JUN 2025 | VOL. 27 • NO. 3 • ISSUE 218

ZTE TECHNOLOGIES

VIP Voices

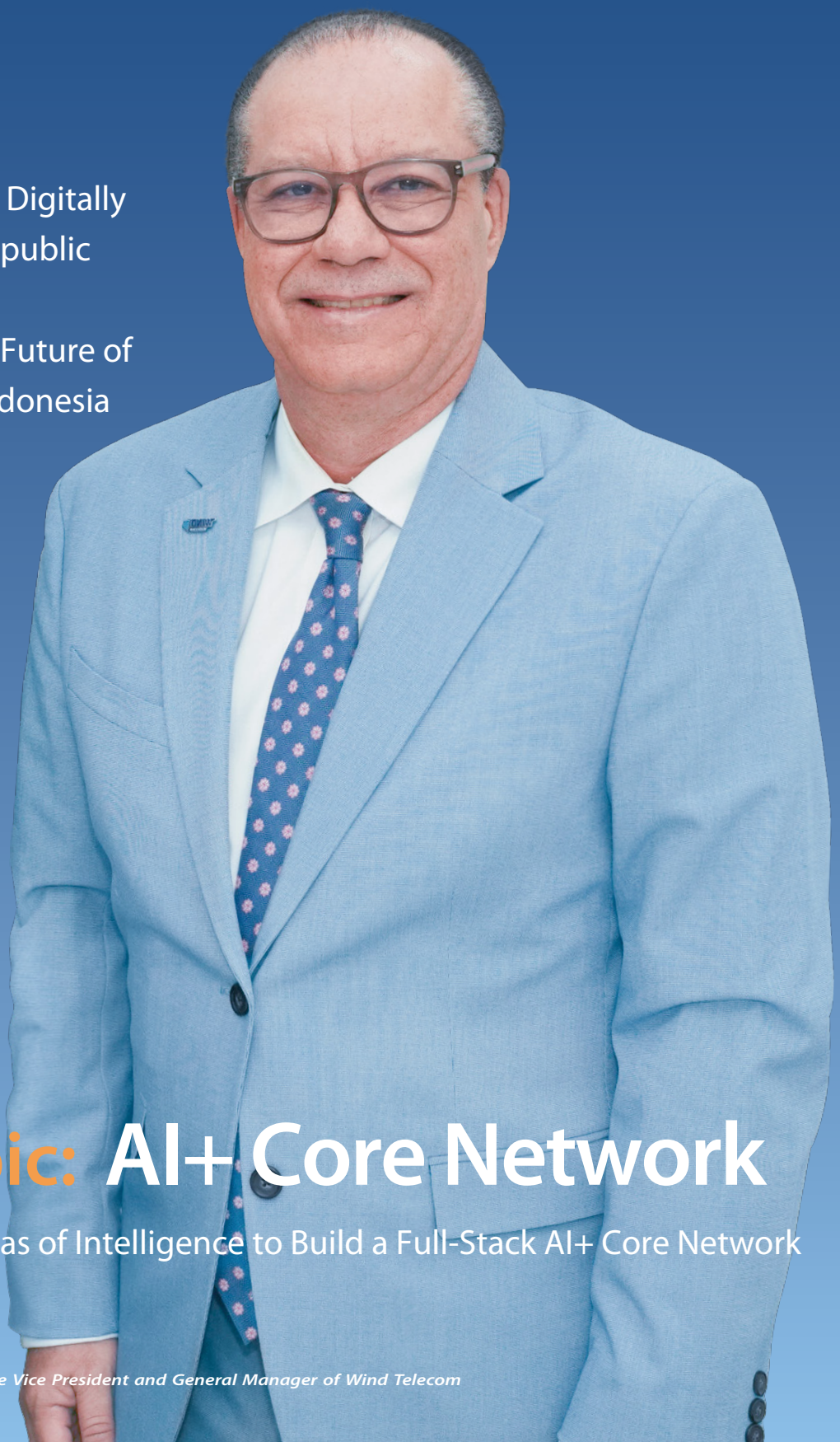
Wind Telecom: Building a Digitally Connected Dominican Republic

MyRepublic: Shaping the Future of High-Speed Internet in Indonesia

Expert Views

AI Ushers in a New Era of Intelligent Communication

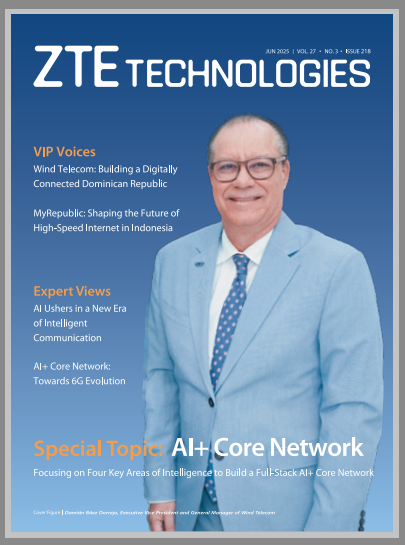
AI+ Core Network: Towards 6G Evolution



Special Topic: AI+ Core Network

Focusing on Four Key Areas of Intelligence to Build a Full-Stack AI+ Core Network

Cover Figure | *Damián Báez Dorrejo, Executive Vice President and General Manager of Wind Telecom*



ZTE TECHNOLOGIES

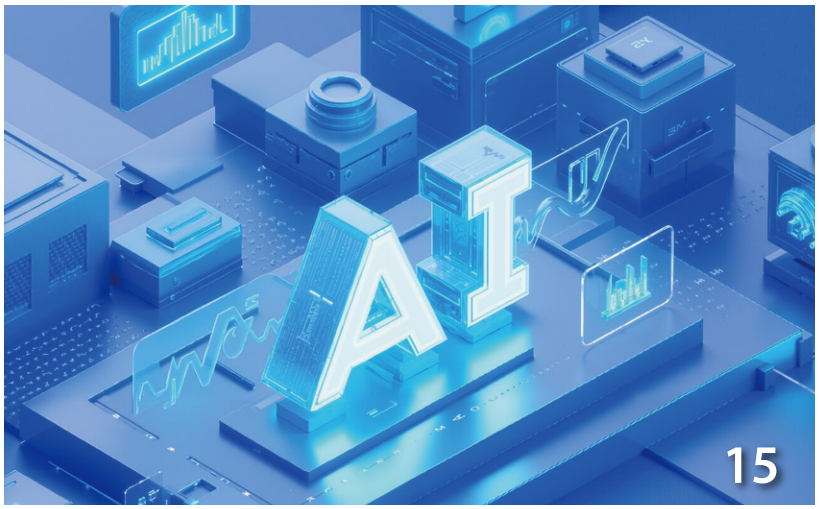
JUN 2025 | VOL. 27 • NO. 3 • ISSUE 218

Sponsor: ZTE Corporation

Published by ZTE Technologies Editorial Office
General Editor: Lin Xiaodong
Deputy General Editor: Lu Dan
Editor-in-Chief: Liu Yang
Circulation Manager: Wang Pingping

Editorial Office Address: NO. 55, Hi-tech Road South, Shenzhen, P.R. China
Circulation Office Address: 12F Kaixuan Building, 329 Jinzhai Road, Heifei, P.R. China
Website: www.zte.com.cn/en/about/publications
Email: magazine@zte.com.cn

CONTENTS



Success Stories

Wind Telecom

Building a Digitally Connected Dominican Republic

Reporter: Radhika Devi

Damián Báez Dorrejo

Executive Vice President and General Manager of Wind Telecom

With the significant growth of the ICT market in the Dominican Republic, new opportunities are emerging to enhance digital connectivity across the country. In this context, Wind Telecom is expanding its role as a key enabler of digital transformation. Damián Báez Dorrejo, Executive Vice President and General Manager of Wind Telecom, shares how the company is helping to build a digitally connected country and discusses Wind Telecom's plans for next-generation technologies at Mobile World Congress (MWC) 2025. Wind Telecom provides integrated internet, online television, and telephone services to individual customers, as well as specialized services to companies, institutions and other service providers in the Dominican Republic.

How would you describe the ICT landscape in the Dominican Republic, considering current trends, opportunities, and gaps affecting both the residential (B2C) and business (B2B) markets?

The ICT landscape in the Dominican Republic is dynamic and rapidly evolving, fueled by a surging demand for high-speed internet. In the residential market (B2C), households increasingly seek digital entertainment and educational resources, while in the business market (B2B), companies require robust connectivity to optimize operations and stay competitive. Opportunities abound in expanding access to underserved suburban and rural areas, where gaps in coverage persist as a key challenge. Wind Telecom, leveraging ZTE's advanced technologies like GPON networks and 4G/5G solutions, plays a pivotal role in bridging these gaps. By delivering reliable connectivity nationwide, this partnership drives economic growth and social inclusion, positioning the Dominican Republic for a digitally connected future.

Wind Telecom has evolved remarkably in recent years. Could you tell us about its journey and how it currently positions itself in the country's digital transformation process?

Wind Telecom has transformed from a traditional service provider into a cornerstone of the Dominican Republic's digital transformation. This evolution is anchored by a decade-long partnership with ZTE, which has equipped Wind Telecom with cutting-edge technologies to diversify its offerings. For B2C customers, this includes GPON, 4G wireless home services, and portable devices, while B2B clients benefit from passive and active fiber solutions and managed services. By prioritizing innovation and customer excellence, Wind Telecom positions itself as a leader in reducing the digital divide, enhancing national competitiveness, and fostering a better-connected future.

Wind Telecom has prioritized fiber optic projects, such as GPON networks with ZTE technology. What motivated this decision, and how does it fit into your strategy for delivering scalable connectivity?

Wind Telecom's focus on fiber optic projects, particularly ZTE's GPON technology, stems from the need for versatile, high-capacity connectivity. GPON supports ultra-fast internet, voice, and streaming TV over a single fiber link, meeting diverse demands efficiently. This decision enables scalable growth, cost reduction, and real-time quality monitoring from a centralized operations center. Aligned with Wind Telecom's strategy, this approach

Damián Báez Dorrejo at ZTE's booth during MWC 2025

builds a robust network that addresses current needs and anticipates future demands, ensuring superior digital experiences for both residential and business clients while closing the digital divide nationwide.

The 10-year collaboration between Wind Telecom and ZTE has been fundamental to the success of your projects. What have been the most notable achievements of this partnership in the past year, and what are your expectations for the future working together?

The 10-year partnership with ZTE has delivered significant milestones for Wind Telecom. Key achievements in 2024 include integrating advanced solutions into GPON networks, implementing a centralized monitoring system for enhanced service quality, and achieving substantial infrastructure expansion. Looking forward, Wind Telecom anticipates ZTE's continued support in deploying next-generation technologies like XGSPON and developing innovative B2C and B2B solutions. This collaboration remains essential to sustaining growth, improving connectivity, and driving digital transformation across the Dominican Republic.

What are Wind Telecom's plans for adopting next-generation technologies and leading digital transformation? How does the collaboration with ZTE reinforce your vision of prosperity for the Dominican Republic?

Wind Telecom aims to lead digital transformation by expanding its infrastructure with next-generation technologies such as XGSPON, 5G, and IoT solutions. For the residential segment, plans include launching MIFI, streaming TV, and video surveillance, while the business sector will gain advanced managed services. ZTE's expertise and state-of-the-art equipment are critical to these initiatives. Wind Telecom's vision is a fully connected Dominican Republic, where every citizen has access to digital tools for prosperity. With ZTE as a trusted partner, Wind Telecom is driving economic and social progress, building a digitally inclusive future. ZTE TECHNOLOGIES

MyRepublic

Shaping the Future of High-Speed Internet in Indonesia

Reporter: Huang Lexuan

Timotius Sulaiman
CEO of MyRepublic

MyRepublic Indonesia, a leading provider of fiber-optic internet and subscription TV services in Indonesia, reached a major milestone earlier this year by surpassing 1 million active customers through its nationwide fiber to the home (FTTH) rollout. Timotius Sulaiman, CEO of MyRepublic, shares insights into the company's digital transformation journey.

First of all, congratulations on achieving the milestone of 1 million subscribers! Secondly, how has the journey been so far, and what have you been working on in the past year?

Thank you! Reaching 1 million subscribers is a truly humbling and honorable milestone for us, reflecting years of dedication, innovation, and commitment to delivering high-quality internet services. Our journey has been both challenging and rewarding as we continuously strive to expand our coverage and enhance our services to meet growing customer demands.

Over the past year, we have focused on expanding our network to 56 cities and 88 regencies, adding 3 million new home passes. We have also implemented initiatives to improve service reliability and strengthen the customer experience through advanced digital solutions. Additionally, we have invested in technological innovations to support our long-term vision of making high-speed internet accessible to more people across Indonesia.

What do you think are the key drivers catalyzing your growth in the industry?

Several factors have driven our growth in the industry. First, the increasing demand for high-speed internet, especially for remote work, online education, and entertainment, has accelerated FTTH adoption. Second, our commitment to providing high-quality, symmetrical internet speeds with fiber-optic technology has set us apart from competitors. Third, strong strategic partnerships, effective marketing campaigns, a strong sales force team, and customer-centric innovations have played a crucial role in driving market penetration, including into rural areas. Lastly, our relentless focus on expanding coverage to urban and suburban areas has helped us tap into new markets and drive sustained growth. Beyond expanding connectivity, these initiatives have also created numerous job opportunities, contributing to local economic growth.

How important is your partnership with ZTE in reaching this milestone? What do you expect for future cooperation?

Our partnership with ZTE has been instrumental in achieving this milestone. Their advanced technology solutions, reliable infrastructure, and strong technical support have significantly contributed to the efficiency and scalability of our network. Through this collaboration, we have been able to enhance our service quality and accelerate deployment in various regions.

Moving forward, we expect to further strengthen our partnership with ZTE by leveraging their cutting-edge innovations, such as AI-driven network optimization and next-generation broadband solutions. We believe that continued collaboration will help us achieve even greater milestones in the future.

What will this milestone lead to, and what are the key challenges you foresee related to it?

Reaching 1 million FTTH subscribers is a

stepping stone toward our larger goal of expanding digital connectivity across Indonesia. This milestone will enable us to further invest in network development, service improvements, and customer engagement initiatives. As we move forward, we are embracing our 2025 theme, "Breaking Limits, Beyond Excellence," which reflects our ambition to push boundaries and continuously enhance our services.

