5G-AxAI New Technology, New Case, New Model White Paper

https://gtigroup.org/

GTI

5G-AxAI New Technology, New Case, New Model White Paper

Version:	V1.0
Deliverable Type	□ **Procedural Document** ☑ **Working Document**
Confidential Level	☑ **Open to GTI Operator Members** ☑ **Open to GTI Partners** □ **Open to Public**
Program	5G-AxAI
Working Group	N/A
Project	Project 1: Network Intelligence Project 2: Digital Twin Network Intelligence Project 3: Application Intelligence Project 4: Sustainability Intelligence
Task	N/A
Source members	China Mobile, DOCOMO Beijing Labs, Huawei, Ericsson, ZTE, Nokia, Intel, Qualcomm, MTK, Unisoc, Xiaomi, OPPO, vivo, HONOR
Support members	
Editor	
Last Edit Date	
Approval Date	

Confidentiality: This document may contain information that is confidential and access to this document is restricted to the persons listed in the Confidential Level. This document may not be used, disclosed or reproduced, in whole or in part, without the prior written authorization of GTI, and those so authorized may only use this document for the purpose consistent with the authorization. GTI disclaims any liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice.

Document History

Date	Meeting #	Version #	Revision Contents

Table of Contents

1 Executive Summary

In the current era of technology innovation, the Artificial Intelligence (AI) technology has been advancing rapidly, bringing new opportunities for the network. The convergence of 5G-Advanced (5G-A) and AI would be an inevitable industrial trend, unleashing the multiplier effect in telecommunications and other industries. On the one hand, 5G-AxAI would meet network demand, improving network performance and efficiency. On the other hand, 5G-AxAI would also provide new services for the industries, accelerating the industry intelligent revolution. Innovative technology of 5G-AxAI has emerged and made effect in 4 major areas: in network intelligence field, it makes the network display exceptionally high quality; in digital twin network intelligence field, it enables the low-cost trial and high-efficiency innovation; in application intelligence filed, it expands the network service scope; in sustainability intelligence, it achieves the goal of low-carbon. All of these technology provides an innovation engine for the new applications, which involve service guarantee, personal AI agent, embodied AI, autonomous driving, etc. And it also lead to new business model, indicating the network service transformation from the connectivity to the connectivity, computing and intelligence fusion.

2 5G-AxAI: New Capabilities meet New Need, unleash New Value

The cutting-edge AI technology brings new capabilities for the society, enabling a novel way of human production and life[1]. The AI development has entered into a new phase and become light-weighted, generalized, and concrete. The optimized algorithms greatly decrease the AI cost, and reduce the barriers for the small business; the multi-modal foundation models unify the heterogeneous human data, and expand the application scope; the robots are integrated with understanding and thinking abilities, and make the AI objectified in real world.

At the same stage, the communication network is going through the period of large-scale deployment and innovation. The number of 5G networks has reached 398 while 5G SA operators reach 154 globally. And the next-phase 5G-A system begins to take shape and is implemented by more than 15 operators.

A new tale happens when two advanced and fast-developing industries meet together. Along with the 5G-A network, the effect and value of AI are further enlarged. The convergence of 5G-A and AI becomes an inevitable trend and brings changes in many aspects of technology, applications, industry and so on. New need of network and industry will be satisfied based on the 5G-AxAI technology development, promoting social efficiency and creating new value.

2.1 5G-A Commercial progress

In the mid of 2024, the 3GPP has already finished the work on Release 18, and its work on Release 19 is expected to be completed by the end of 2025. 5G has officially entered the 5G-A stage, and it provides multiple values for society[2-4]. First, high-speed, low-latency, and almost ubiquitous connectivity capabilities. Second, integration of various innovative information technologies. Third, the ability base for the industry services. In a word, 5G-A has achieved significant upgrades in terms of rate and intelligence capabilities. According to data, the actual 5G-A rate has jumped to 3-5Gbps and is expected to exceed 10Gbps in the future, a 10-fold increase compared with traditional 5G network. Technologies such as Extremely Large Antenna Array (ELAA) are used to improve coverage of high frequency bands, and the multiple carriers coordination technology is used to efficiently aggregate bandwidths of multiple frequency bands, thereby greatly improving spectral efficiency. On the terminal side, novel high-end chips are released in 2024 to support large bandwidth capabilities beyond 6 Component Carriers (CC). Moreover, 5G-A continuous capability upgrading has also opened up a new blue ocean in the industry. For example, The 5G-A Internet of Vehicles (IoV) supports large-scale pilot projects on vehicle road clouds and accelerates the development of the intelligent connected automobile industry.

From a global perspective, more than 60 operators and partners around the world have announced 5G-A commercial plans. 5G-A has become a hot topic in the industry. On December 10, 2024, a special seminar on the theme of "World First 5G-A Region Sets Sail" was launched. Experts from authoritative industry organizations, regulators, leading operators and equipment vendors, such as GSMA, TDRA, DU, e&, Vodafone, Ooredoo, Huawei, Ericsson, and Nokia Delegates shared the innovations of 5G-A. In China, 5G-A networks have been launched in more than 300 cities, and China has set an good commercial example for the world. In other regions, lots of operators have carried out 5G-A verification, such as Asia Pacific HKT, CTM, Malaysia Maxis, Europe DNA, Vodafone, Latin America VIVO, and TIM.

In China, China Mobile actively promotes the implementation of 5G-A network, who takes the lead in adopting innovative technologies. For instance, it deploys 3CC aggregation networks and 5G-A intelligent control plane (NWDAF), improving the transmission efficiency and providing unique assurance capabilities for key customers. China Mobile has also issued 5G-A charge plans for travel, game, and live broadcast, and are subscribed by more than 500,000 5G-A users. For example, the 5G-A business travel plan provides a high-speed experience of up to 3Gbps in the downlink and 200Mbps in the uplink.

In the Middle East, operators have issued wireless broadband packages up to 300Mbps. Wireless home broadband uses the 5G-A high-speed network to provide high-quality experience similar to optical fibers, and solves the problems of difficult fiber deployment and high construction costs.

Overall, 5G-A has great potential for development and monetization. It is in the stage of continuous developing and large-scale implementation, awaiting for the new driving forces to achieve further breakthroughs.

2.2 New AI Capabilities

AI is the technology that enables machines to mimic human intelligence, allow them to tackle tasks like learning, comprehension, problem solving, perception, decision making, and autonomy. The latest big foundation models (e.g., GPT-4o, Deekseek) have demonstrated their remarkable ability in understanding natural language and thinking about the problem. AI Algorithms include supervised learning, unsupervised learning and reinforcement learning. And among them, deep learning is a kind of effective and powerful method, which has been widely used in various fields. Neural networks with multiple different kinds of layers (e.g., MLP, CNN, RNN, Transformers) have demonstrated their great capabilities in perception, prediction, decision-making and optimization. These capabilities have already shown its value in healthcare, finance, transportation, governance and many other industries.

1) AI Perception

AI is able to perceive and understand the surrounding environment through processing the data collected from various sensors. It enables machines to simulate the way humans perceive the world, including vision, hearing, touch, and other senses. It even goes beyond human because it could also directly obtain information from the cyber-space. For example, in hospitals, AI could monitor patients' vital signs and detect early signs of deterioration, in this way the medical staff would be alerted in time to provide emergency help.

2) AI Prediction

AI is also able to forecast future events, trends, or outcomes based on historical data and current information. It has the ability to handle complex data and identify patterns that may not be apparent to humans. In weather forecasting, AI models can predict future weather conditions accurately by analyzing a large amount of meteorological data, including temperature, humidity, wind speed, and pressure. In financial field, AI could predict stock prices and market trends based on historical stock data, economic indicators, and news sentiment.

3) AI Decision-making

AI could analyze data, evaluate various options, and make recommendations or decisions to achieve specific goals. It has the ability to normalize multi-modal input conditions to the unified vector space, extract the common features, and weight them to get the final decision. In autonomous driving, AI system could help analyzes data from different kinds of sensors such as camera, lidar, and radar to detect obstacles, other vehicles, and pedestrians. It then makes decisions about when to accelerate, brake, or turn to avoid collisions and follow traffic rules.

4) AI Generation

AI is capable of creating new content by learning and analyzing vast amounts of related data. This ability could be adapted and expanded to different areas (text, audio, image, video). In creative fields, it provides creative inspiration for artists, designers, writers, musicians, helping them quickly generate creative content. In education field, it generates teaching resources such as teaching animations and personalized learning materials, helping improve the teaching effect.

All of these AI capabilities could be utilized and improved in the network field. Combined with the network capability of low latency, high-reliability, ubiquitous connectivity, etc., communication-intelligence fusion services would emerge and drive the industry to the next phase.

2.3 Opportunities for the Cross-domain Convergence

2.3.1 Meeting Evolved Network Demands

The convergence of 5G-A and AI offers transformative potential in meeting the network evolution demand and enabling superior network performance. AI technique could be used to optimize multiple functional modules across an entire link or system in a data-driven way, thereby improving the entire end-to-end communication system. By combining the advanced capabilities of 5G-A and AI, networks can achieve unprecedented efficiency, adaptability, and reliability, meeting the ever-growing demands of modern connectivity. Below is an analysis of how

AI technologies address critical challenges in network fields.

1) Efficient Resource Utilization

Efficient resource utilization is essential in 5G-A to meet the increasing demand for connectivity, particularly in dense urban areas, industrial Internet of Things (IoT) applications, and high-traffic environments. AI plays a pivotal role in optimizing key network resources such as spectrum, bandwidth, and infrastructure.

Dynamic Resource Management: AI algorithms analyze real-time traffic conditions and predict demand patterns, enabling 5G-A networks to dynamically allocate resources such as bandwidth, power, and spectrum to areas of highest need. This ensures efficient utilization of network assets and seamless adaptation to varying traffic loads.

Network Slicing Optimization: Through 5G-A's network slicing capabilities, virtual networks can be tailored for specific applications (e.g., IoT, enhanced mobile broadband, and low-latency services). AI enhances this by optimizing resource allocation for each slice in real time, adapting to network performance metrics and user requirements.

Predictive Traffic Analytics: AI-powered predictive models leverage historical data and usage trends to forecast network congestion or bottlenecks. This proactive approach allows networks to redistribute resources in advance, ensuring uninterrupted operation even during peak usage periods.

AI-Enhanced Spectrum Efficiency: AI algorithms dynamically assign available frequencies based on real-time demand, minimizing interference and maximizing spectral efficiency. AI can also enable effective spectrum sharing between operators and technologies, boosting network capacity.

Optimizing MIMO Technology: 5G-A's Multiple Input Multiple Output (MIMO) technology significantly enhances data throughput. AI fine-tunes MIMO parameters dynamically, optimizing system performance with minimal power consumption while maintaining high data rates.

2) User-centric Network Operation

User-centric operation is critical for ensuring users receive a consistent and reliable level of service, including key metrics such as bandwidth, latency, jitter, and packet loss. For 5G-A, Quality of Service (QoS) optimization is particularly important due to the wide range of services the network supports, which guarantees the user experience and balances the network load.

Real-Time Monitoring and Adjustment: AI enables continuous monitoring of network performance and user experience. By analyzing metrics like latency, throughput, and packet loss in real time, AI can dynamically adjust network parameters to ensure optimal QoS. For example, it can prioritize time-sensitive traffic, such as autonomous vehicles or remote surgeries, to minimize delays and packet loss while allocating sufficient resources for other services like video streaming.

Intelligent Traffic Shaping: AI algorithms can classify traffic based on the type of service, such as video, voice, or IoT. By understanding the specific requirements of each service, AI can prioritize high-priority traffic and route low-priority traffic efficiently, ensuring seamless performance and avoiding congestion.

Context-Aware Traffic Prioritization: AI leverages contextual data, such as time of day, user location, and network conditions, to dynamically adjust QoS settings. For instance, during peak hours, AI can prioritize applications related to emergency services or healthcare over general browsing or streaming, ensuring critical services remain unaffected.

End-to-End QoS Assurance: AI can enable end-to-end QoS control, ensuring that the QoS standards are met not only at the network level but also at the application level. AI can analyze the entire network path—from the edge to the core—to detect any potential issues that could affect service delivery. This holistic approach ensures a better user experience, especially in critical-use cases like smart cities, autonomous driving, and industrial automation.

Intelligent Traffic Offloading: AI predicts traffic patterns and proactively offloads traffic to alternative, efficient paths. This includes rerouting data to edge computing nodes, neighboring cells, or other optimized network paths. Such proactive measures ensure the core network remains uncongested, improving overall network stability and performance.

3) Network Automation

Intelligent algorithms can significantly enhance the performance of network pipes, reduce Operations&Maintenance (O&M) costs, and increase the efficiency of new service rollouts. Furthermore, AI will drive O&M practices towards levels L4/L5, assisting telecom operators in improving service quality and increasing revenue.

Agile Network: In the current O&M system, functions operate independently of each other. To achieve task goals, O&M personnel must manually orchestrate and integrate these functions, which requires frequent interactions with the system and leads to reduced O&M efficiency.

AI technologies are transforming the O&M mode from being function-centric to task-centric, allowing the system to better cater to human needs. By incorporating a unified knowledge model into the O&M system, O&M personnel can communicate with it using natural language. This enables the system to understand user intent and autonomously orchestrate and integrate functions to fulfill task goals seamlessly.

Greener Network: There is often a contradiction between network energy saving and experience assurance. The biggest contradiction in energy saving is that we do not know when to shut down and to what extent we shut down.

With AI algorithms, network could dynamically balance network resource allocation and energy saving policies to achieve smooth transition from off-peak to peak hours. Through the implementation of the foregoing intelligent strategies, efficient collaboration between experience assurance and energy saving could be achieved, significantly reducing network energy consumption while assuring the user experience.

Service Innovation Empowerment: Telecom industries are actively exploring new network capabilities and services. To meet various requirements, it is necessary to break away from the traditional working mode of service provisioning based on manual experience and operation.

In the new service provisioning phase, operators could directly deliver service requirements in the form of natural language. With the help of AI methods, the network is able to understand the intent by translating the intent to the machine code, perform online simulation in the network twin environment, and automatically generate network configurations to meet new requirements.

2.3.2 Meeting Diverse Industry Demands

The industry demand for 5G-AxAI is growing rapidly due to the convergence of two transformative technologies: 5G-A and AI. This combination is expected to enable a wide range of new applications and services across multiple industries.

1) Telemedicine and Remote Healthcare

5G-A offers low latency and high-speed data transfer capabilities, which are essential for enabling real-time video consultations and even remote surgical procedures. AI enhances these capabilities by supporting diagnostic tools that analyze medical images, sensor data, and patient histories, aiding healthcare professionals in making informed treatment decisions.

Telemedicine and Remote Surgery: 5G-A provides the high bandwidth and low latency necessary for real-time remote consultations and surgeries. AI can assist doctors in making diagnoses or performing surgery with robotic assistance, using AI-powered imaging and data analysis to improve accuracy.

Health Monitoring: AI-powered devices, enhanced by 5G-A networks, enable real-time monitoring of patient health data. These systems can predict potential health issues before they become critical by analyzing data from wearable, sensors, and other medical devices.

AI-Assisted Diagnostics: AI-driven models can analyze medical scans, including X-rays and Magnetic Resonance Imaging (MRI), with remarkable accuracy. When paired with 5G-A's capability to transfer large datasets swiftly, these tools enable faster and more precise diagnoses, particularly in remote or undeserved regions, improving access to quality healthcare.

2) Autonomous Vehicles and Transportation

5G-A playing a crucial role in promoting the development of autonomous vehicles, which provides impressive data transmission and supports massive connectivity. Along with the computing system deployed in the network, the vehicle is able to have a more comprehensive and detailed perception of the surrounding environment and make more accurate driving decisions.

Autonomous Vehicles: 5G-A delivers the high-speed, ultra-low-latency connectivity required for autonomous vehicles to communicate seamlessly with each other and with infrastructure, such as traffic lights and road sensors. AI processes data from multiple sensors, including LIDAR, cameras, and radar, enabling real-time decision-making for safe and efficient driving.

Vehicle-to-Everything (V2X): 5G-A supports V2X communication, facilitating the exchange of data between vehicles, infrastructure, and even pedestrians. AI leverages this data for advanced traffic management, predictive route optimization, and real-time traffic updates, enhancing road safety and travel efficiency.

Fleet Management and Logistics: 5G-A enables real-time monitoring and seamless communication across large vehicle fleets. AI algorithms analyze data to predict maintenance needs, optimize delivery routes, and improve fuel efficiency, ensuring cost-effective and sustainable logistics operations.

3) Manufacturing and Industry 4.0

5G-A network connects a large number of devices and sensors in factories and transmits a large amount of data (e.g., production process parameters, equipment operation status) to the central control system in network or application server for analysis and processing, enabling real-time monitoring and optimization of the production process.

