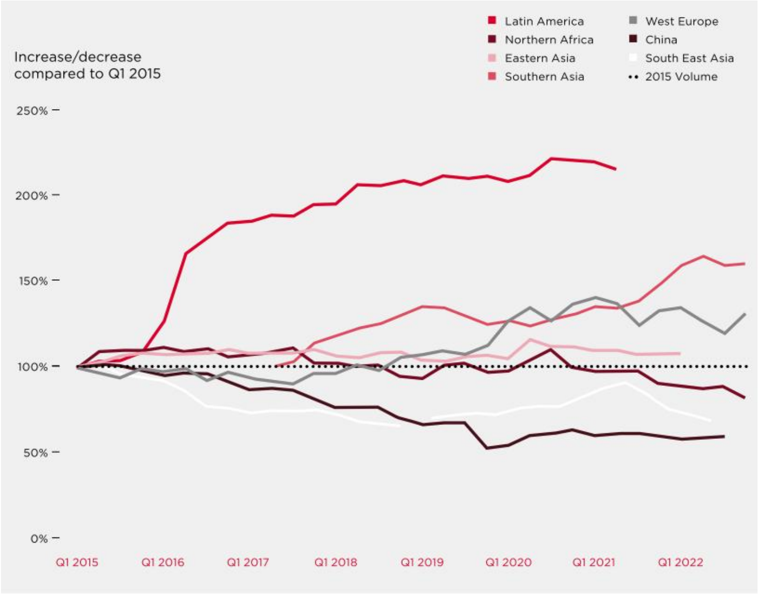
3. Messaging as AI Driving Everything: Based on highly accurate intent recognition, utilizing the unified native messaging entry of terminals and multidimensional AI capabilities, it completes various message commands, entity transactions, and life tasks, where messages are actions.

In the Message Intelligence Drive phase, characterized primarily by ubiquitous messages, ubiquitous intelligence, and ubiquitous intelligence-driven, it constructs Communication Entry 3.0, supporting ubiquitous message intelligence to drive all services and applications.

4.2 "AI+" New Calling: Reconstructing the New Value of Voice Networks

4.2.1 Voice Revenue Continues to Decline, Calling for Real-time Communication Reform

With the proliferation of 4G and 5G technologies, the widespread adoption of intelligent terminals, and the escalating demands of users for enhanced communication experiences, OTT (Over-The-Top) services, renowned for their precise market positioning, rapid innovation, and rich content, have witnessed unprecedented growth. This has significantly diverted traditional communication users, resulting in a gradual decline in operators' traditional voice service volumes year by year. This phenomenon is even more pronounced in regions with developed communication systems.



Source: *GSMA Intelligence*

Fig. 4-6 Percentage Changes in Voice Usage for Selected Operators in Different Regions

With the advent of the AI era, AI technology has experienced explosive growth, reshaping numerous industries. AI smartphones and OTT services, leveraging their respective advantages, have eagerly entered the competition for becoming AI entry points, rapidly shifting their business models towards intelligent, personalized, and synthetic virtual-real experiences. Traditional communication operators face multiple challenges, including business reshaping, transformation of operating models, and a larger-scale diversion of users.

The emergence of New Calling services provides global operators with a significant opportunity to transition from user-centric operations to experience-driven business models. Integrating AI with real-time communications, New Calling has given rise to a variety of intelligent and personalized applications such as lighting up the screen, star calling, business collaboration, digital human operators, AI answering, AI chatting, AI-generated digital human, and intelligent agent AI entry, transforming the traditionally monotonous image of operators' applications.

According to the GSMA "The Mobile Economy 2024" report, it is predicted that by 2030, the global average adoption rate of 5G connections will reach 56%, with this figure exceeding 80% in regions with advanced telecommunications. See the following figure; additionally, with the development of technology and the reduction in costs, AI glasses, AR glasses, and other devices are expected to become widely applicable and more user-friendly by 2027, bringing more opportunities for the sustainable development of New Calling.

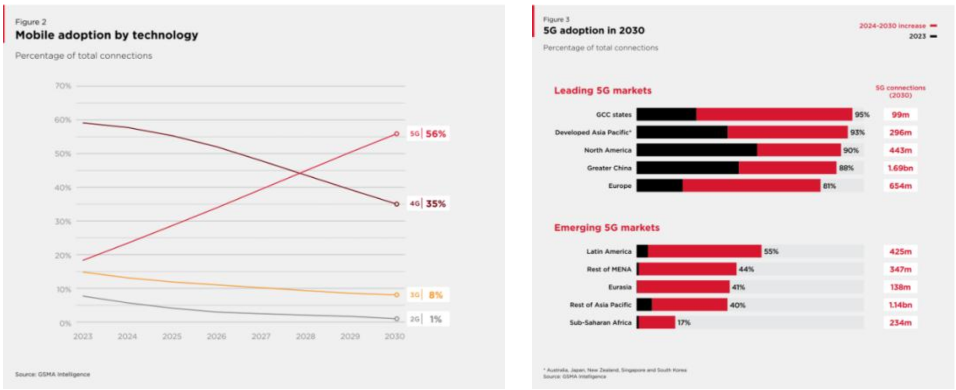


Fig. 4-7 Global 5G Connection Proportion Trend (Image Source: GSMA-The Mobile Economy 2024)

New Calling have a broad development prospect and are currently in the market cultivation stage. While mobile operators are facing significant development opportunities, they will also encounter the following difficulties and challenges in developing New Calling:

- Operators' current innovation capabilities in the new calling ecosystem are low, and they struggle to compete with AI terminals and OTT

New Calling competes with AI terminals and OTT for AI communication entries, forming a tripartite situation. New Calling needs to continuously introduce diverse intelligent and personalized new applications to meet the needs of upgrading user experience. Operators' voice communication ecosystem focuses on network innovation, and their application innovation capabilities are significantly inferior to those of OTT and AI terminals.

- The VoNR network, with its closed architecture and complexity, coupled with a long TTM, cannot deliver the latest AI experiences to users immediately

The VoNR network, known for its numerous network elements, interfaces, and complex processes, operates in a closed environment. Launching applications requires passing through intricate FOA procedures, which prevent dynamic loading and orchestration of new capabilities for quick creation of new applications. It does not support customization for personal applications or loading of private digital assets such as private digital human, belonging to individuals or enterprises. This results in New Calling services built on the VoNR network being unable to launch the latest AI applications right away.

- New Calling services, in the early stages of development, have yet to establish a profit model, facing challenges in achieving a sustainable business cycle

Scenarios for New Calling services, featuring visualizations, interactivity, AI enhancements, and personalization, are currently in experimentation, and the user base is being cultivated. The profit model requires further exploration. Similarly, AI terminal and OTT vendors are also experimenting with different monetization models for AI features.

- The introduction of AI+, visual, interactive, and personalized New Calling services poses new challenges in terms of security, trustworthiness, and privacy protection

The cloning of voices, actions, and images by AI, as well as features such as page sharing and whiteboard data interaction capabilities of H5 applet on Data Channels, and visual New Calling options, have brought communication convenience to users while also introducing significant security risks.

- The widespread adoption of New Calling terminals and the retraining of user habits require a process, which constrains the promotion of New Calling

Interactive New Calling and multimodal AI entries require terminal to support for Data

Channel (DC) communication capabilities. While some brands' terminals already support DC, the transition from market entry to widespread adoption of DC terminals is still ongoing.

Intelligent, personalized, and visually interactive New Calling enriches the experience while also altering habits formed during voice communications, requiring time for users to adjust.

4.2.2 The "Four New and One Universal" Solution Aims to Create "AI+" Real-Time Communications

In 2023, the launch of the OpenAI GPT large language model ushered in the era of AI, marking the AI Year; in the same year, China Mobile officially commercialized the world's first New Calling network. In the face of the explosive development of AI and the many difficulties and challenges encountered in the development of New Calling, ZTE, as a leading provider of New Calling solutions, pioneered the AI+ New Calling based on an open ecosystem. The AI+ New Calling adopts the "Four New and One Universal" solution to create "AI+" real-time communication. The new solution features "New Architecture, New Ecosystem, New Experience, New Security" and "Universal" characteristics, balancing stability and innovation, providing intelligent, personalized diverse experiences, and has excellent terminal universality, compatibility, and network interoperability. It supports network capability exposure, credible security, and endogenous intelligence.

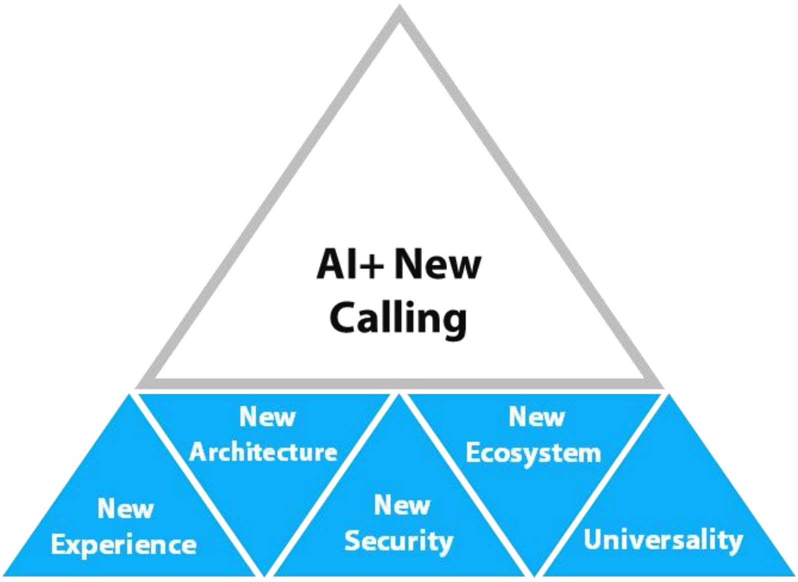


Fig. 4-8 “Four New and One Universal” AI+ New Calling Solution

4.2.3 Bidirectional Open New Architecture, Agilely Updating AI Capabilities

The complexity of the VoNR network results in slow innovation and long TTM of launching new services. The numerous network elements, interfaces, and complex processes in VoNR, as well as the tight coupling between services and media, pose significant challenges for innovation in New Calling applications. Developing a New Calling application not only requires designing and developing service logic but also involves complex process and interface adaptation and testing, leading to a lengthy development cycle. The tight coupling between services and media is characterized by numerous messages and complex processes and service logic. These issues are not apparent in voice communications with single service features but become severe in New Calling with generalized service features. Additionally, VoNR is a closed network that does not support dynamic loading of new capabilities. Adding new capabilities requires upgrading the system version, and the FOA process is complex and time-consuming. These problems not only slow down the innovation speed of AI-powered New Calling but also extend the time required for new features to go live. Therefore, even if mobile operators possess innovative capabilities comparable to those of OTT and AI terminal vendors, they may not be able to provide users with the latest AI service experience in a timely manner.

The separate deployment of multiple network elements in VoNR increases media latency and reduces network reliability. The separation of the media plane in VoNR from the media gateway for New Calling, as well as the separate deployment of the New Calling media gateway and AI server, results in the media stream of New Calling needing to be transferred, copied, and encoded/decoded multiple times between multiple devices. This not only increases latency and east-west bandwidth usage but also hinders the coordination of multi-modal media multi-streams. Additionally, it increases the complexity of networking, which is not conducive to network stability.

Moreover, the closed architecture of the VoNR network hinders the development of New Calling services. Operators need to open up the capabilities of the New Calling network to facilitate the development of visual, interactive, and intelligent New Calling applications within the industry, thereby helping enterprises improve customer service quality and efficiency.

AI+ New Calling adopts a "four layers and five planes" architectural design. The new architecture is simple, open, intelligent, and self-governing, supporting personalized customization, achieving "zero" wait and "zero" risk rapid innovation, empowering all walks of life.

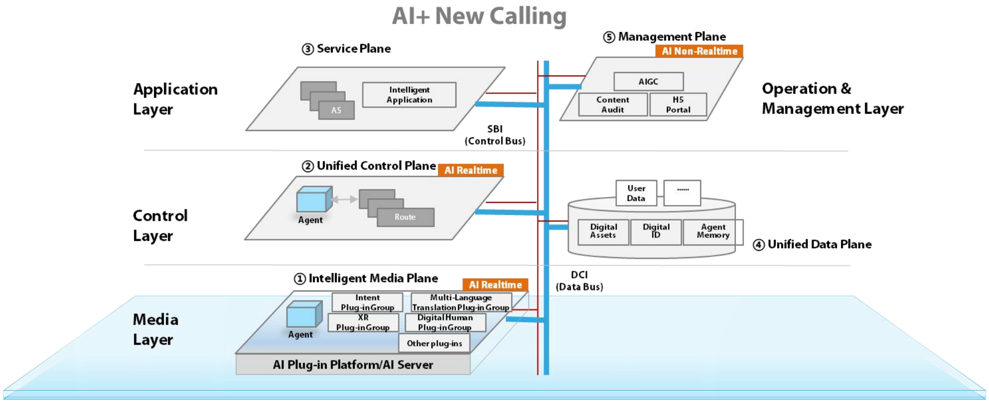


Fig. 4-9 "Four Layers and Five Planes" AI+ New Calling Architecture

- Intelligent Media Plane: Integrates media and AI, supporting the servitization and pluginization of media capabilities and AI capabilities.
- Unified Control Plane: Integrates CSCF/SSS and capability exposure platform, supporting basic session, routing and other service-oriented functions.
- Service Plane: Smartly implements service logic.
- Unified Data Plane: Integrates user data, digital assets, agent long and short-term memory, authentication, digital identity management, and etc.
- Management Plane: Responsible for uploading corporate and personal digital assets, content generation, content auditing, plugin and H5 applet management, etc.

4.2.3.1 Bidirectional Opening

The new architecture supports bidirectional opening of AI+ New Calling networks. The plug-in architecture is used to open the new architecture to the AI industry, open the traditional closed-network, and enable the third-party AI to empower New Calling networks. The service-based architecture opens network capabilities and AI capabilities to developers and applications, facilitating the AI+ New Calling network to provide New Calling services, media, and AI services to empower for the industry.

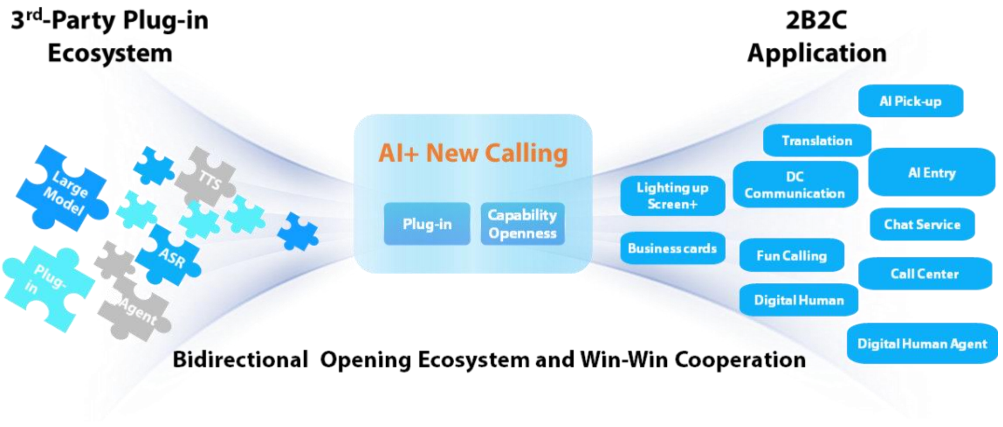


Fig. 4-10 Bidirectional Opening Architecture

- **The Dynamic Orchestrable Plug-in Technology Empowers the Network**

The plug-in architecture is an open framework where devices and capabilities are separated. In the framework, capabilities are encapsulated as plug-ins, which are decoupled from devices and plug-ins. Through the plug-in architecture, a network device is divided into a base device and a capability AI plug-in, and the capability plug-in operates in the base device. The base device is provided by the equipment vendor, and the capability plug-ins are developed by professional third-party suppliers or equipment vendors.

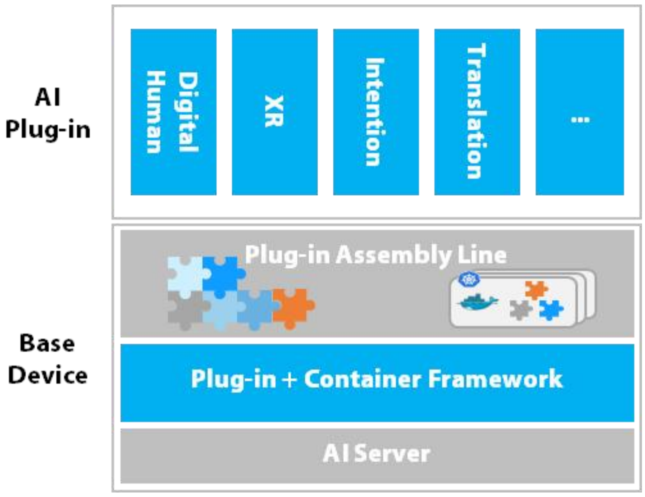


Fig. 4-11 Plug-in Device Architecture

The plug-in architecture is divided into four layers: service layer, framework layer, capability plug-in layer, and support layer. The service layer serves media capabilities and provides media services for AS and capability exposure platform. The framework layer is responsible for managing, orchestrating, and executing plug-ins, thus forming a plug-in pipeline, and combining new media capabilities. The capability plug-in layer is a set of capability plug-ins loaded by the framework layer to the device, including self-owned plug-ins of the base device and third-party plug-ins. The support layer is the real-time operating environment of plug-ins.

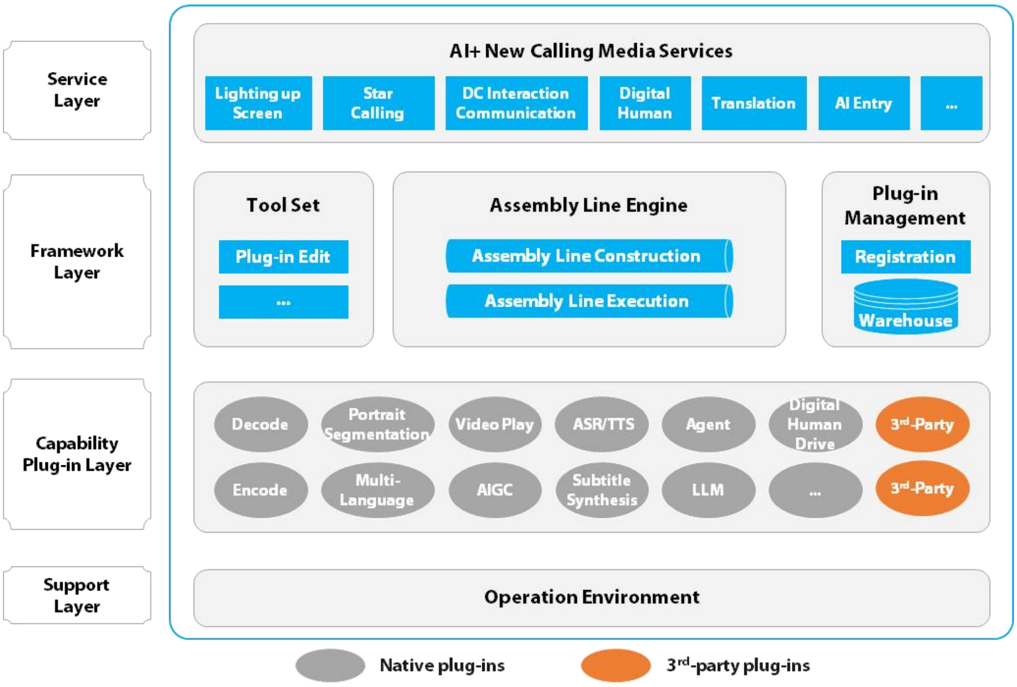
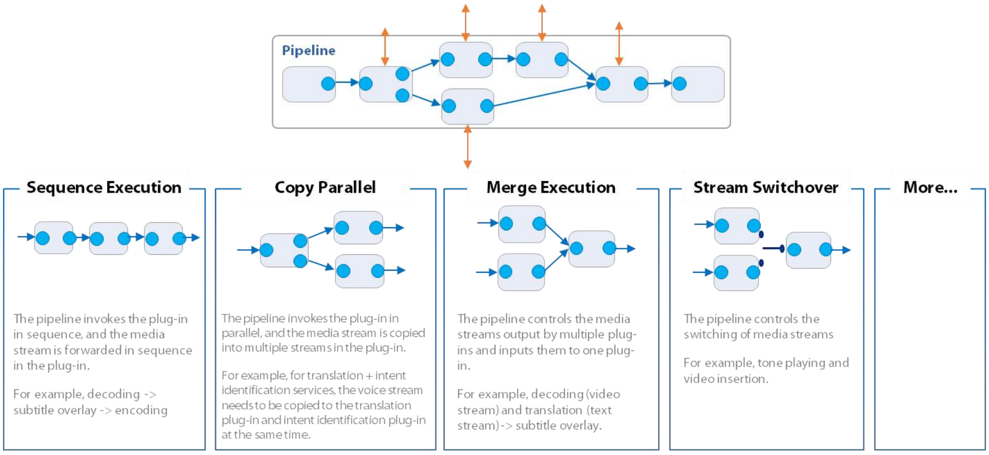


Fig. 4-12 Plug-in Framework

The unified media plane uses a plug-in framework. Various New Calling media capabilities, such as third-party large models, agent, ASR, TTS, digital human driving, audio and video codecs, spatial perception and computing, rendering, multi-language translation, and gesture recognition, are all deployed in media devices as plug-ins. The suppliers in the plug-in ecosystem are strongly combined to promote rapid innovation of AI+ New

Calling.

Dynamic orchestration of plug-ins can generate new media capabilities online. Plug-ins comply with unified interface standards. Standardized plug-ins orchestrate various media processing pipelines in accordance with different sessions and application requirements. The pipeline supports multiple execution modes, such as sequential execution, parallel replication, combined execution, and stream switching. Based on atomic plug-ins and through dynamic orchestration, "zero" coding is implemented to dynamically generate new media capabilities. In addition, the plug-in pipeline greatly simplifies the media control procedure and AS service logic. The original complicated media invocation process is replaced by the media pipeline. The media logic of different New Calling applications is orchestrated into one or more corresponding media pipelines, and encapsulated in a service-based manner. Each service is allocated with a globally unique ServiceKey for the AS to invoke. The complex media control logic is omitted, and the AS logic is lightweight. Processes are simplified and logic is lightweight, making the innovative applications of AS more efficient.



The pipeline invokes the plug-in in sequence, and the media stream is forwarded in sequence in the plug-in.

For example, decoding -> subtitle overlay -> encoding

The pipeline invokes the plug-in in parallel, and the media stream is copied into multiple streams in the plug-in.

For example, for translation + intent identification services, the voice stream needs to be copied to the translation plug-in and intent identification plug-in at the same time.

The pipeline controls the media streams output by multiple plug-ins and inputs them to one plug-in.

For example, decoding (video stream) and translation (text stream)-> subtitle overlay.

The pipeline controls the switching of media streams

For example, tone playing and video insertion.

Fig. 4-13 Dynamic Orchestration of Plug-ins

The system supports dynamic loading of plug-ins to quickly upload new capabilities. After AI+ New Calling is put into commercial use, new capabilities are added to the network without the need to upgrade the system version of base devices. The new capability plug-ins can be dynamically loaded online through plug-in management at the framework layer, and can operate in real time after being dynamically registered to the base device, implementing "zero" waiting and "zero" risk of getting online. The dynamic loading of plug-ins ensures stable and flexible network operation, and can rapidly launch new capabilities, eliminating complicated FOA and greatly reducing OPEX.

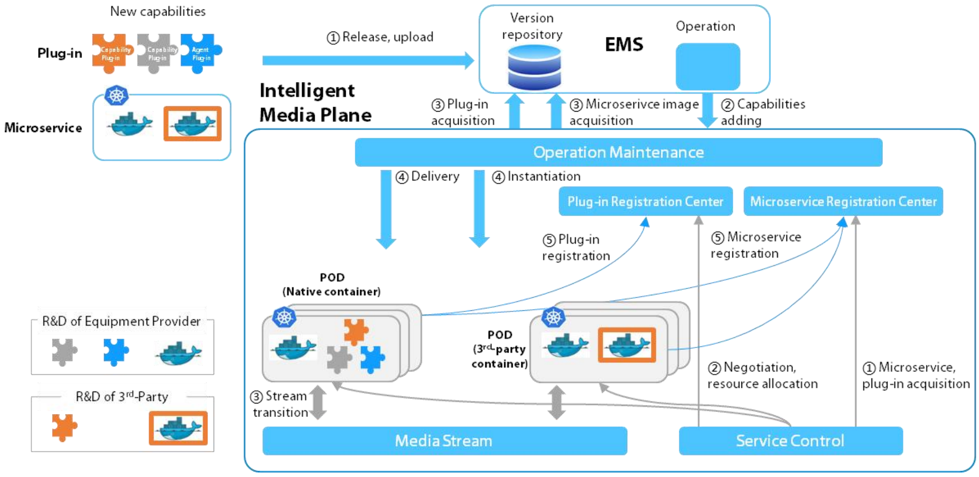


Fig. 4-14 Dynamic Loading of Plug-ins

- **Multi-dimensional unaware capability opening technology to empower the industry**

The multi-dimensional network capability of AI+ New Calling is open to the outside without perception, thus empowering the industry. Through the VoNR+ capability platform and industry gateways, the AI+ New Calling system implements multi-dimensional open network capabilities, including service capabilities and media-plane capabilities， such as audio and video, data channel, digital human, and AI.