At the heart of our journey is MyRepublic Indonesia's vision: "Go Far, Fast, and Beyond." We are committed to accelerating digital transformation, reaching more communities, and delivering world-class internet services at an unprecedented pace.

However, as we scale up, we anticipate several challenges, including the need for continuous infrastructure investments, navigating regulatory requirements, and addressing rising customer expectations for speed, reliability, and affordability. Ensuring seamless service quality while managing rapid expansion will remain one of our key focus areas in the coming years.

What is your outlook for the FTTH market in Indonesia, and how can MyRepublic make a difference in its growth?

Indonesia's FTTH market has tremendous growth potential, driven by increasing digital adoption and government initiatives to boost broadband penetration. We foresee continuous expansion in urban and suburban areas as demand for high-speed internet continues to rise.

Fiber broadband plays a vital role as the foundation and backbone of the nation's Digital Vision 2045, enabling innovation, economic growth, and digital inclusion. At MyRepublic Indonesia, we are committed to building this digital infrastructure—laying the groundwork for a more connected and technologically advanced Indonesia.

With our extensive experience, strong infrastructure, and commitment to innovation, MyRepublic Indonesia is well-positioned to drive this growth. By focusing on accessibility, reliability, and customer satisfaction, we aim to bridge the digital divide and provide seamless internet connectivity to even more households and businesses across Indonesia. ZTE TECHNOLOGIES

AI Ushers in a New Era of Intelligent Communication

Guo Xuefeng
Chief Expert on CCN Product Planning, ZTE

The rapid development of AI technology has become a core driver for a new wave of technological revolution and industrial transformation. From network O&M to new service and capability innovation, AI is reshaping every aspect of the communications field, and is set to have a profound impact on the development of the core network.

Large Models Drive a Leap in AI Capabilities

Large models are driving a qualitative leap in AI capabilities. From natural language understanding and text generation to multimodal interaction, time-series prediction, and complex reasoning, AI has achieved remarkable progress.

Scaling from 100 billion to 10 trillion parameters, large language models (LLMs) demonstrate strong performance in semantic understanding, emotional analysis, text generation, and emergent capabilities, subverting how industries operate. The Transformer architecture has expanded from language processing to more domains such as computer vision, speech recognition, structured data, time-series prediction, and multimodal processing, enabling a huge leap in the capabilities of related domain models. Multimodal LLMs (MLLMs) excel in extending modality and generalizing application scenarios. Time-series models show great potential in power load prediction, energy scheduling optimization, weather prediction, and disaster prediction. Since 2024, mixture of experts (MOE) has developed rapidly, significantly reducing inference costs and making large models more accessible.

At the application framework level, AI agent technology is entering its inaugural year of development. It integrates memory and planning capabilities with LLMs and rapidly builds agents for a variety of application scenarios by linking different tools. Retrieval-augmented generation (RAG) greatly accelerates knowledge injection into large models, driving the deployment of AI applications such as enterprise knowledge management and intelligent Q&A. Emerging architectures and technologies, such as multi-agent coordination, large-small model coordination, and cloud-edge-end collaborative inference, are promoting the integration of large models into complex scenarios.

AI applications are becoming a key business driver for cloud vendors, and an important engine for enterprises to reduce costs and improve efficiency. Large model-based applications are expanding from consumer to business domains, driving digital transformation across industries.

AI Igniting a New Wave of Network Intelligence

In 2024, major operators and equipment vendors worldwide released AI strategies to advance network intelligence, conducting extensive research and practices in domain-specific models, application scenarios, and product solutions.

At MWC 2024, the Global Telco AI Alliance (GTAA) was officially launched by Deutsche Telekom, e& Group, Singtel, SoftBank and SK Telecom. The GTAA aims to develop LLMs for the telecom sector, offering more trustworthy, lower-cost, and more efficient Telco LLMs for members, and enabling AI applications such as virtual agents, fraud filtering, and personal AI assistants.

Deutsche Telekom's app-free AI smartphone unveiled at MWC 2024 is visionary. By applying LLMs, it provides users with a unified portal for trusted access to internet services and the AI world.

Telefonica has formulated a comprehensive AI strategy spanning from strategy formulation to implementation and deployment. The strategy is being implemented at multiple levels, including customer service, service processing, user experience, and content platforms.

The GSMA proposes that AI has brought unlimited potential to the telecom industry but emphasizes the need for responsible AI, advocating for the establishment of industry specifications through principles, strategies, and standards.

In China, the three major operators have released self-developed LLMs for the network realm. China Mobile has launched a series of network LLMs, including a network natural language model, a network structured data model, and a network vision model, to enhance the value of autonomous networks. China Unicom introduced its Yuanjing 2.0 LLM, an updated version to Yuanjing 1.0 designed to empower various sectors. China Telecom unveiled its Xingchen LLM, which includes the Xingchen semantic model, Xingchen voice model, and Xingchen multimodal model. In terms of AI applications, China Telecom was the first to propose using its self-developed Xingchen network model within its enterprise for tasks such as monitoring, troubleshooting, maintenance and optimization processes, improving troubleshooting efficiency by more than 30%.

Standards related to AI in telecom are advancing rapidly. 3GPP defines the network data analytics function (NWDAF), which enables network performance and efficiency improvement on the intelligent plane. AI endogeneity has become a key trend and important research direction in 6G standard development. The TM Forum (TMF) is introducing generative AI into autonomous networks, exploring application scenarios, frameworks, and implementation solutions. CCSA in China is leading the formulation of protocols and specifications to enable network intelligence through multiple large models.

Deep Integration of Core Network and AI, Shaping the Future of Intelligent Communication

In the 5G and 5G-A phases, network capabilities have been greatly improved, yet challenges such as high construction costs, complicated O&M, and insufficient value realization remain. AI technology provides strong support for network intelligence. By incorporating AI and edge computing, the 5G-A core network enhances network intelligence, reduces O&M costs, improves computing network resource efficiency, optimizes network service quality, and expands value-driven service scenarios.

While networks are moving toward multi-factor integrated services, O&M is advancing to higher-order intelligence. Intelligent technologies are required to automate O&M in complex scenarios and tasks. Starting with assistants for interaction, analysis, and generation, the evolution will progress towards decision-making and control scenarios, based on model accuracy and task risk assessments. For example, large models for the O&M field are introduced to link existing AIOps tools to enable the creation of agents for alarm analysis, fault diagnosis, and repair. Multiple agents can collaborate to orchestrate a cascaded flow for complex tasks, achieving automated network fault diagnosis and repair. Time-series models can analyze historical fault data and real-time network operation data to facilitate proactive fault prediction, and automatically take corresponding repair measures or provide optimization suggestions.

> **The development and application of AI and large models drive the evolution of core network intelligence. In turn, the core network provides powerful connectivity and computing support for AI.**

As 5G networks rapidly advance and user requirements diversify, network operations need to shift from scale-based to performance-based—focusing on user value mining, expanding network and service innovations, and providing differentiated and refined services for users. At the product design level, AI and large models are leveraged to support operation analysis, product design, and user retention. In terms of service innovation, large voice models are used to recognize user intent via voice interaction, addressing the issue of poor interaction in the new-calling video mode. LLMs and MLLMs enable real-time translation and interesting calls. Meanwhile, large anti-fraud models perform emotional analysis and intent recognition, solving the problem of low accuracy in traditional rule-based fraud call detection. At the service level, telecom intelligent customer services powered by large language models evolve from menu-based interactions to natural language interaction, from semantic understanding to emotional perception, and from domain experts to encyclopedia-level experts, offering 24/7 high-quality online services.

The vision of 6G is to create a more intelligent network environment. AI for Network will evolve from a plug-in feature to an endogenous capability, with AI integrated into all layers of the network architecture. At the same time, AI applications will become ubiquitous, and Network for AI will become the main theme. The core network should prioritize AI applications as its primary target scenario, enhancing its architecture, capabilities, and performance to support more extensive and complex scenarios.

- **Innovating architecture for network and AI integration:** The 6G core network needs to innovate its network architecture to deeply integrate computing, network, and AI, achieving endogenous AI. For example, by deploying a collaborative computing and network platform at the network edge and in the cloud, the AI model can be trained and inferred in the most appropriate locations. Edge nodes will process AI tasks with high real-time requirements, while the cloud will be responsible for large-scale and complicated AI computing. The network will intelligently allocate tasks to different nodes.
- **Meeting complex AI application requirements:** Based on integrated communication-sensing and communication-intelligence, the network can automatically perceive the computing and network resource requirements of AI tasks—such as large-scale deep learning model training or inference—and dynamically schedules network topology, bandwidth, and computing nodes to better support complex new AI services. With an integrated space-air-ground-sea network architecture, the system can provide seamless connectivity for widely distributed AI sensors and computing devices, providing real-time massive data transmission for immersive services such as holographic communication, immersive communication, and glasses-free 3D, ensuring high-quality information transfer and presentation. It also guarantees sufficient bandwidth, low-latency performance, and data security for large-scale AI model training.

The deep integration and mutual reinforcement between the core network and AI is an inevitable trend. The development and application of AI and large models drive the evolution of core network intelligence. In turn, the core network provides powerful connectivity and computing support for AI, meeting users' requirements for high-quality communication and intelligent services, while providing the foundation for the digital transformation of industries. ZTE TECHNOLOGIES

AI+ Core Network: Towards 6G Evolution

Zhou Jianfeng
Chief CCN Preresearch Planning Engineer, ZTE

The evolution of the next-generation core network is driven by AI. While the AI architecture shifts from centralized external plug-ins to ubiquitous endogenous integration, AI is driving the network architecture change, evolving the 6G core network from traditional data transmission into the neural center of intelligent services.

AI Architecture Transformation

Fig. 1 shows the endogenous intelligent architecture of the core network. The evolution of the AI architecture is reflected in the following aspects:

- **Centralized to ubiquitous:** The system evolves from a centralized AI architecture, centered around the network data analysis function/management data analysis function (NWDAF/MDAF), to a distributed, ubiquitous AI architecture, enabling better adaptation to dynamic user requirements and traffic patterns and the delivery of more flexible and personalized services.
- **AI for Network to Network for AI:** The evolution is shifting from network intelligence that optimizes network performance, efficiency, and user experience through AI, to AI as a service (AIaaS) that provides end-to-end service guarantee and capability services for AI applications.
- **Single-body to multi-body:** AI serves as the core to coordinate cross-layer and cross-domain resources, meeting the connectivity, computing, data, and model requirements of various real-time services such as immersive experiences.
- **High-order intelligence with AI+ digital twin:** The AI agent understands intent and can reflect the real-time behavior and performance of the real physical world through digital twin construction, enabling accurate spatiotemporal inference for the physical world. This is applicable to scenarios such as autonomous driving, robotics, intelligent industrial production, and remote healthcare.

AI Leading Network Architecture Change

To address the challenges of multiple NEs, complicated interaction, and the slow launch of new functions in 5G/5G-A, as well as to meet the future 6G requirements such as immersive communication, endogenous intelligence and integrated sensing and communication, ZTE proposes a unified, intelligent service-based architecture (iSBA) powered by AI and enhanced by dual channels and AI-agent multi-element coordination. Built on iSBA, ZTE has created the intelligent distributed communication network (iDCN) and the AI-agent real-time communication network (ARCN), laying the foundation

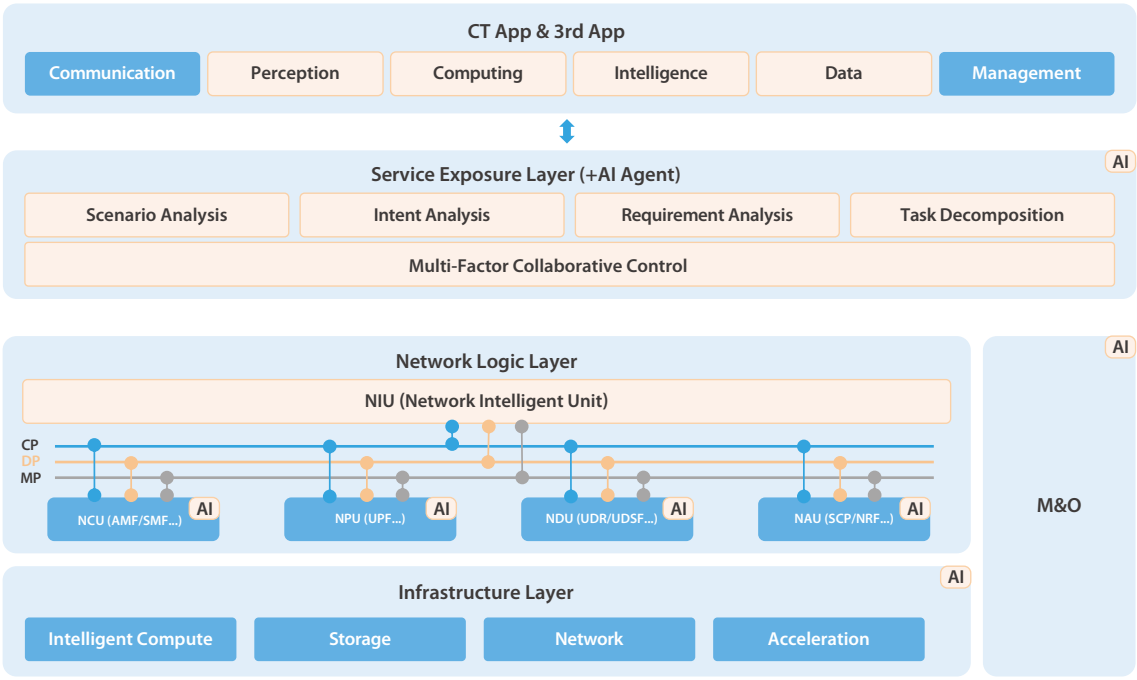


Fig. 1 Endogenous intelligence of the core network.

for the 6G embodied AI core network: AI Core.