Smart Manufacturing: 5G-A delivers the high bandwidth required for real-time monitoring of machines, assembly lines, and robotic systems. AI analyzes sensor data to optimize production processes, identifying inefficiencies or potential issues before they lead to breakdowns.

Predictive Maintenance: AI algorithms leverage sensor data to predict equipment failures before they occur. With 5G-A's continuous data transmission capabilities, systems can send real-time alerts for maintenance, reducing downtime and enhancing operational efficiency.

Robotic Automation: 5G-A's ultra-low latency enables AI-powered robots to perform highly synchronized tasks in real-time, such as assembly and quality control, significantly improving precision and productivity in manufacturing environments.

4) Retail and E-commerce

5G-A and AI has great impact on retail and e-commerce, bringing about changes in various aspects such as shopping experience, logistics, and supply chain management. Real-time data sharing and communication make demand forecasting more accurate and customer experience better, enhancing the overall efficiency, flexibility and attraction.

Personalized Customer Experience: AI analyzes customer behavior to provide personalized recommendations and targeted promotions. 5G-A supports this by enabling real-time data exchange and immersive experiences such as augmented reality (AR) in retail stores.

Inventory Management and Logistics: AI can predict product demand and optimize stock levels in real-time. Combined with 5G-A's ability to track shipments and warehouse operations, retailers can improve efficiency and reduce stockouts.

Smart Stores: 5G-A enables the integration of IoT devices (e.g., smart shelves, sensors) in retail environments. AI analyzes data from these devices to improve product placement, detect theft, and optimize in-store experiences.

5) Smart Cities and Urban Infrastructure

5G-A and AI are the key enablers in the construction of smart cities, facilitating significant advancements in smart governance, smart security, smart traffic, etc. Through the technology development, the city could monitor the real-time status, make comprehensive decisions and manage urban operations.

Traffic Management: 5G-A networks are ideal for smart city infrastructure, enabling real-time data analysis from sensors across the city to optimize traffic flow, reduce accidents, and improve overall urban efficiency. AI is key in processing large-scale data to make intelligent decisions.

Public Safety and Surveillance: AI-powered security systems, including facial recognition and anomaly detection, benefit from the speed and connectivity of 5G-A networks. These systems can process real-time data from surveillance cameras to enhance public safety.

The demand for 5G-A and AI technologies extends across a wide range of industries, each capitalizing on the combination of high-speed connectivity, real-time data processing, and intelligent decision-making. Sectors such as healthcare, automotive, manufacturing, telecommunications, retail, and smart cities are harnessing these technologies to boost efficiency, enhance customer experiences, and unlock new applications. The synergy between 5G-A and AI not only enhances current use cases but also enables the development of innovative, next-generation applications that were once unimaginable, driving digital transformation across these industries.

2.4 Multiplier Effect of 5G-A and AI

The convergence of 5G-A and AI creates a powerful multiplier effect, amplifying the impact of each technology across multiple sectors. This convergence accelerates innovation, boosts efficiency, enhances customer experiences, and fosters the development of new business models[5].

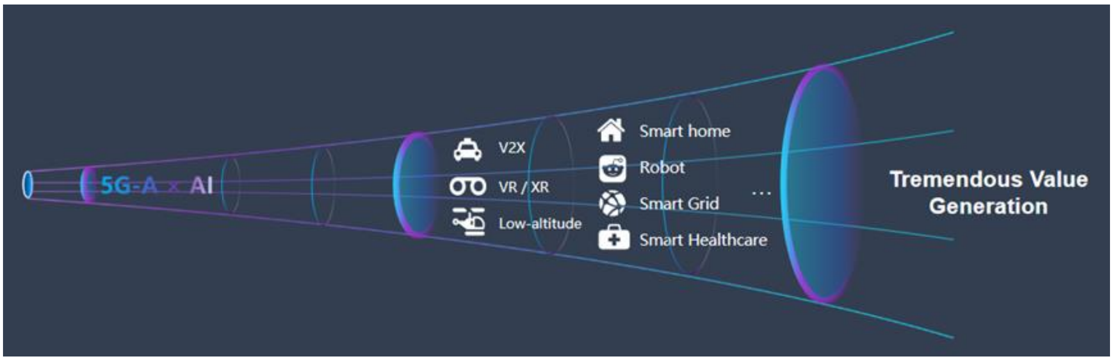


Figure 2-1 The convergence of 5G-A and AI generates tremendous value

The outcome of the 5G-A and AI collision is the fusion effect rather than simple summation. Two parallel universes (5G-A domain and AI domain) collide and expand to higher dimension, and the inner particle generates interaction force with each other, releasing tremendous energy. This energy spreads to different application fields, and unfolds great value for the human society. The value is mainly reflected in three aspects:

1) Boosting Network Performance and Efficiency

Network Optimization: AI enhances 5G-A by facilitating autonomous network management, predictive resource allocation, and fault detection. This leads to more efficient use of network resources (such as bandwidth and power), reducing operational costs for service providers while improving user experience.

Networks Automation: AI enables network automation, allowing 5G-A networks to self-optimize and self-heal. This reduces the need for manual interventions, speeds up network deployment, and enhances scalability to meet the increasing demands of IoT and connected devices.

Low-Latency Applications: AI models help optimize traffic routing and predict network congestion, supporting ultra-low latency services such as autonomous vehicles, augmented reality, and remote surgeries. 5G-A’s ability to provide high data rates and massive connectivity ensures smooth and uninterrupted operation of these AI-powered services.

Predictive Analytics: AI-driven predictive maintenance and traffic management allow 5G-A networks to anticipate demand shifts, detect bottlenecks, and perform proactive maintenance. This minimizes service disruptions, lowers operational costs, and enhances network reliability.

2) Unlocking New Services and Applications

Edge Computing and Real-Time Processing: 5G-A supports edge computing, where data is processed closer to its source, such as IoT devices or sensors. This reduces latency and bandwidth usage, enabling AI applications that demand real-time processing, including autonomous vehicles, smart manufacturing, and smart cities.

Massive Connectivity: 5G-A can accommodate vast IoT ecosystems with millions of devices transmitting small data packets. When combined with AI’s ability to process and make decisions based on this data, industries can optimize operations, improve customer service, and develop new business models. For instance, in smart agriculture, AI can analyze sensor data to enhance irrigation and crop management, while 5G-A ensures seamless real-time data transmission across the farm.

Immersive Communication: In sectors like education, entertainment, and retail, AI-powered AR/VR applications are enhanced by 5G-A ability to deliver high-quality, low-latency immersive experiences. This opens up innovative opportunities for virtual training, remote collaboration, and virtual product try-ons.

3) Accelerating the Industry Revolution

Automation and Optimization: AI drives intelligent automation across industries, including manufacturing (Industry 4.0), healthcare, transportation, and smart cities. In manufacturing, AI-powered robots and predictive maintenance systems, supported by 5G-A's ultra-low latency, boost production efficiency and minimize downtime. In healthcare, AI algorithms analyze patient data in real time to create personalized treatment plans, backed by 5G-A's fast data transfer.

Real-Time Analytics and Personalization: With the help of 5G-A’s connectivity, AI enables industries to personalize services based on real-time data. For instance, in retail, AI analyzes customer behavior to offer tailored recommendations, while 5G-A ensures seamless data transmission from IoT devices, sensors, and mobile apps for an enhanced shopping experience.

Improved Decision-Making: AI’s ability to process complex data sets provides businesses with real-time insights into market trends, customer preferences, and operational inefficiencies. This allows industries to make quick, data-driven decisions, boosting business agility and competitiveness.

Predictive and Prescriptive Analytics: In logistics, AI-powered predictive models optimize routing and inventory management, while 5G-A’s connectivity facilitates real-time monitoring of goods in transit. This results in cost savings, improved customer satisfaction, and greater operational efficiency.

The multiplier effect of 5G-A and AI convergence is profound, creating great value of improved performance, new business opportunities, and innovation. AI drives smarter decision-making and automation, while 5G-A provides the connectivity infrastructure necessary to support these advances with speed and reliability. Together, they propel the digital transformation of industries, improving operational efficiency, fostering new business models, and delivering societal benefits. The outcome is a dynamic, cross-industry ecosystem where growth and innovation are amplified, benefiting businesses, consumers, and society.

3 5G-AxAI Breeds New Technologies

The convergence of 5G-A and AI brings revolutionary innovations. Core innovations focus on four major areas: 1) network intelligence, which aims to enhance network quality through intelligent optimization; 2) digital twin network intelligence, which aims to enable low-cost trials and high-efficiency innovation; 3) application intelligence, which aims to expand network service scope through enhanced capabilities; 4) sustainability intelligence, which aims to achieve important environmental goals through intelligent energy efficiency. Together, these innovations power a new wave of applications spanning diverse areas like service assurance, personal AI agents, embodied AI, and intelligent network infrastructure lifecycle management. Through continuous advancement in these technological domains, 5G-AxAI is actively reshaping network paradigms and accelerating industry transformation while driving the evolution from simple connectivity to an integrated fusion of connectivity, computing and intelligence.

3.1 Network intelligence

3.1.1 Intelligent Real-time Network Closed-loop

By introducing intelligent control plane and user plane, 5G-A could achieve real-time application identification, experience perception, and dynamic service control, thus supporting operators to continuously optimize network resource scheduling, and enabling differentiated service assurance.

1) Real-time experience perception

With coordination of the Network Data Analytics Function (NWDAF) intelligent analytic capability and User Plane Function (UPF) traffic recognition capabilities, the intelligent network could provide precise and real-time user experience evaluation based on massive amounts of network data. The process involves four stages: data processing, AI model structure design, model pre-training, and model fine-tuning.

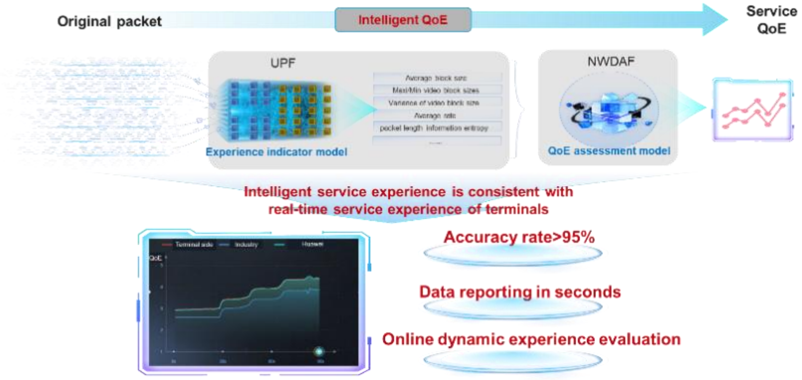

Original packet
Intelligent QoE
Service QoE
UPF
Average block size
Max/Min video block sizes
Variance of video block size
Average rate
packet length information entropy
Experience indicator model
NWDAF
QoE assessment model
Intelligent service experience is consistent with real-time service experience of terminals
Terminal side
Industry
QoE
Accuracy rate>95%
Data reporting in seconds
Online dynamic experience evaluation

Figure 3-1 Framework for real-time experience perception

Data processing aims to provide high-quality, readily usable corpora for model training, laying the foundation for the accuracy and generalization capabilities of subsequent AI models. Several aspects are considered for data processing, including privacy security, feature selection, encoding, tokenization.

The AI model structure design involves defining and designing basic models suitable for network operational state data compression and modeling. It designs domain models based on Transformer and Mamba structures, integrating multi-modal learning to fuse signaling, traffic, and other multi-source data. By combining knowledge distillation and model quantization techniques, complex models are compressed into lightweight versions to meet the real-time data analysis and deployment requirements of UPF and NWDAF.

The model pre-training is to build a basic model with general characteristics through global training on large-scale data, while optimizing the training process to improve training efficiency and effectiveness. Pre-training optimizes multi-task learning performance through adaptive loss weighting algorithms, dynamically adjusting task priorities; it also enhances the efficiency of large-scale model training by combining distributed training frameworks.

The model fine-tuning provides high-precision task capabilities for models through training based on specific service scenarios. In the fine-tuning phase, transfer learning can be used to quickly adapt to different scenarios, and small-sample learning can be used to improve the model performance in scenarios where data is insufficient. Based on the fine-tuning technology, application awareness and Quality of Experience (QoE) awareness tasks can share the same basic model and can be upgraded independently.

2) Differentiated experience assurance

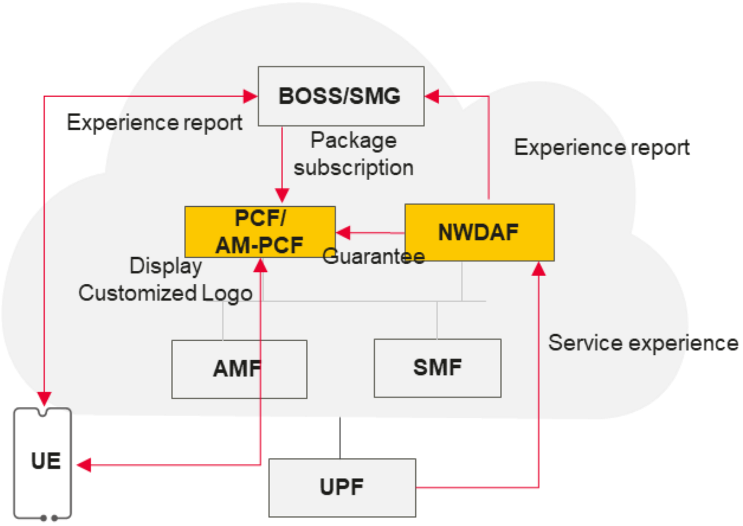


Figure 3-2 Framework for differentiated experience assurance

Differentiated experience assurance technology facilitates operators maximizing resource utilization and enhancing profitability. Specifically, when the network runs at low-load states, the differentiated experience assurance technology could stimulate user traffic, so that users are willing to migrate high-traffic services such as live streaming to the mobile network; when the network runs at high-load states, the differentiated experience technology could provide sufficient assurance for key services of key customers, so that the network states could not degrade their experiences.

In the process of service assurance, the UPF continuously analyzes user service experience and reports the analysis results to the NWDAF. The NWDAF triggers the assurance policy based on poor experience. After the Policy Control Function (PCF) delivers the assurance policy, it triggers dedicated Logo display on the User Equipment (UE). After the assurance is complete, proactively push experience reports so that users can perceive the assurance results and streamline end-to-end experience closure.

3.1.2 Network Foundation Model

Communication Network is the aggregation node of all traffic and could obtain massive valuable data. With the help of Large Language Model (LLM) technology, it becomes feasible to build a foundation model incorporating network knowledge, enabling precise network operations and comprehensive user analysis. The foundation model greatly improves the generality and robustness compared to traditional AI methods, providing stronger capabilities of perception and optimization. Moreover, it makes the task adaptation more easier and avoids redundant development of network functions.

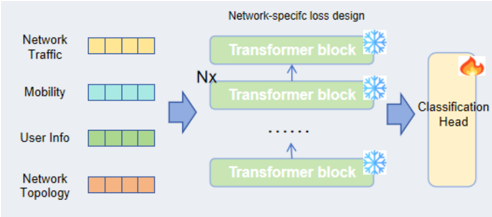


Figure 3-3 Foundation model design for the intelligent network

In general, the LLM mainly focuses on handling with speech, text, and images rather than network data. To reach full potential of LLM in network, the foundation model needs to be re-designed in the following aspects. First, it is important to customize the method of network data tokenization. The network data needs to be tokenized according to its characteristic, which maps the multi-dimensional network data (e.g., traffic, user info, topology) to vector spaces and helps the model extract high-level data features. Second, the model structure and optimization objectives should be re-built. Appropriate setting of layer, parameter, loss function and training policy would guarantee the model convergence when it incorporates so much heterogeneous network data. Third, the cost of model adaptation and deployment should be decreased due to the current status of network, which haven't been equipped with efficient computing power but requires real-time processing. This could be achieved by Parameter-Efficient Fine-Tuning (PEFT) and model compression.

In the area of network traffic recognition, unlike typical methods, the foundation model doesn't depend on plain-text rules and could extract the high-level implicit features to distinguish different kinds of packets and flows. It brings better accuracy especially in the case of encrypted traffic, private protocol and unregistered service type. Moreover, the network inherently supports to track user-related data such as location change, network access and traffic usage, which enables the network to perceive users' preferences, lifestyles and occupations. Given the impressive semantic understanding, contextual information processing, and cross-domain adaption abilities of the foundation model, it could act as an important tool for more accurate mobile user profiling by integrating multi-source heterogeneous user-related data, which further facilitates the development of more individualized marketing strategies.

3.1.3 Intelligent Network Infrastructure Life-cycle

Network infrastructure involves hardware devices, cloud platforms, network functions, and network management systems. To better support increasingly complex upper-layer applications, network infrastructure requires for efficient and high-quality construction and delivery. Extensive and repeated verification is needed during stages such as research and development, testing, deployment, and integration to ensure infrastructure stability. Therefore, introducing AI capabilities into network infrastructure is urgently needed to replace significant labor costs.