Open capabilities are used for agile development of 2B2C New Calling applications in the industry to provide users with visual, interactive, and intelligent New Calling services and improve user experience and service quality, such as lighting up the screen, star calling, call center for New Calling, and digital human agent.

Applications are decoupled from capabilities to achieve interfaces and procedures that are insensitive to capability changes. The AI+ New Calling capability openness deeply integrates the target control process, service-based technology, and ServiceKey media pipeline technology to unify and standardize the interfaces and invoking processes of different 2B2C applications. Different 2B2C applications use the same interfaces and processes to invoke services and media capabilities.

In addition, enterprises have many advantages in developing call centers based on New Calling APIs. The call center uses the media capability of the AI+ New Calling network, and does not need to deploy media devices in enterprises. Media processing is completed in the operator's network, and does not need to be routed to enterprises. This not only reduces latency, but also helps enterprises save at least 50% of transmission line rental fees and avoid one-off investment in media devices in the call center.

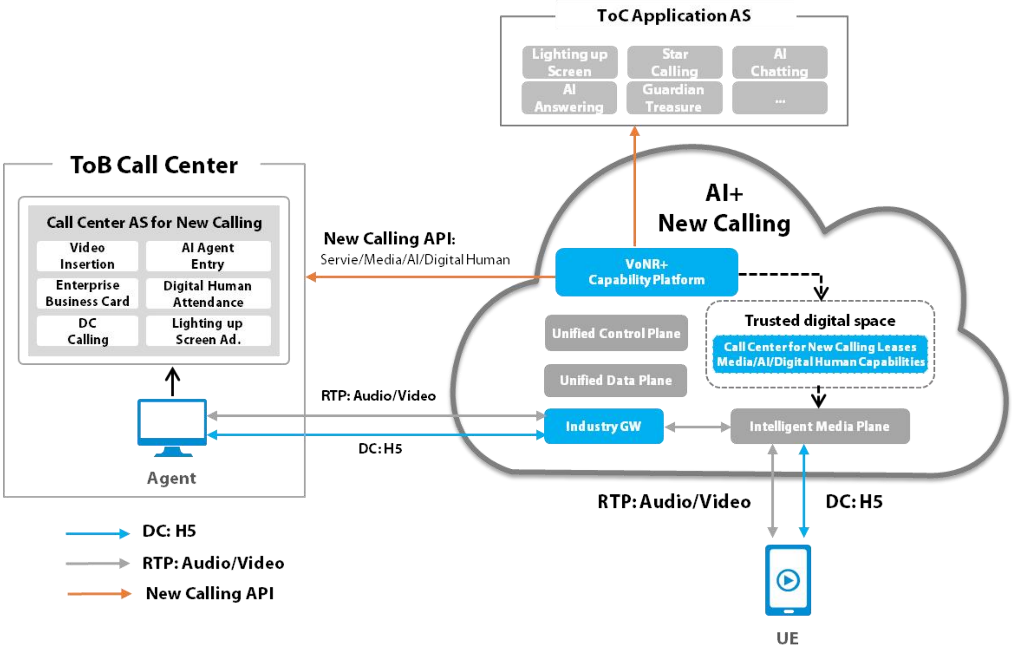


Fig. 4-15 AI+ New Calling Capabilities Openness

4.2.3.2 Endogenous Intelligence

Endogenous Intelligence enables devices to have human-like brains and achieve intelligent device autonomy. With the development of artificial intelligence, communication features are gradually generalized, and are no longer limited to single voice communication and fixed service processes, such as intent communication, XR communication, digital intelligent human, and generative communication. The intelligent device automatically identifies user intentions, automatically completes intelligent orchestration and invocation of capability plug-ins, and provides generic services that cannot be pre-defined by manual pipelines.

The zero-training intent enhancement technology achieves intelligent autonomy and low-cost continuous evolution of devices. The agent, large model, RAG, and plug-in form a zero-training intent engine. The devices based on this engine do not need to be trained frequently, but only needs to add new AI capability plug-ins and knowledge repository dynamically and incrementally. The agent can quickly find a solution through high-precision retrieval and learning, and invoke the new AI capability plug-in to implement user intention. In addition, the knowledge base and other corpus-based fine training agents and large models are regularly used to periodically improve the efficiency and steps of intent identification.

Multi-agent coordination and target control technologies are combined to facilitate simplified processes. Devices in traditional communication networks do not have agents, so they cannot independently make decisions. An agent of an intelligent device has the brain-like capability, and can automatically identify explicit or implicit intentions in multi-modal media and signaling messages, automatically complete decision-making and task decomposition, and independently invoke various tool plug-ins. A simple target control mechanism is used for coordination with other device agents, and complicated procedures of process control and active/standby control of traditional communication devices are no longer required.

AI and media intelligent orchestration technologies integrate multi-modal and multi-streaming media to improve user experience. Media processing and AI are orchestrated and executed intelligently along the same pipeline, and are processed in a single-point and centralized manner. Media coding and decoding are integrated, and data does not need to be duplicated multiple times and transmitted over long distances. With the advantages of low interaction latency, non-detour media, good user experience, and low transmission costs, media processing greatly improves the experience of multi-modal communication and XR communication.

Endogenous intelligence, integrated training and inference, and privacy data is closed loop within network to ensure security. The private data assets processed by the AI, such as intent identification and semantic extraction, do not go out of the network. The inference is completed on the unified control plane, unified data plane, and intelligent media plane. The training and content generation is implemented on the management plane, which can effectively prevent user data leakage and protect user privacy.

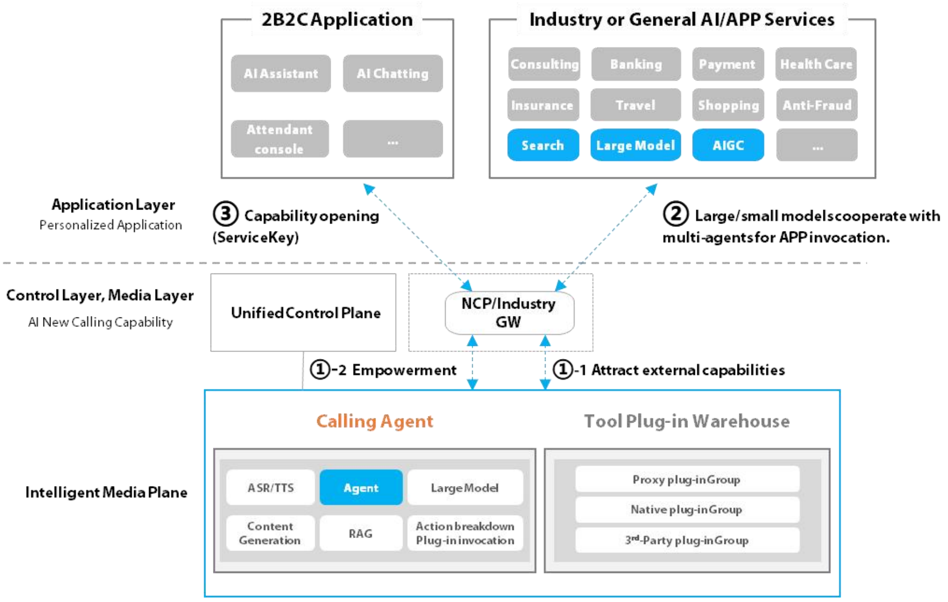


Fig. 4-16 Endogenous Intelligence in the Control Layer and Media Layer of AI+ New Calling

4.2.3.3 Network Simplification

AI+ New Calling adopts the in-depth multi- method integration technology to simplify the network architecture and the processes. The new simplified architecture makes application innovation more efficient and networks more stable.

The stateless concurrent service processing and NFs integration technology unifies control-plane NEs and simplifies the process. The high-frequency service AS, control-plane NEs, and capability exposure gateway are integrated through the NFs technology to simplify networking. Multi-NEs integration shield complicated interfaces and invocation relationships, and provides simple and efficient routing and service capabilities. New technologies such as stateless concurrent service triggering, orchestration, conflict management, and invocation reduce the coupling degree between multiple services and simplify processes. The control plane is unified, and applications are decoupled from the network, making application innovation more efficient.

The intelligent all-in-one computer technology integrates intelligence and media to implement intensive deployment of the media plane. High-throughput, strong computing, low-latency, edge-network collaborative XR, AI, and other AI+ New Calling applications need to be deployed at the edge of the AI+ media device. The shorter the transmission distance between the AI devices and the terminal, the better the system is. The intelligent media plane uses the all-in-one technology, integrates bare metal containers, plug-ins, and service-based technologies, and implements centralized deployment of AI NEs and media NEs as one NE. It has the advantages of simple physical NE networking and flexible resource pool expansion. In addition, the intelligent media-plane device has built-in GPUs, implementing single-point processing and in-depth integration of AI+ media.

The ubiquitous data management technology standardizes and integrates the management and storage of user data, AI data, and digital assets. The standardization of the management and storage of AI data and digital assets is the universal and personalized data basis of AI+ New Calling, and guarantees the sharing and migration of data and New Calling users between different devices, different areas, and different operator networks. Data integration provides one-stop data services. The unified data plane uses distributed computing and storage technologies, vector database and other multi-modal data storage technologies, and RDMA and file transmission technologies. Through the SBI control bus and DCI high-speed data bus, the unified data plane provides the network with full-lifecycle management, storage, and value-added data services such as production and consumption of network-wide data. Data management provides user authentication, digital identity authentication, and service subscription management services. Data storage provides data collection, processing, and storage services for AI data, digital assets, and user information.

The new and simplified architecture supports compatibility and interoperability with the existing network. The integrated and service-based architecture simplifies the network architecture, processes, and service logic, and reduces the number of interfaces. The new architecture simplifies the network and supports compatibility and interoperability with existing traditional networks.

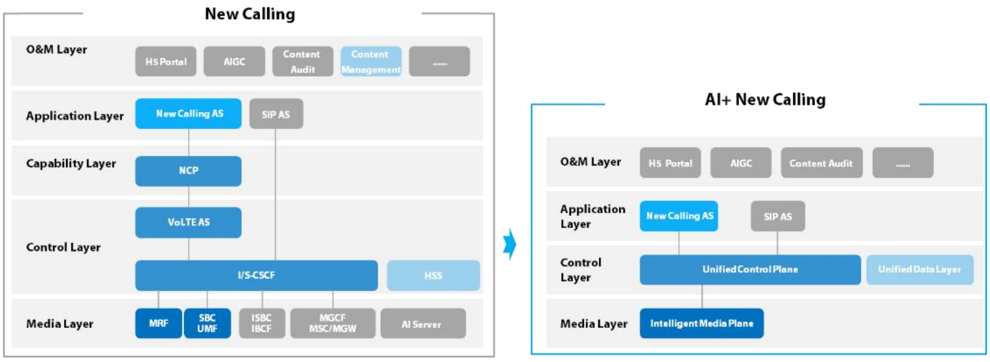


Fig. 4-17 Simplified Network with AI+ New Calling

4.2.4 Two Industrial Chains to Build a Rapid Innovation Ecosystem

The AI+ ecosystem designs two open industrial chains to improve innovation capabilities and efficiency. AI+ New Calling integrates communication and intelligence, involving media processing, ASR, TTS, multi-language translation, agents, NLP/CV/multi-modality large-model inference and training, federation learning, picture/video/text AIGC, digital human driving and generation, avatar, 3D digital human, XR, naked-eye 3D, rendering, AI anti-fraud, content audit, distributed computing, H5 applet, and other high-tech fields. In addition, the AI+ New Calling breaks the boundaries between CT and IT. In addition to providing visual and interactive call functions, the AI+ New Calling can also invoke the IT applications through the AI portal to provide services such as navigation, ticket booking, weather, shopping, retrieval, and consultation. The current voice ecosystem is only a subset of the AI+ New Calling ecosystem. Experience innovation in the huge New Calling ecosystem requires the cooperation of multiple industries, and efficient innovation cannot be guaranteed without scientific mechanisms. Therefore, the AI+ New Calling ecosystem designs two end-to-end open industrial chains: Plug-in industrial chain and H5 applet industrial chain. In addition, AI+ New Calling supports CICD, grayscale upgrade, and provides a digital twin platform to quickly incubate new applications.

- **Plug-in industry chain**

The AI, digital human, XR, multilingual translation, and audio and video capabilities of different vendors in the ecosystem are released in the form of plug-ins or containers. After being assembled and tested by the plug-in factory, the capabilities are released to the operator's operation management center, and deployed in the AI+ New Calling network as required. During the deployment process, no network device version upgrade is required.

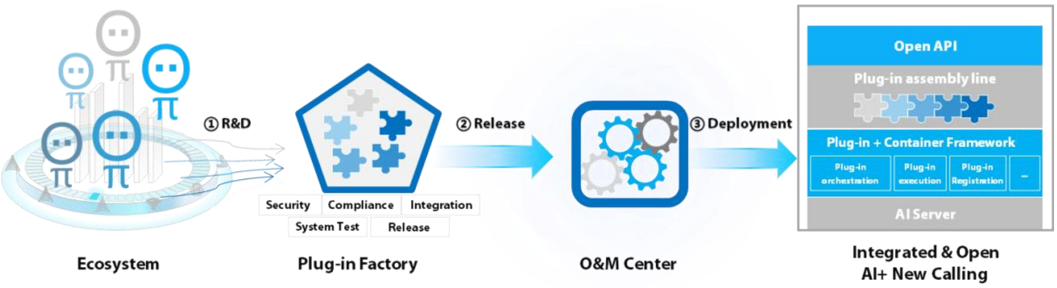


Fig. 4-18 Plug-in Industry Chain

- **H5 applet industry chain**

New calling does not require the terminal to install the APP. Some new applications need the coordination between the terminal and the AS. The new functions on the terminal side are automatically downloaded from the network through Data Channel in the form of H5 applets. AS developers of H5 applets and new H5 applications are located in the H5 ecosystem. New products are launched after passing the H5 factory test.

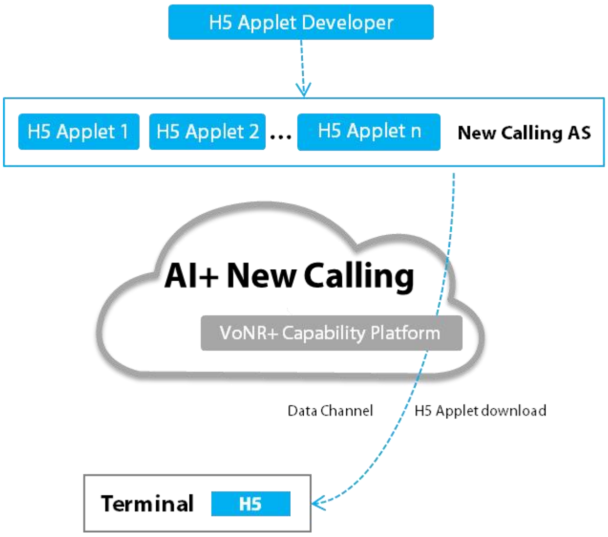


Fig. 4-19 H5 Industry Chain

4.2.5 Three New Intelligent Experiences to Facilitate the Closed-Loop Business of New Calling

Based on the plug-in ecosystem, H5 ecosystem, and new architecture of open intelligence, AI+ New Calling creates a new 2B2C experience that is "visual, interactive, AI+, and personalized". By deeply integrating VC/DC communication with AI capabilities such as digital human driving, digital human generation, agents, and multi-language translation, ZTE has launched end-to-end solutions such as " Unidirectional Video+", "AI Entry", and "AI+ ToB New Calling" to build profit models for content monetization, ToC application intelligence, and interactive commercial communication closed-loop services of ToB agents. In addition, AI applications such as AI translation, AI business shorthand, AI gesture recognition, and fun calling also provide users with interesting and practical experience.

4.2.5.1 Unidirectional Video Enhancement Technology for Integrating Multiple Elements to Intelligently Make Monetization of Video Contents

Both unidirectional video combined with private digital humans and unidirectional video combined with advertising improve users' visual experience and protect their privacy without altering their communication habits. The AIGC of the operation platform generates private digital humans and advertisement materials for individuals or enterprises through images uploaded by the H5 portal. In conjunction with unidirectional video and assembly line, the personalized visual services before and during calls are launched to realize the monetization of video contents, such as enterprise business cards, city business cards, enterprise digital humans, and personal voice-driven digital humans.

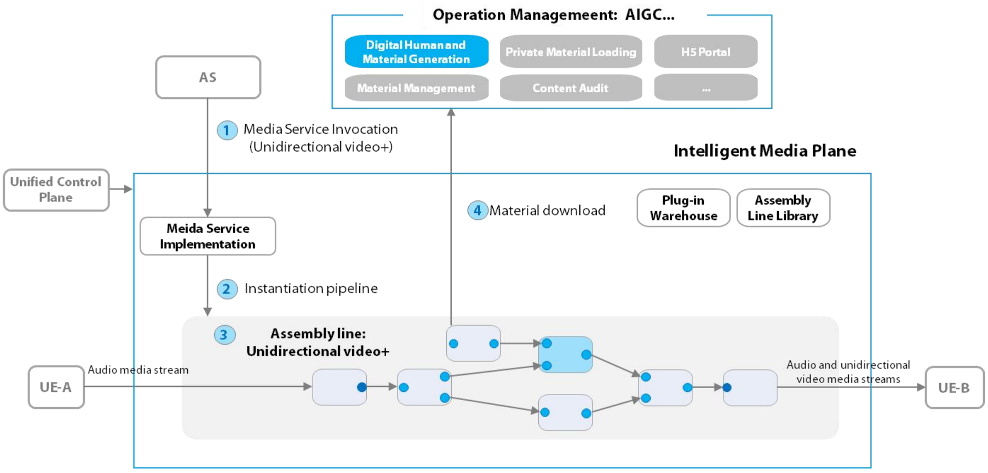

Operation Management: AIGC...
Digital Human and Material Generation
Private Material Loading
H5 Portal
Material Management
Content Audit
...
AS
1 Media Service Invocation (Unidirectional video+)
Intelligent Media Plane
Unified Control Plane
4 Material download
Plug-in Warehouse
Assembly Line Library
Meida Service Implementation
2 Instantiation pipeline
3 Assembly line: Unidirectional video+
UE-A
Audio media stream
Audio and unidirectional video media streams
UE-B

Fig. 4-20 AIGC + Personal Voice-Driven Digital Human Applications

4.2.5.2 AI Entry Multi-modal Multi-Agent Coordination Technology to Reshape Business Intelligence

The AI entry intelligently reshapes services, expands communication boundaries, and provides intelligent and user-friendly communication applications, such as AI assistant, AI chatting service, voice agent portal, APP service portal, and enterprise call center portal.

AI assistants can answer calls and handle advertisements, harassment, express delivery, and fraud calls when users are busy, offline, or do not disturb, thus improving the ARUP value. AI accompanying chat supports communication between people and AI, such as star calling, providing emotional value, and serving for fans and silver hair people.

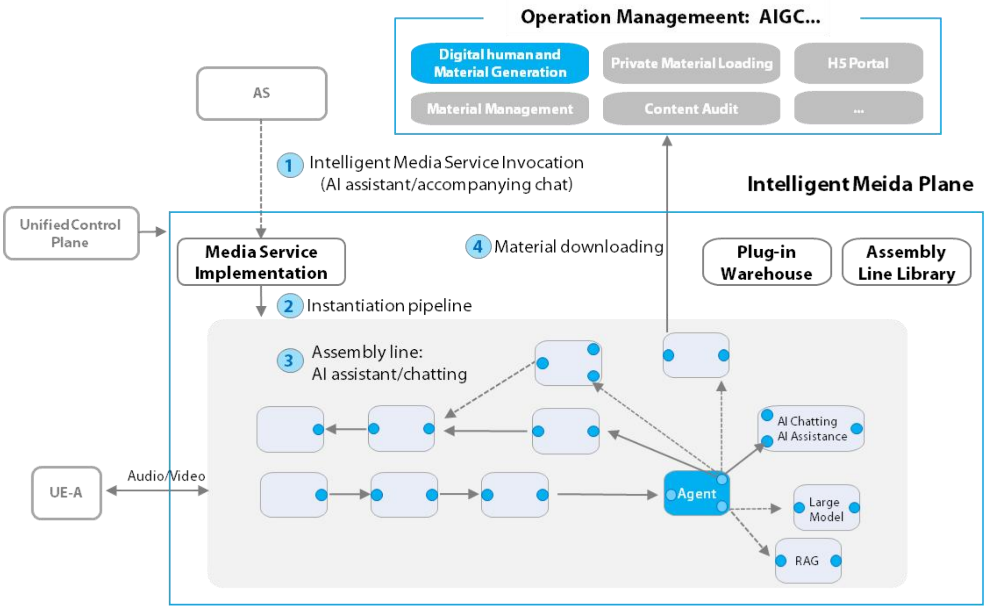


Fig. 4-21 AI Assistant and AI Associate Chat

The APP service portal provides a simple APP mode. After a user enters an intention, the portal automatically outputs the result, shields complicated APP operations, provides the housekeeper service and efficient service, and serves the people who are not willing to use the APP, such as the silver hair people.

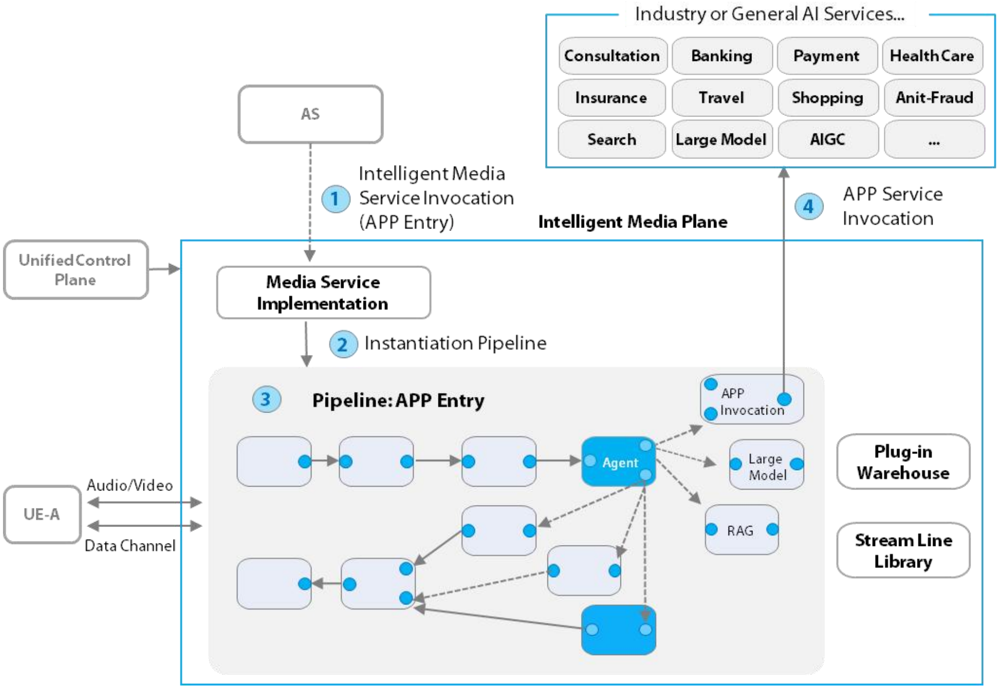


Fig. 4-22 Voice Agent Entry and APP Service Entry

4.2.5.3 AI+ ToB New Calling Digital Intelligence Call Center Technology for Providing Agent Interactive Business Services

Operators occupy a leading position in B2C, working, and stranger communication. In addition, the New Calling service has full AI capabilities, high universality, and good interoperability. Compared with AI terminals and OTT, the New Calling has an absolute advantage in the ToB market. The ToB New Calling has all-round capabilities of "listening, viewing, touching, intent, and digital human", and the DC + H5 applet communication touch-closed loop facilitates sales and services. Digital intelligent agents are online 24 hours a day, reducing service costs. Unidirectional video is provided for improving communication efficiency, product publicity, and user privacy protection.

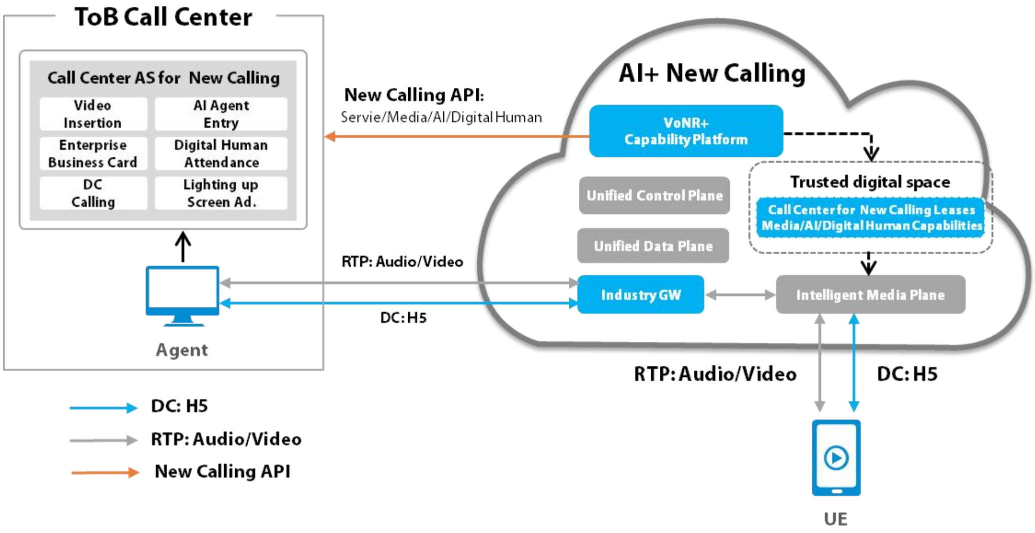


Fig. 4-23 Call Center for AI+ New Calling

4.2.6 New Hexagon Security Mechanism to Build a Profitable and Trusted Anchor Point

Trusted is the anchor point of New Calling. New security creates an end-to-end hexagon security mechanism from six aspects: Trusted space, content compliance, call security, privacy protection, transmission security, and operation security, to protect the security of AI+ New Calling. Compared with voice communication, AI+ New Calling, which is applied in visual, interaction AI+ and individualization, has a large number of private digital assets that need to be protected, in comparison with voice communication. AI generation and AI cloning make it difficult for users to identify true or false, which may lead to an increase in the number of telecom frauds. Interactive DC New Calling facilitates transactions during calls and provides more opportunities for frauds to take advantage of. Video calling may leak user privacy by accident. The ToB application is the key profit model of the New Calling, and trust is the prerequisite of transactions. Contract signing, account opening, signature, and payment all require New Calling to provide an end-to-end trustworthy security mechanism.