Unified Network Architecture Base

iSBA is the cornerstone of the 6G network architecture. While cloudification + SBA/eSBA networks have enabled the success of the 5G/5G-A network architecture, the iSBA introduces new key features such as intelligence, service-based design, and multi-element integration:

- **Intelligence (intelligent base):** iSBA provides a unified intelligent base encompassing GPU, AI OS, and AI cloud full-stack intelligence. The iDCN and ARCN, built on this base, not only saves software and hardware resources, but also reuses AI's basic functions and frameworks—facilitating network-service coordination and network-media integration.
- **Service-based (high-speed channel):** Based on the service-based interface (SBI), SBA/eSBA unifies and simplifies the signaling interaction interface between NEs. The iSBA further introduces a data communication interface (DCI) to establish dual channels together with SBI. SBI and DCI work collaboratively: SBI transfers signaling data, while DCI transfers large data streams, enabling high-performance interaction between NEs. In addition, iSBA integrates and streamlines 5G NEs, proposing an ultra-simple network architecture that supports microservices, plug-ins, agile frameworks, and plug-and-play capabilities.
- **Multi-element (integrated scheduling):** In the future, evolution will shift from single communication to the integration of multiple elements—including communication, sensing, computing, and intelligence—achieving unified scheduling based on AI agents. The system will leverage AI for multimodal intent identification, decomposes complex tasks, and schedules atomic capabilities across multiple elements, orchestrates resources and procedures, and visualizes resource distribution through digital twins.

Intelligent Distributed Communication Network

During the evolution to 6G, the current network faces many challenges, including increased architectural complexity, limited automation in O&M, and growing security and reliability risks. In the future, the 6G network will need to support more types of users and services. The number and types of terminals as well as the number of subnets will increase exponentially, imposing higher requirements on network performance and service experience. These new service requirements, along with existing network issues, drive the evolution of

networks towards iDCN (Fig. 2).

The iDCN network has the following key features:

- **Intelligence:** Evolves from external AI to internal collaboration to meet the end-to-end intelligent collaboration requirements of future networks and services. Within a domain, centralized management and control are combined with distributed coordination to form hierarchical intelligence and enable autonomous intra-domain networks. Across domains, intelligent coordination is achieved via AI agents.
- **Simplicity:** Evolves from NE interconnection to subnet interconnection, providing an architecture basis for fast service and network expansion. Within a domain, inter-NE signaling interactions are minimized through the aggregation of information and functions. Between domains, subnet interconnection and discovery replace NE interconnection and discovery, simplifying the topology and procedures for interconnection discovery and selection.
- **Flexibility:** Evolves from manual configuration to plug-and-play, enabling fast service deployment, dynamic resource sharing, and network automation and intelligence. Within a domain, network functions can be extended through plug-ins, reducing the impact on peripheral NEs and the TTM for new functions and services. Across domains, network-level plug-and-play is enabled through the automatic management mechanism of on-demand subnet creation and deletion.
- **Security:** Evolves from physical protection to a virtual-physical symbiosis to improve both network security and reliability. Within a domain, the network digital twins (NDT) technology is introduced to enhance user, network, and service state awareness, improving active immunity. Inter-domain isolation protection, topology hiding, and secure access are implemented through the network assisted unit (NAU).

Three key technologies in the iDCN include:

- **Hierarchical NAU interconnection:** Mutual discovery, disaster recovery, and trust assurance between NAUs replace the existing inter-NE mesh interconnection architecture.
- **AI agent coordination:** The NAU integrates AI agent capabilities to support autonomous task processing and decision-making, perceive the status and capabilities of each functional service within the domain, and collaborate with other inter-domain network nodes and agent functions to improve the intelligent management and operational optimization capabilities of autonomous networks.
- **NDT virtual-real coexistence:** By constructing a digital twin system for the core network, real-time

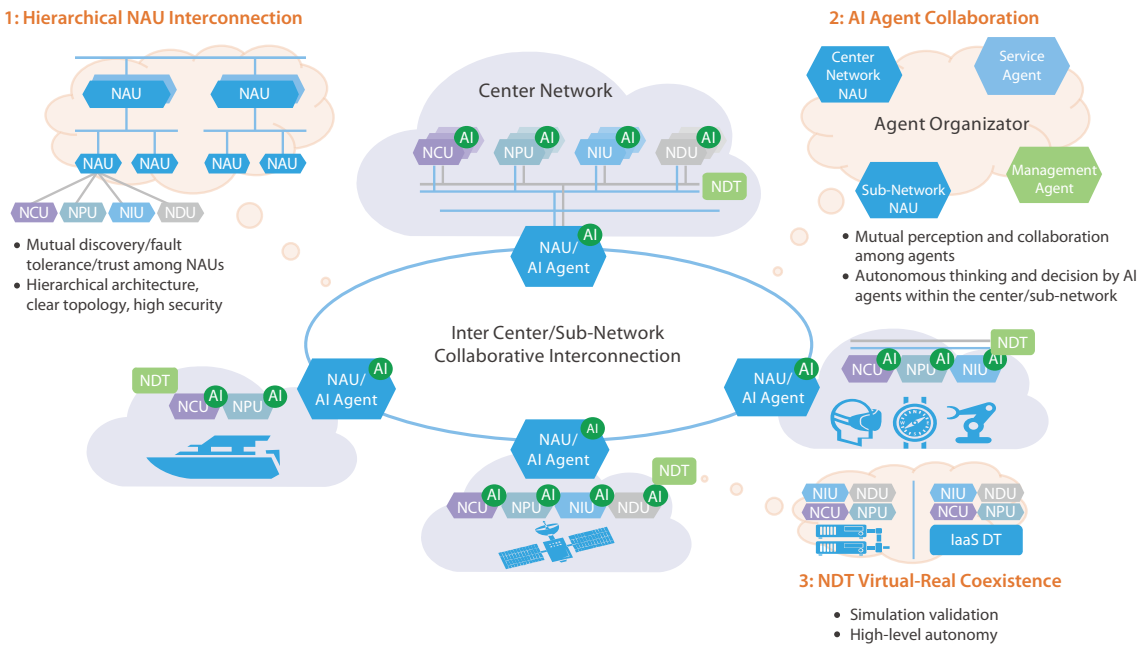


Fig. 2 Intelligent distributed communication network architecture.

AI is driving the evolution of networks toward 6G, simplifying the network and making it more flexible and customizable. It will empower 6G to serve all industries and the intelligent world, supporting sustainable and high-speed development of society.

perception and predictive analysis capabilities are enabled. Simulation-based verification and virtual-real interaction guarantee the network's operating system and improve its intelligent autonomy.

Agent Real-Time Communication Network

In the future, real-time communication will evolve from "audio and video" to multimodal communication involving touch, taste, and perception; from "human-object" communication to "human-object-virtual" communication; and from "real body" to "avatar or digital counterpart" communication, enabling real-time communication services for 6G-native intelligent agents. Based on the unified iSBA, the 6G core network integrates the data, media, and intelligent planes of the 6G network and real-time communications, offering a unified, plug-in-based intelligent architecture that supports multimodal communication between agents.

- **Intelligence empowering network (AI for ARCN):** Through endogenous AI agents, the system provides 6G-native real-time communication services to enable multimodal "human-object-virtual" communication. The AI agent, with LLM/MLLM at its core, performs task decomposition, inference, decision-making, reflection, and self-learning. It uses technologies such as RAG for short- and long-term data storage, inputs data into the LLM/MLLM as context, completes tasks through the tools, and extends real-time communication services from the physical world to the digital world.
- **Network empowering intelligence (ARCN for AI):** The AI agents will be distributed across terminals, wireless networks, core networks, and even third-party networks. The iSBA dual channels (SBI and DCI) provide efficient transport capacity for coordination between agents, as well as for data and model transmission, thereby enabling high-speed and deterministic real-time communication between AI agents and between AI agents and humans.

Multi-Element Coordination

The 6G network needs to solve the challenge of coordinating diverse resources and capabilities, enabling different services to be scheduled to the optimal compute node via the optimal network path, so as to meet the specific QoS requirements and achieve optimal alignment between requirements and resources.

Multi-element coordination refers to the coordination of computing and network resources under diverse QoS constraints. Elements denote the factors being coordinated, while computing and network resources are the objects of coordination. Key technologies include multi-element identification, multi-element perception, multi-element measurement, multi-element orchestration and scheduling, multi-element QoS, and multi-element coordination capability openness.

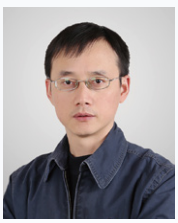
Multi-element coordination spans multiple domains, including cloud, edge, terminal, RAN, and core network, and covers multiple dimensions, such as communication, sensing, intelligence, computing, and security. It is challenging to introduce integrated AI agent capabilities to solve the complexity of multi-element coordination, provide differentiated and customized services for users, and achieve optimal matching between requirements and resources at minimum cost.

AI is driving the evolution of networks toward 6G, simplifying the network and making it more flexible and customizable. It will empower 6G to serve all industries and the intelligent world, supporting sustainable and high-speed development of society. ZTE TECHNOLOGIES

Focusing on Four Key Areas of Intelligence to Build a Full-Stack AI+ Core Network

Wang Weibin

Chief Scientist of Product Planning, ZTE

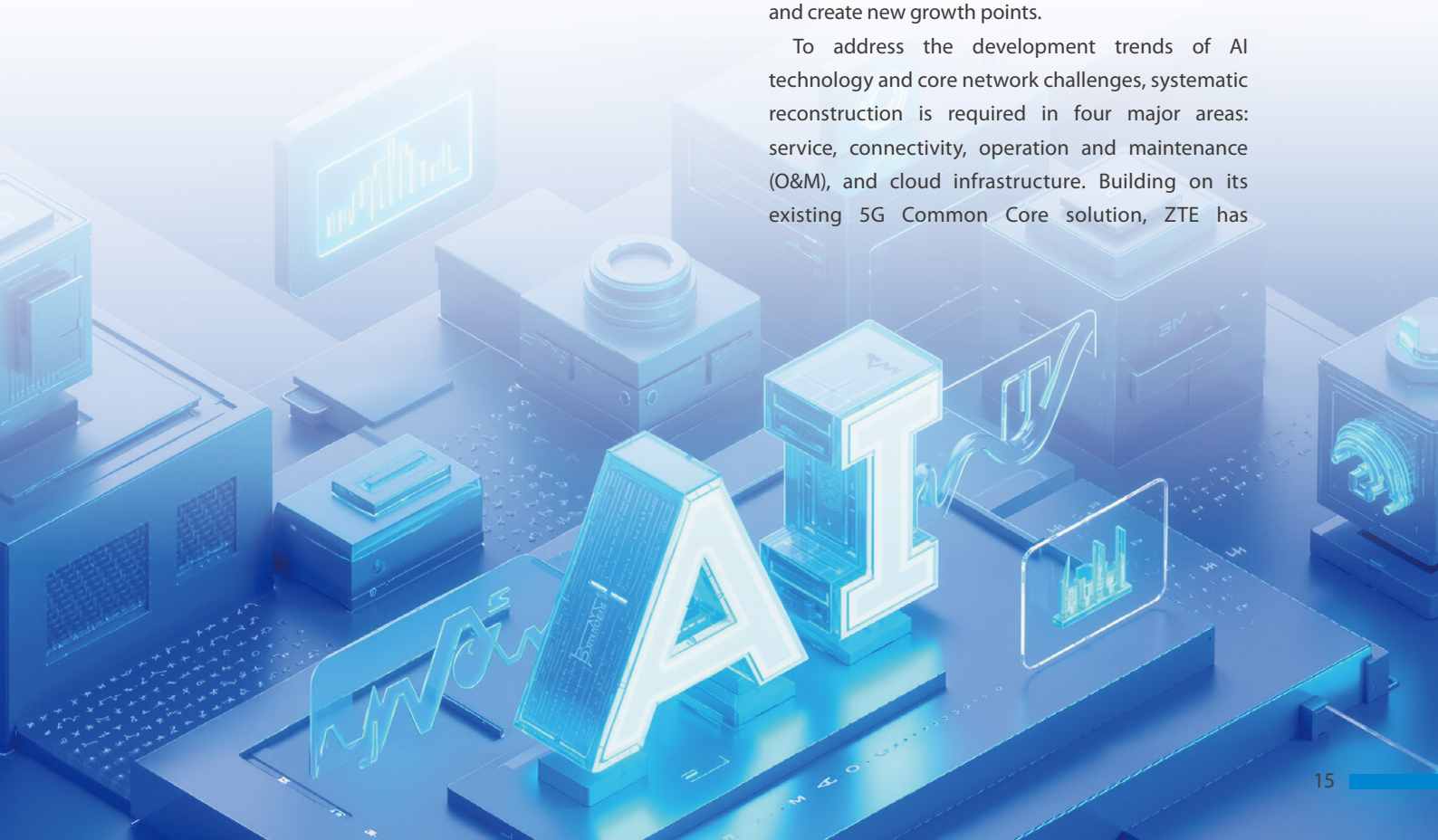


Lu Guanghui

Chief CCN Product Architect, ZTE

The world is entering a new era of digitalization, intelligence, and networking, driving major transformations in networks and AI. In the network field, 5G has entered its second phase, while 6G development has begun. In the AI field, generative AI has made breakthrough progress. AI has become the engine of new-quality productivity, shifting industries from "+AI" to "AI+". As the brain of the mobile network, the core network urgently needs a complete "AI+" transformation to enhance its autonomous and innovation capabilities, enable differentiated services, and create new growth points.

To address the development trends of AI technology and core network challenges, systematic reconstruction is required in four major areas: service, connectivity, operation and maintenance (O&M), and cloud infrastructure. Building on its existing 5G Common Core solution, ZTE has



launched the AI+ core network solution: the AI-reshaped core (AIR Core). In the short term (5G-A), the solution focuses on core network experience and efficiency, while in the long term (6G), it aims at evolving the core network's native intelligent architecture. It creates a full-stack intelligent "new network brain" solution built on an open ecosystem, helping operators accelerate 5G monetization and the deployment of AI applications.

The AI+ core network solution consists of four areas of intelligence, built on a "three-layer, one-domain" architecture (Fig. 1):

- **AI+ cloud infrastructure layer:** Leverages intelligent computing hardware resources to provide AI capabilities for core network applications.
- **AI+ connectivity layer:** Provides the capabilities to stimulate traffic and enhance the experience, enabling the shift from "traffic-centric" to "experience-centric" operation.
- **AI+ service application layer:** Delivers intelligent capabilities for messaging and new calling services.
- **AI+ O&M domain:** Redefines the O&M paradigm, supporting the evolution of the autonomous network towards unmanned operation.