To make the network more intelligent, AI agents could be applied and enhanced in the full lifecycle of infrastructure integration and verification, such as testing, integration, delivery, and operation and maintenance of 5G-A infrastructure. In the testing and validation phase, AI agents can assist in requirement management, solution formulation, acceptance validation, and pilot validation. In the construction and delivery phase, AI agents can automatically complete tasks such as environment setup, delivery validation, configuration, deployment, and issue localization. In the operation and maintenance evaluation phase, AI agents can perform quality assessment, problem prediction, operation and maintenance monitoring, and disaster recovery drills. Accordingly, the intelligent infrastructure integration and verification technology will greatly enhance the level of test automation and management efficiency in network infrastructure.

The architecture if intelligent network infrastructure is divided into four layers, similar to the human senses, brain, nerve signals, and hands. Each layer has its corresponding functions and applications in the 5G-A network.

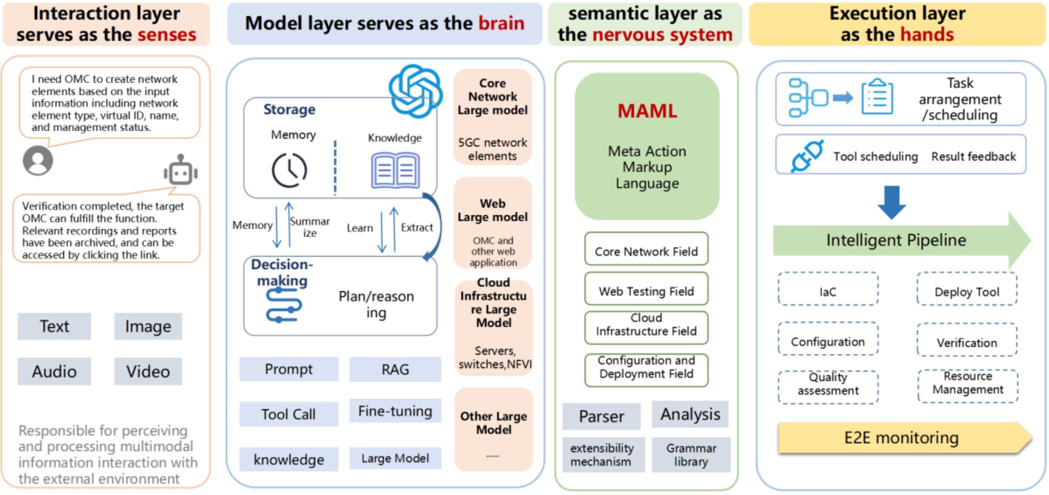

Interaction layer serves as the senses
I need OMC to create network elements based on the input information including network element type, virtual ID, name, and management status.
Verification completed, the target OMC can fulfill the function. Relevant recordings and reports have been archived, and can be accessed by clicking the link.
Text
Image
Audio
Video
Responsible for perceiving and processing multimodal information interaction with the external environment
Model layer serves as the brain
Storage
Memory
Knowledge
Memory
Summar ize
Learn
Extract
Decision-making
Plan/reason ing
Prompt
RAG
Tool Call
Fine-tuning
knowledge
Large Model
Core Network Large model
5GC network elements
Web Large model
OMC and other web application
Cloud Infrastructu re Large Model
Servers, switches,NFVI
Other Large Model
......
semantic layer as the nervous system
MAML
Meta Action Markup Language
Core Network Field
Web Testing Field
Cloud Infrastructure Field
Configuration and Deployment Field
Parser
Analysis
extensibility mechanism
Grammar library
Execution layer as the hands
Task arrangement /scheduling
Tool scheduling
Result feedback
Intelligent Pipeline
IaC
Deploy Tool
Configuration
Verification
Quality assessment
Resource Management
E2E monitoring

Figure 3-4 Architecture of intelligent network infrastructure solution

Interaction Layer: Serving as the "senses" for interaction between the 5G-A AI agents and humans, this layer is responsible for perceiving and processing multi-modal information interactions with the external environment, including text, images, audio, and video. It also feeds back the results of task processing to users through the interaction layer and enables iterative interaction processes.

Model Layer: Serving as the brain of the AI agent, this layer is responsible for task decomposition, planning, reasoning, memory, reflection, and tool usage. For different scenarios in 5G-A infrastructure, the large model is divided into core network models, web models, cloud platform models, etc., for completing specific tasks in different domains.

Semantic Layer: To bridge the gap between task execution and natural language, a semantic layer is established, serving as the neural signal layer for the AI agent, enabling large models to understand and execute various tasks. This semantic layer sets a unified MAML (Meta Action Markup Language) and defines sub-languages for subdomains such as network verification, web verification, cloud platform verification, deployment and configuration.

Execution Layer: Serving as the "hands and feet" of the AI Agent, this layer is responsible for executing, orchestrating, and scheduling specific tasks. Leveraging intelligent pipelines, the execution layer completes tasks such as environment deployment, test case validation, quality assessment, and resource management based on automation tools or scripts, and then feeds back results to the model layer for iteration, reflection, and planning of subsequent tasks.

Through this technology, 5G-A network infrastructure can achieve more efficient and intelligent management and operation to meet the rapidly evolving needs of future network technologies.

3.1.4 AI for Network Mobility

The application of AI communication networks represents a significant breakthrough in 5G-A. One of the critical innovations is AI for network mobility. Addressing the dynamic and complex demands of networks, AI offers new solutions to optimize handovers, predict potential failures, and reduce resource and measurement overhead.

1) RRM Measurement Prediction

Radio Resource Management (RRM) measurement prediction plays a pivotal role in facilitating efficient mobility management[6]. AI-based approach leverages historical and real-time measurement data across clusters of cells to generate predictions that improve handover decisions and reduce the need for frequent measurement and reporting. These AI models could be trained to predict measurements in the temporal, frequency, and spatial domains. The framework diagram as shown in Figure 3-5 illustrates the overall prediction structure, where AI models can be deployed at various locations within the network. Among the proposed approaches, we emphasize that the cluster cell-based method combined with Case 3 (direct prediction) demonstrates superior potential for leveraging AI prediction gains. This is achieved by incorporating richer spatial information and performing joint optimization across multiple modules, enhancing both prediction accuracy and network performance.

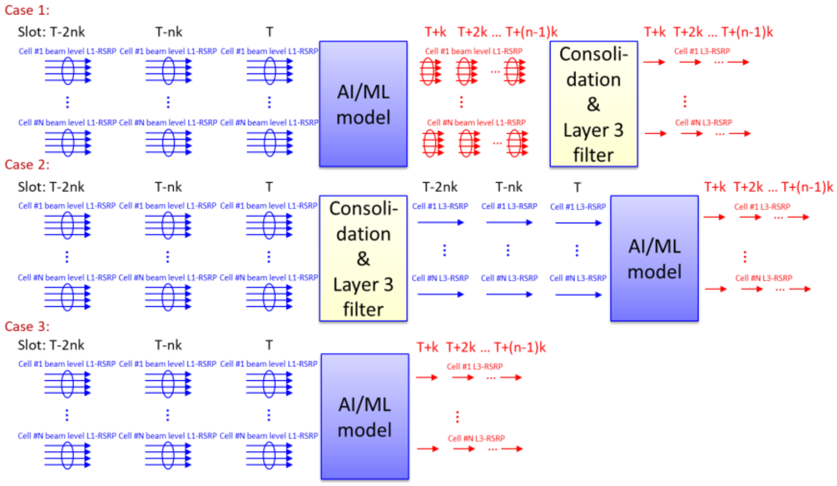


Figure 3-5 RRM measurement prediction framework

2) Failure Prediction

The prediction of Handover Failures (HOF) and Radio Link Failures (RLF) is essential for maintaining seamless connectivity during user mobility[7]. AI models analyze trends in signal quality over time to identify patterns that indicate potential failures, enabling proactive network adjustments to mitigate service interruptions. RLF scenarios associated with T310 expiry—a representative case for failure prediction—can be addressed using both short-term and long-term approaches.

In short-term scenarios, where T310 has already been triggered, AI models estimate the likelihood of failure within the remaining duration of the timer. These predictions provide immediate insights into the risk of failure, supporting real-time decision-making and recovery actions. Long-term predictions, conducted before T310 activation, focus on estimating both the probability and timing of potential failures. By incorporating extended RRM measurement trends across multiple cells, long-term approaches allow for preemptive planning and optimization of mobility strategies.

Direct prediction models, which output failure probabilities directly, are particularly suited to short-term scenarios due to their ability to offer precise and actionable insights. Conversely, indirect prediction approaches, which rely on temporal RRM measurement predictions, are more effective in long-term scenarios. These methods capture the evolving conditions of both serving and neighboring cells over extended time windows, offering a comprehensive perspective for resource allocation and mobility management.

In environments with higher failure risks, such as FR2 deployments, these predictive capabilities are particularly valuable. While direct predictions address immediate risks, indirect predictions leverage multi-cell spatial and temporal data to support strategic planning, ensuring reliable network performance. Together, these methods enhance the ability to reduce HOF and RLF rates, contributing to more robust and adaptive network operations.

By integrating these predictive approaches, AI models play a transformative role in ensuring seamless service delivery, even in the most dynamic and complex mobility scenarios.

3) Event Prediction

By leveraging AI models, the prediction of measurement events, such as Event A3, helps optimize network operations, reducing measurement overhead and enhancing handover performance through more precise and timely event triggers[8]. Similar to failure prediction, measurement event prediction can also be achieved through both indirect and direct approaches. Indirect prediction utilizes RRM measurement predictions as inputs, allowing the network to infer the occurrence of events based on trends in temporal and spatial data from serving and neighboring cells. This method integrates seamlessly with existing mobility frameworks, leveraging multi-cell correlations to maintain prediction accuracy while reducing operational complexity. On the other hand, direct prediction bypasses intermediate steps, with AI models directly outputting event triggers. This approach is particularly well-suited for real-time applications, enabling faster response times and streamlined processing.

In scenarios focused on reducing measurement overhead, such as FR1 deployments, predicted measurement events can replace actual measurements, significantly lowering signaling load while maintaining equivalent performance. For instance, temporal domain predictions allow the network to decrease the frequency of measurement reporting without sacrificing event accuracy. Meanwhile, in FR2 environments, where mobility demands are higher, these predictions enhance handover performance by proactively adjusting parameters like hysteresis and time-to-trigger. This reduces the likelihood of late handovers and improves the selection of optimal target cells, ensuring smoother transitions between network nodes.

Evaluating the accuracy of these predictions is crucial to their effectiveness. Metrics such as missed prediction ratios, false prediction ratios, and time accuracy are essential for assessing reliability. Accurate predictions ensure that event triggers occur at the right time, enabling timely handovers and minimizing disruptions. These capabilities contribute to a more adaptive and efficient mobility management framework, addressing the dynamic challenges of next-generation networks.

4) Charting the Path Forward

AI for mobility in 5G-A represents a transformative step forward, as demonstrated by advancements in measurement accuracy, failure prediction reliability, and event forecasting capabilities. The development of robust frameworks that integrate prediction outcomes into network decision-making processes will be critical to realizing these benefits. Future research will continue to enhance the adaptability of AI models to varied deployment conditions, ensuring consistency and scalability.

By embedding AI-driven insights into the heart of 5G-A, the industry moves closer to achieving its vision of a more intelligent, adaptive, and efficient network. These innovations lay a strong foundation for future capabilities in mobility management, paving the way for a truly connected future.

3.2 Digtal Twin Network Intelligence

Digital Twin Network (DTN) represents a paradigm shift in how we manage and optimize 5G-A network and beyond. By creating a virtual replica of the physical network infrastructure, DTN enables operators to simulate, analyze, and optimize network operations with unprecedented precision. This digital representation serves as both a mirror of current network state and a sandbox for testing future scenarios, offering a powerful platform for innovation and network optimization.

3.2.1 Network Data Mapping Technology

Data mapping forms the cornerstone of network digital twin implementation, providing essential input for twin modeling. The quality and comprehensiveness of collected data directly determine how faithfully the digital twin reflects its physical counterpart. However, modern networks present significant challenges for data collection due to their vast scale, diverse device ecosystem, complex interface configurations, and dynamic traffic patterns. These challenges make high-precision data collection both resource-intensive and potentially disruptive to network performance.

To address these challenges, the focus has shifted toward developing an efficient network data collection mechanism that meets the specific requirements of network twins. This mechanism adopts an on-demand approach to data collection, carefully selecting collection targets, precision levels, protocols, and transmission methods. The goal is to achieve a balanced solution that delivers comprehensive data while maintaining efficiency and energy conservation.

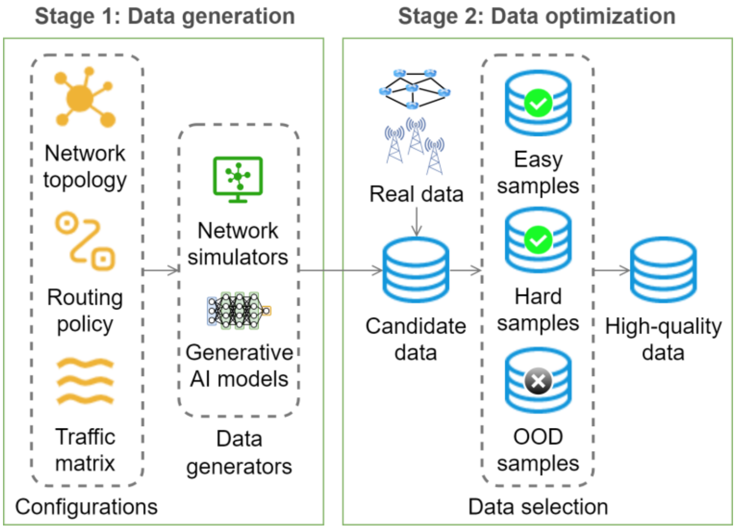


Figure 3-6 The framework of AutoOPT

The AutoOPT framework represents an innovative approach to this challenge by leveraging generative AI models for data generation and optimization. As illustrated in Figure 3-6, this framework operates through two distinct stages. In the data generation stage, it creates synthetic data from small-scale networks using scale-independent indicators, ensuring that DTN AI models[9] can effectively generalize to larger network environments. The subsequent optimization stage automatically identifies and filters high-quality data through seed sample selection and incremental refinement, ultimately enhancing both the accuracy and generalization capabilities of DTN AI models.

This approach not only addresses the immediate challenges of data collection but also creates a foundation for continuous improvement in digital twin fidelity. By using the digital twin itself as a simulated data generation entity, we can enrich machine learning training datasets and create more robust models for network analysis and optimization.

3.2.2 Digital Twin Modeling Technology

NDT modeling technology is pivotal for realizing effective network intelligence in complex 5G-Advanced and future networks. It focuses on constructing high-fidelity virtual replicas of physical network infrastructures, balancing model accuracy with computational efficiency. In the context of increasingly intricate and dynamic network architectures, a robust NDT model is essential for functionalities such as accurate state replication, predictive simulation, and historical event analysis, enabling proactive network management and optimization.

NDT modeling is structured around three fundamental dimensions, synergistically contributing to a holistic digital representation. Firstly, network state digital twinning addresses the real-time and historical representation of network entities, their attributes, and inter-relationships. This dimension encompasses both invariant network configurations and dynamic operational telemetry. Technique employed includes:

- **Network Emulators:** Utilizing platforms that simulate network environments with varying levels of abstraction, from detailed packet-level emulation to higher-level behavioral models.
- **Data Abstraction Expressions:** Employing formalized data models and schema to

represent network state data efficiently and facilitate semantic interoperability. This can involve utilizing data serialization formats and standardized information models.

- **Knowledge Graphs:** Constructing graph-based representations to capture network topology, entity relationships, and state dependencies, enabling knowledge-driven reasoning and inference within the digital twin.

Secondly, service and environment twinning focuses on modeling the service layer and its operational context. This is crucial for understanding service performance and user experience within the network. Key methodologies include:

- **Generative Adversarial Networks (GANs):** Leveraging GAN architectures to synthesize realistic and dynamic service behavior patterns and simulate user interaction profiles. GANs are particularly effective in capturing complex, non-linear service dynamics and generating synthetic datasets for model training and validation.
- **Deep Learning-based Environmental Reconstruction:** Utilizing deep learning algorithms, particularly CNN and point cloud processing techniques, to process sensor data (e.g., LiDAR, camera imagery, deployment blueprints) for automatic identification, extraction, and 3D reconstruction of the physical network deployment environment. This generates spatially accurate context for network simulations and visualizations.