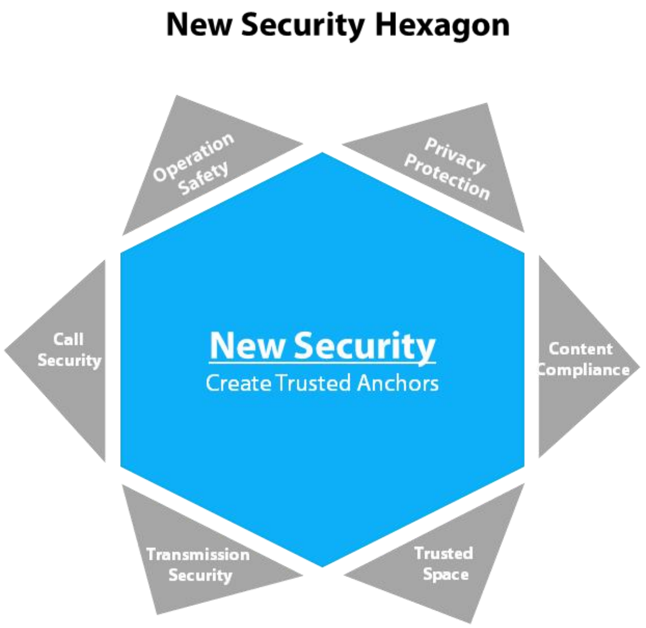


Fig. 4-24 AI New Calling Security Hexagonal

The AI+ New Calling provides end-to-end security guarantee. In new security, private agents, large models, and digital assets operate in dedicated trusted digital space, and content does not go out of the network, effectively preventing personal privacy disclosure. Digital identity authentication and watermark provide visible and trusted services for both communication parties. Unidirectional video is provided to protect user privacy. Private assets, session signaling, and multi-modal media are transmitted in encrypted mode. The content audit platform audits the generated contents and the contents of personal or enterprise materials uploaded through H5 portal. The AI anti-fraud plug-ins are deployed to provide the safe applications and protect the security of New Calling communication. Voice communication is a basic communication service of operators. The AI+ New Calling provides disaster tolerance security mechanisms such as bypass solution, intelligent flow control, and application firewall to ensure the safe operation of New Calling and ensure that the stability of voice communication is not affected.

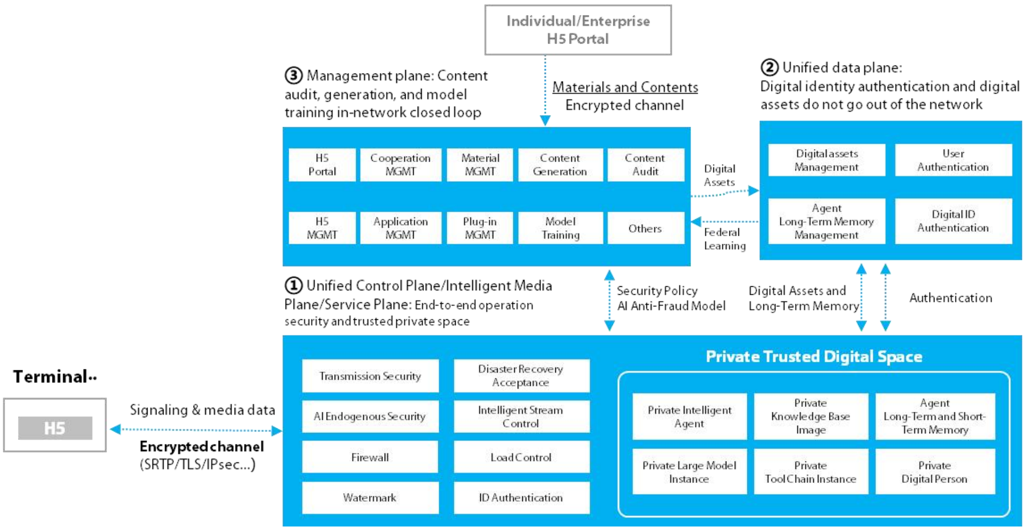

Individual/Enterprise H5 Portal
③ Management plane: Content audit, generation, and model training in-network closed loop
Materials and Contents
Encrypted channel
② Unified data plane:
Digital identity authentication and digital assets do not go out of the network
H5 Portal
Cooperation MGMT
Material MGMT
Content Generation
Content Audit
H5 MGMT
Application MGMT
Plug-in MGMT
Model Training
Others
Digital Assets
Federal Learning
Digital assets Management
User Authentication
Agent Long-Term Memory Management
Digital ID Authentication
① Unified Control Plane/Intelligent Media Plane/Service Plane: End-to-end operation security and trusted private space
Security Policy
AI Anti-Fraud Model
Digital Assets and Long-Term Memory
Authentication
Terminal··
H5
Signaling & media data
Encrypted channel
(SRTP/TLS/IPsec...)
Transmission Security
Disaster Recovery Acceptance
AI Endogenous Security
Intelligent Stream Control
Firewall
Load Control
Watermark
ID Authentication
Private Trusted Digital Space
Private Intelligent Agent
Private Knowledge Base Image
Agent Long-Term and Short-Term Memory
Private Large Model Instance
Private Tool Chain Instance
Private Digital Person

Fig. 4-25 End-To-End Security Architecture of AI+ New Calling

4.2.7 Universal New Calling to Accelerate Value Reshaping of Real-Time Communication

Different applications of the AI+ New Calling have different capability requirements for terminals. Video calling and interactive calling need terminals to support unidirectional or bidirectional video calls, and data channel calls respectively. In order not to change the user habit of "telecom operators' applications do not need to install APP", the popularity of interactive New Calling needs to wait for a large number of DC-capable mobile phones to be launched. The development of the global communication market is unbalanced. In accordance with the popularity of the communication market, user habits, and two types of capability terminals at different development stages, we have designed different development strategies and supporting universal applications to speed up the development of New Calling and reshape the value of real-time communication.

- **Video first and then interaction, universal-video first and then bidirectional-video**

To not change the voice communication habits of users, you can use multiple applications such as unidirectional video+ to cultivate the habits of users of video calls, such as lightening up the screen, calling business card, video insertion, fun calling, and ToB video menu-type call center for New Calling.

To improve the stickiness of video calls, unidirectional video is integrated with AI+, content generation to launch personalized digital human generation, voice-driven digital human, digital human agent, star calling (chatting), and AI answering services.

- **Intention interaction first, and then DC interaction**

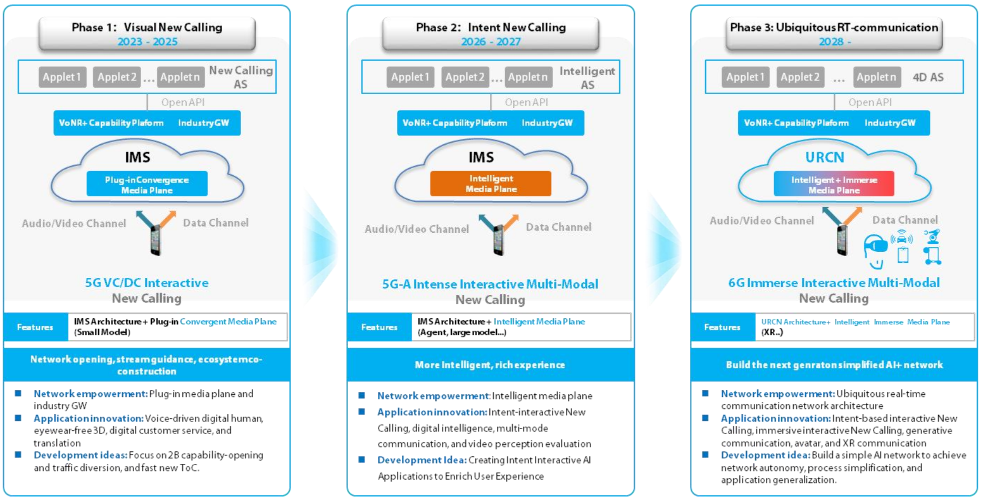
The popularization of DC mobile phones will take time, and the DC-based interactive New Calling applications cannot be widely promoted within a short period of time. Interactive New Calling is an important requirement for ToB communication. In the intent New Calling, the agent that uses the voice AI entry is used to replace the DC. The agent automatically invokes various types of APP in accordance with the user's voice input intention, converts the APP result returned by the web page into a unidirectional video, and sends the video to the user. The interaction is simple and efficient, and complicated APP operations are not required.

With the popularity of DC mobile phones, drag and touch applications can be provided through DC interaction, and ToB New Calling business operations are more convenient. In addition, intention and DC work together to provide multi-modal super-agent AI entries, and multi-modal interaction makes communication more efficient and user-friendly.

4.2.8 Suggestions on Challenges and Development of Next-Generation Real-Time Communication

The integration between AI and communication, immersion and sentimental, ubiquitous connection, and virtual-real symbiosis have become the consensus of 6G development. The next-generation real-time communication oriented to 6G is an intelligent, immersive, virtual-real communication that consists of intelligent applications such as digital intelligent human, avatar, XR, intentional communication, AIGC, and intelligent devices such as AI glasses, AI mobile phones, XR glasses, and body intelligence. The next-generation real-time communication is in the market cultivation stage, facing many challenges. AI+ New Calling is only an intermediate phase in the evolution of voice communication to next-generation real-time communication. At present, the costs of digital human, intent-based communication, AIGC, XR, and intelligent computing are too high, the industrial chain is not mature, the technical division of labor is not clear, and DC terminals are not yet popularized. The key technologies and humanization of XR terminals need to be tackled. Standards such as XR, avatar, AI+ network, and DC terminal SDK are being formulated. New architectures and standards need to be defined for next-generation real-time communication, intelligent application profit models need to be explored, and new habits of users need to be cultivated. This restricts the development process of next-generation real-time communication.

The evolution of next-generation real-time communication can be divided into three phases: visual phase, intention phase, and ubiquitous phase in accordance with the current standard status, industrial chain and technical maturity, universal terminal applicability, and user habit training process. In 2024-2025, the phase of video and New Calling communication focuses on unidirectional video communication, does not change the voice communication operation mode of users, and gradually cultivates the video communication habits of users. In 2026-2027 that is the intention (AI+) phase, we can create intelligent experience, focus on AI applications such as AI entry and digital intelligent human, and reconstruct voice value. After 2028, URCN (Ubiquitous Real-time Communication) focuses on immersive communication and ubiquitous real-time communication.


Phase 1: Visual New Calling
2023 - 2025
Applet 1
Applet 2
...
Applet n
New Calling AS
Open API
VoNR+ Capability Plaform
IndustryGW
IMS
Plug-in Convergence Media Plane
Audio/Video Channel
Data Channel
5G VC/DC Interactive New Calling
Features
IMS Architecture + Plug-in Convergent Media Plane (Small Model)
Network opening, stream guidance, ecosystem co-construction

- Network empowerment: Plug-in media plane and industry GW
- Application innovation: Voice-driven digital human, eyewear-free 3D, digital customer service, and translation
- Development ideas: Focus on 2B capability-opening and traffic diversion, and fast new ToC.

Phase 2: Intent New Calling
2026 - 2027
Applet 1
Applet 2
...
Applet n
Intelligent AS
Open API
VoNR+ Capability Plaform
IndustryGW
IMS
Intelligent Media Plane
Audio/Video Channel
Data Channel
5G-A Intense Interactive Multi-Modal New Calling
Features
IMS Architecture + Intelligent Media Plane (Agent, large model...)
More Intelligent, rich experience

- Network empowerment: Intelligent media plane
- Application innovation: Intent-interactive New Calling, digital intelligence, multi-mode communication, and video perception evaluation
- Development Idea: Creating Intent Interactive AI Applications to Enrich User Experience

Phase 3: Ubiquitous RT-communication
2028 -
Applet 1
Applet 2
...
Applet n
4D AS
Open API
VoNR+ Capability Plaform
IndustryGW
URCN
Intelligent + Immerse Media Plane
Audio/Video Channel
Data Channel
6G Immerse Interactive Multi-Modal New Calling
Features
URCN Architecture+ Intelligent Immerse Media Plane (XR..)
Build the next genraton simplified AI+ network

- Network empowerment: Ubiquitous real-time communication network architecture
- Application innovation: Intent-based interactive New Calling, immersive interactive New Calling, generative communication, avatar, and XR communication
- Development idea: Build a simple AI network to achieve network autonomy, process simplification, and application generalization.

Fig. 4-26 Next-Generation Real-Time Communication Evolution

The sustainable development of next-generation real-time communication requires leading standards, an end-to-end complete industrial chain, an open ecosystem, and a clear division of labor. Communication upgrade cannot be completed overnight, so it needs to be carried out in phases. A large number of CT/IT high and new technologies are enriched, the industry chain is longer, and the ecosystem is larger. The next-generation real-time communication that turns to experience operation requires standards to coordinate the orderly development of industries in the ecosystem, break the current AI+ high-tech barriers, and create universal New Calling. With the popularity of AI/XR and VC/DC terminals, the next-generation real-time communication will replace voice communication as a kind of daily basic communication

5 "AI+" Connection: Mining Data Value

5.1 5G Entering a Stable Period, and the Subsequent Development Slowing Down

Since the start of 5G construction in 2018, 5G construction has entered a high-quality and stable development phase from a high-speed development phase. In the second quarter of 2024, the total number of global 5G users reached 1.87 billion, the largest number of 5G users in East Asia (China, Japan, and Korea) reached 1,057 million, and the number of 5G users in China reached 927 million. The number of 5G users in North America is about 317 million, that in Europe is about 223 million, and that in other countries is about 281 million.

However, the pure traffic based dividend brought by 5G is exhausted. With the development of video technologies, the DOU of a single user will not increase in the future. For example, the bit rate of H266 video codec is reduced by 49% compared with H265 video codec, and the bandwidth requirement is reduced. The 1080P HD service is affected by the return on investment ratio, and the penetration rate is suppressed. At the same time, 5G is entering the 5G-Advanced stage. The introduction of new technologies will empower emerging fields such as low-altitude economy and vehicle-road synergy, and expand the market space. However, it also raises higher requirements for networks in terms of agility and scalability. In addition, as marked by ChatGPT and released by Open AI,, AI enters a new stage, and the application of large models is becoming more and more widely, which deeply affects all industries including the telecom industry, bringing new development opportunities as well as unprecedented challenges.

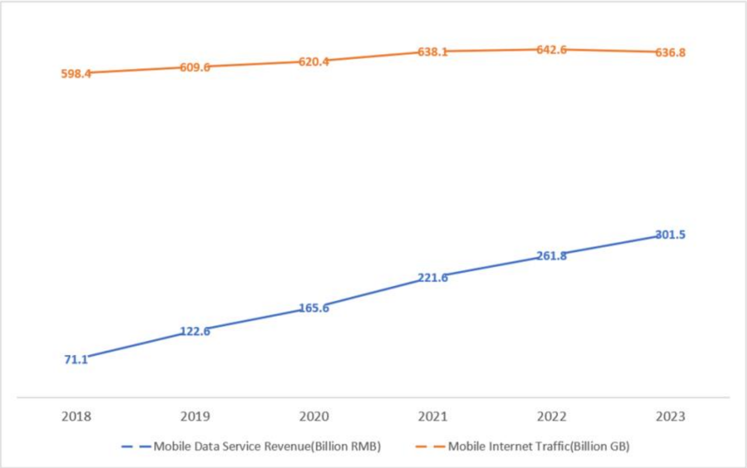


Fig. 5-1 2018 – 2023 Mobile Data Traffic and Revenue Development in 2018 - 2023 (Source: Website of the Ministry of Industry and Information Technology)

- **Revenue growth is stagnant, and new growth points need to be explored**
  - The traffic growth is slowed down, and the revenue model based on traffic encounters bottlenecks. The revenue growth of operators is slow or even declines.
  - New services failed to bring commercial value to operators.
  - End user consumption models are fixed, which is not conducive to the promotion of new services.
- **Differentiated pipeline capabilities are insufficient, and refined operation is difficult**
  - High-value user experience is not guaranteed, and differentiated services are not provided.
  - Operators do not know users' service preferences, so it is difficult to implement precise marketing and the network value is reduced.
  - Operators lack user experience data and cannot actively optimize networks.
- **The general devices are oriented to the public network, and cannot meet the diversified requirements of the industry**

- In the industrial field, dedicated industrial UPF is required to meet deterministic requirements and achieve intelligent orchestration.
- Low-altitude industries require built-in AI algorithms and computing-power UPF to integrate communication and perception.
- The smart transportation field requires an independent IoV Internet of Vehicles (UPF) to support ultra-large uplink bandwidth and ultra-low downlink latency and implement dynamic AI orchestration.
- Low latency and packet loss sensitive services have poor experience in mobile networks, hindering the expansion of 5G into the industry.

- **ICT technology integration is insufficient, and efficiency needs to be improved and consumption needs to be reduced**
  - Low resource utilization and high energy consumption do not meet the green development path, and new technologies need to be introduced to save energy and reduce consumption.
  - The integration of cloud-based and AI technologies in mobile networks is insufficient. Resource orchestration and allocation depend on static planning.
- **Signaling storms occur frequently, causing huge risks**
  - The 5G network is decoupled in layers, and the architecture is relatively complicated. There are many types of terminals and a large number of IoT terminals. New services and applications are emerging, and signaling messages have increased dramatically.
  - Although the operators have deployed preventive solutions, the evaluation result is insufficient and the response result is poor.

5.2 3-Layer Architecture with 4 Capabilities to Empower Intelligent Connections

The introduction of AI in the 5GC can effectively meet many current challenges and inject new vitality into connections. The NWDAF is introduced as the brain, the PCF (AM and SM) is used as the policy support point, and the built-in AI Engine of each NE is used as the executor to work together with the RAN and terminals to establish an end-to-end AI connection.

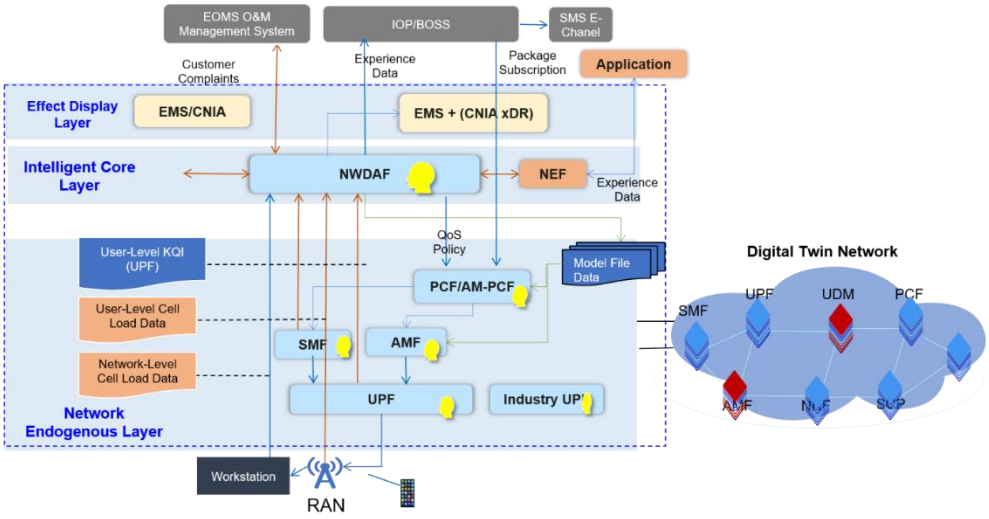


Fig. 5-2 AI+ Connection Intelligent 3-Layer Architecture with 4 Capabilities

The AI+ connection intelligent network adopts a three-layer architecture:

- Network endogenous layer: It corresponds to CN NEs. The intelligent processing units/functions are integrated on the basis of the current CN NEs, and independent industrial UPFs are deployed for functional service application scenarios. In addition, the digital twin technology is used to twin a virtual digital network layer.
- Intelligent core layer: It includes the NWDAF and NEF. The NWDAF acts as the intelligent plane brain to collect the data reported by the NEs at the internal layer of the network, performs prediction and policy decision based on the AI algorithm, and delivers the decision-making policy to the internal layer of the network for execution to form a closed loop.
- Effect display layer: It displays the intelligent execution effect on the network O&M interface.

The four enhanced capabilities include:

- Commercial operation capability: Development from the traditional flow-based operation mode to the experience operation mode.
- Multi-industry support capability: The enhanced network capability can provide

more capabilities (such as deterministic capabilities) and resources (such as computing resources) to better serve the industry.

- Multi-technology integration capability: After various technologies are combined with AI, the advantages of various technologies can be better utilized.
- Network security capability: The combination of security and AI makes the network more resilient.

5.3 Business Mode Innovation, from Traffic Operation to Experience Operation

The mobile network has always provided users with different levels of QoS guarantee through static subscription. However, with the development of technologies and the rapid increase of various applications, on the one hand, users need more targeted service experience guarantee and are willing to pay for it. For example, live users need to perform HD live broadcast without interruption in hot spots. For e-sports users, the latency needs to be reduced and the competition needs to be accelerated. For efficient communication, business travel users need better call quality and data transmission speed. On the other hand, operators need to accurately perceive user requirements and network conditions in different scenarios, dynamically allocate network resources, and realize value realization. Therefore, differentiated experience guarantee is provided for different levels of users, service applications, and network locations in the network, including:

1. Provides high-level experience for high-value users in the network.
2. For Top N experience-sensitive applications (such as games and live broadcast) in the network, as key services for experience guarantee, various guarantee packages are launched for users to purchase.
3. For specific places (such as concerts, gymnasiums, and high-speed railways) where users are densely gathered, specific users (such as security and VIP) are guaranteed.

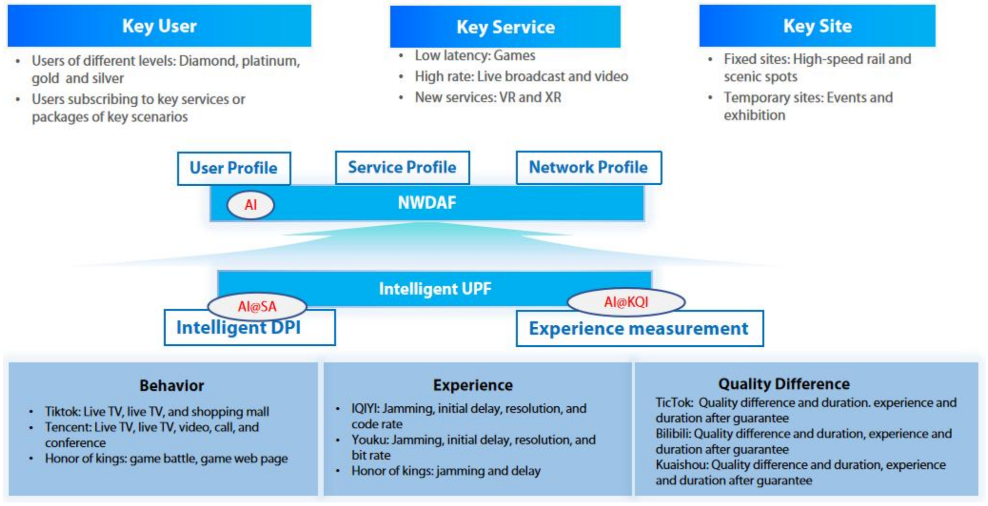

Key User
• Users of different levels: Diamond, platinum, gold and silver
• Users subscribing to key services or packages of key scenarios
Key Service
• Low latency: Games
• High rate: Live broadcast and video
• New services: VR and XR
Key Site
• Fixed sites: High-speed rail and scenic spots
• Temporary sites: Events and exhibition
User Profile
Service Profile
Network Profile
AI
NWDAF
Intelligent UPF
AI@SA
Intelligent DPI
AI@KQI
Experience measurement
Behavior
• Tiktok: Live TV, live TV, and shopping mall
• Tencent: Live TV, live TV, video, call, and conference
• Honor of kings: game battle, game web page
Experience
• IQIYI: Jamming, initial delay, resolution, and code rate
• Youku: Jamming, initial delay, resolution, and bit rate
• Honor of kings: jamming and delay
Quality Difference
TicTok: Quality difference and duration. experience and duration after guarantee
Bilibili: Quality difference and duration, experience and duration after guarantee
Kuaishou: Quality difference and duration, experience and duration after guarantee

Fig. 5-3 Experience Operation Diagram

UPFs provide two key capabilities based on the built-in AI: Intelligent software-hardware coordination service identification technology and holographic KQI experience measurement technology.