AI+ Service: Multimodal Interactive New Services

OTT is eroding tradtional voice and messaging traffic. The AI+ service leverages AI technology to enable multimodal, interactive, and immersive service experiences for real-time voice and messaging, rejuvenating traditional voice and messaging communication services.

AI+ 5G New Calling enables intent-driven interaction between humans and machines by integrating AI capabilities. It intelligently orchestrates multiple service capabilities, allowing for more complex services, such as "switch to a cartoon avatar" and "answer the call on my behalf" within a single interaction. This upgrades a simple voice call to multimodal communication, delivering a more intelligent and diversified calling experience.

AI+ messaging introduces an intelligent plane and integrating it with the messaging plane to create a unified AI+ messaging portal. It provides three categories of services: AI+ messaging services, AI+ information services, and AI+ application services—catering to individual consumers (ToC), industries and enterprises (ToB), households (ToH), and other sectors (ToO), creating a new-generation business model. No modifications or upgrades are required for mobile terminals.

AI+ Connectivity: Experience Monetization & Differentiated Network Operation

5G networks must shift from traffic-centric to experience-centric operation. Traditional user service assurance is achieved through QoS

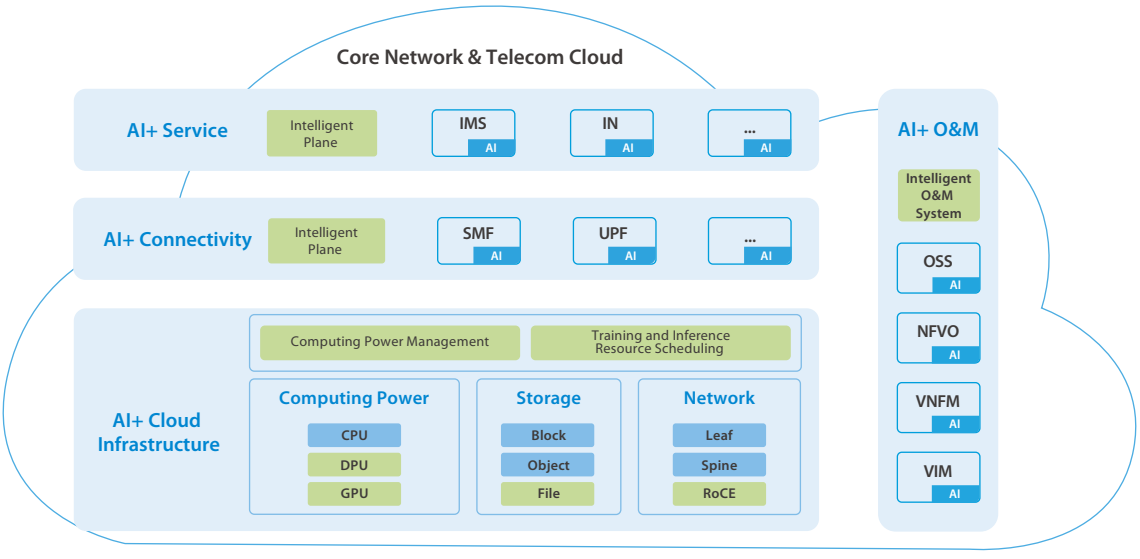


Fig. 1 ZTE's AI+ core network (three-layer, one-domain) solution.

subscriptions, which cannot perceive real-time service quality, resulting in a poor user experience and making experience-based operation difficult to implement.

AI+ connectivity builds an intelligent plane on the 5GC side, with the network data analytics function (NWDAF) acting as the brain, the policy control function (PCF) as the policy anchor, and the AI engine embedded in network elements (NEs) as the executor. Collaborating with the wireless network, it constructs end-to-end connection intelligence that enables all-round profiling of users, services, and networks. This provides operators with a means to explore the value of connectivity, achieving real-time user experience perception, experience monetization, and intrinsic security.

In terms of experience monetization, intelligent service identification and service experience measurement are crucial. The introduction of AI technology has greatly improved service identification and experience measurement, with the update cycle for the deep packet inspection (DPI) library also shortened from weekly to hourly, greatly enhancing dynamic identification, experience measurement, and response capabilities for online services. Operators can utilize intelligent service identification and service experience measurement to accurately analyze user experience, adjust data channel QoS accordingly, and implement differentiated services, greatly enhancing user experience, realizing experience monetization, and enabling refined operations.

Regarding intrinsic security, the solution leverages the built-in intelligent engine, network operation data, and AI algorithms to predict the states of terminals and NEs, enabling efficient node selection and service orchestration, end-to-end load management, energy-saving management, and signaling storm prevention. For energy savings, improved prediction accuracy of idle states in NEs enables dynamic and precise energy management. For example, the power-saving strategy is initiated before an idle period and deactivated before a busy period. For signaling storm prevention, the solution intelligently identifies abnormal terminals, preventing network attacks and signaling storms. Additionally, based on the digital twin network, it simulates signaling storm scenarios and takes preventive measures proactively.

AI+ O&M: A New Paradigm in Advanced Autonomous O&M

The core network features a large variety and quantity of NEs, as well as decentralized deployments. The adoption of cloudification technology, combined with the stringent requirements for high network stability, has intensified the pressure and challenges on core network O&M. AI technology presents unprecedented opportunities to transform core network O&M. ZTE's AI+ O&M combines large AI models with digital twins, promoting the evolution of network O&M towards a higher level of autonomy.

- A large language model (LLM) is used to build a core network signaling model. By analyzing network signaling in real time, the model can detect anomalies, identify potential problems, and automatically adjust signaling policies to optimize network performance and user experience.
- By integrating large-model technology, AI for IT operations (AIOps) implements end-to-end management—from fault detection to repair—through multi-agent coordination during the O&M process. No longer limited to specific application scenarios, AIOps demonstrates powerful generalization capability, allowing it to automatically adapt to different network environments and diagnose and resolve network faults through agent coordination.
- By introducing an intent-based network and integrating it with digital twin technology for intent verification and network interaction, precise execution of operational intents—such as SLA assurance and energy efficiency—can be achieved, paving the way for an advanced autonomous network.

AI+ Cloud: High-Efficiency, High-Stability New Infrastructure

The cloud infrastructure layer is the foundation for running the core network, consisting of cloud operating systems, servers, storage devices, network devices, and other software and hardware. The AI+ cloud encompasses two aspects of

Fig. 2 ZTE's integrated intelligence, computing, and application solution.

intelligence: on the one hand, it provides an efficient and diverse computing power resource pool to empower the development of the core network's AI capabilities; on the other hand, it introduces AI technology into infrastructure management.

- Provide an efficient and diverse computing power resource pool to empower the construction of the core network's scenario-based AI capabilities. With evolution of cloud infrastructure from the traditional CPU-centric general computing to CPU+GPU-centric general-intelligent collaborative computing, it is necessary to support technologies such as computing power pooling, orchestration and scheduling, high-performance parallel storage access, and high-channel lossless networks to ensure the efficiency and stability of resource supply. At the product level, to meet the demands of fine-tuning and real-time inference scenarios, ZTE has launched a training and inference integrated machine that is ready to use out of the box (see Fig. 2).
- Introducing AI technology in infrastructure management enables multi-dimensional intelligent resource scheduling and orchestration from a business perspective, as well as intelligent O&M featuring automatic pre-event prediction and early warnings, intelligent in-event discovery, delineation and restoration, and post-event review and optimization. This leads to high infrastructure stability, smarter O&M, and efficient resource utilization.

ZTE: Empowering Customers to Continuously Create New Value

Through cooperation with leading operators, ZTE's full-stack AI+ core network solution addresses both the rapid monetization of 5G-A networks and the implementation of AI applications, while also laying the foundation for future 6G networks.

In service intelligence, ZTE released "Smart Safeguard", the industry's first SMS anti-fraud system based on a large model, and successfully deployed the world's first modular AI+ 5G new calling network.

In connectivity intelligence, ZTE, in partnership with operators, created the industry's first commercial project for hierarchical VIP user guarantees.

In O&M intelligence, ZTE pioneered a quantitative evaluation system for resource pool switching based on digital twins. By considering the relationship between NEs and resource pools, it simulates and quantitatively analyzes the resource pool switching process.

ZTE will closely follow AI and network advancements, actively innovate, and fully support operators in building AI+ core networks, to develop a future-oriented "new network brain". ZTE TECHNOLOGIES

AI+ Telecom Cloud: Trends and Key Technologies

Zhu Kun

Chief Cloud Computing Planning Engineer, ZTE

With the debut of ChatGPT, AI technologies have accelerated rapidly, making the intelligent transformation of core network an inevitable trend. As the computing infrastructure platform for the core network, the intelligent transformation of the telecom cloud is a key step in this process.

To meet the high-performance, large-scale parallel processing, and low-latency interconnection requirements of AI model training and inference, the telecom cloud is evolving from traditional CPU-centric general computing to heterogeneous computing centered on DPUs, GPUs, and NPUs. It supports technologies such as computing resource pooling and orchestration, high-performance parallel storage access, high-speed lossless network, and compute-native architecture, ensuring efficient and stable resource provisioning. In deployment, the AI+ telecom cloud adopts hybrid pooling of intelligent and general computing resources and a distributed architecture to better support the intelligent upgrade of the core network.

Resource Pooling Significantly Improves Infrastructure Utilization

Computing pooling leverages software-defined hardware acceleration to enable the more efficient and flexible aggregation, scheduling, and release of massive AI acceleration computing power through capabilities such as GPU virtualization, multi-card aggregation, remote invocation, and dynamic release. It ensures precise end-to-end computing allocation for AI model development, training, deployment, testing, and release, maximizing resource utilization and improving the overall efficiency of the intelligent computing center.

Unified memory pooling, based on the computing bus protocol, achieves consistent memory semantics and spatial addressing capabilities. It integrates multiple physical memory devices or memory resources into a logical memory pool, enabling unified scheduling, monitoring, and management. This technology dynamically allocates and releases memory resources, flexibly adjusting them according to application needs, thus avoiding frequent data movement between compute units and memory devices such as cache, HBM, and DDR during large model training. It improves overall system performance while reducing development complexity and error rates.

Intelligent Computing Storage Meets High Performance and High Concurrency Challenges in Training and Inference

During the various stages of end-to-end large model development, there is a growing demand for innovation in storage, particularly in terms of massive, diversified capacity and high-performance concurrency. Therefore, intelligent computing storage must offer the following features:

- **Unified storage platform:** Builds a unified storage platform that supports AI processing flows across different phases, providing diverse data storage capabilities and multi-protocol interoperability.
- **Comprehensive software and hardware optimization for improved performance:**

Hardware-level acceleration include offloading storage interface protocols via DPUs, performing operations such as deduplication, compression, and security, and automatically tiering and partitioning data based on access frequency. Software tuning methods include distributed caching, a parallel file access system, and private clients.

- **Data entropy reduction:** Reduces unnecessary data movement and duplication, and optimizes storage and access policies to reduce the "data entropy tax". Data transmission and storage overheads are reduced through technologies such as deduplication and compression.

Open High-Speed Lossless Network Reduces Parallel Computing Overhead

Parallel computing improves overall computing efficiency in AI large-model training, but it brings synchronization overhead and communication latency. A critical industry challenge is how to achieve high-speed interconnection between GPUs in a super-large-scale intelligent computing cluster to significantly improve GPU utilization.

In scale-up networks, GPU high-speed open interconnection technology based on a switching topology replaces traditional point-to-point GPU communication with a switching-based interconnection mode, enhancing the scalability and communication bandwidth of a single server, breaking the eight-GPU limit, and greatly improving cluster computing power.

The scale-out interconnection network between super-node servers is also important for solving bottlenecks such as communication bandwidth and latency in model training. While RoCEv2 is an open Ethernet-based solution, vendors have generally developed their own enhancements—such as congestion control and end-network coordination—which are often tied to their own switching devices, making decoupling challenging. Therefore, a key industry goal is to provide an open and comprehensive RoCE solution based on RoCEv2.

Compute-Native Technology Enables a Decoupled Ecosystem for Heterogeneous Computing

With the advancement of AI chip technology and manufacturers no longer restricted to a few brands, heterogeneous computing pools based on diverse infrastructure environments and GPU types are becoming the future trend.

Compute-native technology ensures that applications can request computing power based on a

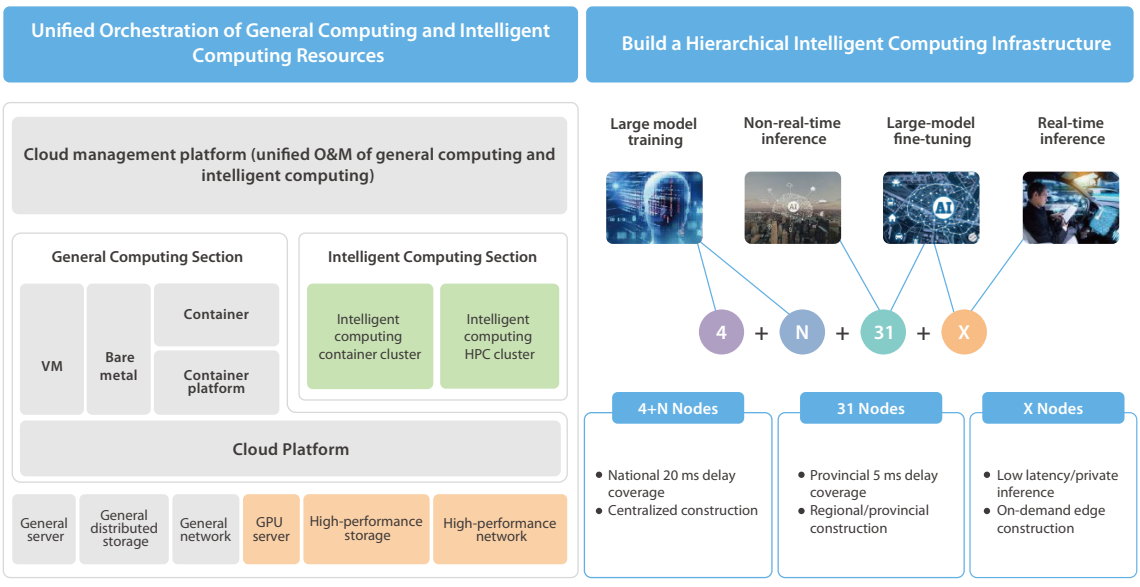


Fig. 1 Distributed hybrid computing pool deployment architecture for AI+ telecom cloud.

uniformly defined intelligent computing value. The compute-native layer provides GPU resources corresponding to the computing value, an interface for resource invocation that shields vendor differences, and an application compilation and runtime environment independent of vendors. This setup shields the underlying heterogeneous GPU resources and fully decouples upper-layer AI framework applications from the underlying GPU type.