Thirdly, network behavior twinning models the functional characteristics of network elements. This dimension employs a spectrum of modeling approaches based on the desired level of fidelity and computational cost:

- **White-box Modeling:** Implementing network protocols and device functionalities directly within the digital twin environment, often through Software-Defined Networking (SDN) and Network Function Virtualization (NFV) principles. This provides high transparency and control over simulated network operations, suitable for detailed protocol analysis and functional verification.
- **Black-box Modeling:** Utilizing data-driven modeling techniques, such as machine learning and statistical modeling, to predict network status transitions and performance metrics based on observed historical data. This approach abstracts away from internal protocol details, focusing on input-output relationships and predictive accuracy, suitable for performance forecasting and anomaly detection.
- **Gray-box Modeling:** Combining abstract network mechanism representations with selected actual network functions. This hybrid approach offers a pragmatic balance between modeling fidelity and computational complexity. Formal methods like Petri Nets are particularly valuable in this context, offering a mathematically rigorous framework for modeling concurrent and asynchronous events in distributed industrial IoT network digital twins, as demonstrated in prior research[10].

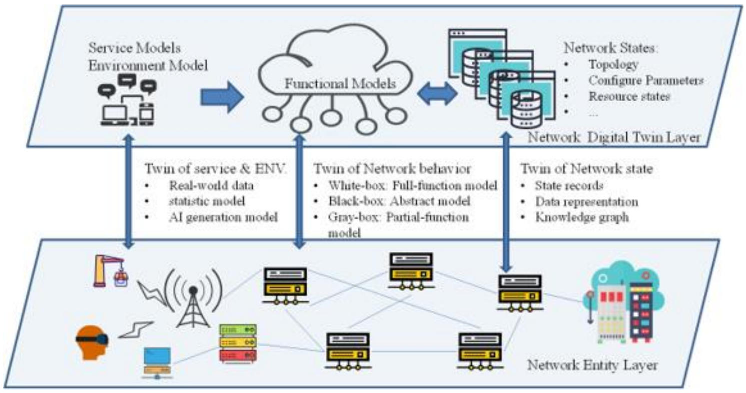


Figure 3-7 The NDT modeling method

These three dimensions, as conceptually illustrated in Figure 3-7, are integral to constructing a comprehensive NDT. This integrated modeling paradigm empowers network operators with enhanced capabilities for network comprehension, predictive analytics, and proactive optimization, ultimately facilitating the evolution towards intelligent, autonomous network management and supporting the advanced service requirements of modern network infrastructures.

3.3 Application Intelligence

Communication networks are evolving from basic connectivity services to intelligent application services. Through the deep integration of AI technology, networks not only provide traditional voice and data services but also support new intelligent services such as multi-modal interaction, real-time translation, and digital humans. To support these innovative applications, networks have undergone technical evolution. These enhancements not only expand the application boundaries of networks but also create new business models for operators, shifting from traffic operations to value-based operations.

3.3.1 IMS Data Channel

The IP Multimedia Subsystem Data Channel (IMS DC) is an advanced feature that enables high-speed, real-time transmission of non-voice data during communication sessions, leveraging the infrastructure of IMS-based voice and video services. It is designed to meet specific requirements related to latency, bandwidth, and reliability, which are crucial for applications such as AR, real-time multimedia, and IoT interactions. This section explores the architecture, function, and advantages of IMS DC in detail.

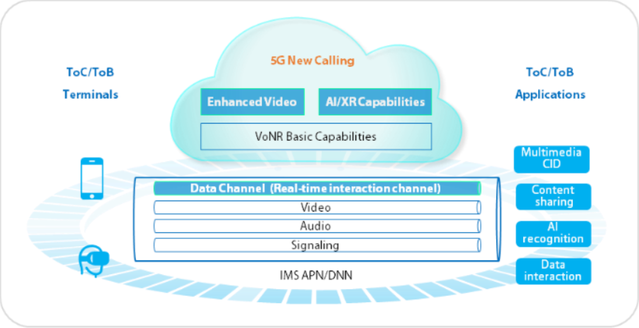


Figure 3-8 IMS DC enables real-time multimedia interaction

As shown in Figure 3-8, the IMS DC is integrated within the IMS framework, building on the voice and video channels to create a seamless, multipurpose communication channel. This integration allows for synchronous transmission of data types alongside traditional voice or video calls, ensuring that data flows remain synchronized with the multimedia content. The IMS DC operates efficiently under the existing IMS architecture, utilizing the inherent advantages of telecom networks, such as global connectivity via telephone numbers, unified authentication, and robust session management.

The IMS DC is divided into two primary components:

- **Bootstrap Data Channel (BDC):** this channel facilitates the download and execution of IMS DC applications on the terminal. These applications may include web content (e.g., HTML5 pages), media elements, and control scripts (e.g., JavaScript) required for real-time interactions.
- **Application Data Channel (ADC):** this is the data conduit through which applications on the terminal transmit their data to other devices or networks. The ADC ensures the real-time transmission of application-specific data in parallel with voice/video traffic, maintaining consistent user experience and minimal delay.

The IMS DC is established between the terminal and the network to manage data exchange efficiently. When a communication session begins, the DC Server establishes a BDC to push DC applications to both calling and receiving terminals. Once these applications are active, the ADC is used to handle data transmission, ensuring real-time synchronization of all multimedia and application data.

The control and media functions are separated in the IMS DC architecture. This approach uses Data Channel Signaling Function (DCSF) to handle signaling control and Media Function (MF) to manage media resources, supporting functionalities like AR rendering and media processing, which greatly improves flexibility and scalability for future applications.

The IMS DC provides essential benefits for modern communications, including:

- **Real-Time, Multi-Modal Interaction:** By combining video, voice, and data streams, it enables immersive and interactive communication experiences, such as AR and virtual meetings, directly within the IMS framework.
- **Enhanced QoS and Security:** Leveraging the robust capabilities of IMS, the IMS DC ensures QoS management and strong security protocols, protecting data integrity and minimizing latency.
- **Flexible Application Integration:** The ability to dynamically introduce new applications through the BDC streamlines the process of extending IMS capabilities, ensuring that future innovations can be rapidly integrated into the network without major

infrastructure changes.

In summary, the IMS DC represents a pivotal step forward in the evolution of telecommunication services, offering a unified channel for data, voice, and video communication, optimized for real-time, multimedia interactions across diverse network environments. The ongoing advancements in IMS DC, particularly with the integration of Augmented Reality (AR) and future communications services, underline its role as a cornerstone for 5G-A and beyond communication systems.

3.3.2 Interactive New Calling

The 5G-A new call services provide users with interactive, intelligent, and immersive innovative new services and new scenarios, such as digital human, intelligent translation, intent communication etc.

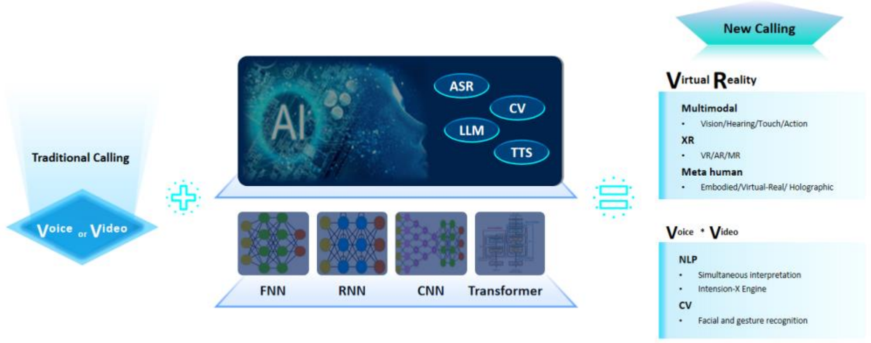


Figure 3-9 New AI Technologies for New Calling

The new AI technologies refresh traditional call services, and promotes the expansion of business models from voice or video to multi-modal communication. With the rapid development of AI technologies, AI has become an indispensable force to promote communication technology innovation. Through the AI empowerment network, continuous innovation is injected into call services. There are many new AI technology used for new calling as following:

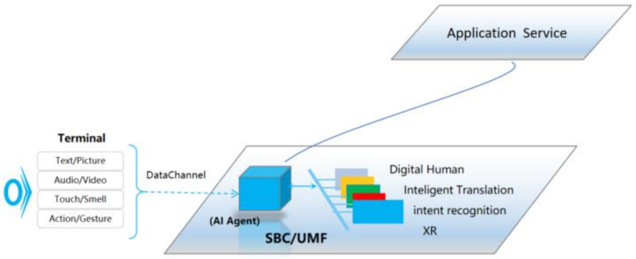


Figure 3-10 Interactive New Calling

- **Intent Recognition:** intent recognition refers to the recognition and understanding of intentions or purposes expressed in human languages through natural language processing technologies like LLM, Natural Language Processing (NLP). Intent-based communication greatly simplifies the interaction procedure of new calls. The system automatically identifies and executes the intent based on only texts and pictures. For example, if you order 1,000 pizzas, the system automatically executes the intent identification and make orders.

- **XR + AGI Technology:** the Extended Reality (XR) interactive technology refers to a real and virtual combination and man-machine interaction environment generated through computer technologies and wearable devices. It integrates virtual information and real scenarios to create a man-machine interaction virtual environment. It can be used in entertainment, education, medical care, and industrial fields to provide more abundant and immersive experience. Artificial General Intelligence (AGI) technology can be used to quickly generate XR videos and digital worlds.
- **Digital human:** digital human technology is a high-tech product that combines computer graphics, motion capture, image rendering, and artificial intelligence. It allows the creation of virtual characters with human looks, behaviors, and characteristics. These virtual characters can exist in digital space and interact with the real world.
- **Intelligent translation:** combined with real-time communication, intelligent translation translates the voice of a subscriber into text information and displays the text information on the subscriber's mobile phone. Intelligent translation can be used for translation between different languages or different dialects based on AI like Automatic Speech Recognition (ASR), LLM etc. It is also applicable to real-time speech/text conversion when voice calls cannot be made, for example, communication between a normal person and a hearing-impaired person.

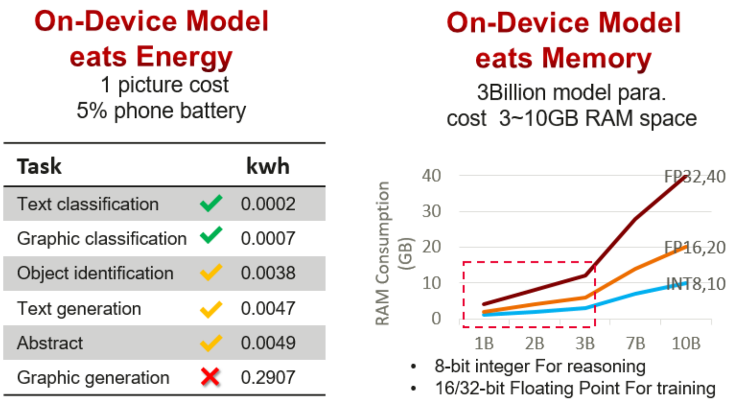
Through the application of AI technologies such as ASR, intent recognition, intelligent translation, and digital human, new calling will benefit from the following:

- **New business model:** in the future, communications will evolve from traffic and time operations to value operations and provide value services, such as digital man, personal assistants, and avatar/ego communications.
- **Shortening the service launch time:** by integrating the AI technology, the service launch time can be simplified. Through the plug-in intelligent AI platform, new service components can be plug-and-play.

3.3.3 Cloud-Edge-Terminal Collaboration

Cloud-edge-terminal collaboration is central to optimizing communication networks, especially for AI-driven applications. As AI models grow in complexity, efficiently distributing tasks across cloud, edge, and terminal layers is vital. Key strategies like computational offloading, resource scheduling, and data collaboration are critical to enabling seamless collaboration across these layers.

1) Computational Offloading



On-Device Model eats Energy
1 picture cost
5% phone battery

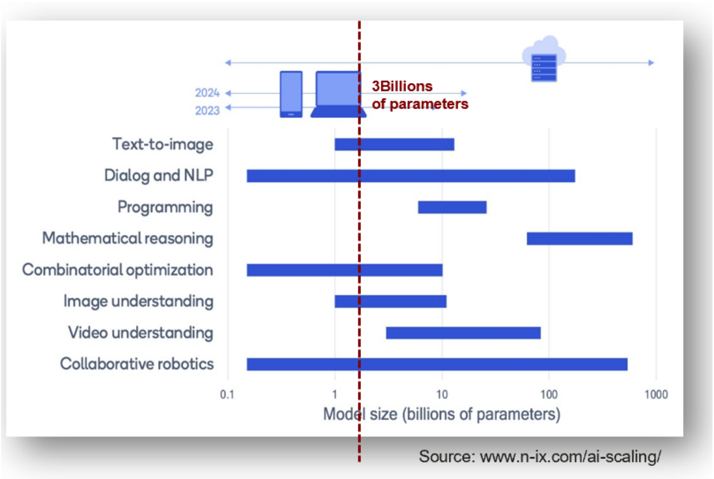
Task		kwh
Text classification	✓	0.0002
Graphic classification	✓	0.0007
Object identification	✓	0.0038
Text generation	✓	0.0047
Abstract	✓	0.0049
Graphic generation	✗	0.2907

On-Device Model eats Memory
3Billion model para.
cost 3~10GB RAM space

- 8-bit integer For reasoning
- 16/32-bit Floating Point For training

Figure 3-11 On-Device AI: energy and memory consumption

With the increasing complexity of AI models, terminal devices are often not capable of handling the computational load on their own. Offloading these tasks to the cloud or edge allows for better distribution of computational workloads, ensuring real-time performance at the terminal. As illustrated in Figure 3-11, energy and memory usage on terminal devices rise significantly as the complexity of AI models increases. This highlights the importance of offloading more resource-intensive tasks to the cloud or edge, thus allowing terminal devices to focus on lighter computations.



Source: www.n-ix.com/ai-scaling/

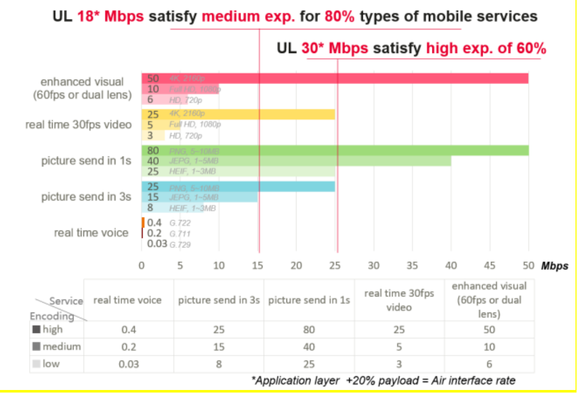
Figure 3-12 AI model size requirements for different tasks

Similarly, Figure 3-12 shows the varying computational resource needs of different AI models. Large-scale models, such as those used in deep learning, are typically offloaded to the cloud, while smaller models that need real-time inference are handled at the edge or terminal. This tiered approach ensures that AI applications can scale effectively, based on the specific computational requirements of the task at hand.

2) Uplink Centric Broadband Communication

With the popularization of AI applications such as AI video calls and AI Agents, AI-human interact with each other in multi-modal, such as images and videos, and uploading of images and videos from terminal to cloud becomes necessary. According to related research, an uplink rate of 20Mbps is required to ensure interaction experience of 80% common applications and 30Mbps is required to ensure interaction experience of 60% high-experience applications like enhanced

Supplementary Uplink (SUL).



Service / Encoding	real time voice	picture send in 3s	picture send in 1s	real time 30fps video	enhanced visual (60fps or dual lens)
■ high	0.4	25	80	25	50
■ medium	0.2	15	40	5	10
■ low	0.03	8	25	3	6

*Application layer +20% payload = Air interface rate

Figure 3-13 Uplink Experience Requirements for Multi-modal Interaction

Uplink Centric Broadband Communication (UCBC) improves the upstream bandwidth capability by 10 times, meeting the upload requirements of multi-modal interaction, machine vision, and massive broadband IoT in different AI application scenarios, accelerating the intelligent upgrade of thousands of industries. Currently, more spectrum and bandwidths are aggregated based on the current uplink channels to improve the uplink capability. The main technologies include flexible spectrum access, SUL enhancement, and uplink carrier aggregation. FA SUL networking could be applied to achieve large uplink bandwidth and meet multi-modal interaction requirements. In addition, the uplink and downlink slot assignment of the TDD frequency band are adjusted to increase the scheduling of uplink time-slot resources, improving the uplink capability.

3) Resource Scheduling

Efficient resource scheduling is crucial for ensuring that tasks are allocated optimally across cloud, edge, and terminal nodes. Dynamic allocation of resources ensures that computational power, memory, and bandwidth are used effectively, minimizing latency and maximizing throughput. AI applications often require significant uplink bandwidth, especially when processing large datasets. Meeting this demand is critical for ensuring smooth data transfer between the terminal, edge, and cloud layers.

In the context of cloud-edge collaboration, the role of resource scheduling becomes even more significant. The core focus is on managing tasks based on the current network load and resource availability, ensuring that computational resources are used as efficiently as possible. This dynamic scheduling helps maintain optimal performance, reducing bottlenecks and enabling faster task execution.

4) Data Collaboration

Data collaboration ensures the smooth flow of information between the cloud, edge, and terminal. Data collected at the terminal is first processed and cached at the edge, reducing the need to transfer large amounts of data to the cloud. This local processing minimizes latency and makes real-time decision-making possible, particularly for applications where time sensitivity is critical.