Based on the collected data (such as user behavior, experience, and poor quality data from the intelligent UPFs, and load from the RAN OAM and NEs), the NWDAF implements the following capabilities through the multi-dimensional cross-layer cross-domain profile technology:

1. Service profile: Helps operators to discover hot applications and experience data of various applications in the network, and assists operators in service package design.

2. User profile: Helps operators to find target customers and accurately recommends user experience their favorite improvement packages.

3. Network profile: Helps operators to analyze the use of network resources, especially the network congestion in key places (areas with centralized personnel), and assists operators to provide better operation experience guarantee packages. For example, if a multi-frequency base station is deployed in a severely congested area, the guaranteed UEs are shunted to a specific frequency band through the RFSP to prevent the establishment of dedicated GBR bearers from further squeezing the resources of common UEs.

5.3.1 Multi-Dimensional Cross-Layer Cross-Domain Portrait Technology

Multi-dimensional portraits can help operators better carry out experience operation. For example, operators expect to carry out an experience improvement package containing a certain Top video service for business travelers. Users in the network can be found on a large-scale mobile travelers based on mobility profiles, and the Top video services and video service experience can be further analyzed. The technology is used to accurately discover that some business travelers like to use a video, and experience is poor due to their frequent access to and from the CBD area where the population is concentrated. For the portrait user group, an invitation SMS message can be triggered to invite the user to subscribe to the experience improvement package to improve the user's experience of watching videos.

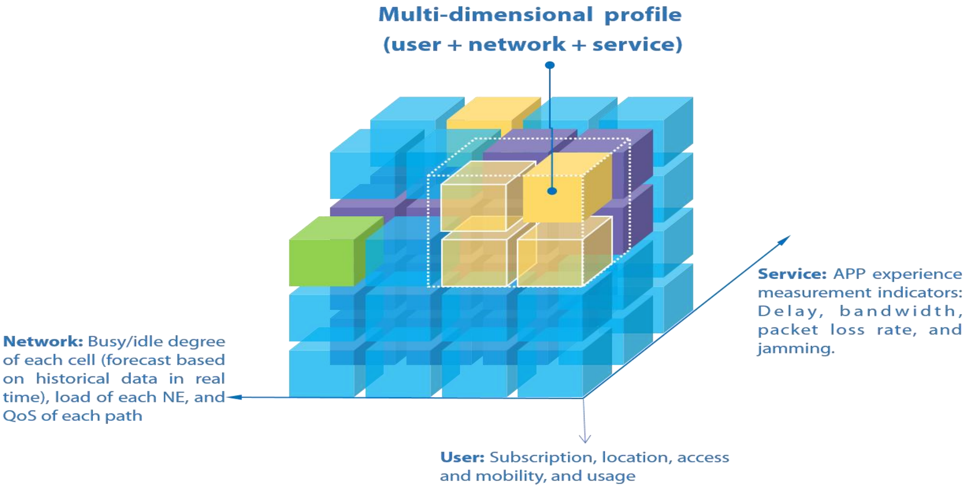

Multi-dimensional profile
(user + network + service)
Service: APP experience measurement indicators: Delay, bandwidth, packet loss rate, and jamming.
Network: Busy/idle degree of each cell (forecast based on historical data in real time), load of each NE, and QoS of each path
User: Subscription, location, access and mobility, and usage

Fig. 5-4 Multi-Dimensional Profile

A multi-dimensional cross-layer and cross-domain profile includes the user dimension, network dimension, and service dimension. The information after the profile reflects the experience of a specific user in a specific place. To implement accurate portrait and profile-based experience package operation, cross-layer coordination is required.

- Collects data between the wireless network and the NWDAF. The NWDAF obtains radio congestion information from the wireless side to assist in portraying the busy and idle dimensions of the wireless network.
- Data is connected between NF and the NWDAF. The NWDAF collects user experience data, user location and mobility data from the network layer to help

complete user and service profiles.

- Through the interface between the NWDAF and the operating domain, based on profile information, the NWDAF pushes information to the operating domain to assist in package operation, such as target customers and package experience results.

5.3.2 Intelligent SW-HW Coordination Service Identification Technology

The proportion of encrypted data streams in a mobile network is increasing. Encrypted data streams need to be identified based on packet characteristics. The traditional method is to obtain data streams from the network, generate a feature library after offline analysis, and then load the generated feature library into the network to identify encrypted data streams. In this way, it takes a long period from the time when a new application appears to the time when the feature library is generated and loaded, and the best time for the application to operate experience services may be missed.

Various Internet applications are frequently updated, and the feature library identification mode is used. Therefore, the feature library needs to be continuously updated to match new version applications. Due to the long period from the analysis and identification of the feature library to the completion of loading, the identification rate of experience operation applications is reduced, and the effect of experience operation is affected.

Facing the challenges of mobile network data encryption and variable application versions, the traditional SA feature library identification method has great challenges in accuracy and timeliness, and cannot meet the requirements of experience operation. The AI-based application identification technology needs to be introduced to meet this new requirement.

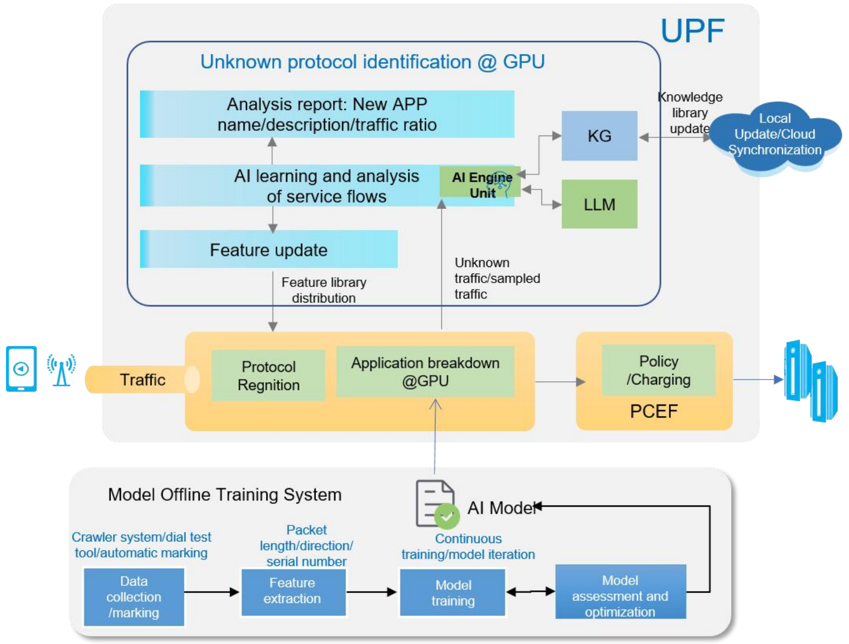

UPF
Unknown protocol identification @ GPU
Analysis report: New APP name/description/traffic ratio
AI learning and analysis of service flows
AI Engine Unit
Feature update
KG
LLM
Knowledge library update
Local Update/Cloud Synchronization
Feature library distribution
Unknown traffic/sampled traffic
Traffic
Protocol Regnition
Application breakdown @GPU
Policy /Charging
PCEF
Model Offline Training System
AI Model
Crawler system/dial test tool/automatic marking
Packet length/direction/ serial number
Continuous training/model iteration
Data collection /marking
Feature extraction
Model training
Model assessment and optimization

Fig. 5-5 Architecture of Intelligent SW-HW Coordination Service Identification

In the above architecture, with the super computing provided by GPU, two key intelligent components operate:

- Unknown protocol identification, including:
  - Intelligent analysis: Automatically analyzes service flow, identifies new applications in the network and updates existing applications, analyzes packet characteristics, and clusters flow characteristics.
  - Intelligent marking: Based on the LLM, different service flows of the same application can be analyzed for service aggregation and marking.
  - Feature generation: Generates protocol features and verifies their validity. The generated new feature library is loaded online to the UPF to implement online update of the feature library. The update period of the feature library is shortened to days.
- The applications are not put into subdivision, including:
  - Offline training-subdivision service identification AI model
  - Completes data stream inference online, and identifies the subdivision

applications that service streams belong to, such as voice calls, video calls, text transfer, and file transfer in social software.

During experience operation, the corresponding GBR guaranteed bandwidth can be allocated to specific applications to implement precise resource configuration and maximize the use of wireless resources.

5.3.3 Holographic KQI Experience Measurement Technology

In the process of promoting experience operation packages, accurately measuring user experience has become the key. This requires us to adopt a measurement technology that can reflect user experience in a three-dimensional and real way – holographic KQI measurement technology. This technology achieves all-round and in-depth measurement of user experience through multi-dimensional KQI indicators.

In the past, experience measurement mainly relies on the comprehensive collection and analysis of transmission data, player message data, and operating system information data. However, in the experience operation scenarios of operators, due to complicated factors such as data stream encryption, it becomes extremely difficult to directly obtain detailed data such as players and operating systems. This limitation poses a huge challenge under the traditional measurement mode.

The holographic KQI measurement technology breaks this bottleneck. It uses the AI algorithm to analyze data streams, extract various features of data streams (such as packet length distribution, delay, and flow), and generate data models corresponding to application data streams with different user experiences through modeling. For real data streams in the network, the characteristics of the data streams are also extracted, and various real experience indicators (KQI) corresponding to the data streams are obtained through model inference. This is equivalent to an all-round extraction of data stream features, and projection of these features into a model. The model restores the real user experience indicators through inference operations.

With this technology, operators can deeply analyze data streams such as video and game, and accurately capture and extract the corresponding KQI experience indicators even when data is encrypted. This technology is not only highly accurate, but also presents multiple details of user experience in a three-dimensional and real manner, providing powerful technical support for operators to accurately measure and optimize user experience.

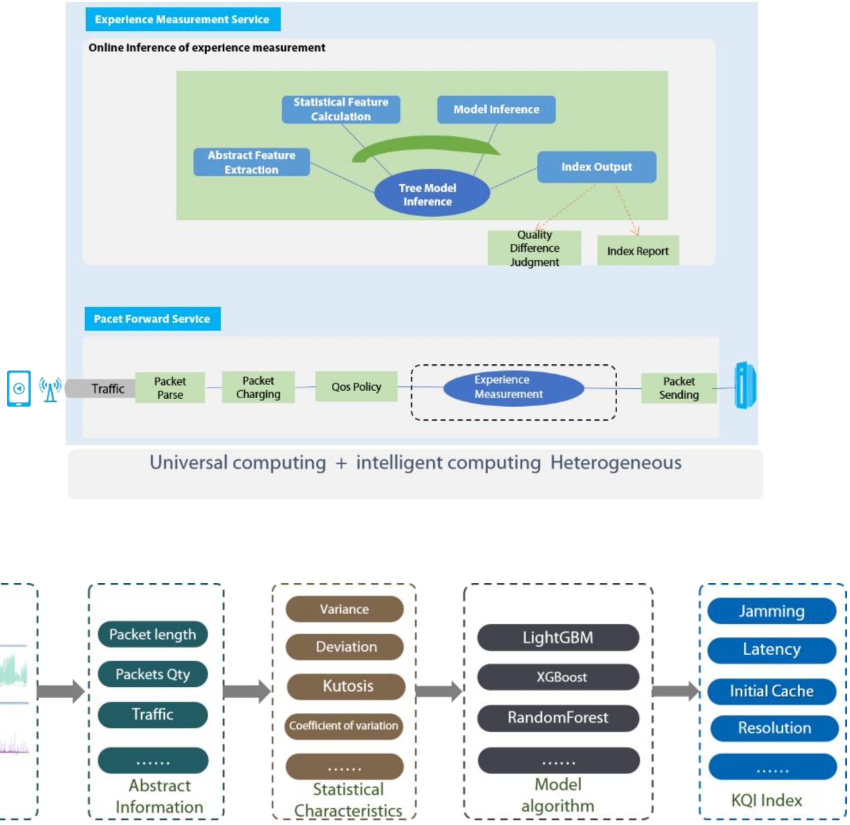


Fig. 5-6 Holographic KQI Experience Measurement Technology

5.4 Empowering Industries and Expanding Connectivity

When wireless connections replace wired connections, it is necessary to meet the requirements of various industries. In conclusion, industrial applications have the following requirements for connections:

- Industry applications need to be connected to provide the endogenous deterministic function, for example, to ensure that the instructions in the industrial production process are executed in order as expected. The wireless network provides deterministic connections in two ways. One is that an external TSN gateway ensures deterministic data transmission, which will bring better costs. The other is that a built-in TSN gateway provides intrinsic deterministic data transmission, which puts forward high requirements for end-to-end network orchestration. If manual configuration is used, a large number of network configuration and maintenance tasks will lose the value of introducing wireless connections to improve network flexibility. It is necessary to introduce AI to implement intelligent orchestration.

- Industrial applications have low latency requirements, such as industrial production, which requires that the UPF should be able to sink to industrial parks. In addition, because the environment of industrial application equipment rooms is limited, more environment requirements are proposed for UPF equipment, and industrial UPF devices that are applicable to industrial environments need to be deployed. In addition, image detection and other processing in industrial detection fields require AI computing. Therefore, industrial UPF devices that integrate AI computing need to be provided in limited industrial equipment rooms.

5.4.1 AI for Deterministic to Broadening the Service Breadth in the OT Domain

In the industry 4.0 era, the traditional network architecture cannot meet the urgent requirements for low latency, high reliability, and deterministic transmission in industrial production. To solve this problem, we can innovatively deploy OT-UPF on industrial production sites to provide enterprises with convenient nearby access services. With the built-in AI capability, OT-UPF breaks the "best-effort" transmission limitation of traditional 5G networks, and customizes deterministic transmission channels for industrial production. This transformative measure not only avoids the dilemma of complicated configuration and complicated O&M of traditional TSN networks, but also takes production tasks as the guide and intelligently optimizes end-to-end forwarding and scheduling, ensuring the deterministic of the entire transmission path, laying a solid foundation for the digital transformation and efficiency improvement of industrial production.

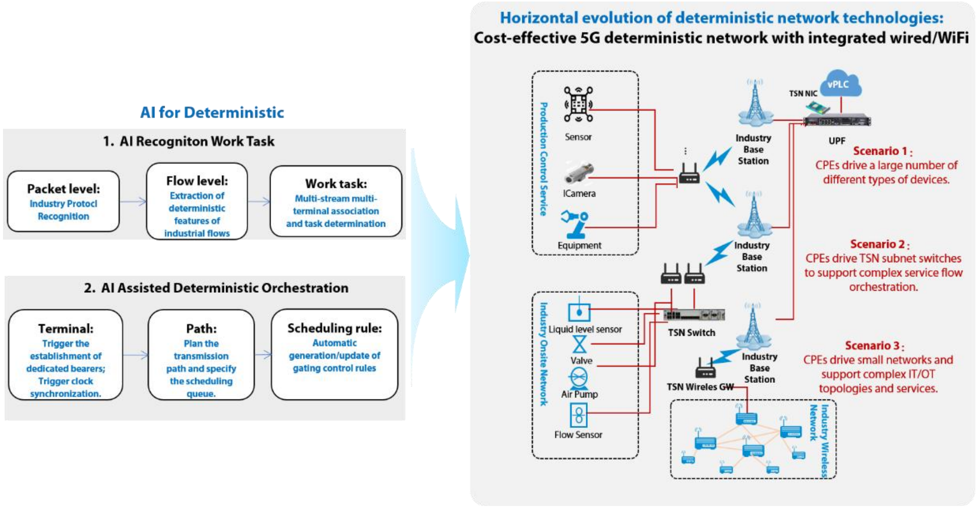

AI for Deterministic
1. AI Recogniton Work Task
Packet level: Industry Protocl Recognition
Flow level: Extraction of deterministic features of industrial flows
Work task: Multi-stream multi-terminal association and task determination
2. AI Assisted Deterministic Orchestration
Terminal: Trigger the establishment of dedicated bearers; Trigger clock synchronization.
Path: Plan the transmission path and specify the scheduling queue.
Scheduling rule: Automatic generation/update of gating control rules
Horizontal evolution of deterministic network technologies:
Cost-effective 5G deterministic network with integrated wired/WiFi
Production Control Service
Sensor
ICamera
Equipment
Industry Base Station
TSN NIC
vPLC
UPF
Scenario 1 : CPEs drive a large number of different types of devices.
Industry Base Station
Scenario 2 : CPEs drive TSN subnet switches to support complex service flow orchestration.
Industry Onsite Network
Liquid level sensor
Valve
Air Pump
Flow Sensor
TSN Switch
Industry Base Station
TSN Wireles GW
Scenario 3 : CPEs drive small networks and support complex IT/OT topologies and services.
Industry Wireless Network

Fig. 5-7 AI for Deterministic Network

- Deploys the OT-UPF on the industrial production site to provide nearby access for enterprises.
- The built-in AI capability in OT-UPF provides enterprises with deterministic transmission channels. The traditional 5G network is a best-effort transmission channel, which cannot meet the requirements of industrial deterministic indicators. To meet the requirements of industrial deterministic indicators, TSN needs to be introduced and end-to-end configuration planning needs to be performed, which brings huge complexity to network operation and maintenance. After AI is introduced, industrial production tasks are taken as the main line, and forwarding scheduling is automatically performed on each main line (gating scheduling rules are set) to ensure end-to-end coordination and achieve determinism on the entire transmission path.

5.4.2 Convergence of Intelligence and Network to Fully Support Edge Applications

When we face with decentralized computing, insufficient model generalization, and potential data security problems in edge applications, the edge intelligent computing UPF emerges. As a new-generation network infrastructure, it integrates AI computing, provides a one-stop computing network base for intelligent services and equipment, and attracts and expands the ecosystem of intelligent industries/parks and other 5G private networks. From the dedicated AI model to the general large model, the AI capability of the edge intelligent UPF is continuously enhanced. With the improvement of the AI card capability, the integration of offline and online training is realized, ensuring real-time dynamic adjustment and avoiding data outflow, and providing a solid guarantee for the intelligent transformation of the park.

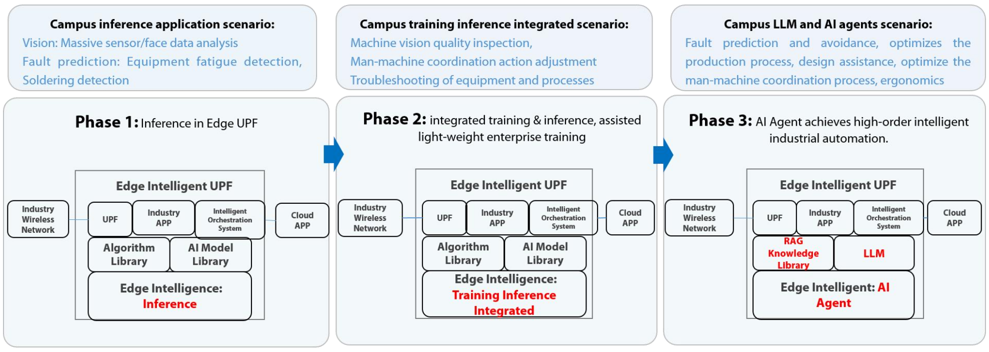

Campus inference application scenario:
Vision: Massive sensor/face data analysis
Fault prediction: Equipment fatigue detection, Soldering detection
Campus training inference integrated scenario:
Machine vision quality inspection,
Man-machine coordination action adjustment
Troubleshooting of equipment and processes
Campus LLM and AI agents scenario:
Fault prediction and avoidance, optimizes the production process, design assistance, optimize the man-machine coordination process, ergonomics
Phase 1: Inference in Edge UPF
Phase 2: integrated training & inference, assisted light-weight enterprise training
Phase 3: AI Agent achieves high-order intelligent industrial automation.
Edge Intelligent UPF
Industry Wireless Network
UPF
Industry APP
Intelligent Orchestration System
Cloud APP
Algorithm Library
AI Model Library
Edge Intelligence: Inference
Edge Intelligence: Training Inference Integrated
RAG Knowledge Library
LLM
Edge Intelligent: AI Agent

Fig. 5-8 Edge Intelligent Computing UPF

- The edge intelligent computing UPF integrates AI computing, provides a one-stop computing network base for intelligent services and equipment, and attracts and expands the ecosystem of intelligent industrial 5G private networks.
- The edge intelligent computing UPF provides a dedicated scenario (such as machine vision model), and gradually becomes an AI model that supports general scenarios. The generalization capability of the AI model is gradually enhanced.
- With the improvement of the embedded AI card capability, the edge intelligent computing UPF is gradually enhanced from offline training and online inference to online training and online inference (integrated with training and inference) to implement real-time dynamic adjustment, and effectively prevent data from leaving the park.

5.5 Multi-technology Integration to Improve Connection Efficiency

5.5.1 AI+ Cloudified Technology for Energy Saving

After the virtualization technology is introduced to core network, energy can be saved in idle hours through frequency reduction or dormancy. The introduction of AI can improve the accuracy of network idle prediction, achieve dynamic and precise energy saving, and improve the energy saving effect.

Scenario: The NF load fluctuates in different time periods. In the low-load range, energy can be saved through frequency reduction/dormancy. In the high-load range, frequency reduction/dormancy needs to be canceled.

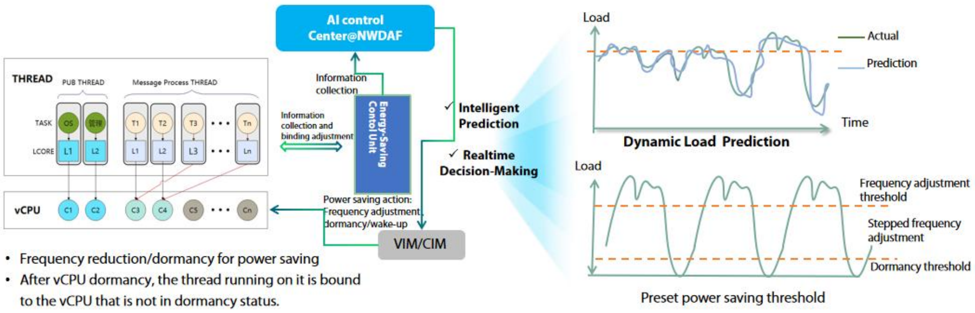

AI control Center@NWDAF
THREAD
PUB THREAD
Message Process THREAD
TASK
OS
管理
T1
T2
T3
Tn
LCORE
L1
L2
L1
L2
L3
Ln
vCPU
C1
C2
C3
C4
C5
Cn
Information collection
Information collection and binding adjustment
Energy-Saving Contol Unit
✓ Intelligent Prediction
✓ Realtime Decision-Making
Power saving action: Frequency adjustment, dormancy/wake-up
VIM/CIM
• Frequency reduction/dormancy for power saving
• After vCPU dormancy, the thread running on it is bound to the vCPU that is not in dormancy status.
Load
Actual
Prediction
Time
Dynamic Load Prediction
Load
Frequency adjustment threshold
Stepped frequency adjustment
Dormancy threshold
Preset power saving threshold

Fig. 5-9 Intelligent Power Saving

AI application: Based on the historical load data collected from the NF, the AI control center uses the AI algorithm to calculate the load data of the NF in the subsequent time period.

In dormancy mode, the threads that originally run on multiple vCPUs are migrated to a small number of centralized threads. Because the number of running threads is not reduced, the original load sharing is not changed, services are basically not perceived, and user experience is not affected.

5.5.2 AI+ Network Technology to Optimize Network Signaling Load

Paging signaling accounts for a large proportion in network signaling. Therefore, paging signaling is continuously optimized during network development. During this process, the balance between reducing paging signaling and reducing paging success time must be considered. In most cases, the paging success time can be reduced through large-scale paging, and the paging success time can be reduced through small-scale paging first and then step-by-step paging. Therefore, reducing paging signaling and reducing paging success time is a pair of contradictions. It is difficult to achieve the optimal balance through the traditional fixed policy. The AI technology is introduced to achieve optimal balance based on user track profiles.

In the intelligent paging solution, AMF has a built-in AI, which calculates the base station where the user is most likely to stay in the current paging time based on the user's historical activity track, and then initiates paging to the base station or base stations with the highest probability to achieve the optimal balance between paging efficiency and paging time.

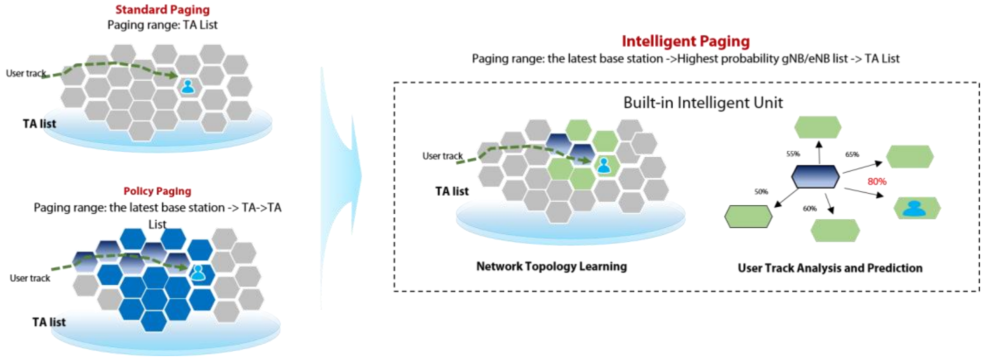

Standard Paging
Paging range: TA List
User track
TA list
Policy Paging
Paging range: the latest base station -> TA->TA List
User track
TA list
Intelligent Paging
Paging range: the latest base station ->Highest probability gNB/eNB list -> TA List
Built-in Intelligent Unit
User track
TA list
55%
65%
80%
50%
60%
Network Topology Learning
User Track Analysis and Prediction

Fig. 5-10 Intelligent Paging

5.6 Native Security to Enhance Connectivity Resilience

With the development of mobile networks, connections are increasingly large in scale, and massive connections easily trigger signaling storms, bringing huge risks to network security. For example, Operator R in Canada found that its networks are disconnected on a large scale for about 19 hours. Network-wide services of Operator K in Japan were down for 62 hours. Service stability is the basis of the communication industry. Network breakdown not only causes direct revenue loss and huge compensation, but also affects the brand reputation and even survival and development of operators. Therefore, it becomes the consensus for operators to prevent and respond to signaling storms. The AI technology can effectively prevent and respond to signaling storms.