Distributed Hybrid Deployment Meets Core Network Applications' Comprehensive Resource Requirements

Core network NEs require both general and intelligent computing infrastructure resources, while training and inference applications also demand distributed deployment. Therefore, hybrid pooling and distributed deployment of general and intelligent computing resources have become key features of AI+ telecom cloud deployment (see Fig. 1).

The telecom cloud is seamlessly transitioning from general computing to intelligent computing resource pools, with unified orchestration and management of both as a key feature. A centralized cloud platform typically manages general and intelligent computing infrastructure resources, such as computing, storage, and network, while the orchestration of general and intelligent computing resources is handled by the telecom cloud management platform.

Pre-training of foundational large models, precision tuning of industrial large models, and fine-tuning of large models for customer scenarios all demand varied computing power and deployment locations. In line with the hierarchical architecture of operators' telecom clouds, the AI+ telecom cloud also adopts a three-level deployment model: a hub-level large model training center, regional resource pools for integrated training and inference, and edge-level all-in-one training and inference machines.

The transformation of the telecom cloud across computing, storage, network, orchestration, and deployment provides an infrastructure foundation for the intelligent evolution of core network services and O&M. With a full range of intelligent computing products and extensive experience in end-to-end intelligent computing center deployment, ZTE is well-positioned to help operators drive the intelligent transformation of their core networks. ZTE TECHNOLOGIES

Building Trusted Multidimensional Agent Services with AI+ Messaging

Huang Xiaobing

Chief Engineer of Messaging Products, ZTE

Gartner predicts that by 2025, generative AI will be embedded in 80% of conversational AI offerings. As a basic telecom service centered on message-based dialog and interaction, 5G messaging is a typical entry point for conversational AI. In China, there are already more than 1.3 million ToB Chatbot applications, a figure estimated to grow to tens of millions in three years. These applications are characterized by strong demands for general AI dialogs, industrial knowledge, enterprise applications, and domain agents, offering great development value. For ToC, the focus is on building high-frequency, must-have personal agents. All of these trends present important opportunities for AI+ messaging.

Three Layers and Two Planes: Creating Three Types of Portals for AI+ Messaging

By introducing an intelligent plane and seamlessly integrating it with the messaging plane, the network-side messaging platform forms a three-layer, two-plane AI+ messaging architecture. It creates three service portals—AI+ messaging service, AI+ information service, and AI+ application service—serving individual (ToC), business (ToB), home (ToH), and other (ToO) domains, while guaranteeing the reliability and trustworthiness of message content (see Fig. 1).

The architecture uses the intelligent control layer as its brain, aggregating and scheduling capabilities across all layers and planes.

- **Data layer:** Collects raw CDRs, data, and corpus from the daily operation of the network, platform, and BOSS. After data processing and cleaning, it provides core, long-term corpus for the intelligent layer.
- **Intelligent layer:** Includes large and small models and algorithms. Through corpus training, fine-tuning, or continuous learning, it natively generates various large AI model applications, which are tightly integrated with the messaging plane to complete the service process.
- **Intelligent control layer:** Acts as the front-end brain, offering capabilities such as intent recognition, task orchestration and distribution, template adaptation, and service subscription. It is the control center of AI+ messaging.
- **Messaging plane:** Provides mobile communication messaging services, capabilities, and channels, offering services such as SMS, MMS, 5G messaging, and industrial messaging.
- **Intelligent plane:** As a new capability plane, it includes the intelligent control layer, intelligent layer, and data layer, integrating with the messaging plane to enable the AI+ capability upgrade, and providing trusted multimodal agent services for end users through three main service portals.

MLLM-Enhanced Anti-Fraud Service for Trusted Messaging Communication

According to the 2024 report from the Global Anti-Scam Alliance (GASA), consumers lost about US$1.03 trillion to to online scams in 2023. As scams continue to evolve and fraud incidents increase, they cause huge economic losses and social impacts that traditional governance solutions cannot address. The multimodal anti-fraud model application for AI+ messaging enable the upgrade from the traditional governance model to an AI-driven approach, guaranteeing the credibility of new communications.

Built on large language models (LLMs), computer vision (CV) models, and multimodal LLMs (MLLMs), this anti-fraud solution can identify and process

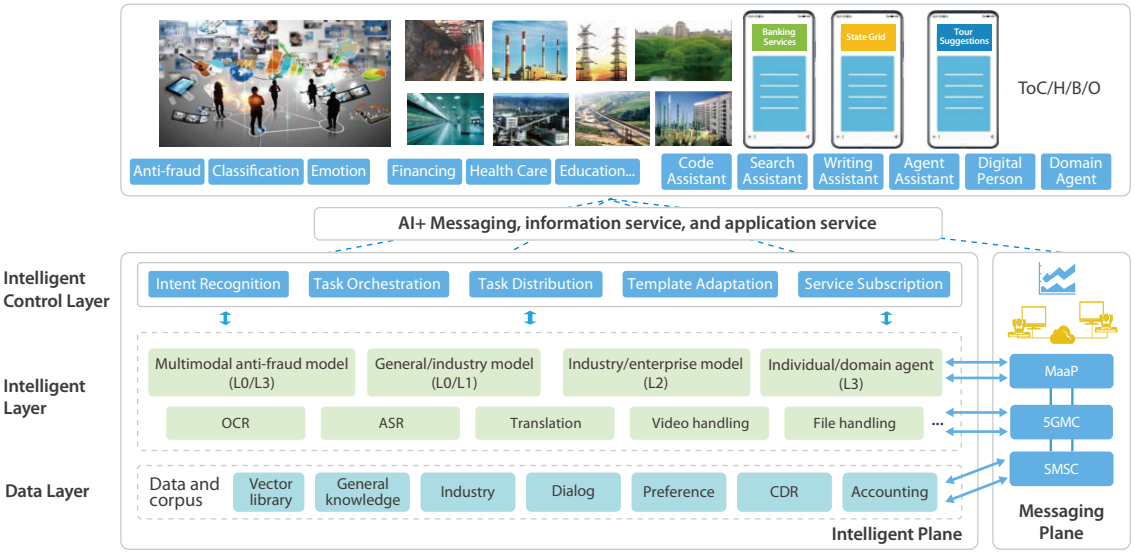


Fig. 1 The three-layer, two-plane architecture of the AI+ messaging platform.

various media types, including text, images, audio, video, and files. It features powerful semantic analysis, emotional recognition, intent analysis, and logical reasoning capabilities, greatly improving the accuracy of identifying malicious messages.

The multimodal anti-fraud LLM is highly efficient, precise, and comprehensive. With a well-curated corpus, fine-tuning, and prompt design, its accuracy consistently exceeds 95%, far surpassing traditional network governance solutions. This greatly improves the user experience, creating reliable, trusted, and worry-free communication services.

Endogenous Multimodal Agent: Messaging as an AI Service (MaaAIS)

To address the poor user experience of current 5G messaging services, as well as the high threshold, high cost, and long cycle of enabling AI applications in industry chatbots, the AI+ messaging solution, based on the platform's endogenous multimodal agent, implements the MaaAIS to enable ubiquitous AI services.

For ToB services, network-wide chatbots and domain agents offer the following features: seamless activation with no need for development or reconstruction; instant and batch commissioning; QoS guarantees; high security with data remaining within the network; and minimal cost.

For ToC services, long-term accumulation and learning from platform-level corpus and data allow the creation of personal agents that understand users best.

The endogenous multimodal agent also supports the AI-powered dialog functions across various types of messages, including SMS, MMS, and 5G messaging, bidirectional multimodal interaction, automatic web search generation, document summarization, high-precision intent recognition, and automatic card template adaptation, providing an optimal user experience.

Future Prospects

The upgrade and transformation of AI+ messaging will accelerate the development of multimodal agent entry services that are terminal-native, accessible via phone numbers, and highly reliable, driving advances in technology, services, and value:

- **Technology upgrade:** Integrates and develops the intelligent plane to complete the technological evolution and build next-generation messaging infrastructure.
- **Service upgrade:** Transitions from basic messaging communication to AI+ information service and AI+ application service.
- **Value upgrade:** Shifts from the traditional per-message charging model to a differentiated AI+ service-based charging model, creating new and incremental business models.

With the evolution of communications networks to 6G, AI+ messaging will pave the way for a a smarter world through messaging. ZTE TECHNOLOGIES

AI+ New Calling: A New Start for Voice Service

Ni Ming
Chief IMS Planning Engineer, ZTE

The development of 5G New Calling is driven not only by users' demand to move beyond single forms of call content and interaction, but also by operators' need to respond to external challenges. The rise of AI technologies has intensified the competition from terminal companies and OTT players, pressuring operators to provide better services and gain a competitive advantage.

By the end of 2024, China had more than 1 billion 5G mobile phone users. Voice service remains fundamental, and the large user base provides operators with a strong entry point into voice service offerings. By integrating AI technology into the audio, video, and data channels of New Calling, multimodal interaction becomes possible. The native dialer is evolving into a new entry point for intelligent applications, helping operators to shift from managing call duration to operating multimodal content.

AI+ New Calling Solution and Application Exploration

To develop innovative services for New Calling, ZTE has continuously enhanced the IMS architecture of the foundational real-time communication network by incorporating data channels, capability exposure, applets, and AI technologies.

In terms of AI technology support, ZTE and China Mobile jointly launched media-plane plug-in technology to meet the low latency, high reliability, and easy scalability requirements for intelligent multimodal conversion and processing of innovative services. This enables the online deployment of AI capability through an agile base solution, facilitating the rapid rollout of innovative services.

As shown in Fig. 1, the media-plane plug-in architecture is divided into four layers: service, framework, capability plug-in, and support. The service, framework, and support layers provide the system's basic functions, while the capability layer is an optional function that can be loaded as required.

- **Service layer:** Handles media service interfaces and converts them into plug-in orchestration.
- **Framework layer:** Contains the pipeline engine, loads plug-ins according to the orchestration, and executes the pipeline.
- **Capability plug-in layer:** Supports both self-owned and third-party plug-ins, include basic media processing plug-ins (e.g., codec), multimode media processing plug-ins (e.g., XR), and AI processing plug-ins (e.g., face and intent recognition).
- **Support layer:** Provides the runtime environment for plug-ins.

The architecture decouples plug-in capabilities from the system. With system stability as a prerequisite, new capabilities and functions can be rapidly deployed and launched through flexible plug-in orchestration. Additionally, based on standardized plug-in architecture and interfaces, ZTE promotes collective innovation and ecosystem co-construction of AI and new media processing capabilities, facilitating the development of AI-powered New Calling.

The innovative agile base solution integrates calling and AI to deliver a brand-new calling experience, expand the scope of communication services, and create a new commercial closed loop.

AI-generated content (AIGC) greatly reduces the threshold and cost of content creation, offering rich and personalized displays for individual and enterprise users through the New Calling feature

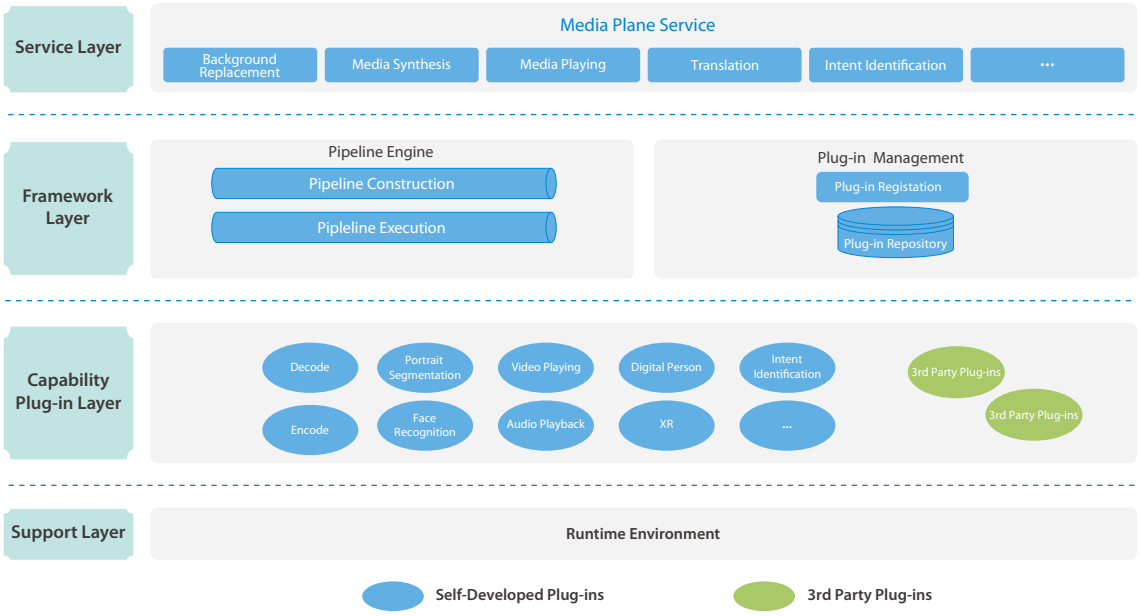


Fig. 1 Media-plane plug-in architecture

"lighting up screen".

The AI agent-based call assistant identifies user intent during a call and generates relevant images and videos to enhance the interaction. For example, when a user talks about a tourist attraction or a product feature, the assistant can display related visuals on the terminal screen in real time. The assistant can also provide intelligent responses to avoid missing calls. For calls from relatives, friends, and colleagues, it can create personalized effects for emotional engagement. For unknown callers, it helps prevent harassment and fraud.

Moreover, the call assistant can offer services such as food ordering, travel booking, and financial assistance via third-party applications, enabling closed-loop value-added services during calls.

For industrial applications, such as service marketing, customer care, and after-sales guarantee, AI+ New Calling offers greater efficiency and lower costs. For example, AI intelligent customer service provides users with faster, more personalized experiences, while remote guidance such as screen sharing and AR marker greatly improve service efficiency and reduce service costs.