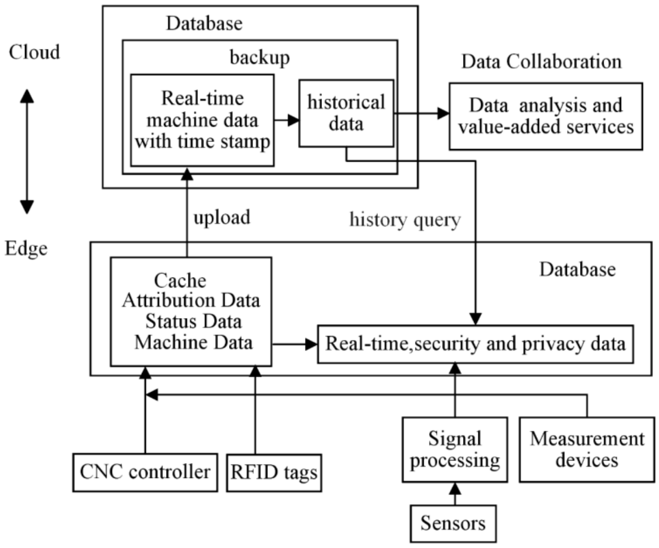


Figure 3-14 Cloud-Edge Data Collaboration Scheme

Figure 3-14 illustrates how data is processed and managed across these layers. Data is pre-processed at the edge before being uploaded to the cloud, which then stores and further analyzes the data. This collaboration ensures that only the most relevant data is transmitted to the cloud, optimizing bandwidth usage and ensuring efficient model training. It also enables secure handling of sensitive data, especially in domains like healthcare and finance.

The synergy between computational offloading, resource scheduling, and data collaboration enables cloud-edge-terminal collaboration to operate efficiently. By leveraging these strategies, the overall performance of AI applications improves, with computational power distributed appropriately, latency minimized, and data flows optimized.

As AI-driven applications continue to evolve, the ability to dynamically allocate resources and manage data across the cloud, edge, and terminal will become even more important. This collaboration not only improves the efficiency of AI processing but also allows applications to scale effectively, meeting the growing demands of next-generation services.

3.4 Sustainability Intelligence

With the rising global electricity prices, power conservation and energy efficiency have become critical priorities in the telecommunications industry. For mobile networks, reducing energy consumption to achieve net-zero emissions has become a core goal for operators, with over 85% of mobile network operators committed to this objective. As both the Radio Access Network (RAN) and Core Network contribute significantly to overall energy consumption, enhancing energy efficiency in both areas is crucial.

The integration of AI plays a pivotal role in improving energy efficiency. By analyzing real-time traffic patterns and resource utilization, AI enables optimized resource scheduling and power management, significantly reducing energy consumption while maintaining service quality.

3.4.1 Enhancing Equipment-Level Energy Efficiency

At the equipment level, the primary energy consumer is the Power Amplifier (PA) in the radio unit, which amplifies signals for transmission. Optimizing the efficiency of the PA is essential for improving overall energy savings. Existing technologies improve the PA’s linearity and efficiency under low load conditions; however, most solutions show limited optimization effects during periods of low load, which are common in network operations.

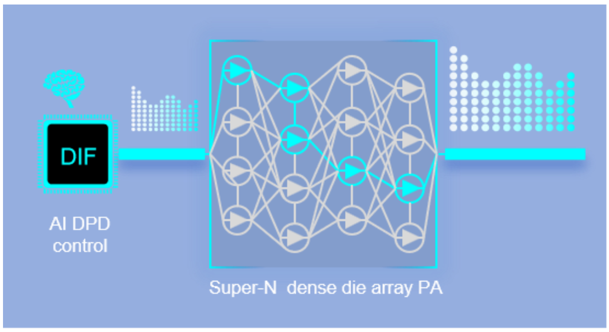


Figure 3-15 Intelligent PA control

To address this, AI-driven Digital Pre-distortion (DPD) techniques have been employed to improve the linearity of the PA. Traditional DPD algorithms have been upgraded with machine learning models that dynamically adjust to different environmental and signal conditions. Deep Neural Networks (DNNs) model the nonlinear characteristics of the PA and adjust the input signal in real time to achieve better power efficiency and signal quality.

AI-driven DPD systems are trained on extensive real-world signal data, taking into account various frequency bands, power levels, and environmental conditions. These systems can autonomously adjust parameters in complex network environments, reducing signal distortion and enhancing performance.

3.4.2 Optimizing Network-Level Energy Efficiency

At the network level, AI predicts traffic load fluctuations by analyzing historical traffic patterns, weather data, and local events, allowing for dynamic adjustment of network resources to reduce unnecessary energy consumption. AI-powered shutdown mechanisms can predict low-traffic periods and automatically deactivate components such as power amplifiers and transceivers, reducing energy consumption without affecting service quality.

AI models also allow for intelligent control of network equipment by analyzing real-time feedback from both terminal devices and network components. This enables the network to dynamically adjust power distribution and transmission strategies. For instance, PA and Base-band Units (BBUs) in base stations can enter deep sleep during low-load periods and be reactivated only when traffic increases.

Furthermore, AI optimization extends beyond the RAN and includes energy efficiency improvements in the Core Network. In the Core Network, intelligent scheduling and resource management mechanisms analyze traffic patterns, user demands, and network topology to dynamically route data flows and minimize unnecessary energy consumption.

Through multi-dimensional energy-saving strategies, including time, space, frequency, and power domain optimizations, AI enables precise energy optimization without compromising performance. This capability is expected to evolve further in R18 and beyond, incorporating network digital twin technology for offline energy-saving predictions, enabling more accurate energy-saving strategies for future network deployments

4 5G-AxAI Enables New Cases

4.1 Differentiated Experience Assurance

As mobile internet develops, users' demands for network are becoming increasingly personalized. From 4G to 5G, operators primarily manage data services through traffic management. In the era of 5G-A, utilizing AI technology to dynamically adjust network resource allocation is crucial for meeting the differentiated experience requirements of various user levels (such as diamond, platinum, gold, silver, and standard cards), different types of services (like short videos, games, and live streaming), medium to high-speed mobile scenarios (such as high-speed rail and subway), and high-capacity scenarios (like concert venues and tourist attractions). Technologies such as network perception, experience assurance, and experience evaluation can significantly enhance user satisfaction.

Network perception: Currently, the intelligent network perception module supports the identification of tens of thousands of service categories, including mainstream domestic and international service traffic. It also finely identifies various sub-services within super apps, such as voice calls, video calls, live streaming, video conferencing, and cloud gaming, with an overall sensing rate exceeding 95%.

Experience Assurance: Based on the results of network perception, targeted assurance strategies are provided for corresponding services. Key technologies for experience assurance adopted on the wireless side include precise service assurance prediction, multi-objective optimization service assurance strategies, and intelligent multi-frequency coordination. On the core network side, network intelligence based on NWDAF is introduced to provide dynamic Guaranteed Bit Rate (GBR) assurance according to user needs.

Experience Evaluation: Experience evaluation involves developing a comprehensive set of QoE metrics based on the characteristics of different services. For video services, this includes assessing clarity, smoothness, and timely interaction through multiple dimensions such as content quality, transmission quality, and interaction quality, which are then integrated into a unified standard score [1, 5]. For instant messaging services, the focus is on factors like service type and response timeliness, with metrics including text, voice, image, and video transmission types, along with service volume, duration, speed, and performance indicators under TCP/UDP protocols (such as packet loss rate, delay, jitter, and TCP retransmission rate), all fitted to calculate the Mean Opinion Score (MOS) for instant messaging services. This comprehensive approach ensures that the evaluation accurately reflects the quality experienced by users across various types of services.

Test results on the existing network are exhibited as follows:

- **On the wireless network side:** After deploying the precise service assurance strategies, the latency for various types of services, such as short video, QR code payment, and web browsing, was reduced by approximately 20%.

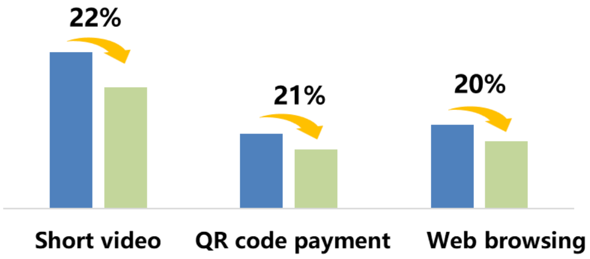


Figure 4-1 Reduction in latency for short video, QR code payment, and web browsing services.

- **On the Core Network side:** After introducing NWDAF-based intelligence, when a decline in user experience is detected, the GBR assurance is then employed to enhance user's experience.

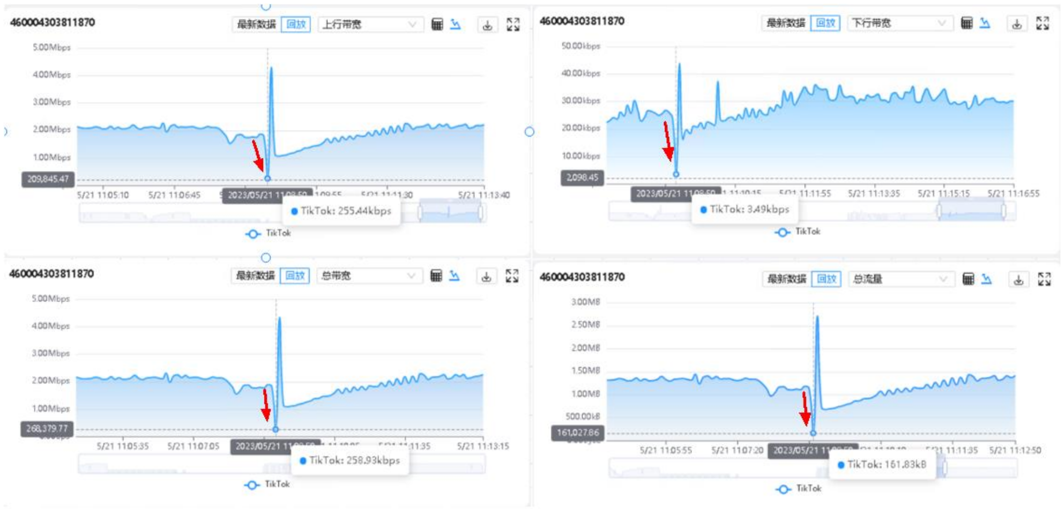


Figure 4-2 For VIP users, upstream speed was 2Mbps before congestion, dropped to 255Kbps after congestion, and restored to 2Mbps after establishing GBR.

Additionally, during the service assurance process, the logo on the user's mobile device can dynamically display the text "VIP Assurance Active" in real-time, matching the enhanced experience and elevating the user's perception. After the assurance process ends, users receive an experience report providing real-time feedback on the measurable assurance effects. This completes the closed-loop experience management, achieving a higher level of care for the customer.

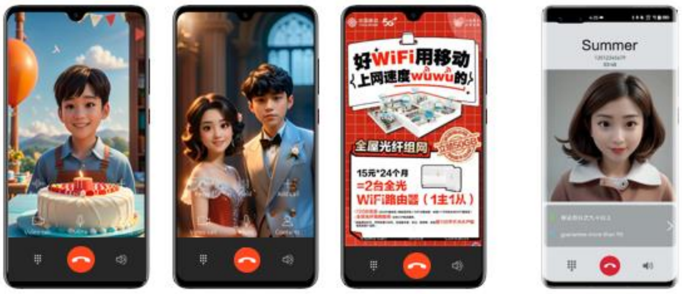
Figure 4-3 Zhejiang Mobile Test Results

4.2 New Calling Service

In recent years, the development of multimedia large models, AI Agent, and DC technologies is driving the communication industry from traditional audio and video to multi-modal, interactive communication. This provides users with a series of rich and colorful services such as intelligent video calls, interactive video calls, and call intelligent assistants, significantly enhancing the communication experience and efficiency for individuals and enterprises. It promotes the prosperous development of the communication industry and opens up new growth space for operator businesses. In 2023, China Mobile, in collaboration with partners, launched the 5G New Calling service and completed the construction of the first phase of the New Calling network covering 31 provinces nationwide by the end of the year. Relevant New Calling network elements were deployed in the resource pools of the eight major regions, involving the upgrade of existing IMS network elements and the deployment of new calling capability elements. This supported the release of six New Calling service scenarios: lighting up screen, fun calling, celebrity calls, AI shorthand, speech to text, and real-time translation.

1) **Lighting up the screen: from voice to video to content, opening up a new space for content operation**

In 2023, China Mobile Jiangsu worked with Huawei and other partners to develop a new call lighting up screen service to meet consumers' emotional expression requirements. The service was released in Q4 of that year.



Personal Visualized Voice Calling

Enterprise Visualized Voice Calling

Avatar Calling

Figure 4-4 New call lighting up screen service

Lighting up the screen is a product centered on user communication social attributes. Users can use the product to set personal virtual images. During a voice call, users can transmit and display preset pictures, videos or virtual digital human images to the calling party without switching video calls, so as to convey emotions and express individuality. Enterprise users can customize corporate images, spread corporate value, or conduct product marketing. During calls, they can trigger introduction videos or pictures based on different keywords to improve communication efficiency and effect. Carriers can publicize anti-fraud and security knowledge to assume more social responsibilities. In less than a year, Jiangsu Mobile has more than 5 million users of the screen-lighting service, and more than 100 million content displays per month. It has received good market feedback and has become a new call service and is popular with consumers. China Mobile Guangdong enables the ToB application for all installation and maintenance personnel. During communication with customers, keywords are used to trigger Fiber to The Room (FTTR) brand promotion and package promotion videos, facilitating service promotion and efficient communication. In the second half of 2024, China Mobile upgraded its screen-lighting business, implanting Artificial Intelligence Generated Content (AIGC) self-creation capabilities, allowing users to create their own stylized images and show them to each other during calls. Jiangsu Mobile also explores the "Another Me" function of digital human calls. Based on the user's appearance characteristics, the stylized digital human image is generated by AI. During a call, the digital human is driven by the network AI based on the user's voice. Real-time synchronization of my voice and digital person's expression and lip shape, infusing more fresh experience into the call. By the end of 2024, more than 15 million users had subscribed to the light-up screen service. Turning on the screen enables carriers to upgrade the call duration operation to an average of 90 seconds, which brings greater business and social value to calls and becomes a new mobile media.

2) **Real-time translation & Speech to text: Technology for Good, Building a Bridge for Real-Time Communication**

In September 2023, at the 19th Asian Games in Hangzhou, China Mobile launched the "Smart Translation" service for the Asian Games. When cross-language communication is required, the "Smart Translation" service can automatically identify and translate voice content during native video calls without relying on translation software. Bilingual subtitles are displayed. During the Asian Games, real-time translation between English, Korean, Japanese, Arabic and other languages can be supported, allowing communication to cross the language gap. The Call Caption service can identify the voice content of calls and display the voice content in large fonts on the mobile phone screen to reduce communication obstacles and greatly improve the call experience of the hearing impaired.

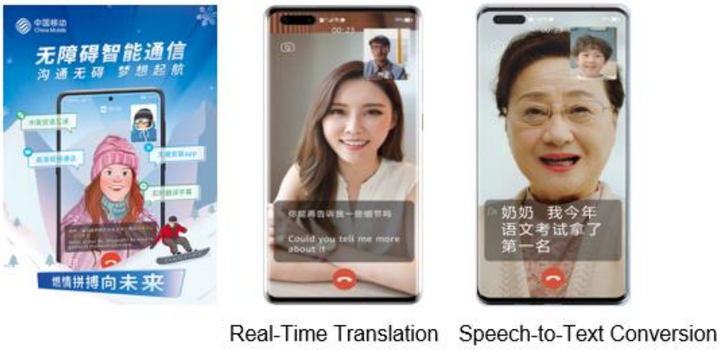


Figure 4-5 Real-time translation and call captioning service

In 2024, the real-time translation and call captioning service was launched in all provinces of China Mobile. In less than one year, more than 5 million users subscribed to the call captioning or call captioning service. The new call truly fulfills the tenet of technology for good. Help users to move towards the goal of "barrier-free communication".

By the end of 2024, the number of new 5G call users on the live network has reached 40 million, greatly improving user call experience and communication efficiency, and bringing corresponding business and social values to operators.

In addition to the original voice and video channels, DCs are added to transmit data during audio and video calls. Interactive calls are introduced to implement multimedia content such as image and video sharing, message communication, screen sharing, AR annotation, and file transfer during calls. Improve the experience and efficiency of remote communication and expand the industry value boundary. In the fourth quarter of 2024, China Mobile launched the trial commercial use of interactive video users on the live network. It launched six DC applications, namely, content sharing, message box, fun call+, real-time translation+, screen sharing, and digital person, to implement quick setting, sharing, and convenient interaction during calls.