5.6.1 Identifying Abnormal UEs Intelligently to Prevent Signaling Storms

With rapid development of 5G networks and increasingly complicated network attack methods, the frequent occurrence of abnormal terminals becomes a major potential risk to network security. Therefore, the intelligent identification technology can detect and isolate abnormal terminals in a timely manner to effectively prevent network attacks. It is a key measure to ensure stable network operation and avoid signaling storms.

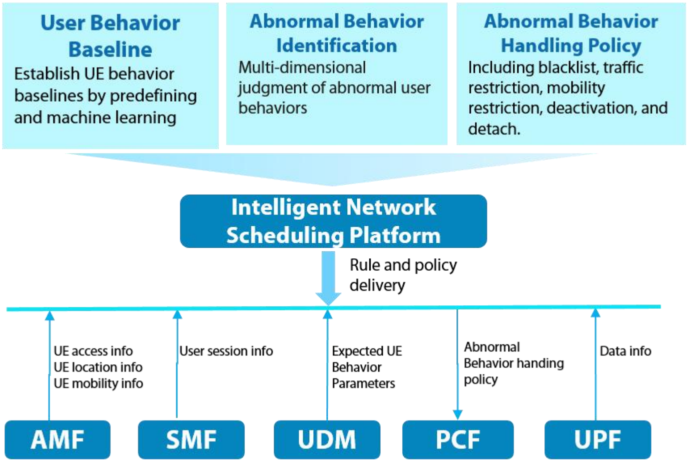

User Behavior Baseline
Establish UE behavior baselines by predefining and machine learning
Abnormal Behavior Identification
Multi-dimensional judgment of abnormal user behaviors
Abnormal Behavior Handling Policy
Including blacklist, traffic restriction, mobility restriction, deactivation, and detach.
Intelligent Network Scheduling Platform
Rule and policy delivery
UE access info
UE location info
UE mobility info
User session info
Expected UE Behavior Parameters
Abnormal Behavior handing policy
Data info
AMF
SMF
UDM
PCF
UPF

Fig. 5-11 Identification and Protection for Abnormal Terminals

1. NWDAF has a built-in intelligent scheduling platform. It collects the control-plane and user-plane data reported by each NE, trains and generates a baseline for normal user behavior through the AI algorithm.
2. Based on the behavior baseline, NWDAF further determines whether a current active user exceeds the baseline, and then determines that this user is an abnormal terminal. For the determined abnormal terminal, NWDAF forbids the user from accessing the network for a period of time (parking user), forbids the user from using data services (service forbidden), and separates the user based on the predefined processing policy.

5.6.2 Generating a Signaling Storm Response Plan Based on Digital Twin

Driven by AI technologies, digital twin networks provide unprecedented network problem detection and prevention capabilities. By accurately replicating the signaling processing capability of the production network and using the AI algorithm to predict flow growth and signaling changes, the digital twin network identifies network bottlenecks in advance, simulates signaling shocks in abnormal scenarios, provides a scientific basis for network performance optimization and signaling storm prevention, and ensures that the network can respond quickly and provide targeted optimization suggestions in the face of network pressure.

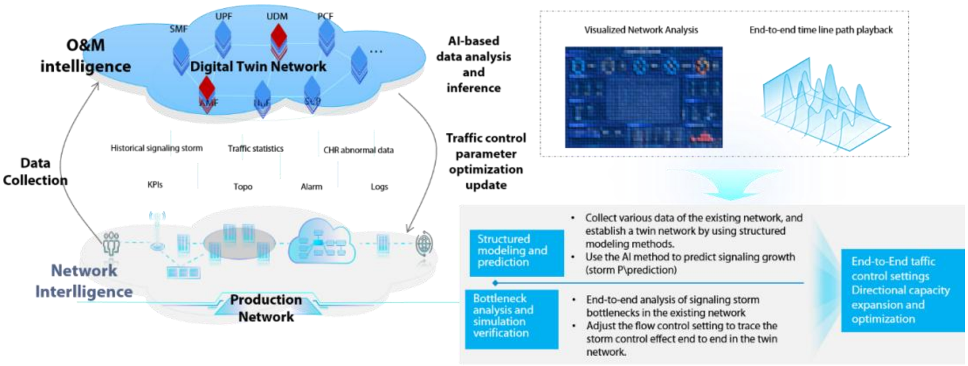

O&M
intelligence
SMF
UPF
UDM
PCF
...
Digital Twin Network
AMF
NRF
SCP
AI-based
data analysis
and
inference
Visualized Network Analysis
End-to-end time line path playback
Historical signaling storm
Traffic statistics
CHR abnormal data
Data
Collection
KPIs
Topo
Alarm
Logs
Traffic control
parameter
optimization
update
Network
Interlligence
Production
Network
Structured
modeling and
prediction
Collect various data of the existing network, and establish a twin network by using structured modeling methods.
Use the AI method to predict signaling growth (storm P\prediction)
Bottleneck
analysis and
simulation
verification
End-to-end analysis of signaling storm bottlenecks in the existing network
Adjust the flow control setting to trace the storm control effect end to end in the twin network.
End-to-End taffic
control settings
Directional capacity
expansion and
optimization

Fig. 5-12 Signaling Storm Simulation and Protection Based on Digital Twin

1. When a digital twin network is deployed, the signaling processing capability of each twin node in the network complies with the capability of each node in the production network.

2. Collects flow data from the production network, and predicts the end-to-end signaling growth by using the AI method. For example, when the number of 5G registrations increases, the corresponding interaction signaling messages on the UDM side also increase. According to the analysis, with the increase of the number of users and the flow model, the nodes and interfaces with bottlenecks first appear in the production network. Countermeasures can be deployed in advance.

3. Simulates a sudden signaling impact when the network is abnormal (for example, a node is faulty). The network recovery time and the recovery time of each node can be simulated under the signaling impact. The system can provide optimization suggestions for the node with the largest impact. For example, when the network is faulty, the UDM suffers a huge impact due to a large number of re-registration users. To reduce the impact, it is recommended to limit the number of messages that the

AMF sends to the UDM, and provide a specific recommended value in accordance with the UDM capability.

5.7 Continuous Evolution to 6G, from Scenario-Based to All-round AI

AI is introduced on the basis of existing network connections to meet the challenges facing some current networks. The advantages of introducing AI in this way are as follows: (1) Applying AI in specific scenarios makes the model more targeted and efficient. (2) The impact on the existing architecture and processes is small, so it is easier to deploy. However, the disadvantages of the overlay mode are also obvious: (1) The model has poor universality or generalization capability. In each new requirement scenario, the model needs to be re-developed, and the commissioning period is long. (2) There are a large number of models, and model management is complicated. Therefore, after the network evolves to 6G network, an internal AI is required to collect data from each unit of the 6G system and complete system-level large-scale model training. The large-scale communication model is applied to each decision point of the 6G system, so that network decision making becomes more intelligent and efficient, such as resource allocation, load balancing, and path selection.

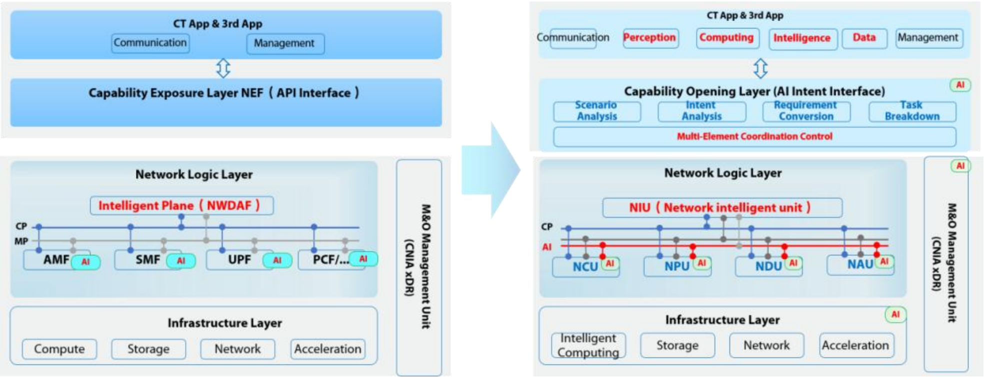
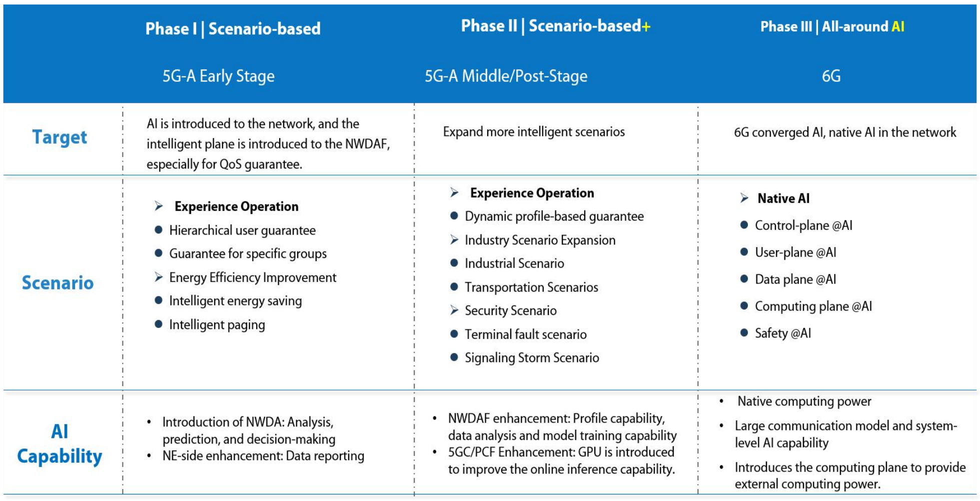


Fig. 5-13 Evolution from the 5G Intelligent Architecture to the 6G Intelligent Architecture

The evolution from 5G AI+ connectivity to 6G intrinsic connectivity is a process of continuous evolution. During the evolution process, the application scenarios of AI are continuously expanded. For example, from a public network to an industrial network, AI technologies are continuously enhanced, such as from simple machine learning to federated learning in horizontal and vertical fields, from offline training and online inference to online training and inference.



	Phase I | Scenario-based 5G-A Early Stage	Phase II | Scenario-based+ 5G-A Middle/Post-Stage	Phase III | All-around AI 6G
Target	AI is introduced to the network, and the intelligent plane is introduced to the NWDAF, especially for QoS guarantee.	Expand more intelligent scenarios	6G converged AI, native AI in the network
Scenario	➢ **Experience Operation** ● Hierarchical user guarantee ● Guarantee for specific groups ➢ Energy Efficiency Improvement ● Intelligent energy saving ● Intelligent paging	➢ **Experience Operation** ● Dynamic profile-based guarantee ➢ Industry Scenario Expansion ● Industrial Scenario ● Transportation Scenarios ➢ Security Scenario ● Terminal fault scenario ● Signaling Storm Scenario	➢ **Native AI** ● Control-plane @AI ● User-plane @AI ● Data plane @AI ● Computing plane @AI ● Safety @AI
AI Capability	• Introduction of NWDA: Analysis, prediction, and decision-making • NE-side enhancement: Data reporting	• NWDAF enhancement: Profile capability, data analysis and model training capability • 5GC/PCF Enhancement: GPU is introduced to improve the online inference capability.	• Native computing power • Large communication model and system-level AI capability • Introduces the computing plane to provide external computing power.

Fig. 5-14 Network Intelligence Evolution

6 "AI+" O&M: Reshaping the O&M Paradigm

The digital transformation of communication and industries is accelerating, new networks, new services, and new technologies are emerging, network structures are becoming more complex, service applications are diversified, O&M management is more complicated, and the importance of network security and privacy protection is becoming increasingly important. Traditional manual O&M cannot meet the current O&M requirements due to limited efficiency and capabilities. In this context, automation and intelligence of O&M become the consensus of the industry, and intelligent O&M becomes the key to maintaining competitiveness.

The emergence of emerging technologies such as intent network, AI large model, and digital twin brings new opportunities for intelligent O&M. These technologies have many advantages, such as humanized man-machine interaction mode, massive formatted data processing, high-precision analysis and prediction, and high-fidelity simulation verification. By introducing the AI large model, intent network, and digital twin technologies, we can build “AI+” O&M, reshape the O&M model of core network, and promote operators to evolve from L3-level O&M of TMF Autonomous Networks to L4-level advanced autonomous O&M.

6.1 Network O&M Challenges: 3-Multiple, 3-New, and 3-Cross

With the development of 5G networks and the introduction of virtualization and cloudification technologies, communication networks are becoming more and more complicated, communication services are becoming more and more diversified, and infrastructure and service systems are facing various complicated situations. The main changes are as follows:

- 3-Multiple: Multiple access, multiple types, and multiple NEs resulting in high maintenance costs. For most operators, 2G/3G/4G/5G multi-access networks, data and voice networks, and public and private networks coexist, multiplying the number of NEs and interfaces, making O&M more difficult and increasing OPEX costs.
- 3-Cross: Cross-layer, cross-domain, and cross-vendor result in complicated O&M, and difficult manual O&M. The virtualization architecture includes the hardware layer, virtual layer, and network function layer. The products at each layer can be provided by multiple vendors. The end-to-end network also has multiple domains such as wireless domain, bearer domain, transmission domain, and core network domain. To achieve end-to-end O&M and fault delimitation, cross-layer, cross-domain, and cross-vendor cooperation is required, resulting in exponential increase in the complexity and difficulty of O&M management. The manual operation period is long, the efficiency is low, and the error rate is high. It is difficult to deal with O&M in complicated scenarios.

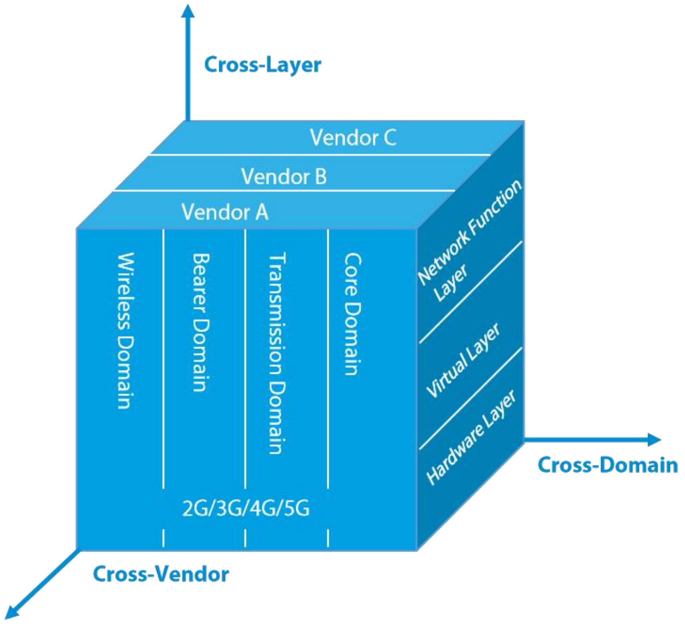


Fig. 6-1 Cross-Network O&M Challenges

- 3-New: New services, new architectures, and new technologies need to reshape the O&M model.

New services: New services are constantly emerging for 5G networks, such as New Calling, XR, Metaverse, industrial interconnection, and drone. These services are launched in a short time, and need to be updated frequently. They are often accompanied by massive data, and have extremely high requirements for real-time and availability. In addition, customers in different industries have different service requirements and SLA requirements. Therefore, intelligent O&M is urgently needed to precisely manage each network and carry out intelligent O&M in accordance with the requirements of different networks to ensure high stability and security, ensure continuous service operation, and improve user satisfaction.

New architecture: The Service Based Architecture separates network functions into multiple independent services, and each service needs to be monitored, managed, and troubleshot respectively. When a service procedure changes, for example, a new service scenario requires different service combinations, the service orchestration policy can be adjusted in a timely manner. Traditional manual monitoring methods cannot meet these requirements. Automatic monitoring methods are required to obtain the status information of each NE in real time, such as performance indicators and fault alarms. In addition, intelligent methods are required to analyze the associations between NEs, so that the root cause of a fault can be quickly determined when it occurs.

New technologies: After the emergence of emerging technologies such as AI large models, intent networks, and digital twins, industry standard organizations and operators are actively exploring how to apply these new technologies to network O&M to achieve intelligent O&M of the core network. For example, you can use the powerful analysis capability of the AI model to predict network faults, or use the digital twin technology to simulate and optimize the network.

6.2 "4-Layer 1-Entity" AI+ O&M Architecture, Evolving to High-Level Autonomous Network

To actively meet the challenges of core network O&M and meet the industry's requirements for automation and intelligence, a new “4-layer 1-entity” intelligent O&M architecture based on intent-based networks, AI large and small models, and digital twins needs to be built to achieve full open decoupling of networks, data, models, and applications and push networks into the high-level autonomous phase. As shown in the following figure, through decoupling, coordination, and smooth evolution of large and small models and heterogeneous models, this architecture creates a digital and intelligent technology base for flexible orchestration to enable efficient O&M. Digital twins integrating data, models, and applications, empower planning, construction, O&M, and optimization, and precisely build highly stable networks. The unified O&M portal covers the monitoring center, troubleshooting center, emergency center, and network change center, empowers high-value O&M scenarios such as fully closed-loop network change, fully closed-loop monitoring and troubleshooting, and fully closed-loop complaint handling, promotes the transformation of network O&M from passive emergency mode to active and efficient mode, and continuously reduces OPEX O&M costs.

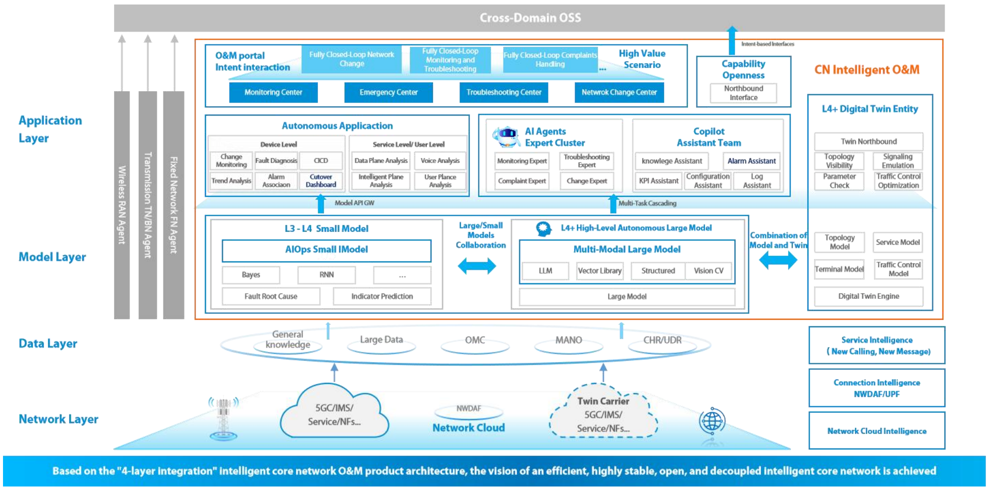


Fig. 6-2 Intelligent O&M Architecture of Integrated Core Network

This system architecture includes four layers: Network layer, data layer, model layer, application layer, and digital twin entity.

1. Network layer: Existing atomic NEs in the core network are included in the 2G/3G/4G/5G/IMS. These atomic NEs are the basic elements for building the entire core network to provide the most basic physical or logical entity support for network operation.

2. Data layer: This layer plays a key role in providing high-quality corpus data in the entire architecture, including the general knowledge base, big data, OMC, MANO, and CHR/UDR. These NEs collect, sort, and store various types of data, providing abundant data resources for upper-layer intelligent analysis and O&M decision-making.

3. Model layer: It is one of the core parts of the intelligent O&M architecture of the core network. It uses AI large and small models as powerful intelligent engines to build a digital and intelligent technology base that can be assembled, orchestrated, and iterated independently. By introducing AI large and small models, a new enabling O&M system with a AI large model as the core and a small model agent as the intelligent application is built. Through the accumulation of multiple years of experience of product experts, ZTE can provide large model services oriented to interaction, analysis, and generation.

4. Application layer: A digital and intelligent technology base built based on the

intelligent layer can intelligently orchestrate and generate autonomous applications, AI Agent expert clusters, and Copilot assistant teams. Through the combination of intent interaction, content generation, and multi-task concatenation, the capability requirements of core network in multiple complicated O&M scenarios such as planning, construction, maintenance, optimization, and operation can be easily met. For example, in the network planning phase, a network construction solution can be intelligently orchestrated in accordance with service requirements and resource conditions. In the O&M phase, faults can be rapidly located, and solutions can be provided.

5. Digital twin: A combination of "service models and twin applications" is built through the digital twin simulation platform. This architecture facilitates rapid deployment of twin applications, agile service innovation, and precise construction of highly stable networks. For example, you can simulate an NE ontology in a digital twin environment to build an NE twin, simulate a service process, discover potential problems in advance, and optimize solutions.

Relying on the "4-layer 1-entity" architecture, the core network intelligent O&M system supports AI Native, and has the capabilities of intelligent network perception and monitoring, intelligent network analysis and diagnosis based on the AI large model agent. It has the capabilities of intelligent network evaluation and decision-making based on digital twins, intelligent network change and execution, and forms a fully closed-loop control mechanism to help networks implement self-configuration, self-optimization, and self-healing.

6.3 Large and Small AI Models Coordination for Efficient O&M

The large AI model can perform high-precision data analysis and prediction based on massive data, rapidly answer various user questions in natural languages, improve work learning efficiency, and improve man-machine interaction experience. It can also efficiently process texts and inspire innovative solutions, such as document generation and analysis, to create convenience and value for users in multiple aspects.

As the communication network and services become more complicated, core network O&M personnel need to master the professional knowledge of multiple products and have strong fault analysis and solution capabilities. However, in the face of massive data and many fault points, manual O&M is difficult to locate and troubleshoot faults quickly.

To reduce the learning and O&M costs of O&M personnel and improve the intelligent O&M analysis and solution generation capabilities of the system, the "4-layer 1-entity" core network intelligent O&M system builds large model services such as intelligent interaction, intelligent analysis, and intelligent generation by relying on the AI large language model capability in the communication field and the AI agent service, combined with general communication knowledge and the refinement material set of years of experience of product experts, to meet the requirements of complicated O&M scenarios such as planning, construction, maintenance, optimization, and operation, and promote the evolution of the network to L4 and higher-level autonomous network.

This system builds a unified intelligent interactive portal through the Agentic RAG technology, and implements communication knowledge FAQ and network operation status check. Through the AI large and small model collaboration technology that combines generalized intent understanding and precise O&M, the system implements fast and efficient intelligent analysis, and provides fault diagnosis, network optimization, and routine check functions. The intelligent generation technology based on reinforcement learning provides efficient text processing capabilities, and can quickly generate documents such as major operation solutions and observation reports.

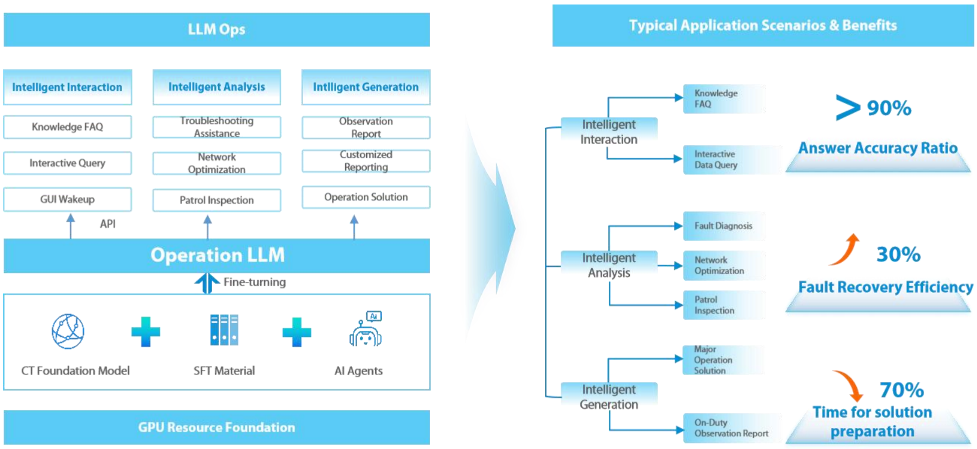


Fig. 6-3 High-Value Application Scenario Based on the LLM Large Model

- **Agent-Coordination Retrieval Enhancement Generation Technology (Agentic RAG)**

First, this technology builds a unified knowledge acquisition portal through the general RAG capability, and enables users to quickly master the professional knowledge of 2G, 3G, 4G, and 5G products through the LLM large model based knowledge FAQ.

Second, the enhanced Agentic RAG integrates the AI agent into the RAG process, and coordinates components to perform more operations beyond simple information retrieval and generation. Therefore, the answer accuracy can reach more than 95%, and the answer contents are more abundant in terms of elements, pictures, and texts. This facilitates retrieval, improves retrieval efficiency, and reduces the learning costs of O&M personnel.

In addition, the Agentic RAG technology can automatically identify the KPI diagnosis and analysis intentions of O&M personnel, invoke the API of different O&M systems in self-adaptation, and quickly obtain data through the NL2SQL. This efficient, flexible, and personalized mode helps create an intelligent interactive portal, so that everyone can instantly become a "communication expert."