Practice and Prospects

Leveraging advancements in 5G-A and AI technologies, China Mobile, together with ZTE and other vendors, has taken the lead in researching, testing, verifying, and promoting New Calling services, effectively enhancing call engagement and service quality, and driving steady user growth and sound business development.

While mainstream Chinese operators are actively conducting pilot projects and commercial deployments of New Calling services, international operators are also showing strong interest and have begun pilot verifications. The global expansion of New Calling services is underway.

Demand for new services and technological advancement are driving the continuous evolution of real-time communication networks. Leading Chinese operators have carried out research on the next-generation real-time communication network, and initiated related requirements projects in ITU-T. ZTE is actively participating in discussions on requirements and architecture for operators' next-generation real-time communication networks, contributing to the preparation and release of white papers.

The next-generation real-time communication network architecture will feature native AI capabilities, enabling intelligence across content production, network control, and media processing, supporting immersive and virtual-real integrated service experiences, and ultimately helping operators speed up commercial monetization. ZTE TECHNOLOGIES

AI-UPF:
A Tool for Mining Traffic Value

Wang Chaoying

CCN Product Planning Architect, ZTE

Liu Xiliang

Chief CCN Product Planning Engineer, ZTE

Since early 2024, 5G has entered the 5G-Advanced (5G-A) era. As user traffic growth slows and the value derived from basic connectivity reach its limit, networks are shifting from rapid expansion to a phase of high-quality and stable development. Consequently, operators are transitioning from traffic-based to experience-based operation.

Experience-Based Operation Solution

ZTE provides an experience-based operation solution powered by "AI+" connectivity intelligence. It provides customized experience guarantee based on user levels, services, and scenarios, unlocking the commercial value of 5G-A. By introducing AI, the user plane function (UPF) enables deep service identification and accurate experience measurement, providing a basis for differentiated service operation at the upper layer. The network data analytics function (NWDAF) conducts in-depth analysis of user services and service measurement data, and generates user assurance policies based on the real-time load of both wireless and core networks, ensuring end-to-end user experience. Fig. 1 shows the solution.

This article focuses on the introduction of AI-UPF in the 5G-A core network to build key intelligent capabilities, enabling precise application identification and accurate service experience perception, providing a data foundation for differentiated services, and ultimately supporting personalized, end-to-end experience assurance.

AI-UPF Unlocks Traffic Value

The AI-UPF is an NE that incorporates AI technology to optimize and enhance the traditional UPFs. With a built-in AI engine and model invocation support, it delivers two key capabilities: intelligent service identification and intelligent experience measurement, unraveling the data pipeline to fully extract network traffic value and support experience-based operation.

Intelligent Service Identification with Greater Scope and Accuracy

As 5G-A networks and generative AI rapidly advance, applications are constantly evolving and new ones are emerging at a fast pace. Traditional signature-based identification requires rule updates with each application change, making timely identification of new or updated applications difficult. In addition, with growing data security concerns and the widespread use of new encryption technologies, traditional methods are increasingly inadequate for experience-based operation. AI-UPF intelligent service identification effectively solves these challenges.

The AI-UPF automates service analysis through its built-in AI engine, shortening the signature library update cycle from monthly to daily. In traditional service identification, once key features such as domain names are extracted from unknown traffic, manual analysis is required to determine the correlation between the features and their corresponding applications, slowing the update cycle. The AI-UPF invokes a large model via its built-in AI engine and uses the model's semantic understanding capabilities together with local knowledge graph entries to analyze correlations—such as similarity and keywords—among key features of service flows. It then infers the applications behind unknown traffic

and cluster the traffic accordingly. Based on the common characteristics within each cluster, the model obtains the application name, type, and description, enabling intelligent tagging and near real-time updates of the signature library.

The AI-UPF uses AI to subdivide services and accurately identify encrypted traffic or private protocols. When sub-services within an application are encrypted, plaintext features are unavailable. Different types of service flows appear as packet sequences with different spatiotemporal features (e.g., packet length, number of packets, and timestamps), making simple rule-based identification methods ineffective. By building an AI model that learns the spatiotemporal characteristics of encrypted traffic, the AI-UPF can identify sub-services (e.g., video, audio, and live broadcast) within applications. To ensure model generalization and identification accuracy, model training and fine-tuning are critical. Training samples are generated through an automatic sampling system that monitors version changes of mainstream applications on the network, triggers service dialing tests and labeling, and accumulates tens of millions of samples. With continuous training and optimization, the model achieves an accuracy of 95% or above in identifying encrypted services, supporting personalized and specific services assurance.

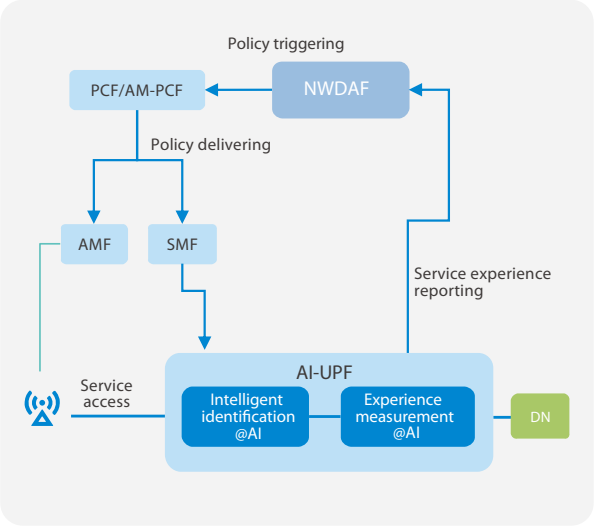


Fig. 1 Experience-based operation solution architecture.

Intelligent Experience Measurement, Closer to Real User Experience

Traditional service quality evaluation relies on network transport-layer KPIs (such as packet loss, jitter, and delay). With the evolution of application transmission technologies—such as encryption and dynamic bitrate—these KPIs no longer effectively reflect the service quality at the application layer.

To accurately predict user-perceived quality, the AI-UPF builds multiple measurement models, such as jamming detection, delay detection, and bitrate detection, to extract data that closely reflect real user experience from weak indicators such as packet length, time intervals, and uplink/downlink characteristics in encrypted applications. For model training, more than 20 degraded network quality scenarios are simulated in the lab using network impairment instruments. Real terminals are used to automatically test more than 100 mainstream applications, generating massive sample data for targeted training of various measurement models, thereby ensuring measurement accuracy. Once deployed in live networks, these models achieve over 90% accuracy in user experience measurement. When degradation occurs, the issue is promptly reported to the NWDAF, triggering end-to-end service experience guarantee mechanisms.

To meet the high performance and low latency requirements of intelligent service processing, the AI-UPF introduces GPU hardware to speed up model inference. It achieves millisecond-level inference latency through data parallelism and multi-GPU parallel processing, increasing performance by tens of times compared to general-purpose CPUs. It also provides service-based model inference with unified inference interfaces, simplifying the inference process and enabling the deployment of more intelligent applications in the future.

Built on the AI-UPF and network intelligence plane, ZTE's 5G-A core network experience-based operation solution enables real-time collection and analysis of user service quality and refined service assurance, boosting operator revenue during the stable development phase of 5G-A networks. ZTE TECHNOLOGIES

AI+ Intelligent O&M: Reshaping the Paradigm

He Wei

Chief Planning Engineer of Core Network Intelligent O&M Products, ZTE

With the accelerating digital transformation in communications and various industries, networks are becoming more complex, service scenarios more diverse, and O&M more challenging. Meanwhile. reducing network O&M costs has become essential, and traditional methods are no longer adequate. Automated and intelligent O&M is widely recognized as the way forward.

The emergence of intent-driven networks, large language models (LLMs), and digital twins opens up new opportunities for intelligent O&M. These technologies offer many advantages, including intuitive user interfaces, and capabilities for large-scale structured data processing, accurate analysis and prediction, and high-fidelity simulation testing. The integration of AI models, intent-driven networks, and digital twins can revolutionize network O&M models and drive operators from the L3 O&M phase to the higher-level L4+ autonomous network phase.

"Four-Layer, One-Entity" Architecture for High-Level Autonomous Core Network O&M

To cope with core network O&M challenges and meet the industry's growing demand for automated and intelligent O&M, ZTE has built a new intelligent O&M architecture— "four-layer, one-entity"—based on large and small model agents and digital twin technologies (Fig. 1). This architecture aims for full openness and decoupling across networks, data, models, and applications, driving the evolution towards higher-level autonomous networks.

Through the coordinated decoupling and smooth evolution of large, small, and heterogeneous models, this architecture builds a flexible, scalable digital technology foundation to enable efficient O&M. It also uses digital twins that deeply integrate data, models, and applications to provide powerful support for network planning, construction, operation, and optimization, ensuring highly stable networks.

Furthermore, by integrating the O&M interfaces for four centers—monitoring, troubleshooting, emergency response, and network change—this architecture enables high-value closed-loop O&M scenarios such as network change management, monitoring and troubleshooting, and complaint handling, driving a shift in network O&M from traditional passive emergency response to a proactive and efficient model with reduced OPEX.

Bidirectional Integration of AI Agents and Digital Twins Empowering Closed-Loop, Efficient O&M and Highly Stable Networks

Relying on LLMs tailored for the communications field and agent services, ZTE's intelligent core network O&M system, based on the "four-layer, one-entity" architecture, has advanced and developed agentic retrieval-augmented generation (Agentic RAG) technology, large- and small-model collaboration technology (which combines generalized intent understanding with precise O&M), and reinforcement learning-based intelligent document generation technology. These technologies enable high-value O&M scenarios featuring intelligent interaction, analysis, and generation, facilitating the network evolution to L4+ high-level autonomy.

In addition, leveraging digital twin technology, the "four-layer, one-entity" intelligent core network O&M system integrates LLMs, AI agents, and multimodal LLMs (MLLMs) to construct digital twins and deliver digital twin models of core network systems, devices, and components. These models

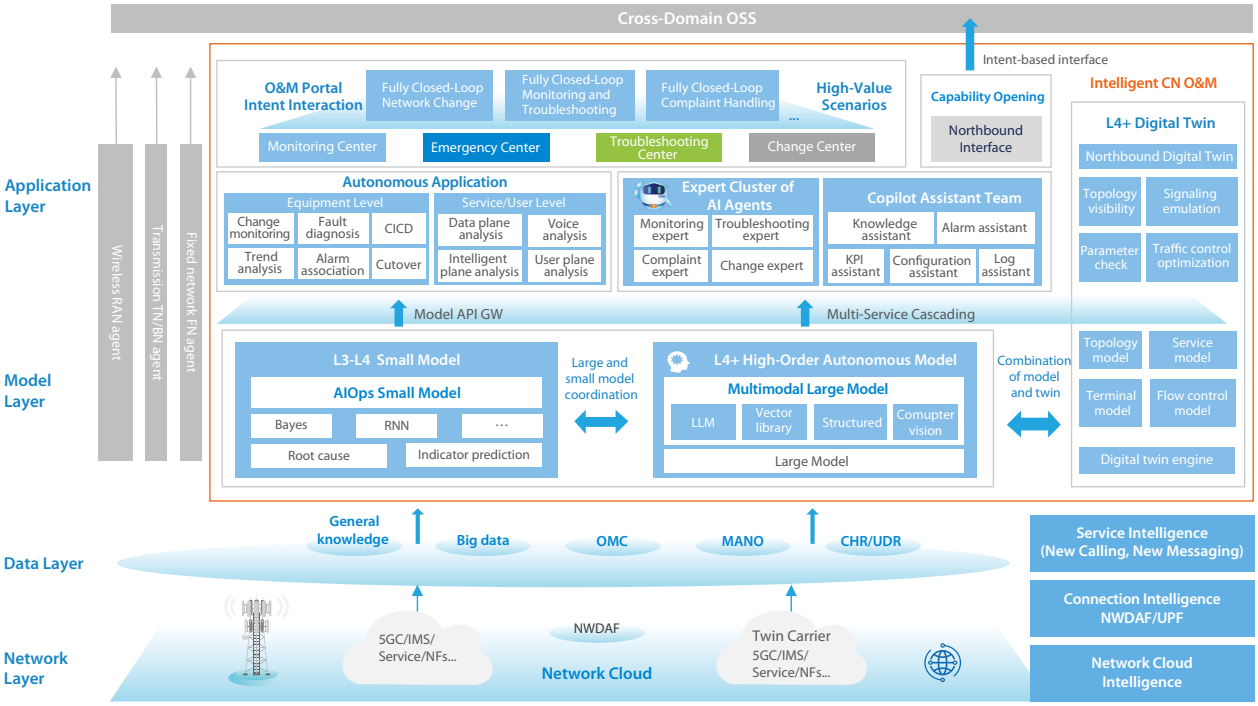


Fig. 1 ZTE intelligent O&M architecture for core network.

provide multidimensional support for O&M, including visualization, simulation, prediction and analysis, and policy feedback, and enable both qualitative and quantitative analysis at low costs, facilitating the shift from manual to machine-assisted and even machine-independent decision-making. This accelerates the evolution towards high-level autonomous O&M and the realization of fully closed-loop automation, effectively reducing O&M costs.

Fully closed-loop O&M management is a comprehensive approach that spans problem discovery, analysis, solution, and optimization. It not only handles faults immediately but also identifies potential risks in advance through data analysis and prediction, enabling proactive and intelligent O&M. In the core network O&M scenario, fully closed-loop O&M management supports three typical application scenarios during actual production processes: network changes, monitoring & troubleshooting, and complaint handling. Currently, manual steps are involved, such as manual reversal and review as well as manual operations such as work order dispatch and receipt. The core value of the fully closed-loop O&M management mode lies in building a self-learning, self-optimizing O&M ecosystem by integrating monitoring, automation tools, AI algorithms, and other technical means.