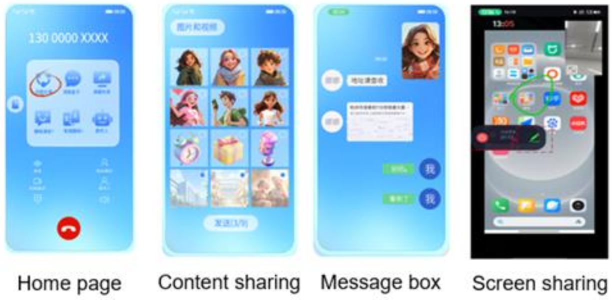


Figure 4-6 Interactive video users

With the collaboration of new communication industry partners, GSMA released TS.66 in June 2024, which defines the DC API standard on the device side. Mainstream chips from chip vendors such as Qualcomm, MediaTek, and UNISOC also support DC capabilities. Some mainstream terminal models from Vivo, OPPO, Xiaomi, Samsung, and Huawei support the DC function. The DC interactive call industry ecosystem is preliminarily established. It is estimated that 2025 will be the first year of large-scale commercial use of DC series applications.

3) AI Agent Enable New Calling, Helping Carriers Build AI Service Entrances

AI comprehensively upgrades the native call experience. As an intelligent call assistant, AI provides services such as intelligent pick-up and intelligent chat, and consolidates the personal call entrance. Upgrade the enterprise customer service hotline and build the enterprise intelligent application portal. In the second half of 2024, China Mobile began to gradually build calling agent capabilities on the network. In Zhejiang, Guangdong, and Jiangsu provinces, China Mobile conducted service tests and trials in service scenarios such as intelligent pickup, AI-assisted chat, and star call. The goal is to build a digital assistant for each user. Build a personal AI service entry.

Intelligent call pilot Personal calling assistant

Figure 4-7 Intelligent call assistant

In scenarios such as unreachable power-off, express delivery, and promotion, the call intelligence helps users intelligently answer calls. Based on call content analysis, the interception conditions can be flexibly set to reduce misjudgment. The interception content is also notified to users and can be called back. To achieve the purpose of anti-harassment, anti-leakage and anti-fraud; It can also help users sort out massive information during calls, remind important information in real time based on call scenarios, reduce users' memory burden, and implement intelligent chat. The call agent can also invoke external vertical domain agents and search tools to help users' complete meal booking and booking, improve call experience and efficiency, and make calls a unified entry for personal AI services. In June 2024, China Mobile Jiangsu streamlined the business processes of vehicle insurance claims and loan customer service scenarios of Company A's customer service hotline. AI was used to help enterprise customer service upgrade intelligently, evolving from traditional voice interaction to multi-modal agents of natural language, video, and data channels. The enterprise digital human customer service personnel obtain key information at a time through natural language interaction. Before the large-scale commercial use of native terminals of the Data Channel, the multi-modal agent capability solves the problems of complex Dual Tone Multi Frequency (DTMF) interaction, multi-layer nesting, and time-consuming. Enterprises can quickly close transactions during calls, reduce transaction costs, achieve win-win among users, enterprises, and carriers, and help carriers expand the enterprise market space.

Intelligent customer service

Figure 4-8 Intelligent customer service

Users can invoke third-party vertical domain agents through code numbers based on actual requirements. Carriers can use code numbers to build a unified entry for call applications, implement One Number, One Agent, and build an enterprise application ecosystem. The calling Agent technology will develop rapidly in 2025, gradually reshaping the call industry ecosystem. As the saying goes, "Calling make AI everywhere, and AI makes calling omnipotent."

4.3 Industrial Deterministic Service

With the rapid development of 5G technology and the gradual improvement of network infrastructure, significant progress has been made in leveraging 5G to empower the digital transformation of industries. Information technologies such as 5G and artificial intelligence have been widely applied in production management and auxiliary production processes. Solutions like data collection, video surveillance, and Automated Guided Vehicle (AGV) have achieved large-scale deployment. However, for production control scenarios, which require extremely high levels of determinism due to short packet transmission cycles (in milliseconds), stringent packet-level requirements (three consecutive packet anomalies can cause system downtime), and a variety of protocols with significant parameter configuration differences, current 5G technology has not yet been fully applied in this domain. Two main challenges remain:

- Traditional 5G QoS Identifier (5QI) mechanisms only guarantee average network performance and cannot provide the required packet-level determinism. This makes it difficult to meet the long-term stable operation needs of industrial control systems.
- Even with advanced networking techniques such as network slicing, differentiated service strategies (such as DS frame structures and PDCP out-of-order delivery), and link redundancy (dual transmission with selective reception), some highly demanding scenarios, especially those requiring extremely low network jitter, still find it challenging to meet service requirements. The existing network solutions are insufficient to fulfill these stringent demands.

To ensure deterministic service experiences, an "1+2+3" deterministic assurance solution is introduced by leveraging AI technology, enhancing the network's deterministic capabilities.

"1" refers to an AI-based service sensing algorithm, focusing on industrial scenarios where AI algorithms are used for self-learning of service characteristics. This continuous iterative optimization accurately identifies packet-level service types, message lengths, transmission intervals, arrival times, and key words in industrial control packets.

"2" refers to two intelligent computing infrastructure solutions, including industrial smart cards for base stations and edge intelligent UPF.

1) **Industrial Smart Cards for Base Stations:** These cards can quickly upgrade traditional base stations into intelligent industrial base stations through plug-in methods, providing a foundational computing platform for AI capabilities such as intelligent service sensing. With forward-compatible design, they flexibly adapt to existing outdoor and indoor mainstream station models without requiring complete station replacement, effectively reducing costs.
2) **Edge Intelligent UPF:** Based on this infrastructure, an intelligent "sensing-inference-orchestration" closed-loop assurance technology system is built for deterministic services. It enables intelligent inspection of deterministic protocol packets, intelligent promotion of deterministic service scenarios, and intelligent management of deterministic orchestration strategies. This addresses challenges such as difficult inter-system coordination, high orchestration complexity, and long deployment cycles. Additionally, it integrates 5G computing power with industrial control and industrial AI capabilities, achieving hybrid real-time virtualized operating systems and base computing power orchestration, promoting deep integration of 5G with industrial

application ecosystems, thus providing more efficient, intelligent, and deterministic production and management methods for enterprises.

"3" refers to three service-based scheduling enhancement solutions, optimizing packet-level scheduling based on characteristics by service sensing to achieve precise matching of network resources with service needs. This enhances deterministic performance while increasing the capacity of deterministic users, ensuring long-term stable operation of services. The three levels of assurance include:

1) **Fixed Configuration Assurance Based on Service Type:** Adjust user-level protocol parameters, such as disabling power-saving features, inactivity timers, BWP switching, and UM mode, to provide differentiated assurance between industrial control terminals and ordinary terminals.
2) **Precision Scheduling Assurance Based on Service Characteristics:** Optimize resource allocation strategies for service flows. On one hand, reserve resources based on packet periods and sizes, making resource allocation more precise and reducing transmission waiting delays. On the other hand, coordinate multiple industrial control flows for staggered transmissions to avoid conflicts, significantly enhancing the ability to ensure bounded low latency under multi-user concurrency.
3) **Key Packet Scheduling Assurance Based on Service Logic:** Optimize scheduler queuing strategies and reliability assurance strategies for key packets within service flows to ensure timely and accurate transmission, preventing watchdog resets and system downtime, thus achieving long-term stable operation of industrial control services.

Figure 4-9 Example of intelligent industry

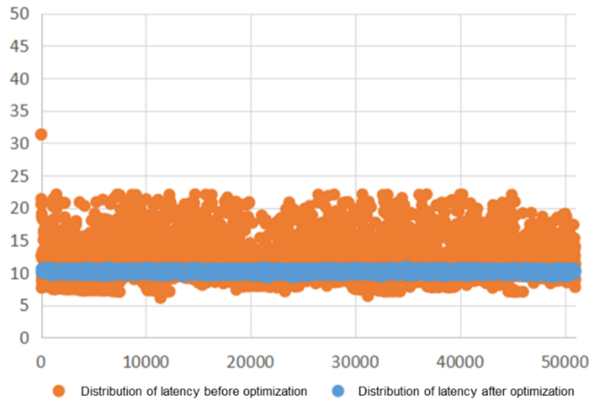


Figure 4-10 End-to end latency distribution

Currently, the deterministic assurance technology solutions and products have been validated in over 40 pilot projects at factories such as Lynk & Co Automotive, Ansteel, and Panda Electronics. Network performance: the deterministic network latency has reached 12ms@99.99%, and the deterministic latency jitter has reached 8ms@99.99%. The effectiveness of deterministic assurance has improved by more than 30%, significantly enhancing latency and jitter performance as well as the reliability of one-time successful transmission. Practical implementation: this technology is widely applied in various scenarios including full-directional wireless Programmable Logic Controller (PLC), remote control of overhead cranes, and collaborative control of AGVs. By enabling wireless industrial control, these solutions facilitate intelligent production in factories, leading to increased production efficiency, improved production quality, reduced labor costs, lower maintenance costs, and accelerating the digital and intelligent transformation of industries.

4.4 Green Energy Saving

Driven by global green development trends and sustainable development goals, the deployment of low-carbon mobile communication networks has become an industry consensus. With the rapid expansion of 5G base stations, electricity costs now constitute a significant portion of operators' overall operational expenditures, making energy conservation and efficiency improvements crucial for enhancing operational efficiency. Digital technologies play a vital role in achieving carbon neutrality and are essential in helping the world address climate change. Actively exploring innovations and applications in energy-saving technologies, and utilizing digital and intelligent methods to enhance network energy efficiency, is not only an effective way to reduce operational costs but also a critical step in promoting sustainable development within the industry.

Energy saving technologies for base stations have already achieved comprehensive multi-dimensional energy consumption management across time domains, frequency domains, power domains, and spatial domains. However, given the complexities of multi-standard and multi-frequency networking, diverse service scenarios, and varying user perception needs, traditional broad-brush energy saving strategies struggle to meet these varied demands. The key challenge lies in precisely predicting service needs, dynamically triggering energy saving mechanisms, and differentiating energy saving strategies under the complex network conditions. Achieving this precision and efficiency is essential for the successful application of energy saving.

To address the limitations of single-target base station energy saving strategies that cannot adapt to complex network structures, diverse scenarios, and service needs, big data analysis and AI algorithms are introduced. These technologies enhance the real-time performance and precision of energy saving strategies across time, frequency, spatial, and power domains, thereby expanding energy saving potential. The schemes are stated as follows.

1) Precise Identification of Energy saving Cells

By conducting joint regional analysis to identify co-covered areas, automatically configuring energy saving compensatory cells and discovering more energy saving opportunities, we can improve network energy efficiency. To achieve this, AI algorithms analyze large amounts of base station data based on differentiated characteristics of different types of base stations. This allows for automatic identification and classification of base stations. Using a data relationship model between base station energy efficiency and traffic, problematic base stations are identified, and energy saving technologies are employed to enhance overall network energy efficiency.

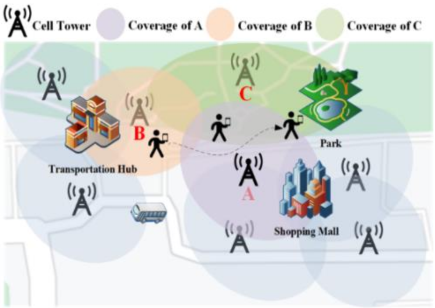


Figure 4-9 Energy-saving cell identification

2) Accurate Prediction of Energy saving Durations

Analyze historical network carrier/cell group data using regularity analysis and train optimal load prediction algorithms with time series models. Identify periods where low load conditions meet criteria for triggering carrier shutdown or deep sleep functions. Utilize AI's self-learning capabilities to iteratively adjust threshold parameters while monitoring fluctuations in performance metrics (basic KPIs, service KPIs, sensing KPIs). This helps find the best balance between energy savings and system performance by optimizing energy saving thresholds.

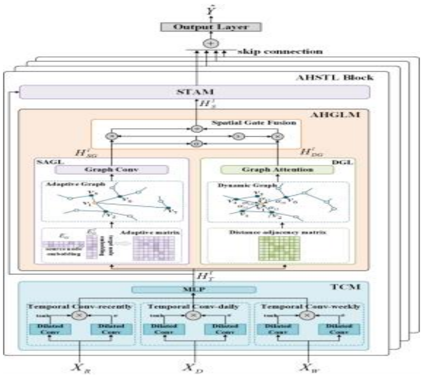


Figure 4-10 Energy-saving duration prediction model

3) Intelligent Optimization of Energy saving Strategies

For multi-scenario, multi-standard, multi-frequency, and multi-energy saving technologies, develop coordinated energy saving strategies. Collect network status data such as Measurement Reports (MR), Performance Measurement (PM), and engineering parameters. Use Gaussian

Mixture Density-Based Spatial Clustering of Applications with Noise (DBSCN) clustering algorithms to determine cell energy saving states. Train models for different cells and optimize multi-dimensional network performance indicators under various parameter configurations. Through joint self-learning within regions, continuously iterate and optimize to generate the best combination of energy saving technologies and network parameter configurations that match current coverage scenarios and service requirements. This achieves intelligent dynamic energy saving strategies tailored to “one site-one time-one strategy”, aiming for maximum energy savings.

4) Network Energy Saving Based on Service Sensing

Introduce service sensing capabilities to optimize energy saving strategies based on different service types and requirements. For example, for latency-sensitive services, develop differentiated centralized scheduling strategies to balance the impact of energy saving strategies on network performance. This enables the network to activate energy saving features without compromising service quality and extend the duration of these features. Additionally, dynamically adapt and optimize energy saving strategies based on changes in service characteristics and network monitoring indicators, ensuring network performance and meeting the needs of latency-sensitive services.

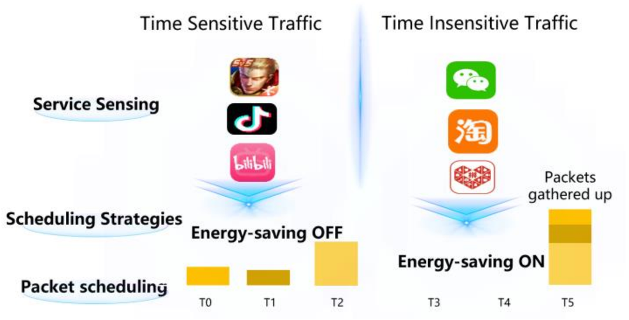


Figure 4-11 Differentiated centralized scheduling based on service sensing

The intelligent schemes such as service forecasting, energy saving cell identification, and energy saving optimization strategy have been deployed across more than 25 provinces in China Mobile’s networks, serving approximately 4.5 million base stations. Service-differentiated energy saving schemes have also been piloted nationwide. On top of conventional energy saving features, intelligent energy saving strategies can achieve an average daily energy saving gain of over 5%, ensuring a balance between energy savings and performance. This has significantly promoted the green and sustainable development of wireless networks.

4.5 High-Reliability Network

Network stability involves multiple aspects, including site-level and data center-level operations. This white paper discusses the challenges, solutions, and practical outcomes of network operation assurance from both site and data center perspectives.

1) Site Operation Stability Assurance

With the large-scale deployment and development of 5G, network structures are becoming increasingly complex, and application scenarios and service requirements are diversifying, placing higher demands on network O&M efficiency and effectiveness. Traditional site O&M methods face the following challenges:

- **Delayed Fault Response:** Relying on manual inspections and passive responses, fault detection and handling are delayed, leading to the escalation of problems.
- **Inadequate Hidden Hazard Prediction:** Lack of real-time monitoring and data analysis capabilities for equipment status makes it difficult to predict potential hidden hazards in advance.
- **Data Silos and Information Opacity:** Data related to equipment, boards, power supplies, etc., is scattered, making it difficult to integrate and share information. The lack of a unified visualization platform complicates rapid problem localization.
- **Low O&M Efficiency:** Manual inspections and fault troubleshooting are time-consuming and labor-intensive, requiring high technical expertise and heavily relying on expert experience.

The deep integration of 5G and AI, along with the adoption of big data, machine learning, digital twins, and other technologies, enables the construction of intelligent digital sites, significantly enhancing the automation level of network O&M. This not only meets the current needs of communication networks but also provides strong support for the future intelligent development of networks. The key technologies are stated as follows:

- **Building Digital Sites:** Through digital technology, equipment, components, operational status, and fault information of base stations are integrated into a unified model, achieving precise understanding of sites, improving site perception accuracy and reliability.
- **Automatic Fault Diagnosis:** AI models intelligently analyze phenomena and error data during fault occurrences, quickly infer and match issues, rapidly identify root causes, significantly reduce downtime, and improve O&M efficiency.
- **Fault Prediction and Early Warning:** Using AI algorithms to establish fault prediction models, automated analysis of site operation data, including hardware and software performance metrics, environmental factors such as temperature and humidity, and historical fault records, can accurately predict the probability of faults and issue timely warnings, effectively preventing service interruptions caused by site failures.
- **O&M Visualization:** Utilizing digital twin technologies to present complex O&M data, equipment status, fault information, and system operation conditions in an intuitive manner. Visualization tools transform abstract data into visual graphics, helping O&M personnel quickly understand system status, identify problems, and make decisions.