- **Large and Small Models Coordination Technology With Integration of Generalized Intent Understanding and Accurate O&M**

By providing a lightweight algorithm engine, iterative data learning, and service experience injection, a visual knowledge map is built to improve the efficiency of establishing O&M event association rules. Based on AI iterative learning and expert experience, one million alarm or event rules can be mined within one hour. Based on the aggregation of associated alarm events, the number of alarms can be greatly reduced, and the alarm aggregation compression rate can reach 80% or above.

Large models are usually used to understand generalization intentions, while small models are used to accurately solve actual O&M breakpoints. Through the collaboration between large and small models, the system can achieve full closed-loop fault detection, fault location, fault escape, and recovery, and assist in network optimization and routine checks. Practice has proved that the fault detection time can be shortened to one minute, the fault location time can be reduced by 50%, and the fault recovery time can be reduced by 50%.

- **Reinforcement learning-based intelligent generation technology**

The intelligent generation technology based on reinforcement learning solves the problem of document editing for beginners. Editing documents such as major solutions and intelligent reports is a tedious and challenging task for beginners.

Based on the LLM (large language model) generation capability, this technology can archive global network planning and construction solutions to data lakes. The large model is then used to learn historical planning data, and then generate planning model parameters and solutions in the telecom field for on-site communication with customers to improve user experience.

It is verified that the time to accurately generate a solution by using this technology can be reduced by 70%, and the generated solution has high accuracy and reliability.

6.4 Integrated Digital Twin to Build A Highly Stable Network

With the digital twin technology, the "4-layer 1-entity" core network intelligent O&M system integrates the large language model LLM, the AI agent, and the Multimodal Large Language model (MM-LLMs) to build a digital twin, provide a digital twin model of core network systems, devices, and components, and have the cross-vendor model management capability. It builds a twin model through orchestration or low-code development, obtains calculation results by invoking the vendor's model information, and provides the presentation capability (visualizing the twin calculation results on the user portal) and northbound capability (invoked by the operation support system). It also provides visual, simulation, prediction, and policy feedback capabilities for O&M, provides qualitative and quantitative analysis capabilities at low costs, supports the transformation from manual decision-making to machine decision-making, and empowers high-level autonomous network evolution to achieve automatic closed-loop O&M.

In addition, to build a highly stable network, the digital twin builds a system architecture integrating data, models, and applications, empowering application scenarios such as planning, construction, O&M, and optimization. In a mobile communication network, the core network, which is located at the center, is a necessary channel between terminals and service servers. Its stability and reliability directly affect the availability of network services. If a signaling storm occurs in the core network, the network cannot be used for a long time, and services are interrupted for a long time, affecting a wide range of areas. To enhance the capability of the core network to resist signaling storms, the digital twin technology can be used to twin the status and NE behaviors of the core network, simulate network faults and events, and rehearse the signaling storm process, so as to discover and optimize weak points in the network and improve the anti-risk capability of the network.

- **Multi-Service Simulation Verification and Optimization**

The digital twin uses the built twin simulation network to simulate and verify multi-service scenarios and solutions. For example, it is feasible to simulate the network performance in a high-traffic service scenario, or verify whether the new network optimization solution.

The AI algorithm and statistical method are used to analyze and process the existing network data, and obtain key network information models such as NE mechanism, network signaling, network topology, route weight, network traffic, terminal recovery behavior, and NE recovery behavior. Based on the twin technologies, digital simulation modeling of CN networking status and configurations is completed.

Through flexible orchestration of fault events such as NE operation status and network signaling traffic of the twin core network, the network signaling storm impact caused by multiple abnormal scenarios can be simulated, and the network bottlenecks and their impacts in the signaling storm and service recovery process can be analyzed and located to implement the twin core network.

Because the NEs and terminals of different vendors in the existing network have different operation mechanisms and recovery behaviors, it is necessary to perform personalized learning and modeling for the network information of different vendors in the existing network.

- **Hierarchical Model Construction and Mapping**

The digital twin model is an important part of the digital twin network. Based on perceived network data, it maps the network of a physical entity to virtual space, and constructs a twin digital network that is the same as the physical entity.

The construction of the digital twin model focuses on the classification and merging of multi-source heterogeneous data. You can build digital twin models at the corresponding network level for different network domains. These models include basic attribute information and scenario function information, and can receive instructions and feed back events.

To adapt to the sharing and difference of models in different scenarios, the hierarchical model construction and mapping method is used. This method divides the digital twin model into different layers, each of which has its own functions and characteristics. Together, these models implement comprehensive mapping and simulation of the physical world, helping to clearly understand and construct each part of the digital twin.

Digital twin model construction includes basic model construction and service model construction, as shown in the following figure. The basic model layer defines a model of a single physical network entity, and builds a digital twin model from multiple dimensions, such as an operation rule model, an attribute definition model, and a data model. This is closely related to the types of physical network entities. Service models assemble and integrate basic models for specific twin application scenarios, and build or extend digital twins from different dimensions by using collected network data. In accordance with the functions implemented, the models can be divided into topology rule model, traffic fidelity model, network path model, event detection model, network quality model, and twin management and control model.

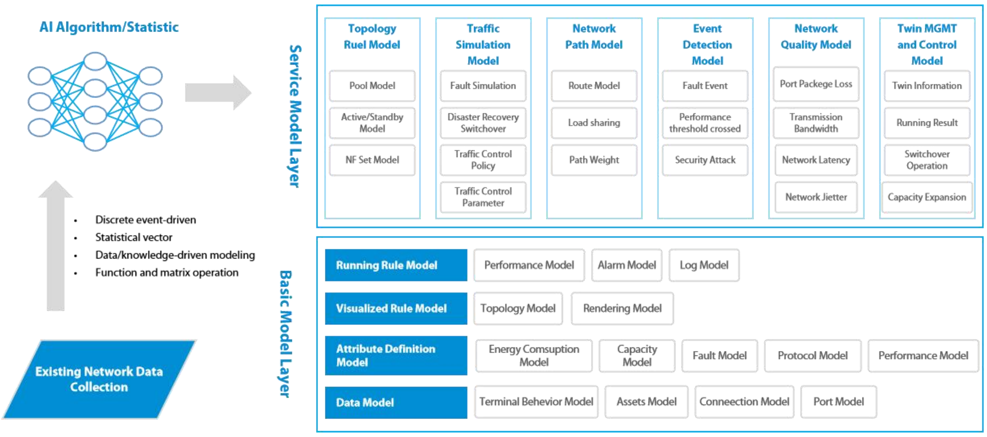


Fig. 6-4 Target Solution of the Digital Twin Model

- **Twin Self-Evaluation and Decision-Making Mechanism**

The achievement and implementation of digital twin network objectives are not achieved overnight, but need continuous objective pulling and continuous iterative promotion.

In the digital twin technology, the decision-making and coordination mechanism of the twin entity is an important sign of the digital twin system reaching a high-level phase, which reflects the development of the digital twin in a more intelligent direction on the basis of simulation, monitoring, and analysis.

In the closed-loop evaluation of the capabilities of the digital twin network, a quantitative evaluation indicator system and a closed-loop mechanism of "evaluation, analysis, improvement, and re-evaluation" shall be established. It is an effective means to test the effectiveness of capability building, promote the improvement of the capabilities of the digital twin network, and ensure the implementation of the top-level design of system planning.

In the scenarios of intelligent disaster-recovery switchover and signaling storm simulation and evaluation, the qualitative and quantitative methods are used. First, the characteristics, capabilities, and scenario requirements of each level of signaling impact simulation grading criteria of the digital twin network are decomposed, objective and quantifiable key performance indicators are sorted out, and evaluation methods are formulated. Then, the network impact twin simulation capability construction of each scenario is evaluated separately. Finally, the operator's digital twin network capability construction result is comprehensively evaluated based on the coverage of intelligent network operation scenarios. Through the closed-loop mechanism of "assessment, analysis, improvement, and re-assessment," the system will identify weaknesses, establish a problem list, and follow up problem solving. In this way, it will continuously promote the capability improvement of the digital twin network and the intelligent operation level in all scenarios.

6.5 Fully Closed-Loop O&M Management to Reduce O&M Costs

Fully closed-loop O&M management is an all-round and no-dead-angle management mode from problem discovery, analysis, and solution to feedback and optimization. It not only focuses on the immediate handling of faults, but also on identifying potential risks in advance through data analysis and prediction, so as to achieve proactive and intelligent O&M. The core value of this model is to build a self-learning and self-optimization O&M ecosystem by integrating monitoring, automatic tools, AI algorithms, and other multi-dimensional technical means.

In view of the challenges facing the core network and the key value scenarios recommended by TMF AN L4 and the actual production requirements of operators, the current fully closed-loop O&M management includes three high-value scenarios: Fully closed-loop network change, fully closed-loop monitoring and troubleshooting, and fully closed-loop complaint handling. The following figure shows the process and performance objectives.

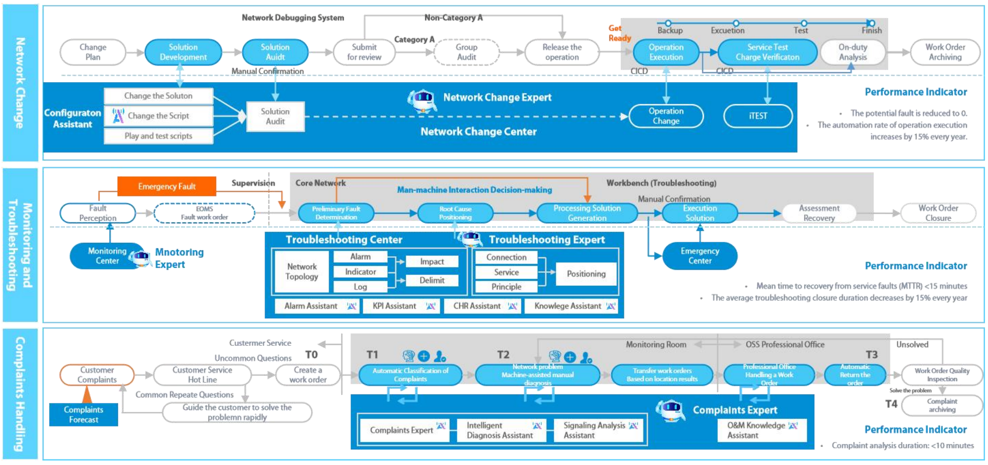


Fig. 6-5 High-Value Scenarios of Fully Closed-Loop O&M Management

- **Fully Closed-Loop Network Change**

In the O&M process of communication equipment, a fully closed-loop network change is a critical end-to-end scenario, involving the adjustment and optimization of the network architecture, equipment, configuration, and security.

Network change starts from the formulation of the plan, and goes through many steps before, during, and after the change, including solution formulation, submission for review, release operation, operation execution, service verification, on-duty observation analysis, and finally archiving of change work orders. In this end-to-end process, there are steps such as manual change and manual review, which affects the timeliness of change operations.

Intelligent network change and execution can achieve process orchestration and operation automation. For example, major operations such as upgrade, cutover, and capacity expansion can achieve full-process automation, minute-level template design, and scheduled and batch task execution. Each NE operation has a large monitoring screen. The process can be defined and monitored. Service indicators can be displayed in minutes. Preventive maintenance tasks can be quickly customized to implement one-click preventive maintenance. With the intelligent network change flow, the potential faults of the entire operation can be reduced to zero, and the automation rate of the operation steps can increase by more than 15% year by year.

- **Fully Closed-Loop Monitoring and Troubleshooting**

Full closed-loop monitoring and troubleshooting covers the entire data transmission process from the source to the terminal. During this process, the monitoring system traces the operational status of communication devices in real time, collects and analyzes operation data, discovers potential problems or faults in a timely manner, and takes corresponding measures to ensure the stable operation of communication devices. Generally, network monitoring and troubleshooting begin with the monitoring process, and carry out multi-dimensional and multi-level health monitoring to implement intelligent indicator abnormality detection and intelligent identification of faults/hidden dangers. Through the analysis of the fluctuation, periodicity, and trend of indicators, an indicator abnormality identification model is automatically built. Through intelligent fault analysis, alarm aggregation and KPI abnormality analysis capabilities are provided to implement intelligent fault analysis and location and cross-layer network fault diagnosis in multiple scenarios and in multiple dimensions, and automatically output diagnosis diagrams and root causes.

- The alarm aggregation solution based on the large alarm model provides the lightweight algorithm engine, data AI iterative learning, service experience injection, and associated alarm aggregation rules. Alarms are aggregated and associated through aggregation rules and spatial relationships to improve alarm rule mining efficiency, greatly reduce the number of alarms, and improve fault location efficiency.

- Based on the exception analysis of the large KPI model, the system automatically understands the KPI analysis intention of O&M personnel, and automatically adapts and invokes different O&M system APIs. The system assists O&M personnel in analyzing abnormal KPI exceptions within several minutes, implementing efficient self-service exception analysis and O&M, and improving efficiency by 50%. In personalized fault analysis scenarios, any scheduling or concatenation can be performed, and the coverage rate can be doubled.

- The AI safe production solution provides NE-level and network-level security protection solutions, and establishes an intrinsic security mechanism.

Through the security policy center, intelligent micro-isolation of resources, active intrusion detection, virus protection, and asset security management are implemented to achieve intelligent micro-isolation and fast security isolation for hosts with abnormal traffic. Intrinsic security implements active defense and detection for 5G network access and operation, enhances automatic deployment of security policies, and enhances the intrinsic security capability of the VNF.

– The one-click fault escape mechanism of the emergency center can preset multiple fault emergency plans, implement unified management and centralized control, and implement automatic and visual emergency operations to shorten the fault recovery duration.

- **Fully Closed-Loop Complaints Handling**

In the O&M process of communication devices, complaints handling is an important end-to-end scenario. It usually includes complaints receiving and analysis, complaints problem sorting and dispatch, complaints content analysis, and complaints result receipt. In the communication field, CHRs and related signaling involved in user complaints need to be extracted and analyzed to accurately locate user complaints.

However, due to the complicated procedure, complicated system invocation, and difficult problem location, it takes a long time to complete problem analysis, resulting in poor user experience. To improve this situation and improve the complaint handling efficiency, it is necessary to introduce a better handling mechanism in key steps.

– With intent-driven complaints handling, with the help of relevant static experience knowledge, complaints case retrieval, complaints associated data query, intelligent signaling analysis, single-domain atomic capability invocation, and large-model comprehensive inference, the capability of a professional office is moved forward to the monitoring room, breaking the limitation of professional office skills and shortening the duration of complaints analysis and handling.

– Use the time range, service type, specific interface, and user number as the query conditions to check whether the data collection in the existing network is complete and whether the clock is synchronized, so as to ensure that the data quality meets the requirements for further analysis.

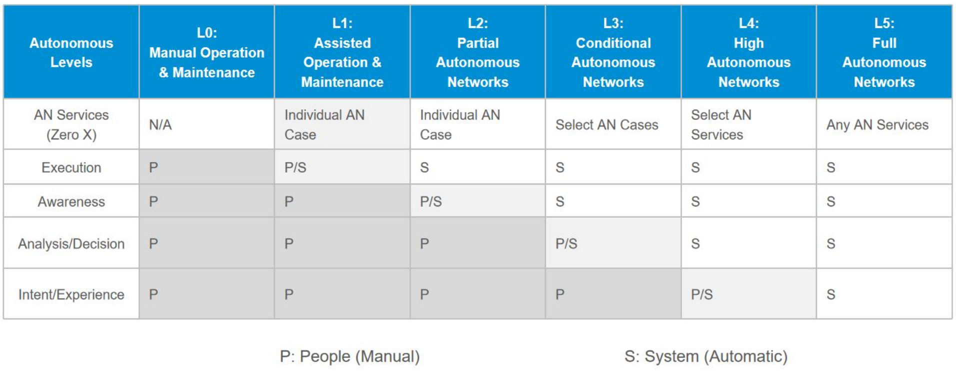
– Based on CHRs, the system generates failure signaling in accordance with the failed CHR association, parses the failure cause code, translates the failure cause, and then provides a solution.

In the above typical fully closed-loop O&M scenario, the monitoring center, troubleshooting center, emergency center, and change center in the four-layer integrated architecture are fully utilized. These four systems strongly support the "AI+" intelligent O&M capabilities of the core network, and implement four core capabilities: Intelligent network perception and monitoring, intelligent network analysis and diagnosis, intelligent network evaluation and decision-making, and intelligent network change and execution. They promote network self-healing, self-optimization, and operation automation, thereby reducing Opex O&M costs.

6.6 Continuous Evolution for "Unmanned" AI+ O&M

The TM Forum defines the autonomous network level to guide the automation and intelligence of networks and services, evaluate the value and gain of autonomous network services, guide the intelligent upgrade of operators and vendors, and describe in detail the grading evaluation methods and processes, task evaluation standards, and scoring methods of autonomous networks. All operators are actively promoting the practice of autonomous networks. The goal is to achieve L5-level full-process system automation, and all scenarios will be automatically completed by the system to achieve an ideal "unmanned" intelligent O&M model.

However, in actual applications, "unmanned" O&M will encounter challenges such as technological complexity, security and privacy, the necessity of human intervention, and supervision and compliance. Therefore, in the future, the evolution of O&M can be continuously enhanced in key technologies such as multi-dimensional agent coordination and orchestration, multi-modal large-model-based adaptive agent coordination, and dual-wheel-drive foundation based on AI+ digital twin, to achieve more efficient O&M, build more stable networks, and continuously reduce O&M costs.



Autonomous Levels	L0: Manual Operation & Maintenance	L1: Assisted Operation & Maintenance	L2: Partial Autonomous Networks	L3: Conditional Autonomous Networks	L4: High Autonomous Networks	L5: Full Autonomous Networks
AN Services (Zero X)	N/A	Individual AN Case	Individual AN Case	Select AN Cases	Select AN Services	Any AN Services
Execution	P	P/S	S	S	S	S
Awareness	P	P	P/S	S	S	S
Analysis/Decision	P	P	P	P/S	S	S
Intent/Experience	P	P	P	P	P/S	S

P: People (Manual) S: System (Automatic)

Fig. 6-6 Autonomous Network Level Defined by TM Forum

- **The multi-dimensional agent collaboration orchestration technology is an innovative guarantee for "unmanned" O&M applications**

Intelligent integration and intelligent connection are important visions of the future network. In addition to serving as a connection infrastructure, the future network should also support AI based on the native design at the architecture layer to provide AIaaS services for users. Its service scope includes not only connection services, but also intrinsic computing, data, and AI services.

In the O&M process, key technologies such as AI, digital twin, and large model must be combined and orchestrated. After sending service requests to the network, multiple agents orchestrate service procedures in accordance with requirements, deploy them to network nodes that meet capability requirements, and finally output global decisions to indicate agents. These technologies can provide support for a large number of intelligent and physical coordination requirements (such as self-discovery, self-healing, and automatic reporting) in intelligent O&M application scenarios.

In terms of joint optimization of multi-dimensional resources, the enhanced learning technology is introduced to perceive the dynamic changes of production networks and resources and achieve the optimal match with user requirements. This is also the innovative guarantee of "unmanned" O&M applications.

- **The multi-modal large-model-based adaptive agent collaboration technology is a key factor for improving the capability of "unmanned" O&M models**

This technology integrates large multi-modal models and adaptive agents for collaborative control. With the capability of a large multi-modal model, this technology enables an agent to understand and process information in multiple modes, and automatically adjust policies and behaviors in accordance with the information and environment changes to achieve coordination between multiple agents.

The large multi-modal model can process and understand various types of information, such as text, picture, audio, and video. It also covers the communication O&M field, including tables, logs, and graphic code streams. This model can perform more complicated and intelligent tasks, such as visual FAQ, image and text generation, speech recognition and synthesis, and video understanding and generation. In combination with an adaptive agent, the agent can automatically adjust policies and behaviors in accordance with environment changes to achieve better coordination. In adaptive collaborative control, each agent has its own objectives and behavior strategies. When affected by other agents, each agent can adjust its own strategies to adapt to new situations, making multi-agent systems more flexible and adaptable.

This technology can be applied in multiple fields, such as robot control, smart home, and smart city. In these scenarios, the agent needs to understand and process information in different modes, and performs self-adaptive adjustment and coordination in accordance with information and environment changes. In O&M scenarios, there are increasing requirements for the input and output of multiple intelligent interfaces, regardless of network changes, troubleshooting, or complaint handling. By using the multi-modal large-scale model and self-adaptive agent collaboration technology, an agent can fully understand different environments and task requirements in O&M scenarios, and automatically adjust policies and behaviors in accordance with environment changes to achieve better coordination effect.

However, this technology also faces many challenges, such as data alignment and integration, model selection and training, and computing resource requirements, which need to be gradually optimized and enhanced during the continuous evolution of underlying system technologies.

- **The integrated dual-wheel-drive foundation based on AI+ digital twin is an intelligent engine for "unmanned" O&M.**

First, a single AI (such as neural network) has some limitations. When a neural network is used, the training data and test data must follow the same distribution, and the data set must be comprehensive and balanced. In this case, it is usually applicable to scenarios such as single-indicator prediction and network error detection. However, once the data distribution changes or the faulty data is insufficient, the effectiveness of the neural network is greatly reduced. Digital twin can solve this problem. By creating a high-fidelity dynamic twin model in the virtual twin space, the high-fidelity environment can generate simulation fault data without damaging the actual network. In network resource management and capacity optimization scenarios, the organic combination of AI+digital twin enables simulation, verification, and prediction capabilities to work together and support each other.

Second, digital expression of digital twins involves feature selection. In this case, an unsupervised/self-supervised deep learning method can be used to extract representative features from a large amount of unmarked data without prior knowledge, and the AI algorithm will assist in the operation of this entire process.

In future networks, the goal of "unmanned" O&M evolution focuses on intelligent, automated, and efficient data processing and analysis capabilities to achieve overall optimization and upgrade of O&M. The new “unmanned” O&M paradigm aims to use intelligent and automated methods to minimize manual intervention, thus improving O&M efficiency, reducing costs, and improving system reliability and stability. In this way, its huge potential in improving efficiency, reducing costs, and enhancing system stability is fully utilized

7 "AI+" Cloud Infrastructure: Reshaping the Computing Foundation

With the advent of ChatGPT, the emergence of Artificial Intelligence (AI) technology in a short period of time has become an inevitable trend in the intelligent transformation of core networks. As the computing infrastructure platform of the core network, the intelligent transformation of the cloud infrastructure is a critical link.

AI training tasks and inference applications require high performance, large-scale parallel computing, and low-latency interconnection. As a result, the cloud infrastructure evolves from traditional CPU-centric general computing to DPU/GPU/NPU-centric heterogeneous computing. Computing pooling, orchestration and scheduling, high-performance parallel storage access, and high-channel lossless network technologies are supported, and it is important to ensure efficient and stable resource supply. In addition, shielding the details of heterogeneous resources of underlying GPU, decoupling upper-layer AI framework applications, and computing native technologies of underlying GPU types are also the direction of future evolution.

In terms of deployment, core network NE applications require both general computing and intelligent computing resources. Therefore, the mixed deployment of intelligent computing and general computing resources of the AI+ cloud infrastructure is an important feature. In addition, the distributed deployment and coordination mode of central pre-training, regional refinement, and edge inference of cloud infrastructure intelligent computing resources is exactly the same as the traditional distributed deployment architecture of the central, regional, and edge computing. Therefore, the intelligent and smooth upgrade of the distributed cloud infrastructure architecture can fully meet the intelligent requirements of the core network.

7.1 Resource Pooling to Improve Infrastructure Resource Utilization Ratio

AI resource pooling refers to the pooling of computing and memory resources and the connection and access technologies of these pooling resources. Intelligent computing resources pooling is the key to build an efficient, flexible, and scalable intelligent computing center. The following is a detailed analysis of these pooling technologies and their access technologies:

1. Computing pooling

With the rapid development of AI、 big data technologies, GPU, as an important computing resource, is increasingly widely used in data centers. However, traditional GPU use modes have problems such as low resource usage and poor scalability. According to public statistics, the average GPU usage of the intelligent computing center in traditional mode is lower than 30%. In addition, the software and hardware between GPU devices of different manufacturers are bound to the vertical shaft barrier, further aggravating the problem of low GPU usage.

Therefore, the GPU resource pooling technology emerges. In essence, it abstracts the physical GPU resources of multiple manufacturers into a unified virtual GPU resource pool through software-defined hardware acceleration. Through GPU virtualization, multi-card aggregation, remote invoking, dynamic release, and other capabilities, it implements more efficient and flexible aggregation, scheduling, and releases massive AI acceleration computing. It precisely guarantees the AI model development, training, deployment, testing, and release of end-to-end computing allocation. The enabling resources can be fully utilized, the probability of fragmentation can be reduced, the overall effective computing can be increased, the computing service cost of the intelligent computing center can be reduced, and the overall efficiency of the intelligent computing center can be improved.