Continuous Evolution to Unmanned Autonomous O&M

In the TM Forum, the autonomous network level is clearly defined to guide the automation and intelligence of networks and services, evaluate the value and benefits of autonomous network services, and guide the intelligent upgrade of operators and vendors. To achieve full-process intelligence with the goal of L5, all scenarios must be independently completed by the system, ultimately realizing the ideal "unmanned" intelligent O&M model. Achieving "unmanned" intelligent O&M requires careful considerations of architecture, technologies, and capabilities. The integrated foundation, driven by AI and digital twins, serves as the intelligent engine behind "unmanned" O&M. Adaptive agent collaboration technology based on MLLMs plays a key role in improving the capabilities of "unmanned" O&M models. Multi-dimensional agent collaboration and orchestration technology provides an innovative solution for "unmanned" O&M applications. Only by combining these technologies can we achieve more efficient, secure, and compliant O&M management, building a new O&M model. ZTE TECHNOLOGIES

Unified Data Plane: A Cornerstone of AI+ Core Network

Zheng Guobin

Chief CCN Product Planning Engineer, ZTE

By introducing AI capabilities, the AI+ core network enables operators to transform from traditional traffic-centric operation to a differentiated, experience-centric model, while supporting the development of a more efficient and secure 5G/5G-A network. Unlike traditional 5G core network, it must support large AI model training, analysis, and inference, which require massive amounts of data—hundreds of times more than before. These data are scattered and isolated, covering user-level information (subscriptions, mobility tracks, service experience histories), network-level O&M information (NE topologies, performance statistics, alarms, and logs), and wireless-side data (cell loads and resources). Efficiently collecting, processing, storing, and managing these data silos has become a key challenge. This article introduces the unified data plane as a solution to this challenge.

The unified data plane refers to the construction of a network-intelligence integrated, massive, multimodal data storage and management system in the AI+ core network. It provides unified services for data collection, preprocessing, storage, and analysis, as well as training, inference, management, and data security for AI large models, facilitating the sharing of both data and models. Evolving from the integration of the unified data repository (UDR)/unstructured data storage function (UDSF) in 5GC and the analytics data repository function (ADRF) from the intelligent network storage system, it offers a new solution for data processing in the AI+ core network.

Unified Data Plane Architecture

The unified data plane architecture consists of six layers: data collection, data processing, data storage layer, data analysis, data management, and data bus (see Fig. 1).

- **Data collection layer:** Collects data from 5GC, RAN, and other network entities in real time or periodically via the data collaboration interface-southbound (DCI-S), covering user-level, network-level, and cell-level data.
- **Data processing layer:** Performs data cleaning, aggregation, conversion and normalization to improve data quality and usability, supporting subsequent data analysis, AI training, and inference.
- **Data storage layer:** Stores preprocessed data in a large-capacity, scalable, and reliable multimodal storage system that supports efficient and highly concurrent access to differentiated data.
- **Data analysis layer:** Provides data training, inference, retrieval, and analytical functions, serving as a core component of the unified data plane. It uses the stored data to train AI models, for example, training an LSTM time-series prediction model using user mobility, NE load, and cell load data. The layer handles inference requests, applies the trained AI model for real-time inference, and predicts user behavior characteristics. It also supports vector search based on semantic similarity.
- **Data management layer:** Handles data orchestration, metadata management, data backup and recovery, and model management. It orchestrates data access procedures according to requests, establishes indexes, optimizes query parameters, performs regular backups to ensure data recoverability, and manages AI model updating, loading, and distribution.
- **Data bus layer:** Provides northbound data access and model invocation capabilities based on the DCI-northbound (DCI-N) interface for data-plane

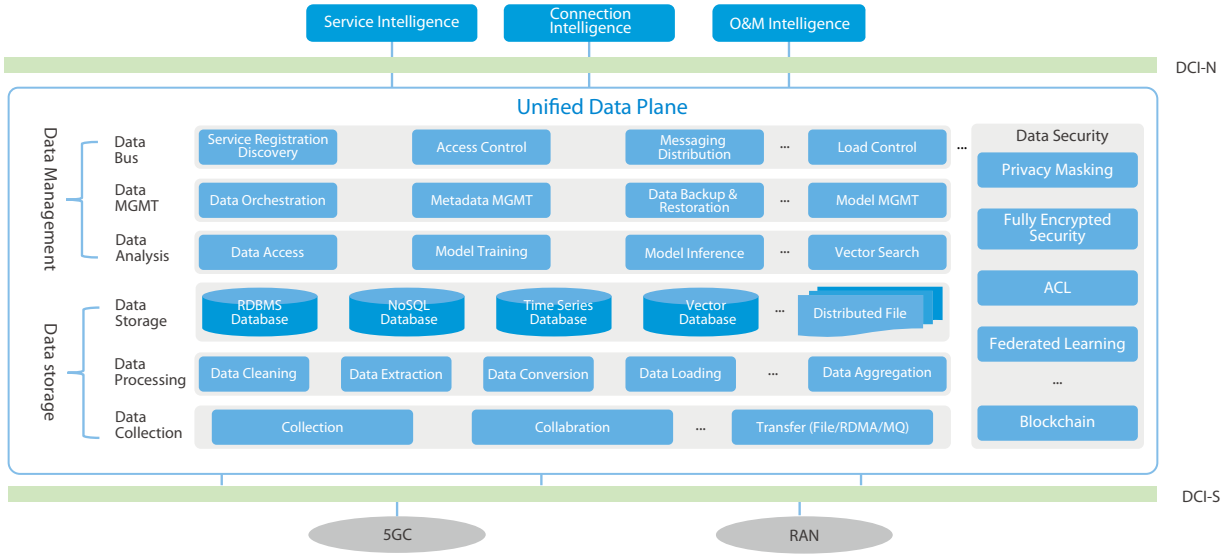


Fig. 1 Unified data plane architecture.

service registration and discovery. It also supports data and model access control and system load balancing and shunting management to ensure that only authorized requesters can access controlled data, improving network operation efficiency.

Key Technologies of Unified Data Plane

The unified data plane provides full-lifecycle management services for collecting, preprocessing, storing, analyzing, and opening massive data, as well as data security and compliance governance capabilities. The key enabling technologies include multimodal database engine technology, distributed computing and storage technology, and security and privacy protection technology.

- **Multimodal database engine technology**

The data stored and managed on the data plane varies in scale, read/write frequency, access performance, and persistence. Diverse database engines and file storage methods need to be adopted for multimodal storage. These include real-time transactional database engines such as RDBMS and NoSQL database engines; real-time analytical database engines such as time-series/columnar database engines; vector database engines; and distributed file or object storage—all aimed at maximizing both storage performance and capacity.

- **Distributed computing and storage technology**

The AI+ core network imposes high requirements on the data plane's capacity, concurrent performance, and response latency, necessitating the use of distributed data storage and computing technologies. Distributed data storage technologies include distributed file storage and object storage systems such as HDFS, MinIO, and Ceph; distributed NoSQL databases and time-series databases such as MongoDB, Redis Cluster, Clickhouse; and ZTE's cloud unified data repository (CUDR). Distributed data computing technologies include distributed computing frameworks like MapReduce and Apache Spark, along with distributed message-queuing and stream-processing platforms such as Apache Kafka, RabbitMQ, and FLink.

- **Security and privacy protection technology**

To ensure data security and prevent leakage, it is necessary to encrypt, store, and mask sensitive data; support access control lists (ACLs) to prevent unauthorized data access and model invocation; and support both vertical and horizontal federated learning. In addition, distributed trusted security management technologies, such as blockchain, should be gradually introduced to improve the security of data and models.

The unified data plane enhances the efficiency and reliability of data collection, management, and storage, while providing abundant data to improve AI model accuracy and generalization capabilities. At the same time, effective security and privacy protection measures are essential. As 5G/5G-A and AI evolve, the unified data plane must also advance to meet the growing demands of network intelligence. With ongoing innovation, the unified data plane will provide strong support for the AI+ core network. ZTE TECHNOLOGIES

AI Opening a New Horizon of Energy Saving for Core Networks

Jin Youxing

Chief CCN Product Planning Engineer, ZTE

With increasing global focus on environmental protection, the telecom industry faces growing pressure to transform. Industry organizations agree that green, low-carbon development is key to future networks. The core network, as the network's brain, plays a vital role in energy saving and emission reduction. Given the large number and diversity of core network NEs, it consumes substantial resources and is significantly affected by service tidal effects. This presents both opportunities and challenges for energy conservation. The introduction of AI into the core network for energy savings is a key enabler of green transformation.

The AI-driven energy-saving architecture of the core network (Fig. 1) centers around the AI green brain. The AI green brain collects data from infrastructure, cloud-based networks, and O&M systems, analyzes it, and dynamically generates energy-saving policies to support global, collaborative energy savings. Through continuous feedback, it evaluates and adjusts the energy-saving effects, ensuring that network energy savings are achieved while meeting service level agreement (SLA) requirements.

Infrastructure: AI-Driven Optimization and Energy Saving for Resource Pools

For a core network based on NFV architecture, the infrastructure has evolved from dedicated equipment and dedicated platforms to various servers and cloud platforms. Control-plane NEs are centrally deployed on universal servers, while the user plane is deployed at various edge nodes as required. At the infrastructure level, energy-saving efforts focus on managing server and cloud platform energy consumption. Current server hardware adopts energy-saving technologies such as efficient heat dissipation, efficient power supplies, and heterogeneous acceleration. In the future, technologies such as integrated storage and computing will be introduced to further improve the computing energy efficiency ratio.

At the software level, AI technology can improve energy efficiency in multiple dimensions.

- **Intelligent prediction:** By collecting performance indicator data and applying AI models, the intelligent prediction module predicts service trends and identifies busy/idle periods from a global perspective, guiding capacity planning and resource pool scaling.
- **Intelligent optimization:** The intelligent optimization module leverages intelligent analysis, capacity prediction, and user trend prediction to proactively adjust resources. Dynamic resource scheduling consolidates fragmented resources, reducing fragmentation. In addition, based on service tidal pattern predictions, different NEs are deployed in a hybrid manner to enable unified scheduling of multiple services within the resource pool, thus reducing peak loads and improving overall resource utilization.
- **Intelligent consumption reduction:** The intelligent consumption reduction module dynamically manages energy consumption based on AI-driven optimization policies. For example, it automatically shuts down unnecessary hardware devices and idle cores, and reduces CPU clock speeds for platform-level energy conservation.

Green Network: Intrinsic Intelligent Algorithms Drive Resource Optimization and Energy Saving

With intrinsic intelligent algorithms, core network NEs can evaluate device status based on service load, user

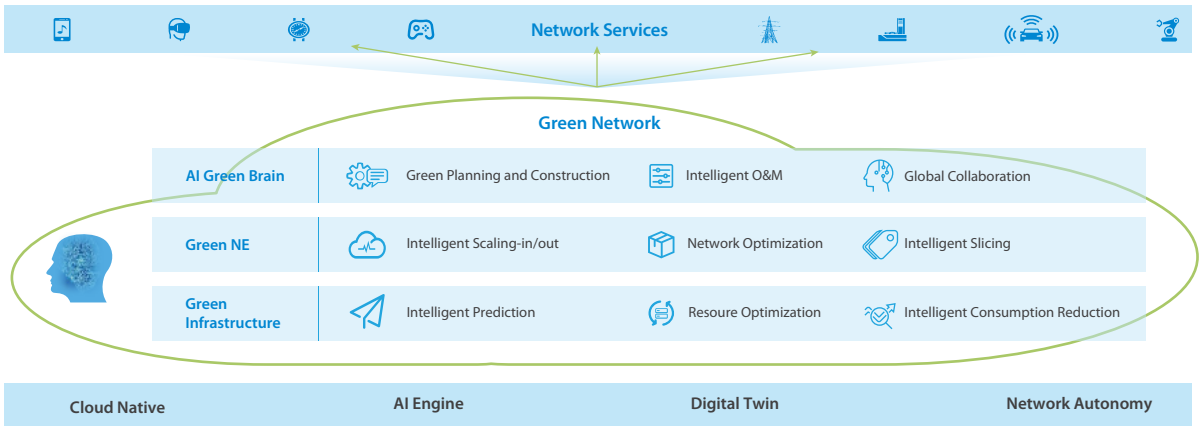


Fig. 1 AI-driven core network energy-saving architecture.

online rate, data throughput, and the status of surrounding NEs. This enables the implementation of policies such as automatic scaling, dynamic service scheduling, and automatic CPU frequency adjustment to optimize resource utilization.

- **Intelligent scaling:** During network operation, the system evaluates and predicts the resources required by the current bearer services of NEs based on historical traffic and capacity expansion requirements, and adjusts capacity accordingly. When traffic increases, computing resources are automatically added; when it decreases, excess resources are reclaimed. In addition, the resources occupied by service components in the core network are dynamically adjusted. When the service is busy, the number of CPU cores occupied by the components is increased; when idle, the CPU core frequency is decreased first, with further adjustments made until the cores are shut down to save energy.
- **Intelligent network optimization:** AI optimizes the core network's topology and routing policie to reduce redundant data transmission and unnecessary signaling interactions. For example, an intelligent routing algorithm is used to select the optimal data transmission path, reducing hops and delays, improving transmission efficiency, and lowering energy consumption.

Green Brain: SLA-Based Intelligent Energy Consumption Evaluation and Optimization

Energy savings must be aligned with service SLA requirements. By monitoring real-time operational status and simulating historical data, the AI green brain builds a resource usage model that accurately reflects actual conditions and establishes a resource consumption simulation system, enabling the prediction of service and energy consumption trends and intelligent evaluation of resource and energy usage. A resource trend model is established to align optimal resource allocation with current service demands, complete the optimized deployment of network services, ensure SLA compliance, and balance resource consumption.

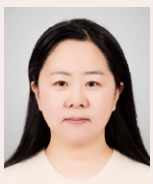
- **NE-level tuning:** NE-level microservice components adopt strategies such as self-sleep, intelligent NE scaling, and dynamic pool migration to reduce system energy consumption. Intelligent switching between energy-saving modes during busy and idle states is enabled through NE traffic-based energy consumption awareness.
- **DC-level tuning:** At the data center level, energy-saving policies include equipment frequency reduction, sleep mode, power-off, and power control based on resource status. Dynamic and transparent service migration across the entire DC network enables intelligent resource defragmentation, preventing the continuous emergence or spread of resource holes.
- **Slice-level tuning:** Energy consumption of core network sub-slices is optimized through resource-saving policies of the core network, and can be coordinated with radio and transmission sub-slice optimization to achieve end-to-end energy efficiency at the slice level.