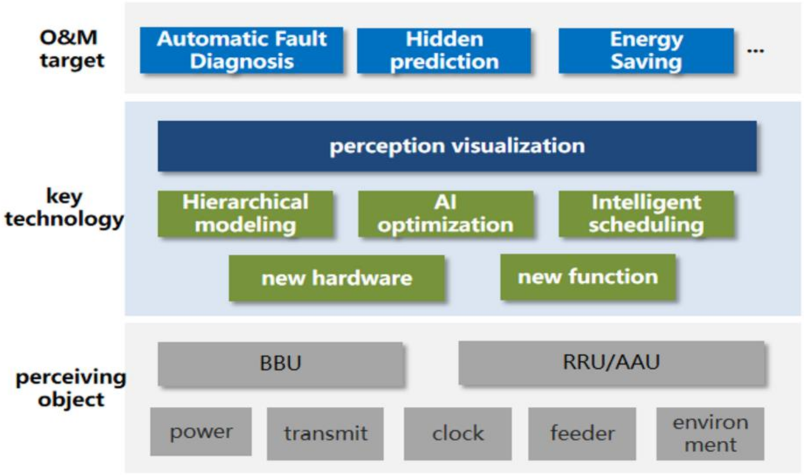


Figure 4-12 Framework of intelligent network O&M

The application has been implemented in the following scenario: digital and intelligent front-haul fiber optic links. Alarm and detection of front-haul fiber optic links are pain points in existing networks, characterized by high fault rates. Moreover, the links primarily consist of passive devices that lack monitoring capabilities, relying on manual on-site troubleshooting. There is an urgent need for intelligent and automated detection methods to improve O&M efficiency. By introducing smart optical modules equipped with reflective loop-back self-test capabilities, combined with intelligent base station fault recognition algorithms and hierarchical data processing architecture, faults can be precisely identified and accurately located automatically. This achieves intelligent detection of front-haul optical paths and significantly enhances O&M efficiency.

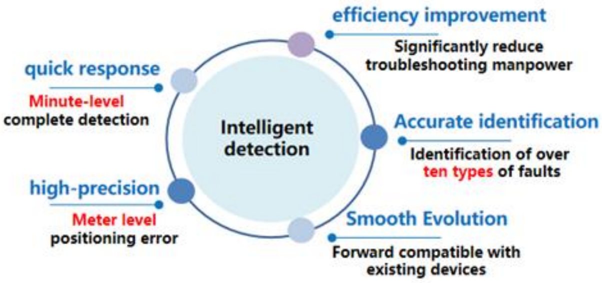

Intelligent detection
efficiency improvement
Significantly reduce troubleshooting manpower
quick response
Minute-level complete detection
Accurate identification
Identification of over ten types of faults
high-precision
Meter level positioning error
Smooth Evolution
Forward compatible with existing devices

Figure 4-13 Digitalisation and intelligence transmission

2) Data Center Intelligent Protection

When a Data Center (DC) level disaster occurs in the network, users originally residing in the failed DC will attempt to reconnect to the backup DC within a short period. This is highly likely to cause a sudden surge in network load at the backup DC, potentially leading to network overload and further expanding the impact of the service outage. Although the current network has a predictive mechanism to handle signaling surges, it complicates the situation and poses significant challenges to the actual switchover process. This is because large-scale disaster recovery switching involves the simultaneous switchover of multiple types of services, rather than a direct transition of a single service.

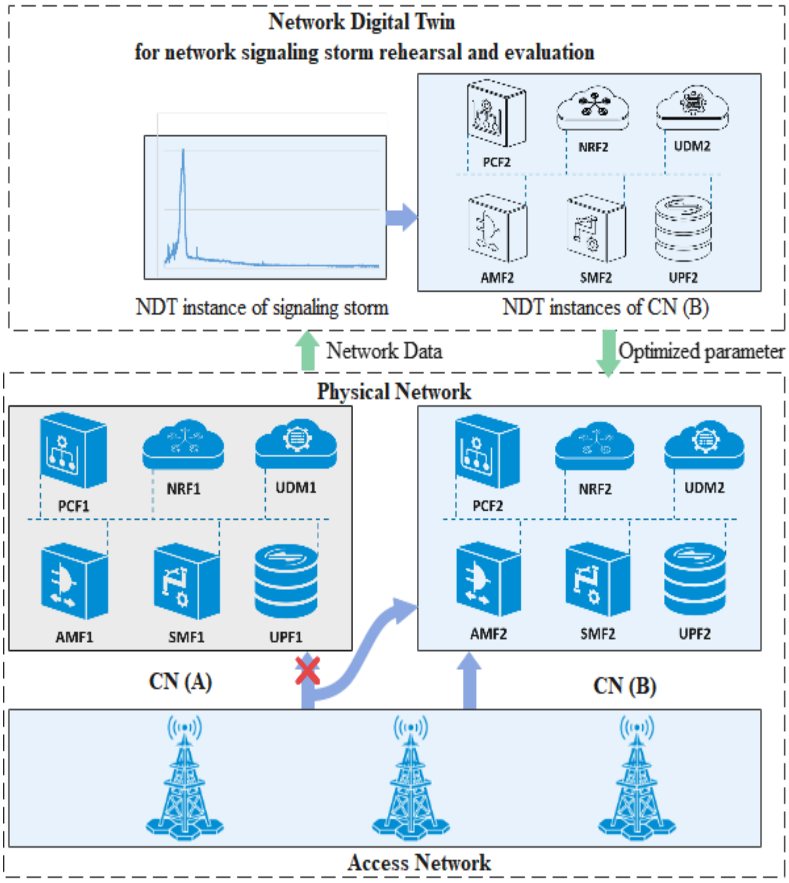


Figure 4-14 Architecture of digital twin network

To better prevent and control signaling storms, China Mobile collaborated with ZTE to conduct a signaling storm prevention and control solution verification for disaster recovery scenarios in September 2024 on the live network. This solution leveraged AI and NDT technologies to predict potential risks and bottlenecks in the network, and to enable proactive remediation and prevention measures to reduce the risks associated with DC disaster recovery switching. Additionally, the verification tested cloud-network disaster recovery capabilities under extreme scenarios, including the switchover of more than 3.4 million online users and handling 240,000 concurrent signaling messages per second.

In this solution verification, a resource load impact model was established by studying the performance components of the resource pool, the performance iterative relationships among virtual machines/physical machines/resource pools/storage pools within the resource pool, and integrating the signaling impact model of cloud Network Functions (NF). Based on the NF signaling impact model of and the resource load impact model, a network digital twin for DC disaster recovery was built to enable quantitative assessment and analysis of cloud network performance impacts. The digital twin system has already been commercially deployed in the live network system through capability exposure and interfacing with the network.

a) A digital twin system for DC disaster recovery was deployed, achieving full automation of the entire process for assessing signaling impact on cloud network elements and load impact on the network cloud. The assessment duration has been reduced to within 10 minutes.

b) A precise assessment model for network function signaling impact and network cloud load impact was established, achieving an assessment accuracy of over 95%, thereby enhancing the reliability of disaster recovery.

c) KPIs of network functions and resource pool equipment were collected in real time during the switchover process. By comparing actual metrics with simulation results, the entire switchover process became fully visible and controllable, improving emergency command efficiency and communication security while eliminating potential secondary risks from the switchover.

The key technologies primarily adopted are listed below:

- **Simulation and Assessment of Signaling Storms:** The digital twin system is utilized to simulate and evaluate the prevention and control effects of disaster recovery signaling storms. It automatically identifies network performance bottlenecks and abnormal flow control parameters, thereby enabling proactive risk prevention.
- **Synergistic Simulation for Realistic Impact Assessment:** A collaborative approach is adopted combining core network signaling storm impact simulation with resource layer traffic simulation. This ensures that the impact simulation waveform more closely resembles actual conditions, effectively enhancing the reliability of disaster recovery.
- **AI and NDT:** The assessment of core network signaling impact utilizes a combination of network digital twin models and AI technologies. This method corrects computational deviations caused by differences in terminal behavior and regional service variations, achieving model self-adaptation.
- **Visualization of Real-time Monitoring and Decision:** Leveraging the real-time monitoring capabilities of network digital twin dashboards, the switching process is made visible and controllable. This supports real-time status perception and rapid decision-making for DC switching.

4.6 Multi-Modal Personal Assistant

In recent years, the application of AI Agent technology has gradually permeated various aspects of daily life. The AI Agent is an intelligent entity capable of perceiving its environment, making decisions, and executing actions. AI assistant services represent a typical application of AI Agent technology. For example, when ordering coffee, users only need to give a command, and the system can automatically complete tasks such as selecting the type of coffee, processing payment, and arranging delivery. Behind this convenient experience lies the rapid advancement of AI Agent technology in areas like natural language processing, machine learning, and big data analysis. AI Agent technology not only understands user intentions but also makes personalized recommendations based on historical data and real-time context.

In October 2024, Zhipu Company, in collaboration with Tsinghua University, launched the AutoGLM tool. Through voice commands, users can employ it to complete tasks such as e-commerce shopping, food delivery ordering, train ticket booking, social media interaction, and sending WeChat messages. In January 2025, OpenAI released the AI Agent application Operator, which can automatically perform various complex operations, including writing code, booking travel, and automated e-commerce shopping. Beyond mobile devices, in the smart homes, AI assistant services can control home appliances; in vehicle systems, they can achieve functions like voice navigation and music playback. As the technology continues to mature, AI assistant services are rapidly integrating into all aspects of people's lives, with the market size continuously expanding and the number of users steadily increasing, showcasing a promising development trend.

Figure 4-15 Various AI assistant services

The implementation of AI assistant services mainly involves three mainstream solutions: edge AI, cloud AI, and hybrid AI. 1) The edge AI operates on local devices, where data processing and model inference are completed on the terminal itself. This approach offers fast response times and high privacy but is limited by device performance, storage capacity, and power consumption, making it less capable of handling complex tasks. 2) The cloud AI places all computations and data storage in the cloud, leveraging powerful cloud computing resources to handle complex tasks. However, it depends on network connectivity, which can introduce latency and pose privacy and security risks. 3) The hybrid AI combines the advantages of both approaches. Simple tasks are executed on the edge, while complex tasks are offloaded to the cloud, balancing performance, privacy, and network dependency. Overall, hybrid AI effectively compensates for the shortcomings of edge AI and cloud AI, making it the dominant implementation solution for current AI assistant services.

Based on the hybrid AI solution, AI assistant services first analyze data on the edge side before uploading it to the cloud for further processing. The business model shifts towards primarily uploading data, placing new demands on upload speeds. Currently, edge-side data uploaded to the cloud mainly consists of key feature fields or single images after analysis. As AI assistant services become more widespread, there will be an increasing demand for continuous image uploads from the edge side, requiring high upload speeds. For instance, transmitting high-definition images may require upload speeds of up to 20 Mbps. The high speed, low latency, and massive connectivity features of 5G-A networks ensure that edge-side data can be quickly uploaded to the cloud for inference, while also enabling timely feedback of cloud processing results to users. Low latency ensures real-time interaction, and massive connectivity capabilities meet the needs of numerous smart devices accessing simultaneously. These features of 5G-A provides strong support for the large-scale application and development of AI assistant services.

Digital humans are another typical application of AI Agent technology, with real-time and stream-based multi-modal AI interaction being key to their functionality. Empowered by multi-modal AI models, human-machine interaction systems can provide users with near-human-like interaction experiences across various domains such as lifestyle information, children's education, work assistance, and healthcare for the elderly. In complex scenarios like industrial production and smart city management, multi-modal AI will break through existing interaction paradigms, offering smarter and more comprehensive AI experiences. For example, SenseTime has launched China's first "what you see is what you get" native multi-modal large model—Riri Xin 5o+ Ruying Digital Human. This product features human-like real-time visual capabilities, enabling smooth video interactions with people. It can not only listen, speak, and see but also do so without any delay. The system leverages both edge and cloud computing resources to quickly generate highly realistic digital avatars and supports fully natural, real-time interactive conversations with users.

Multi-modal AI combines inputs from various sensory modalities such as voice, vision, gestures, and even tactile feedback, making human-machine interactions more natural and human-like. In the past, real-time processing of these large-scale data sets was challenging due to limitations in network speed and stability. The advent of 5G-A networks enables multi-modal AI to acquire, transmit, and analyze vast amounts of data in real time, rapidly generating responses. In the future, real-time mobile interactions will require dual support from edge-side models and communication networks. On the model side, inference latency should reach human-level or near-human-level response times, while the system needs to effectively integrate and transmit information from different sensory channels instantaneously. Multi-modal data uploaded from user terminals must be transmitted quickly through wireless channels to cloud-side servers, placing higher demands on upstream network transmission.

4.7 Embodied Artificial Intelligence

Robotics technology, since its development in the mid-20th century, has evolved from simple automated devices into highly intelligent and autonomous embodied intelligent robots. Embodied artificial intelligence refers to a system where a robot's physical entity interacts with its environment, enabling environmental sensing, information cognition, autonomous decision-making, and action execution. This system can also grow smarter and adapt their actions based on feedback from experiences. Traditional robots operate based on rules and fixed programs, suitable for specific scenarios with single functions and long development cycles, but they lack adaptability. In contrast, embodied artificial intelligence leverages AI large models, offering versatility across tasks and environments, with capabilities to understand and influence the world actively, along with continuous learning and iteration.

The application scenarios of embodied artificial intelligence span multiple domains from daily life to industrial production, showcasing significant market potential and technical advantages. In household environments, embodied intelligent robots can handle household chores, accompany the elderly and children, greatly enhancing convenience. In healthcare, these robots assist doctors with surgical operations, patient care, and rehabilitation training. In manufacturing, they perform complex assembly, inspection, and logistics tasks, significantly improving automation levels and product quality on production lines. In agriculture, embodied intelligent robots autonomously monitor crops, apply pesticides, and harvest fruits. In logistics, warehouse robots efficiently sort and transport goods. In security, patrol robots monitor environments in real-time and warn of potential risks. For example, Leju Robotics has collaborated with multiple automobile factories to deploy multi-purpose embodied intelligent robots on production lines. Quadruped robots, as one form of embodied artificial intelligence, have taken on numerous tasks in harsh environments. For instance, Unitree's robotic dogs are used for garbage collection on Mount Tai.

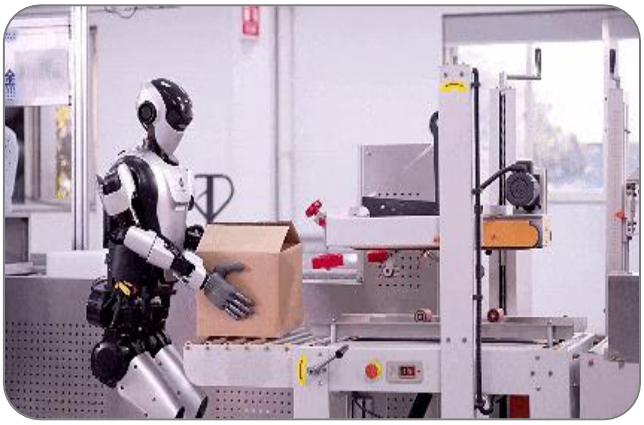


Figure 4-16 Example of a warehouse robot sorting goods

The widespread application of embodied artificial intelligence places higher demands on communication technologies, particularly highlighting the need for 5G-A communication. Leveraging the large bandwidth and low latency characteristics of 5G-A networks, combined with AI capabilities, can enhance the precision positioning, low latency, real-time control, data collection, and incremental training of embodied intelligent systems through cloud collaboration. This reduces the computational power requirements and energy consumption of embodied intelligent systems themselves. For example, in applications requiring high real-time performance and precise control, such as robot parkour, the low latency feature of 5G-A networks ensures that robots can quickly respond to commands, achieving smooth action execution. The large bandwidth feature of 5G-A networks ensures that large volumes of sensor data from robots can be transmitted and processed in real time, further enhancing the perception and decision-making capabilities of embodied intelligent robots.

Embodied artificial intelligence has expanded from indoor, small-scale applications to outdoor, long-distance scenarios, which demand extensive signal coverage and mobility. 5G-A communication fully meets these needs. For instance, Huawei has collaborated with Leju Robotics, and ZTE has partnered with Unitree Technology, to integrate 5G modules into embodied intelligent robots. These collaborations not only enhance the communication capabilities of the robots but also provide technical support for their application in complex and dynamic outdoor environments. Through 5G-A networks, embodied intelligent robots can achieve remote control, real-time data transmission, and multi-robot collaborative operations, significantly improving work efficiency and reliability.