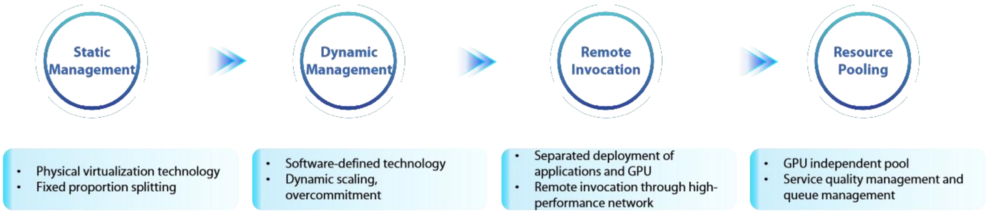

Static Management
Dynamic Management
Remote Invocation
Resource Pooling
• Physical virtualization technology
• Fixed proportion splitting
• Software-defined technology
• Dynamic scaling, overcommitment
• Separated deployment of applications and GPU
• Remote invocation through high-performance network
• GPU independent pool
• Service quality management and queue management

Fig. 7-1 Computing Pooling Capability Hierarchy

As shown in the above figure, from the perspective of maturity and flexibility of heterogeneous computing, the current computing pooling technology can be divided into the following capability levels:

- Static management: A single physical GPU is divided into multiple virtual GPUs in a fixed proportion, such as 1/2 or 1/4. The memory of each virtual GPU is equal, and computing is polled. This technology can solve the problem of sharing and using GPU resources on virtual machines. Typical examples include the MIG technology provided by NVIDIA on some Ampere series GPU in 2021. A100 can be split into a maximum of 7 portions.

- Dynamic management: A single physical GPU can be used as the target, and the physical GPU can be flexibly divided into two dimensions: Computing and video memory. The customized size (the minimum granularity of computing is 1%, and the minimum granularity of video memory is 1MB) can be implemented to meet the differentiated requirements of AI applications. In addition, this technology can fully adapt to the current trend of application cloud nativization, respond to the changes in resource requirements of upper-layer applications in real time, implement dynamic scaling of vGPU resources based on Scale-Up/Scale-Down, and implement GPU resource over-commit through dynamic resource mounting and dynamic resource release.

- Remote invocation: The AI application is deployed separately from the GPU server, and GPU resources can be remotely invoked through a high-performance network. AI applications can be deployed anywhere in the data center, and GPU resources can be invoked as long as the network is reachable, regardless of whether there is GPU on the application deployment node. In this case, the resource management scope is extended from a single node to the entire data center.
- Resource pooling: Supports independent pooling of CPU general computing and GPU intelligent computing. The converged resources in the two resource pools are independently invoked on demand, dynamically scaled in and scaled out, and released after being used up. With the pooling capability, AI applications can invoke GPU resources of any size in accordance with load requirements, and can even aggregate GPU resources of multiple physical nodes. After a container or VM is created, you can still adjust the number and size of virtual GPUs. In addition, the QoS management technology can be introduced into the pooling management technology. Local resources are allocated preferentially in accordance with the task priority, and are invoked remotely. When task resources are insufficient, AI tasks are managed in a queue, and then run when sufficient resources are released.

2. Memory pooling

Large-model training tasks bring great challenges to memory and video memory. Data needs to be frequently moved between cache, memory, and video memory devices. The lack of unified addressing of memory space causes complicated programming models and restricts collaboration between devices. Data transmission and duplication must be managed manually, increasing development difficulty and error rate. In addition, data needs to be converted between memory and video memory for multiple times, and different CPU/GPU heterogeneous devices cannot directly share data or give full play to their own advantages. These factors restrict the overall system performance improvement.

To reduce the impact of the above problems on the overall operation efficiency of the AI+ cloud infrastructure, the unified memory pooling technology based on the computing bus protocol needs to be introduced. The unified memory pool technology implements consistent memory semantic and spatial addressing capabilities, integrates multiple physical memory, video memory devices, and resources into one logical memory pool, and implements unified scheduling, monitoring, and management of memory resources. This technology dynamically allocates and releases memory resources, and flexibly adjusts them in accordance with application requirements, thus optimizing the system response speed and data processing capability.

CXL (Compute Express Link) is an open and standard high-speed interconnection protocol designed for high-performance computing, data center, and storage applications. The CXL technology connects multiple processors, accelerators, storage devices, and other devices through high-speed channels to provide higher bandwidth and lower latency and eliminate transmission bottlenecks of compute-intensive workloads. In the intelligent computing center, the CXL technology can be used to build a memory pool to share and consistently access memory between CPU and accelerators (such as GPU and FPGA), and maintain memory consistency. This means that data does not need to be frequently copied or synchronized during transmission between different devices, thus improving performance. However, CXL needs to be enhanced and optimized in the following aspects to implement the industry:

- Improving the implementation of computing bus protocols and sub-protocols that meet the memory pooling technology.Fully and efficiently implement the CXL.io and CXL.mem protocols to provide channels for I/O communication and memory access between devices, optimize data transmission and replication mechanisms, reduce the additional performance loss introduced by memory pooling, and ensure efficient system operation.
- Accelerating GPU support for implementing a memory consistency mechanism based on CXL. Introducing memory pooling technology will reduce the frequency of protocol conversions between computing and storage devices. By implementing a memory consistency mechanism and optimizing the consistency algorithms among memory, video memory, and cache, ensure synchronized updates of data in shared memory, making the data between devices consistent and available. At the same time, implement a robust error correction mechanism to ensure stable and reliable operation of the memory pool system.
- Expediting the development of a unified interface between multiple heterogeneous devices and the memory pool, with isolation and protection

capabilities. Provide interfaces for collaborative work among multiple heterogeneous devices, focusing on efficient collaboration and shared computing capabilities between devices, reducing the latency and energy consumption caused by data transmission and replication. At the same time, strengthen security measures to ensure that only authorized processors can access the memory pool, preventing access conflicts.

The computing and memory resource pooling technology dynamically allocates and reclaims resources as required, rapidly adapts to changing workloads, shares and reuses resources, improves resource utilization, improves system performance, and reduces hardware investment and maintenance costs. In addition, heterogeneous computing (for example, GPU of different brands and models) can be flexibly, dynamically, and efficiently invoked and allocated. This helps build an open and cooperative ecosystem of computing of multiple manufacturers, and build a more flexible and scalable AI acceleration computing resource pool to adapt to the ever-evolving technological advancements., workload, and application requirements. These advantages make the resource pooling technology one of the important technologies in the artificial intelligence field.

7.2 Intelligent Computing Storage to Meet the Key Challenges of High-Performance and High-Concurrency Training Tasks

In multiple end-to-end links of large model development, innovative requirements are proposed for storage, including:

- Multi-modal storage: Multi-modal data sets such as video, image, and voice bring multi-modal storage and protocol interworking requirements such as block, file, object, and big data.
- Mass storage: To ensure the precision of large-scale model training, the data set is usually 2-3 times of the parameter value. In the current era when large-scale models are rapidly developing from a hundred billion to a trillion, the storage scale is an important indicator.
- High concurrent performance: In a large-model parallel training scenario, multiple training nodes need to read data sets at the same time. During the training process,

the training node needs to periodically save the CheckPoint to ensure the system's ability to resume training from breakpoints. The high performance of these read/write operations can greatly improve the efficiency of large model training.

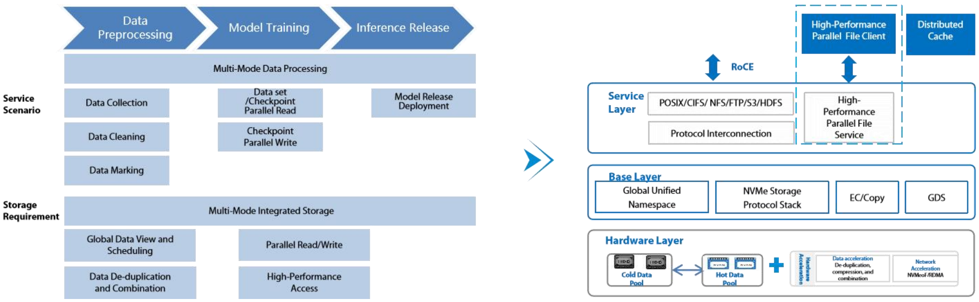


Figure 7-2 Intelligent Computing Storage Requirements and Architecture

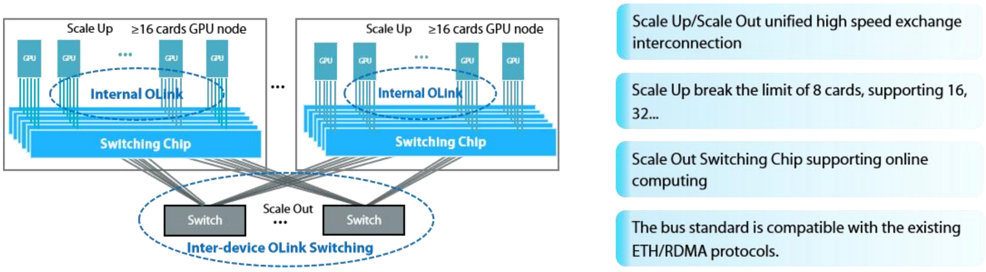
Therefore, as shown in the above figure, the following capabilities are required for intelligent storage:

- Unified storage: Constructing a unified storage system meets the needs of different stages in the AI pipeline, provides diverse data storage capabilities and multi-protocol interoperability for block (iSCSI), file (NAS), object (S3), and big data (HDFS).
- Hardware acceleration: It includes DPU storage interface protocol unloading, de-duplication, compression, and security operations, as well as automatic tiering and partitioned storage of data based on popularity.
- Software acceleration: It includes distributed caching, parallel file access system, and private client technologies. In addition, the NFS over RDMA and GPU GDS technologies can greatly reduce the data access delay.
- Data entropy reduction: Unnecessary data movement and duplication are reduced, storage and access policies are optimized, and "data entropy tax" is reduced. Data transmission and storage overheads are reduced through technologies such as de-duplication and compression.

By providing a high-performance, highly scalable, and multi-dimensional unified storage solution, the intelligent storage system can easily meet large-scale data processing requirements, improve AI model training and inference efficiency, optimize AI system costs and power consumption, and accelerate AI innovation and application implementation. In addition, the distributed intelligent storage system can perfectly support the deployment and operation of the distributed AI architecture. For example, by using a distributed storage system, distributed storage and access of data can be implemented, and parallelism and scalability of cross-domain training task data processing are improved. In addition, intelligent storage provides cross-node data replication and backup to ensure data security and reliability.

7.3 Open A High-Channel Lossless Network to Reduce the Parallel Computing Communication Overhead

With the rapid development of artificial intelligence technologies, the parameter scale of large AI models is expanding rapidly at a speed beyond Moore's Law, and the training of large AI models poses an unprecedented challenge to computing capabilities. In response to this demand, enterprises are building intelligent computing clusters and introducing parallel computing technologies to accelerate model training. However, while parallel computing enhances overall computational efficiency, it also introduces synchronization overhead and communication latency issues. In this context, exploring how to achieve high-speed interconnection within servers (Scale Up) and between servers (Scale Out) in ultra-large-scale intelligent computing clusters, thereby significantly improving GPU utilization, has become a significant challenge faced by the industry.



Scale Up/Scale Out unified high speed exchange interconnection

Scale Up break the limit of 8 cards, supporting 16, 32...

Scale Out Switching Chip supporting online computing

The bus standard is compatible with the existing ETH/RDMA protocols.

Fig. 7-2 High-Channel Lossless Network Architecture

1. Scale-up network interconnection trend

With the increasing requirements for computing in large-scale model training, the traditional intra-device scale-up network point-to-point full mesh interconnection architecture gradually shows its disadvantages of insufficient scalability. Although the full mesh architecture can provide high-bandwidth and low-latency communication capabilities, its extension capability is limited. Especially when the number of GPUs increases, point-to-point communication mode cannot achieve linear extension. In most cases, the full mesh architecture supports a maximum of eight GPU cards in a single server, greatly limiting the training efficiency of large models.

To overcome the scalability limitations of hardware systems and address the closed nature of existing private bus protocols for GPU interconnection, achieving interoperability among multi-vendor chips, we innovatively propose the OLink technology, which is a high-speed, open interconnection for GPUs based on a switched topology.Through this technology, communication between GPUs shifts from the traditional point-to-point interconnection mode to a switched interconnection mode, significantly enhancing the scalability and communication bandwidth of a single machine, breaking through the limitation of 8 GPUs per machine. With the OLink technology, a large-scale high-bandwidth domain (HBD) can be created to greatly improve the computing of clusters. In addition, it promotes the prosperity of the multi-manufacturer ecosystem and provides enterprises with more flexible choices. The openness of this technology brings greater flexibility and sustainability to the industry, and helps promote the diversified development of intelligent computing technologies.

2. Scale-out network interconnection trend

The scale-out interconnection network between super-node servers is also important to solve technical bottlenecks such as communication bandwidth and delay in model training and improve the overall efficiency of model training. Currently, there are two mainstream technologies in the industry: IB and RoCE. The IB network, namely the Infiniband network, is a closed network solution exclusively provided by NVIDIA, with excellent performance but high price. RoCE is an open solution based on the standard Ethernet protocol. However, each manufacturer has its own enhancement solution. Different manufacturers anchor their own switching devices for congestion control and end-network optimization, which makes it difficult to decouple from network devices.

The collaboration between the intelligent computing resource management platform and the RoCE network management and control system to achieve automated deployment of the parameter plane network, along with enhancements based on the open RoCE protocol, offers a general, open, and cost-effective high-performance lossless solution. This is an effective approach to address the aforementioned challenges. However, the construction of this ecosystem faces significant difficulties and challenges, requiring long-term joint efforts from the industry to promote.

The objective of the industry is to provide an open and perfect RoCE solution based on RoCE. At present, in addition to the basic protocol for RoCE congestion control, decoupling is mainly considered on the NIC side and RDMA network side, that is, decoupling between servers and network devices. At present, it is difficult to decouple the RoCE network. For a large-scale network, a more complicated congestion control algorithm or traffic scheduling policy is required. At present, there are two solutions in the industry: One is to implement more elaborate congestion control through enhanced end-network coordination, for example, congestion control algorithms such as HPCC, which can be implemented only through coordination between the NIC side and the switch side. The second type is that the network side of the switch provides better traffic scheduling capability to avoid traffic congestion in the switching network. The above two solutions are required to better solve the congestion control problem in a large-scale lossless network. At present, OLink is pushing the industry to achieve standardization based on the above ideas.

7.4 Native Computing to Build An Ecosystem of Heterogeneous Computing Decoupling

With the development of intelligent computing technologies and the emergence of new applications, while traditional industry giants such as intel, NVIDIA, and AMD launch AI chips, some innovative chip manufacturers have also launched AI chip solutions. Different manufacturers build their own software ecosystems around their own chip architecture, resulting in fragmented and vertical software ecology of each manufacturer. The vertical closed ecosystem of multiple manufacturers brings high costs of cross-architecture application optimization deployment and development, reasonable planning of heterogeneous computing, and difficulties in dynamic migration of applications, resulting in low resource utilization and difficulties in building a healthy ecosystem.

Therefore, heterogeneous open environments based on multiple infrastructure environments and multiple GPU card types are the direction of future evolution. In the early stage, the heterogeneous resource pooling technology can be used to improve resource utilization. The same resource pool supports the management and orchestration of GPU resources of different vendors. The resource pool classifies, manages, and orchestrates the cards of different vendors. When an application requests GPU resources, the applications of different frameworks are scheduled to compatible manufacturers' GPU resources as required.

In the next phase, an open and flexible development and adaptation platform can be constructed by building a standard and unified computing abstract model and programming paradigm interface, effectively interconnecting various heterogeneous hardware resources and computing tasks, adapting heterogeneous computing and service applications on demand, and flexibly migrating, fully releasing the collaborative processing effectiveness of heterogeneous computing, accelerating the innovation of intelligent computing application services, and implementing the native computing architecture that integrates heterogeneous computing resources, migrates applications across architectures in a senseless manner, and facilitates the development of industrial ecology. In this case, the details of underlying GPU heterogeneous resources are shielded, and the upper-layer AI framework applications are completely decoupled from the underlying GPU resources.

Specifically, computing native includes two parts: Computing pool layer and computing abstraction layer.

The computing pooling layer integrates various types of hardware resources, builds a unified abstract model of underlying heterogeneous hardware, and redirects applications to invoke requests of underlying computing-power resources, so that applications can apply for computing through the unified defined abstract intelligent computing-power value, thus shielding the differences between heterogeneous hardware. To cope with the tidal effect of intelligent computing services, the computing pool lay er can provide elastic scaling of computing resources in accordance with service requirements and computing load.

The computing abstraction layer consists of a native stack and a native interface. The native stack mainly includes a unified programming model, cross-architecture compilation, and native operation. The unified programming model and cross-architecture compilation can translate application programs based on specific chip programming into a native intermediate representation of computing (Intermediate Representation) irrelevant to the underlying hardware architecture. The native runtime can perceive and control underlying computing resources, load and parse native programs, and ensure instant mapping between computing tasks and local computing resources and on-demand execution. Based on the abstraction interface of the native computing and the multi-modal hybrid parallel programming model, the native interface builds the unified API of the native computing, the paradigm of the native programming model, and the native compilation and optimization deployment tool, forms a development environment that can be embedded into user services, and assists users in generating native computing programs that can be transferred across architectures, migrated without perception, and executed through task-based mapping.

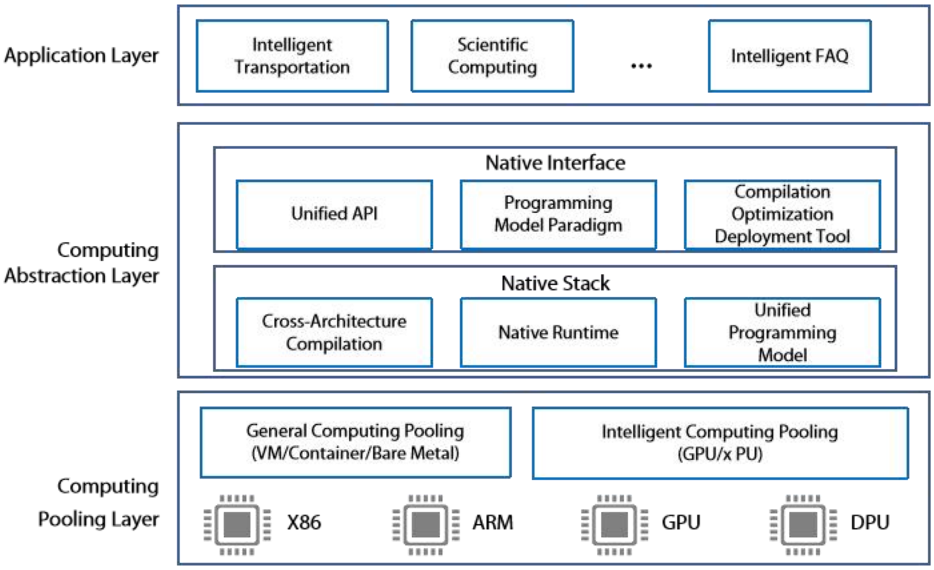

Application Layer
Intelligent Transportation
Scientific Computing
...
Intelligent FAQ
Computing Abstraction Layer
Native Interface
Unified API
Programming Model Paradigm
Compilation Optimization Deployment Tool
Native Stack
Cross-Architecture Compilation
Native Runtime
Unified Programming Model
Computing Pooling Layer
General Computing Pooling (VM/Container/Bare Metal)
Intelligent Computing Pooling (GPU/x PU)
X86
ARM
GPU
DPU

Fig. 7-3 Native Architecture of Computing

7.5 Distributed Deployment of Mixed Pools to Meet the Comprehensive Resource Requirements of Core Network Applications

Core network NEs have requirements for computing and intelligent computing infrastructure resources, and training applications also need distributed deployment. Therefore, the hybrid pool deployment and distributed deployment of universal and intelligent computing become the features of AI+ cloud infrastructure deployment.

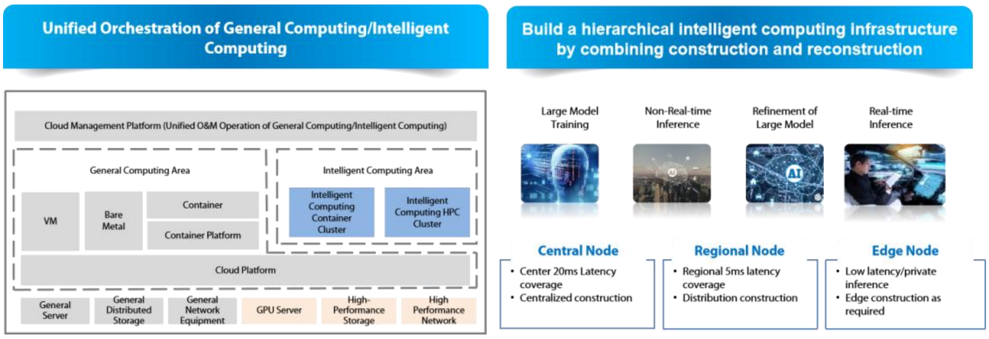

Unified Orchestration of General Computing/Intelligent Computing
Cloud Management Platform (Unified O&M Operation of General Computing/Intelligent Computing)
General Computing Area
VM
Bare Metal
Container
Container Platform
Intelligent Computing Area
Intelligent Computing Container Cluster
Intelligent Computing HPC Cluster
Cloud Platform
General Server
General Distributed Storage
General Network Equipment
GPU Server
High-Performance Storage
High Performance Network
Build a hierarchical intelligent computing infrastructure by combining construction and reconstruction
Large Model Training
Non-Real-time Inference
Refinement of Large Model
Real-time Inference
AI
Central Node
• Center 20ms Latency coverage
• Centralized construction
Regional Node
• Regional 5ms latency coverage
• Distribution construction
Edge Node
• Low latency/private inference
• Edge construction as required

Fig. 7-4 AI+ Cloud Infrastructure Deployment Mode

The cloud infrastructure is smoothly upgraded from the general computing resource pool to the intelligent computing resource pool, and the orchestration management of the general computing and intelligent computing mixed pool is a key feature of the cloud infrastructure. In most cases, a centralized cloud platform is used to manage computing, storage, network, and other infrastructure resources. In the early stage, computing resources are orchestrated by the traditional computing cloud infrastructure management platform, and intelligent computing resources are orchestrated by the intelligent computing resource O&M operation platform. After the system is mature, the upgraded cloud infrastructure management platform orchestrates computing and intelligent computing resources in a unified manner.

The pre-training of basic large models, refinement of industrial large models, and refinement of large models in customer scenarios have different requirements for computing features and deployment locations. Based on the hierarchical distribution architecture of operator cloud infrastructure, AI+ cloud infrastructure deployment also presents three-level deployment modes: Hub large model training center, regional training and promotion integrated resource pool, and edge training and promotion integrated machine. The hub large-scale model training center is a recently-built ultra-large-scale GPU clusters, which meets the requirements of basic large-scale model pre-training, improves the computing efficiency and energy efficiency of large-scale clusters, and improves training reliability. The regional training and integration resource pool is usually implemented by expanding and upgrading the existing network computing resource pool on demand, meeting the requirements for refinement inference, and application deployment of large models. It is necessary to focus on the challenges of improving the efficiency and resource utilization of diversified resource management, opening and decoupling capabilities, and application ecology. The all-in-one edge training machine mainly meets the requirements of refinement of enterprise private-domain data, deployment of inference and applications, rapid deployment of government and enterprise customers, and integrated intelligent computing services.

8 Key Elements of AI Core Deployment

At present, the core network in the industry mainly uses a cloud-based construction solution, which is a physical network architecture based on three-level data centers (central, regional, and edge). The cloud-based core network functions can be deployed in corresponding locations in the network in accordance with scenarios. After "AI+ " is introduced to the core network, several key elements need to be considered, such as selection of large models, scenario-based deployment, and end-to-end compliance and security.

8.1 360° Evaluation System for Selecting the Optimal AI Model

AI large model has strong capabilities in natural language processing, image recognition, speech recognition, and other fields. However, for operators, large models emerge endlessly. In the face of so many choices, how to select a proper large model in accordance with core network requirements? It is recommended that a 360° large-model evaluation system be established to comprehensively evaluate model evaluation indicators, large-model frameworks and platforms, and large-model empowerment application scenarios to help operators select appropriate large models for deployment.

In terms of model evaluation indicators, the complexity, generalization capability, stability, and controllability of the model are the main indicators for evaluating and selecting the model. The model complexity refers to the forward calculation amount and total parameters of the model. The higher the complexity, the stronger the expression capability and fitting capability of the model. Generalization capability refers to the performance of a model on new data outside the training data, and is usually measured by generalization errors. Stability refers to the performance capability of a model under abnormal data. A stable model will not fluctuate greatly due to fluctuation of input data. Controllability refers to the interpretability of a model, the alignment of human will, and the alignment of the law. The selection shall be based on the complexity of the task and the actual requirements. In addition, whether the model is open-source, commercially available, supported languages, and open evaluation results can also be used as evaluation indicators.

At the large model framework and platform level, the end-to-end process of high-quality large model pre-training and refinement needs to be introduced, including the entire process of data processing, pre-training, refinement, evaluation, and inference deployment. The data management and labeling platform provides high-efficiency cleaning of large-model pre-training data, and fine-tuned data labeling and evaluation functions. The pre-training platform provides the one-click creation of multi-device and multi-card pre-training operations and the automatic planning capability of the 3D parallel solution. The model refinement platform provides data processing, presets mainstream large models, and supports multiple refinement methods. The full-process automation is started in one click, greatly reducing the complexity of refinement and shortening the period. The model evaluation platform provides an evaluation pipeline platform with multiple built-in benchmark test data sets, establishes an automatic evaluation system for large models, and outputs evaluation reports in one click. The model-based inference deployment platform, which provides quantitative and sparse methods to reduce inference costs and speed up inference, is compatible with multiple back-end platforms, and supports large-model distributed inference deployment.