ZTE has applied green, energy-saving design to core network products across planning, construction, and maintenance and has steadily improved energy efficiency. Recently, ZTE cooperated with a Chinese telecom operator to successfully complete the industry's first commercial pilot of intelligent 5G UPF power saving, achieving a 7%–15% power reduction without affecting service KPIs or user experience. Looking ahead, ZTE will leverage new AI model capabilities to drive further innovation in core networks, explore new energy-saving solutions, and support the dual-carbon goal. ZTE TECHNOLOGIES

Zhejiang Mobile and ZTE Build 5G Messaging in the 5G-A × AI Era

Zhang Shumin

Project Manager of Planning Technology Dept., Zhejiang Mobile



Liu Hong

System Maintenance & Management of Service Platform Dept., Zhejiang Mobile

Wang Liangqin

Senior Engineer of Service Platform Dept. of Network Management Center, Zhejiang Mobile

At the end of October 2024, Apple released iOS18.1, officially announcing support for 5G Messaging. This allows users to exchange high-definition pictures, videos, real-time locations, and other media forms between Apple and Android devices, and also obtain personalized information services through chatbot interactions. Apple's entry signifies the removal of terminal-level obstacles to the development of 5G Messaging service. This milestone is expected to further expand the global influence of 5G Messaging based on the GSMA rich communication service (RCS) standard and unlock huge market growth opportunities. According to Juniper Research, the number of active RCS users is projected to increase from 1.2 billion in 2024 to 2.1 billion globally in 2025. Additionally, global operator revenue from RCS business messaging (RBM) is forecasted to reach US$ 8 billion in 2025.

As early as April 2020, China Mobile, together with China Telecom and China Unicom, released a 5G Messaging White Paper, announcing an upgrade of basic communication services. Traditional SMS messaging could no longer meet users' diversified requirements and must be upgraded to 5G Messaging. 5G Messaging delivers a brand-new messaging experience to both individual and industrial users and presents new development opportunities for global operators, terminal manufacturers, platform vendors, industrial customers, and related sectors. The three major operators in China have called on industry partners to jointly build a new 5G Messaging ecosystem.

As a pioneer in China Mobile's 5G development, the Zhejiang Branch of China Mobile (Zhejiang Mobile) was one of the first batch of 5G Messaging pilot units under the unified guidance of its Group. The 5G Messaging center system for the Southeast China region was exclusively built by ZTE. The project was launched in February 2020. Within a short three-month construction period, Zhejiang Mobile, together with ZTE, has achieved the pilot goals set by the Group, securing four leading achievements in 5G Messaging construction: the first to complete the first call (April 3), the first to launch 5G Messaging application (May 15), the first to carry SMS service (May 19), and the first to carry 5G Messaging service (June 17). In addition, Zhejiang Mobile released applications covering six major sectors, including government, banking, power, personal entertainment, 5G integrated media, and education, laying a solid foundation for the construction and commercial use of 5G Messaging in China.

Since the commercial launch of 5G, China Mobile has not only built the world's largest and most widely covered 5G network, but also established the world's largest commercial 5G Messaging network, setting a global benchmark for RCS development. To

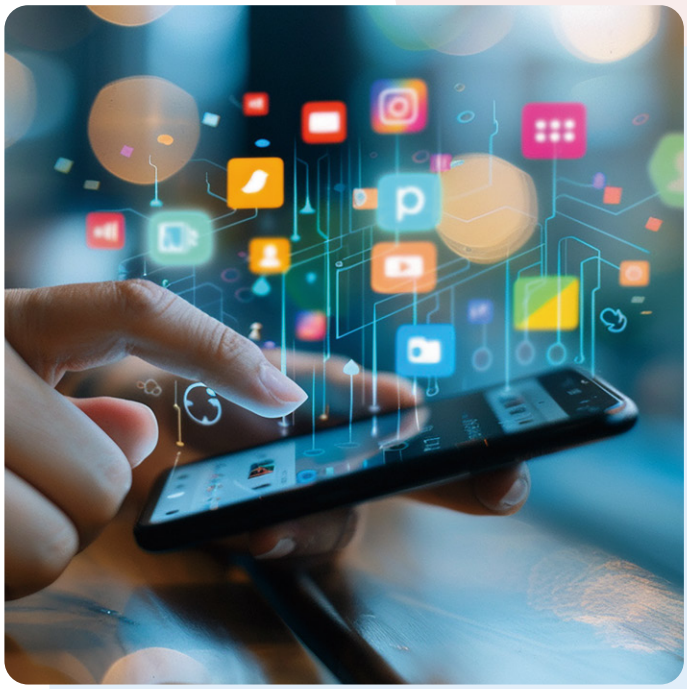
date, China Mobile has deployed a network-wide platform spanning 16 regions, fully promoting 5G Messaging, and launching more than 1.3 million chatbot applications. The number of 5G Messaging users in the Southeast China region exceeds 20 million, and the MaaP platform delivers more than 30 million messages daily on average.

As 5G enters its second phase, 5G-A has been put into commercial use. China Mobile has introduced 5G-A x AI to usher in the "A era", characterized by A-level speed (faster speeds), A-level experience (better experience), and A-level equity (better service). Zhejiang Mobile has collaborated with industry partners to accelerate 5G-A innovation. In China, Zhejiang Mobile has built the world's first 5G-A 10 Gigabit site, the world's first premium 10 Gigabit 5G-A experience route, and the first full-scenario 5G-A application demonstration area for the Asian Games.

In collaboration with ZTE, Zhejiang Mobile also launched the world's first pilot of large-model-based 5G Messaging application—"Intelligent and Trusted AI"—built on the Qingzhou low-code platform, marking the official debut of intelligent chat services in the "A" era. By integrating AI, 5G Messaging is evolving into a unified AI service entry point for mobile networks, covering three key domains: AI+ messaging service, AI+ information service, and AI+ application service. Unlike conventional large models that require additional app downloads, web pages or mini programs, "Intelligent and Trusted AI" is simple and native. Users only need to send a 5G message or SMS to the chatbot application to enjoy a intelligent messaging experience powered by AI.

Zhejiang Mobile collaborated with ZTE's 5G Messaging team to propose an end-to-end AI implementation solution, covering feasibility study, model selection, service processes, and NE interfaces. During model selection, the project team thoroughly compared each model's comprehensiveness, accuracy, compliance, and abnormality, while analyzing the existing network testing latency and user experience. ZTE's Nebula large model solution was ultimately selected for pilot implementation.

With full commitment, the project team completed the first call of the large-model-based 5G Messaging application within just two weeks. Upon its initial launch, the application supported native AI-driven intelligent conversation capabilities across 5G Messaging, SMS, and web modes, and was rapidly iterated to add multimodal dialog capabilities, including text-to-image, image-to-text, voice-to-image, and adaptive generation, covering semantic understanding, speech recognition, visual processing, and bidirectional generation for multimodal interaction. After testing and verification, it outperformed Baidu's Ernie Bot and iFLYTEK's SPARK, earning recognition from the Group and being recommended for wider adoption across other branches.



As 5G Messaging continues to mature and its application scenarios expand, it is poised to play a key role in driving the digital and intelligent transformation of industries and bridging the "last mile" in information services. Zhejiang Mobile, together with ZTE and other partners, is committed to building a leading digital service system centered on new infrastructure, new network connections, new capabilities, and new services. ZTE TECHNOLOGIES

Henan Mobile: 5G-A × AI Innovation Reshaping New Business Models

Liu Rui

CCN Product Technology Manager, ZTE

Zhang Yinran

Director of CCN Product Marketing for China Mobile, ZTE

As the AI wave surged forward and the commercialization of 5G-A began in 2024, the deep integration of these two technologies ushered in a new chapter in the technological transformation of mobile communications. China Mobile Henan branch (Henan Mobile) and ZTE collaborated to integrate AI, big data, and network data analytics function (NWDAF) technologies, building an intelligent assurance system for 5G-A and launching 5G-A packages. They innovatively introduced a "Try & Buy" marketing model, allowing users to try personalized services before paying. By leveraging a "network-to-user" strategy, they precisely identify high-value users and enhance user benefits and perception through personalized packages and dynamic service assurance. The two parties are jointly exploring a new system for the intelligent business operations of 5G-A networks, providing solid support for Henan Mobile to expand value-driven operations and strengthen its brand reputation.

Precision Targeting of Potential Users & Intelligent Package Recommendations

Building on the 3GPP standard architecture, Henan Mobile and ZTE have integrated the NWDAF with the integrated operations platform (IOP), combining the network assurance system with the business operations framework. This enables real-time reporting of user experience data, which correlates user, service, and location factors. The business operations system utilizes this data for in-depth analysis, accurately identifying potential user groups.

By leveraging 5G-A networks, AI, and NWDAF capabilities, Henan Mobile utilizes both subscription-based experience assurance data and sampled experience data reported by NWDAF, integrating them with B-domain user package and tariff information to predict user needs and preferences. This enables precise identification of user segments such as gaming enthusiasts, video streamers, and avid drama watchers on platforms like iQIYI, Tencent Video, and Youku. Tailored packages are then intelligently recommended to these users. This approach replaces traditional manual marketing, which relied on customer group reports and manually configured products and channels, with an AI-driven recommendation matrix that addresses "who to sell to," "what to sell," "how to sell," and "when to sell," transforming conventional business growth paradigms.

In addition, Henan Mobile has innovatively launched the "Try & Buy" marketing model. Under this model, Henan Mobile introduced the 5G-A package, offering users personalized services through a try-before-you-buy approach. For identified high-value users, Henan Mobile sends targeted invitations for "zero" yuan add-on packages, such as low-latency, high-bandwidth mobile game acceleration add-ons for gaming enthusiasts and dedicated live streaming add-ons for influencers. Currently, this system covers 29 popular apps, including TikTok, Huya, iQiyi, WeChat, Honor of Kings, and Tencent Meeting, spanning the live streaming, video, cloud gaming, and online office sectors. This innovative achievement also provides valuable experience for intelligent business development in other provinces.

Dynamic Experience Assurance & Enhanced User Perception

In high-traffic areas such as bustling business districts, university towns, and tourist attractions, Henan Mobile and ZTE have successfully established commercial pilot projects for experience-driven operations. By deploying advanced wireless AI and NWDAF-powered intelligent core networks, new network assurance features have been introduced to

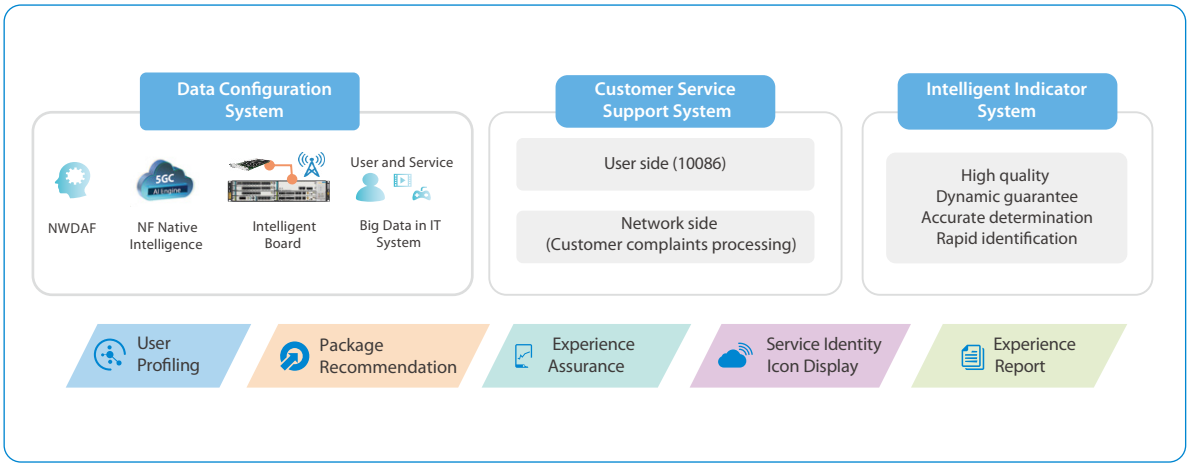


Fig. 1 The 5G-A intelligent assurance system for exploring new package operations.

dynamically guarantee service quality based on real-time user experience (see Fig. 1). These assurances are activated only when service quality deteriorates, effectively enhancing the experience for affected users while avoiding unnecessary resource allocation for those already enjoying optimal network performance.

In Xinyang, Henan, the AI-powered 5G-A network can effortlessly handle peak tourist traffic in popular scenic spots. 5G-A subscribers will receive a "China Mobile VIP Service" icon. During network fluctuations, the system intelligently allocates network resources to ensure high-definition and smooth livestreaming for subscribed users. Compared to ordinary users, video stuttering rates decreased by 15.6% on Huya, 7.6% on TikTok, cloud gaming latency dropped by 89.6% on Migu Play, and large file upload speeds in WeChat improved by 42.3%.

To further enhance user experience and perception, Henan Mobile has introduced dynamic service identity icon displays based on event, time, location, and user group, for example, China Mobile VIP Service, Marathon Event Support, Livestream Protection Active, and Happy Birthday. After the completion of service assurance, users receive customizable experience reports that currently cover four dimensions: number of accelerations, total acceleration duration, improvement percentage, and comparative improvement over ordinary users.

Monetizing Differentiated Experience & Expanding Business Models

As a pioneer of innovation within China Mobile, Henan Mobile continues to lead in network intelligence, operational transformation, and business model innovation. In collaboration with ZTE, it has successfully established an intelligent 5G-A assurance system, achieving early success in expanding personalized services, enhancing high-end user engagement, monetizing value-driven operations, and transforming business models.

From a service perspective, this initiative has significantly enhanced user loyalty and satisfaction. The user-specific icon display feature increases user stickiness and Henan Mobile's market competitiveness through highly personalized services.

From a financial perspective, Henan Mobile has implemented tiered and differentiated experience assurance to meet users' demand for premium network services. Since the trial launch of 5G-A packages, high-value users have shown notable improvements in satisfaction, with DOU up 10% and ARPU rising by 5%. This transition from traffic-based to experience-based operations marks a breakthrough in business model innovation, delivering substantial economic benefits.

From a societal perspective, intelligent differentiated services ensure more efficient network resource allocation, reducing network congestion by 10% and significantly enhancing the public's network experience. By integrating AI into the 5G-A network, Henan Mobile has optimized both user experience and energy efficiency, and reduced network energy consumption by approximately 8% during off-peak periods, supporting sustainable network development.

Looking ahead, Henan Mobile and ZTE will continue to strengthen their collaboration, driving further innovation to deliver even greater value. ZTE TECHNOLOGIES

ZTE

To lead in connectivity and intelligent computing, enabling communication and trust everywhere