At the same time, the application layer of the core network model in the communication field should enable intelligent services, intelligent networks, intelligent O&M, and intelligent cloud infrastructure. In terms of service intelligence, message service anti-fraud, message service multi-modal industry model, new data channel interesting interaction, and commercial assistant can all bring innovative service value. In the field of network intelligence, large models can be applied in precise service identification, key service guarantee, and service operation support analysis. On the core network and cloud infrastructure layer, large-scale models can empower complicated O&M scenarios. Each agent needs to be flexibly orchestrated, implement man-machine interaction in LUI mode, use the planning and tool invocation capabilities, fully integrate the small-scale AI model capabilities and network atomic capabilities of the existing network, and implement smooth evolution of the O&M system.

8.2 Combining Construction and Transformation to Build Hierarchical Intelligent Computing Infrastructures

AI application scenarios include training and inference scenarios. Intelligent computing data centers are required for training and inference. Intelligent computing data centers are different from general computing data centers. Compared with the general computing data center, the intelligent computing data center mainly consists of accelerated computing based on AI chips such as GPU, NPU, and ASIC. It needs high-power racks, air cooling, and liquid cooling for heat dissipation. It has high requirements for network delay and packet loss, and generally requires no convergence network. Resource management needs to support new management mechanisms such as tasks and clusters. Therefore, after "AI+" is introduced to the core network, it is necessary to consider how to introduce intelligent computing resources in the existing three-level general computing data center architecture, including the central, regional, and edge, to support AI applications in all scenarios of the core network.

To meet the AI introduction requirements of the core network, it is recommended that the intelligent computing center be integrated into the general computing data center based on the existing data center layout of the core network to build a hierarchical intelligent computing infrastructure.

Central node: This type of node is mainly used for pre-training and inference of large models of the core network. The combination of new construction and reconstruction can be considered to support the upgrade of intelligent computing. In the large-model pre-training scenario, the intelligent computing data center solution is used. The physical location of the intelligent computing data center should be the same as that of the computing data center, that is, the physical location of the intelligent computing data center should be the same as that of the control plane of the core network. Because large-scale model pre-training has high requirements for intelligent computing centers, especially the heat dissipation mode and networking solution are different from those of the general computing data center, large-scale intelligent computing clusters can be built in a centralized manner, and the liquid cooling heat dissipation mode and RoCE networking mode can be used to meet the requirements of large-scale model pre-training of the core network. For the inference scenario of the central node, considering that there is no big difference between the solution vendors of the inference pool and the general computing pool in networking, storage, and management, especially the intelligent inference computing can be supported by the GPU of the PCIE type on the existing computing server. Therefore, the intelligent inference capability of the control plane of the core network can be supported by reconstructing the existing general computing data center, including capacity expansion, upgrade, and new construction of the general computing data center.

Regional node: This type of node is mainly used for fine adjustment and inference of large-scale models on the control plane of the core network. It features distributed on-demand expansion and diversified computing applications. Therefore, the existing general computing data center can be reconstructed, including capacity expansion, upgrade, and concurrent construction of computing pools, to support the fine debugging and reasoning capabilities of the control plane of the core network. In addition, the orchestration tool can be used to implement flexible network service deployment and dynamic network resource coordination, so that various services can be rapidly developed.

Edge node: It provides network edge computing and inference capabilities for the user plane, media plane, or vertical applications of the core network. In the scenario where the user plane of the core network is used, the general computing is provided in dedicated hardware mode. Therefore, after the inference computing is introduced, the intelligent computing machine can be used to provide integrated services of data, models, and applications for the user plane of the core network. In the MEC scenario, the system provides industry customers with fine-tuned private-domain data, inference, and application. Therefore,the system can provide users with all-in-one training machines, standardized software and hardware configurations, and fast access to government and enterprise customers' equipment rooms for end-to-end integrated services. In this way, data and industry models do not leave the park, differentiated experience is provided, and operators' value-added revenues are increasing.

8.3 Building A Hierarchical Defense-in-Depth Security System for Secure and Compliant AI

The in-depth integration and risk combination of the AI technology and the core network put forward higher requirements for AI security. With continuous expansion of the application scope of the AI technology, threats such as ethics risks, content security risks, and privacy data leakage of the AI technology are becoming increasingly prominent. Security is the focus of attention during the deployment of the AI technology. As a key infrastructure, the core network bears important public communication and provides information services. Once damaged, lost functions, or leaked data,it will affect the national security, national economy, people's livelihood, and public interests.

The following hierarchical defense-in-depth security systems are built to provide protection for the AI core system:

- The infrastructure layer creates a trusted environment. For the security risks of sharing intelligent computing resources among multiple tenants, multiple border security facilities can be deployed to prevent common security attacks in the general resource pool and meet the security protection capability requirements in laws and regulations. The isolation capability based confidential computing is used to build a secure operation environment required by the intelligent computing capability, ensuring that key intelligent computing information is available but invisible.
- The data layer ensures that the entire process is legal and compliant. To prevent security risks such as disputes over intellectual property rights or copyrights, personal privacy or commercial data leakage, and training process of data pollution and toxic hazards, complete data security check measures shall be established to ensure the full-process compliance of data transfer: For the data source, legal check shall be conducted to ensure that high-risk data is blocked and illegal data collection is prevented. In terms of data contents, the data that violates regulations, has privacy, and has copyright risks shall be removed through special content check. In terms of data audit, a traceability and integrity check mechanism is established for data to ensure that data risks can be audited and checked, and harmful data source can be handled in a timely manner.
- The AI model itself provides security guarantee. For the risks of poor interpretability,

stealing and tampering, and countering sample attacks that exist in models and algorithms, the security governance specifications and risk control measures throughout the entire process of AI are used for security prevention. In terms of the security governance process, security development specifications are established and implemented during the design, R&D, deployment, and maintenance processes, and the security defects and discrimination tendencies of model algorithms are eliminated as much as possible. In terms of security management of the supply chain, the AI model tracks software and hardware product vulnerabilities and defects involved in AI, and takes timely repair and reinforcement measures to ensure the security of the supply chain. Model risk management and control provides clear descriptions for internal system construction, reasoning logic, technical interfaces, and output results, correctly reflects the process of generating results in the system, and continuously improves the interpretability and predictability of AI.

- Content security guarantees valid input and output. For the risks of illegal query and generation and illegal content output in AI, multiple security mechanisms shall be used to ensure content compliance and legality. In terms of content management and control, a fence mechanism shall be established, focusing on two key points: Input and output. A customizable and continuously updated illegal inquiry method and illegal content set shall be established to ensure that users cannot inquire about illegal content and obtain hazardous replies. In terms of content detection, the system can detect multiple types of content such as texts, pictures, and videos, and can detect the content security between multiple sessions of users and the same session context to dynamically evaluate users' security reputation and prevent continuous damage.
- Continuous assessment and audit to enhance security: In order to inspect and measure the actual security performance of AI products, based on the existing security risk management system framework, the accuracy and reliability of AI products are evaluated by using CE val, HumanEval, and other data sets. Based on multiple attack and defense models and penetration tests such as MITRE ALTA, the system performs security tests and measurement on data security control capability, attack resistance capability of algorithm models, and content input and output detection capability, and continuously enhances the default security capability through rapid iterative improvement

9 AI Core Practice

9.1 World's First Assembled "AI+" 5G New Calling Network

With rapid development of AI technologies, terminal vendors, OTT (Over-The-Top) service providers and operators have added AI applications to compete for intelligent entry and promote industrial value reshaping. In this competition, with a large number of users and a large number of calls, operators have a unique basis for becoming AI entry. Unlike the OTT application, the call service, as a traditional and common way of communication, naturally has obvious advantages such as not relying on APP installation, real-time interaction, and low latency. This gives the call service great potential and innovation space in the AI era.

Through close cooperation with operators, ZTE has deeply integrated the AI technology and call services, and successfully deployed the world's first Assembled "AI+" network. Six AI applications such as translation and fun calls have been launched by operators to provide users with more intelligent and convenient communication experience and greatly enrich the functions and scenarios of call services. These applications not only improve user experience, but also provide enterprises with more efficient and intelligent customer service solutions. AI applications such as real-time translation, subtitle translation, AI stenography, gesture effects, emoticon, background replacement, and virtual avatars have been enabled for the project. Voice-driven digital human, lightening screens, portrait styles, New Calling agents, and AR tags are enabled one by one.

Through in-depth cooperation with operators, ZTE has achieved in-depth integration of AI and communication services in the world's first assembled "AI+" 5G new calling network. With an open ecosystem architecture, intelligent native capabilities, innovative mechanisms for intelligent orchestration and dynamic loading, and abundant AI applications, operators can enhance user experience, empower enterprises and industries, and promote the intelligent development of the entire communication industry.

9.2 Industry's First Commercialized Hierarchical VIP User Service Assurance Solution

As user requirements become more personalized and diversified, traditional traffic operation models cannot meet the requirements of modern operators in fierce competition. Operators should pay attention to not only traffic growth, but also user experience guarantee, especially in key service scenarios (such as live streaming, gaming, and high-speed railway private network). How to improve the differentiation and quality of network services becomes a key challenge to determine the brand value and user satisfaction of operators. Together with operators, ZTE has successfully verified the service experience guarantee solution for “AI+“ connections, and implemented live streaming service guarantee pilots and high-speed railway private-network camping guarantee pilots, solving problems such as insufficient experience guarantee and inaccurate service priority scheduling in the existing network.

High-priority user guarantee: For high-value VIP users, the system automatically increases their network guarantee levels to ensure that they can enjoy the best network service experience at any time. For key services (such as live broadcast, game, and video conference), the solution sets different 5QI levels for each type of service to implement targeted network scheduling such as rate guarantee and delay optimization. For example, the bandwidth of live streaming services with high rate requirements is preferentially guaranteed, while the low latency of latency-sensitive game services is preferentially guaranteed.

Camping guarantee in the high-speed railway private network: In the high-speed railway scenario, the solution implements user profile in the low-speed area, identifies the high-speed railway VIP user, and delivers the wireless guarantee policy to the high-speed railway user to ensure the camping of the VIP user in the high-speed private network and ensure the stability of its data service. In addition, VIP user experience is guaranteed in transit platform scenarios and high-speed parallel connection scenarios. Platform parking scenario: The solution can automatically identify the residence of VIP users on platforms, preferentially guarantee high-priority scheduling of data and voice communication, and avoid interference to user experience. High-speed parallel connection scenario: The solution ensures that VIP users in the high-speed rail private network reside stably, quickly identifies and clears unnecessary invasion users, and ensures that the network resources of the high-speed rail private network are not wasted.

9.3 Autonomous Network L4 Troubleshooting Scenario (in Cloud Infrastructure) Implementation Practice

- Use case 1: Quantitative evaluation solution for resource pool switchover based on digital twin

The disaster recovery backup mechanism at the resource pool or data center (DC) level plays a critical role in ensuring user service continuity and network operation security. However, during disaster recovery at the resource pool or DC level, the users served by the NEs in the original resource pool must be quickly migrated to other available resource pools or DCs. During the migration, a large number of users reconnect to the network within a short period of time, inevitably causing signaling shocks, and resulting in a sharp high load on the CPU and network resources.

Therefore, ZTE and operators proposed the quantitative evaluation system of resource pool switchover based on digital twin for the first time. Based on the relationship between NEs and resource pools, the resource pool switchover process is simulated and quantified.

– **Automatic data collection and processing for fast evaluation**

Through the fully automated data collection, analysis, and processing flow, the disaster recovery evaluation time is reduced to less than 10 minutes. This technological innovation not only greatly improves the speed of decision-making response, but also effectively reduces the complexity of manual intervention and significantly improves data processing efficiency. Especially in emergency situations, O&M personnel can quickly obtain evaluation results and make decisions to ensure that the network can be restored within the shortest time, improving network stability and emergency response capabilities.

– **Second-level simulation for accurate decision-making on disaster recovery**

Simulation calculation with second-level granularity is introduced to accurately simulate the load impact between NEs and resource pools. Compared with the traditional method, this simulation method can more truthfully restore the actual shock waveform in the network, thus greatly improving the evaluation accuracy to more than 95%. This high-precision simulation technology improves the reliability of the switchover process, provides a more accurate decision basis for emergency fault recovery, and enhances the disaster recovery capability of the system.

- **High-precision adaptive model that can be interpreted transparently throughout the process**

The CN and resource pool impact linkage evaluation solution uses white-box simulation to transparently model the NE layer. This model has powerful expandability, and can adjust and optimize evaluation methods in accordance with different switching scenarios to ensure that the model can adapt to various network topologies and flow changes. The resource pool traffic simulation algorithm based on the number of signaling messages and topology mapping improves the accuracy and adaptive capability of fault evaluation, and further improves the intelligent level and application scope of the disaster recovery system. Based on the combination of theoretical models and AI learning, this solution solves the differences between terminal behaviors and provincial services in signaling impact assessment of the core network. Through adaptive model adjustment, this solution significantly improves the assessment accuracy and ensures that the network assessment accuracy is higher than 95%. This method not only corrects the deviation of traditional calculation, but also improves the adaptability of the network in a changing environment, and provides reliable data support for decision-making.

Through capability openness embedded in the production process, a three-dimensional defense system of "prevention, perception, emergency, traceability" is built, with the capability of preventing problems beforehand, quickly detecting and recovering problems in the event, and accurately tracing problems after the event.

This case is deployed and verified in the operator's existing network. Through the DC disaster-recovery switchover verification, the accuracy of the simulation calculation and the simulation evaluation duration of the twin system reach the design objectives. The twin system interconnects with the upper-level network management system through capability openness to embed capabilities into the production process. The commercial system develops the switching model and framework based on the standard signaling interaction procedure of the core network and the northbound data reporting specifications of the cloud-based resource pool. The system can be widely used in disaster-recovery switching scenarios of the cloud-based core network, and has good replication and conversion capabilities.

- Use case 2: Network fault diagnosis based on large models and multiple agents

Telecom networks are becoming increasingly complicated. With continuous development of 5G and AI technologies, the early warning, location, and handling of cloud Infrastructure faults have faced unprecedented challenges. The existing fault management system cannot cope with the large-scale and dynamic network environment, resulting in the following major O&M pain points:

In the process of telecom network O&M, the efficiency and response speed of fault management are key factors. However, the current O&M system faces many challenges. First, faults cannot be warned in advance. Most problems can be discovered only after customer complaints or alarms are reported. As a result, decisions are delayed, and the best time to handle faults is missed. Second, the fault transmission link is long and the transmission efficiency is low. In a complicated network environment, fault location requires multiple links and usually takes a long time to analyze and locate the problem. This not only increases the time-consuming fault location but also increases the risk of major service faults.

Even if it is a known problem, the current system still needs to rely on a large number of experts to handle it, resulting in repeated efforts of 80% problems. Due to the lack of the intelligent man-machine interactive cause diagnosis function, these problems cannot form a closed loop quickly, resulting in a large waste of time and energy. At the same time, daily O&M reports of the existing network often require 2 to 3 hours of manual operations in each environment, further increasing the pressure on human resources and affecting O&M efficiency.

On this basis, the major operation solution is usually formulated with strong professionalism, involving many details, and the compilation and modification period is long. If any details are omitted, the operation may fail, seriously affecting network operation. Therefore, the entire fault management process faces multiple problems, such as delayed early warning, low positioning efficiency, waste of expert resources, and low report generation efficiency. Therefore, intelligent measures are urgently needed to improve the overall O&M efficiency and accuracy.

Therefore, ZTE proposes a network fault diagnosis solution based on large models and multiple agents, which improves the overall O&M efficiency and accuracy through intelligent means, and ensures that faults can be responded in real time and repaired quickly.

- **Multi-layer and multi-dimensional data perception for holographic network status visualization**

The operational status of a cloud infrastructure involves data from multiple dimensions, including the physical layer and virtual layer. Through the multi-layer and multi-dimensional data integration technology, the large model can integrate data sources from different layers to form a more accurate and comprehensive network status view.

- **Rapid fault root cause diagnosis and precise repair suggestions**

On the general language model, the domain corpus is generated together with the collected inference corpus for FT and SFT, and a large domain model is trained to improve the accuracy of domain downstream tasks by 6%. With the intelligent analysis capability based on large models, the system can rapidly diagnose the root causes of faults, greatly reducing the number of false alarms and ensuring that O&M personnel focus on real faults. Through the ZTE Nebula large model, multi-agent technology, and RAG enhancement, the system provides accurate fault information and targeted repair suggestions, helping site engineers quickly locate and solve problems, greatly shortening the work order processing duration, and improving the fault response efficiency.

- **Closed-loop fault self-healing to reduce manual operation errors**

Based on the ZTE O&M agent, through self-learning and self-adaptive adjustment, the can independently invoke related tools or interfaces to automatically complete closed-loop restoration. This self-healing mechanism can not only rapidly restore system functions without human intervention, but also ensure the integrity and accuracy of the repair process, improving the automatic network O&M level and reducing manual operation errors. The industry's first multi-agent solution can dynamically break down more than seventy tasks, with an accuracy rate of more than 90%. This solution solves complex task intelligence in the field and accurately completes and implements tasks.

9.4 Anti-Fraud LLM for Messaging to Reduce the Number of Fraud-Related SMS Cases by 64%

With the continuous changes of fraud forms in telecom networks, traditional anti-fraud governance solutions cannot flexibly deal with new fraud methods, and there are great bottlenecks in the interception and identification efficiency of fraudulent SMS messages. The increasingly complicated contents and forms of fraudulent SMS messages make the traditional keyword matching-based technology difficult to meet real-time and accurate requirements. In addition, because many anti-fraud systems are not intelligent and adaptive, the identification of new fraud methods is delayed and the interception effect is not satisfactory. To meet this challenge, it is urgent to introduce advanced artificial intelligence technologies, especially large model technologies, to improve the identification accuracy and interception efficiency of fraudulent SMS messages and reduce the damage of fraudulent information to the society from the root causes. In cooperation with Shanghai Mobile, ZTE has proposed and implemented an innovative SMS anti-fraud governance solution centered on the anti-fraud model. This solution introduces advanced technologies such as generalized feature neural network and SNS social feature analysis to analyze the intentions and semantics of fraudulent SMS messages, providing an intelligent and all-round anti-fraud governance solution for telecom operators.

- **Build a large anti-fraud model based on real data, so models are credible**

The dependability of a large anti-fraud model first comes from the quality and breadth of its training data. To ensure the efficiency and accuracy of the model in actual applications, ZTE builds the core data set of the anti-fraud model by collecting and marking a large amount of real data. The data includes millions of fraudulent SMS message samples from different channels, including various types, techniques, and forms of fraud. Through in-depth analysis of the contents of fraudulent SMS messages, the model can learn and capture the subtle features of fraudulent SMS messages, and establish accurate classification and recognition capabilities on this basis.

Diversified and multi-channel data collection: To ensure data representativeness, fraudulent SMS samples are collected from multiple sources, including real user feedback, operator anti-fraud database, and public reporting platform. This ensures that the model covers as many fraud scenarios and methods as possible.

During the development of the large anti-fraud model, the accuracy of the model is verified through cross verification and real scenario tests. In particular, the performance of the model under different operator networks, regions, and cultural backgrounds is tested to ensure that the model can adapt to various variations and regional differences.

Real-time feedback and optimization: In actual applications, the anti-fraud system continuously collects data from user and operator feedback to dynamically optimize the model and further improve its reliability.

- **Deep refinement to reduce hallucinations of large models**

A large anti-fraud model for large-scale pre-training usually has powerful semantic understanding and prediction capabilities. However, some "hallucinations" may occur during application, that is, the model generates inaccurate or irrelevant judgments in accordance with the learning mode. To reduce these hallucinations, the anti-fraud big model uses the in-depth refinement technology to improve its accuracy in specific anti-fraud tasks. The accuracy and output format of inference are optimized, and the hallucination probability is lower than 1/1 million.

- **Semantics mining to improve recognition**

The text content and expression of fraudulent SMS messages are highly variable, which means that the traditional anti-fraud technology based on keyword matching is easy to be deceived. Therefore, to improve the identification capability of the anti-fraud system, the large anti-fraud model uses the semantic mining technology to deeply analyze the deep semantics of SMS messages, so as to identify potential fraud information more accurately.

After the system goes online, the number of overseas fraud-related cases is obviously reduced, contributing to reducing property losses and maintaining social harmony. In the future, ZTE will continue to strengthen research on new technologies, deepen cooperation and application practices, further enhance anti-fraud capabilities, and help operators build intelligent and highly secure communication networks.

10 Evolution Prospect of Core Network Intelligence

Oriented to 6G, ZTE adheres to the concept of AI+, aiming at innovation leadership, cost reduction, and efficiency improvement. Focusing on Service, Connectivity, O&M, and Cloud Infrastructure, ZTE continuously researches service innovation technologies. On the one hand, the AI Core helps operators innovate and reduce costs and improve efficiency. On the other hand, it acts as an AI enabler to meet the intelligent transformation requirements of industry users and help countries achieve the goal of New Quality Productive Forces. At present, innovative AI technologies are emerging. In addition, in the process of in-depth integration of 5G networks and AI technologies to achieve intelligent transformation, industrial partners are required to work together to explore a smooth evolution path of intelligence and jointly promote the solution of intelligence problems.

- Continuously promote the evolution of the NWDAF function

Under the 3GPP standard, NWDAF is continuously developed from R15 to later versions. The NWDAF function of R15 is single, and only network slice load analysis is supported. However, R16 and later versions (R17, R18, and R19) are optimized at different layers. For example, R16 defines a centralized architecture to meet basic data analysis requirements. R17 implements a separated training-push architecture, defines the Analysis Logic Function (AnLF) and the Model Training Logic Function (MTLF), builds a hierarchical intelligent architecture that supports multi-NWDAF coordination, and introduces the data management framework to improve data collection and analysis efficiency. In the future, NWDAF may further expand its analysis capabilities and service scope to better meet the growing service requirements.

- Promote the further development of autonomous networks

Intelligent O&M uses big data and AI technologies to achieve intelligent, automated, and efficient network O&M. At present, network data can be collected and analyzed in real time, and model recognition and prediction analysis can be performed by using machine learning algorithms. In the future, further optimization may be performed in improving O&M efficiency and network stability. For example, improving O&M efficiency may further reduce manual operation dependency and implement O&M tasks more accurately and automatically. To improve network stability, advanced real-time monitoring and intelligent analysis technologies may be used to handle network faults more quickly.

- Adapt to the intelligent development of diversified services.

With more and more service scenarios, for example, different mobile Internet data services (such as live video, video-conferencing, and game) have different requirements for network resource occupation and experience guarantee, the core network needs to accurately identify and meet the requirements of different services. For example, when wireless network resources are heavy, the core network can accurately learn about the customer experience QoE of guaranteed services in real time. In the future, more intelligent resource allocation and guarantee mechanisms may be developed to adapt to more diversified service scenarios. The core network will strengthen cooperation with vertical industries to meet the needs of different industries and promote digital transformation and social progress. The intelligent core network can use AI+ big data technologies to process and analyze massive data, explore potential values, optimize resource configuration, and support decision-making. In the future, more intelligent network services may be customized for the special requirements of different vertical industries.

- AI Native evolution of 6G core network

The connection plane of the 6G core network will be further enhanced on the basis of the existing control plane and user plane of the 5G core network to achieve the programmable capability and the native AI capability of 6G core network. The evolved control plane uses the service-based architecture to further decouple network functions and flexibly invoke network service capabilities on demand. In addition, 6G network will support the distributed network architecture, and the network function discovery and selection mechanism will be further extended. In addition to the NE level, the network function discovery and selection capability will also be supported between networks. Based on the traditional connection plane, 6G core network will introduce a new data plane and a new computing plane to support the intrinsic evolution requirements of the 6G core network capabilities, and implement multi-element collaborative service capabilities oriented to data, computing, intelligence, and other service resources, so as to meet differentiated requirements of industries and integrate intelligent computing, promote cloud network edge-end industry coordination, and facilitate the prosperity and development of the industrial ecosystem.

In conclusion, in the future, based on the intelligent development results of 5G-A core network, the AI core network will evolve towards the AI Native development of 6G core network, and will continue to develop in architecture, functions, and O&M in order to meet the increasingly diversified service requirements and vertical industry requirements, achieve more efficient, intelligent, and flexible network services, and bring more value to operators

11 Abbreviation

Table 11-1 Abbreviation

Abbreviation	Full Name
2B	To Business
2C	To Customer
3GPP	3rd Generation Partnership Project
5G	5th Generation Mobile Communication Technology
5GC	5G Core
6G	6th Generation Mobile Communication Technology
AI	Artificial Intelligence
AIGC	Artificial Intelligence Generated Content AI-Generated Content
AGV	Automated Guided Vehicle
AMF	Access and Mobility Management Function
API	Application Programming Interface
AR	Augmented Reality
B2B	Business-to-Business
B2C	Business-to-Customer
CV	Computer Vision
GSMA	Global System for Mobile communication Association
FOA	First Office Application
FPGA	Field - Programmable Gate Array
FTP	File Transfer Protocol
GDS	GPU Direct Storage GPU
H5	HTML5
HBD	High Band Domain
HDFS	Hadoop Distributed File System Hadoop
IB	InfiniBand
IMT-2030	International Mobile Telecommunications for 2030
I/O	Input/Output
iSCSI	Internet Small Computer System Interface

Abbreviation	Full Name
OTT	Over The Top
POSIX	Portable Operating System Interface for UNIX
HPCC	High Precision Congestion Control
NFS	Network File System
NAS	Network - Attached Storage
NPU	Neural - Processing Unit
NWDAF	NetWork Data Analytics Function
MIG	Multi-Instance GPU
RDMA	Remote Direct Memory Access
RoCE	RDMA over Converged Ethernet
ToC	To Consumer
ToB	To Business
ToH	To Home
ToO	To Other
vGPU	Virtual GPU
VoNR	Voice over New Radio
VR	Virtual Reality