4.8 Immersive Experience

Immersive real-time communication integrates various advanced technologies such as Virtual Reality (VR), Augmented Reality (AR), Mixed Reality (MR), and naked-eye 3D, breaking the limitations of traditional display technologies and offering users a new, highly immersive sensory experience. XR and naked-eye 3D, as high-definition immersive services, serve as the initial core carriers of the metaverse and represent one of the primary future directions for 5G services. Their characteristics of ultra-high definition, high interactivity, and strong computing demand pose challenges for 5G networks in handling such services.

Typical immersive service applications feature ultra-high definition and strong interactivity, requiring 5G networks to enhance service support from single-dimensional to multi-dimensional improvements. There are three major pain points: 1) Inadequate Utilization of Network Capabilities: Network utilization rates are low, and service traffic growth does not meet expectations. 2) Mismatched Coordination Mechanisms: Inefficient transmission mechanisms lead to poor service experiences and low transmission efficiency. 3) Insufficiently Refined Network Feature Assurance: The network's ability to support users and services is not fine-grained, resulting in high complexity in multi-objective optimization and low resource utilization efficiency.

The end-to-end solution for immersive real-time services is a comprehensive approach designed for highly immersive, strongly interactive, and multi-dimensional multi-modal communication services. It aims to enhance user experience, ensuring real-time, stable, and high-quality service delivery during transmission.

China Mobile has introduced a core philosophy centered on "information cross-layer sharing, multi-dimensional connection enhancement, and quantified perception assessment." Based on this philosophy, they have proposed several landmark technical solutions, including millisecond-level service perception assurance, frame-level multi-dimensional connection enhancement, real-time network information openness, and user experience quantification evaluation. These innovations aim to build an original network-service collaboration technology system tailored for XR immersive experiences.

1) Addressing Real-Time Network-Service Interaction

On one hand, for service sensing based on network status, millisecond-level service sensing assurance technology is proposed: To address the inadequacy of service sensing demands, a real-time fine-grained sensing scheme for service frame-level characteristics is proposed, achieving a leap from non-real-time to millisecond-level real-time service quality assurance. Frame-Level QoS Optimization: By perceiving characteristics such as frame cycle and frame importance, optimize scheduling based on frame-level QoS requirements; Frame Identification and Protection: Base stations identify incoming packet patterns to recognize frame cycles, frame tails, frame sizes, and other service features, ensuring frame-level service protection;   Cross-Layer Information Sharing: Critical service information is passed from the application layer to the UPF via RTP extension headers, then forwarded to base stations through GTP-U header fields, facilitating seamless information transfer from the application layer to the network layer.

On the other hand, for networks states sensing based on services, real-time network information openness technology is introduced: To tackle insufficient sensing of network states by services, user plane network information openness technology is proposed, enabling real-time sharing of congestion, throughput, and other network information with the application layer. This addresses shortcomings in previous network capability openness schemes related to comprehensiveness and node diversity, thereby enhancing network capacity and user experience.

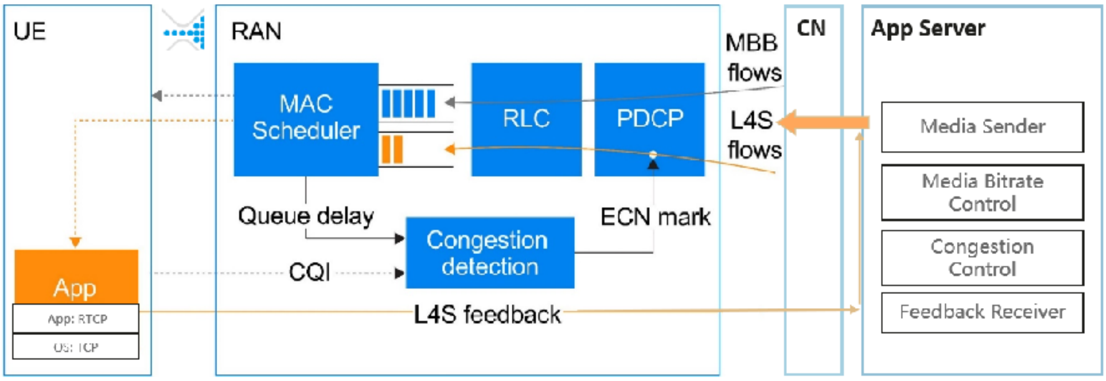


Figure 4-17 Example of millisecond-level service awareness assurance technology

2) Maximizing Connection Capabilities

Proposing frame-level multi-dimensional connection enhancement techniques:

- Guidance with Frame Delay and Error Rate: Define parameters like frame delay and frame error rate to guide fine-grained frame-level protection;
- Optimized Frame Scheduling and Integrity: Implement frame delay scheduling and frame integrity transmission to optimize frame-level QoS metrics.
- Multi-User Staggered Transmission: Maximize system capacity by staggering transmissions among multiple users.
- Switch Control and MCS Selection Optimization: Optimize handover latency using frame boundary-based switch control and Modulation and Coding Scheme (MCS) selection, significantly improving transmission performance and user experience in mobile scenarios for immersive real-time communication services.

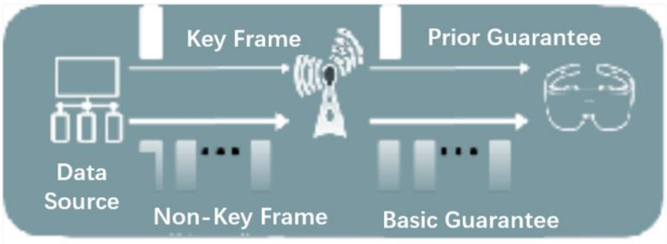


Figure 4-18 Example of multi-stream differentiated transmission

3) Standardizing Experience Assessment

For standardized evaluation of XR service experiences, a frame-level network indicator system, testing methods, and equipment requirements for XR service experiences are proposed. frame-level network Key Performance Indicators (KPIs) such as transmission delay, air interface transmission delay, transmission rate, and reliability are defined to quantitatively evaluate XR service experiences.

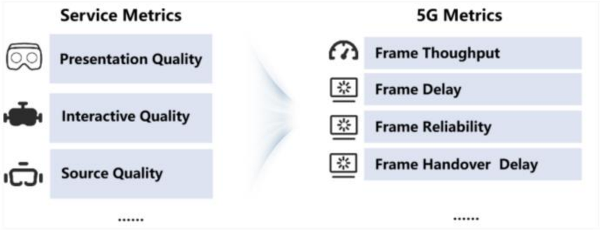


Figure 4-19 5G network metrics system for XR immersive experience

Application Case 1: Beijing Workers' Stadium "Gongti Metaverse"

For the sports sector, China Mobile has partnered with Zhonghe Group to deploy the world's first metaverse application based on a 5G live network at the new Beijing Workers' Stadium. This innovative application integrates network-service integration, cross-layer information sharing, base station service perception, and intelligent data multi-stream transmission technologies. It achieves real-time experience speeds of up to 100 Mbps and frame-level network transmission latency below 20 ms. These capabilities meet the high traffic, strong real-time interaction, and high customer mobility requirements of the Gongti Metaverse application. On-site viewers can enjoy an immersive HD XR football experience at 4K resolution and 90 fps. This is the first deployment of a metaverse application on a 5G live network, setting a demonstration for accelerating large-scale metaverse applications. It showcases how technological innovation drives business innovation and vice versa.

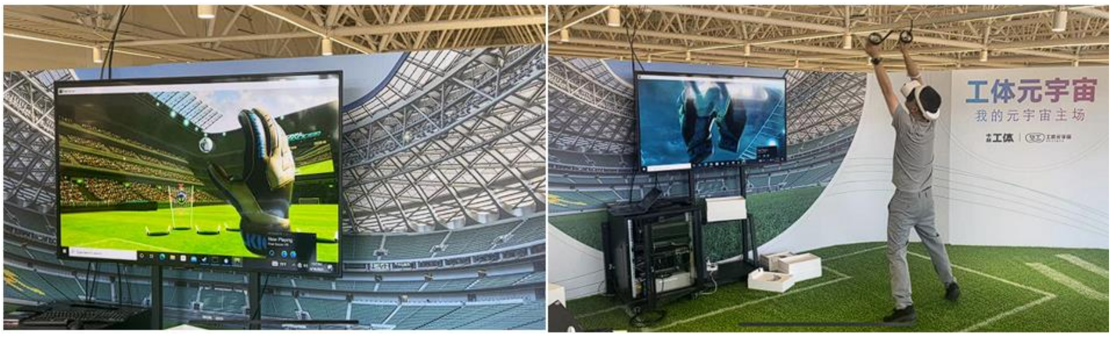


Figure 4-20 Beijing Workers' Stadium "Gongti Metaverse"

Application Case 2: Shougang No.1 Blast Furnace SoReal Sci-Fi Park

In the entertainment sector, China Mobile has collaborated with Danghong Qitian to create the industry's first 5G-A indoor large-space ultra-dense multi-user XR pilot application. For indoor ultra-dense immersive XR scenarios, challenges such as heavy backpack devices, poor heat dissipation, short battery life, and performance demands like high capacity, low latency, and mobility for dozens of concurrent users were addressed. Technologies like 5G-A intelligent service perception, frame-level multi-dimensional connection enhancement, joint intelligent beam management, and a "no backpack" and "cable-free" distributed rendering solution were employed. Leveraging the large bandwidth and low latency provided by the 5G-A network, it ensures real-time 100 Mbps experience rates and average air interface transmission latencies below 10 ms for dozens of users. The service image clarity reaches industry-standard 4K @90 fps quality. In a 1,000 square meter space, users can experience a boundaryless virtual world more freely.

Figure 4-21 Shougang No.1 Blast Furnace SoReal Sci-Fi Park

Application Case 3: New Audio-Visual Experiences for Asian Games Viewing

Based on our company's Migu business platform and the advantages of the 5G-A network, we have created new services such as VR esports, VR live streaming of events, and naked-eye 3D viewing of the Asian Games. Compared to conventional video services, XR and naked-eye 3D technologies allow users to enjoy realistic three-dimensional visual effects, freely switch viewing angles, and experience high-definition immersive audio-visual content. This provides a rich, smooth, and lifelike 3D immersive viewing experience of the Asian Games. The Migu mobile cloud VR business platform customizes XR viewing scenarios for the Asian Games, enabling multiple matches to be viewed on one screen through XR technology. This redefines the "on-site" experience for audiences, allowing them to feel the "immersive, exciting, and interactive" Asian Games metaverse. With the large bandwidth and low-latency assurance capabilities of the 5G-A network, it achieves wireless transmission latency below 20 ms and 125 Mbps frame-level guaranteed speed in multi-user and multi-service concurrency scenarios, ensuring smooth experiences for various types of high-definition interactive 3D cloud services.

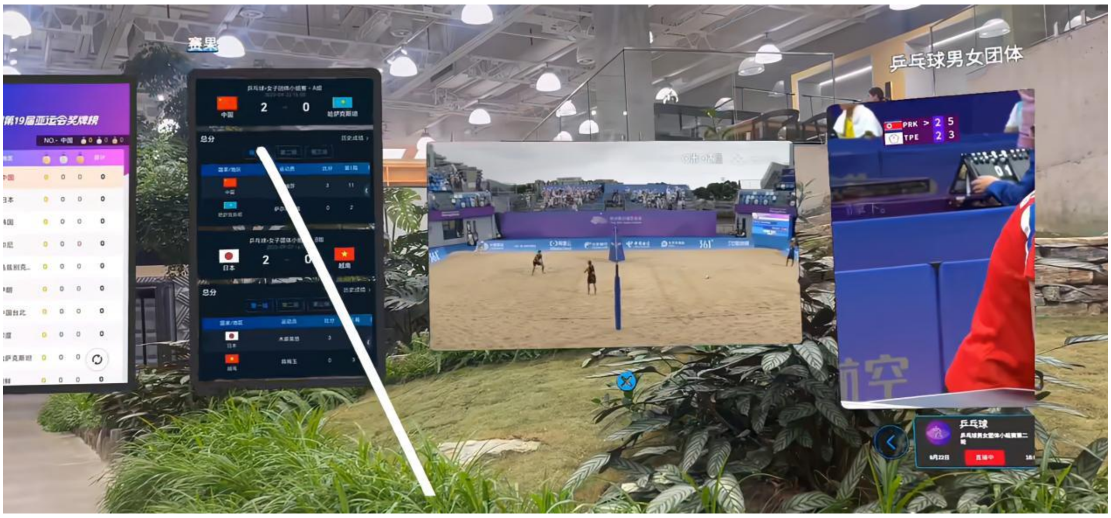


Figure 4-22 New Audio-Visual Experiences for Asian Games Viewing

4.9 Networked Smart Driving

In recent years, vehicles have become increasingly intelligent, transforming cars from mere modes of transportation into "mobile third spaces" for people's travel. 5G has become a must-have for new automotive manufacturers and the preferred choice for premium vehicles, offering enhanced cabin and smart driving experiences. This allows cars to serve as rest areas, offices, and gaming rooms for multiple passengers simultaneously. Connected vehicles feature three major types of applications: intelligent cockpits, vehicle-cloud collaboration, and vehicle-road collaboration. The large multi-screen displays and real-time interaction enhancements in intelligent cockpits, cloud-assisted smart driving in vehicle-cloud collaboration, and road condition sensing in vehicle-road collaboration all demand high-rate and low-latency capabilities from 5G-A networks.

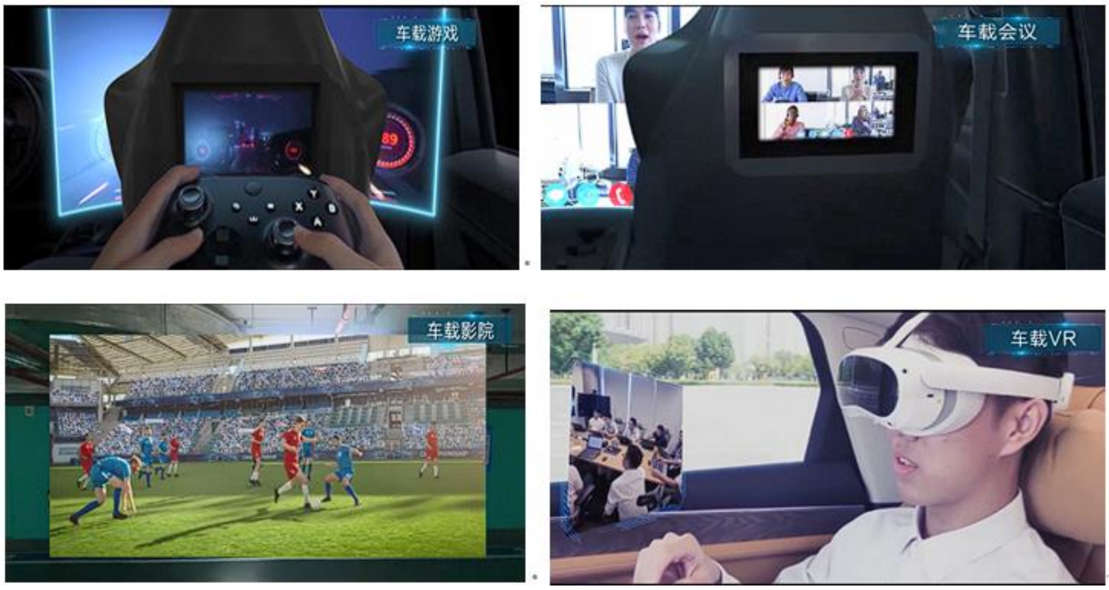


Figure 4-23 Diverse in-vehicle services in intelligent cockpits.

As vehicle intelligence continues to advance, driving methods are gradually evolving from assisted driving to autonomous driving. The five-level standard for automotive autonomous driving, ranging from L0 to L4, is outlined below. Depending on different development strategy paths, the commercial pathways for autonomous driving can primarily be divided into two categories: Gradual Evolution from L2 (Partial Automation): This approach involves progressively enhancing capabilities starting from L2, such as that pursued by manufacturers like Tesla, Li Auto, and NIO; Direct Entry into L4 (Full Automation): This approach involves directly targeting L4 full automation, as seen with autonomous driving solution providers like Waymo, WeRide, Pony.ai, and Baidu's Apollo.

Autonomous driving classification standards

Autonomous driving classification	definition
L0：Fully human control	The driver is solely responsible for operating the vehicle under all road and environmental conditions.

L1：Assisted driving	The driver and the vehicle share control. Under specific road and environmental conditions, the vehicle is equipped with one or more specialized automated control functions.
L2：Partial Automation	The vehicle is primarily in control. Under specific scenarios, the vehicle can perform two or more automated driving functions. However, the driver must remain aware of the surroundings and be ready to intervene if necessary.
L3：Conditional Automation	The vehicle is in full control under specific road and environmental conditions, capable of autonomous driving without requiring the driver to monitor the driving process. However, human intervention is still required in emergency situations.
L4：Full Automation	The vehicle is fully in control and capable of autonomous driving under all road and environmental conditions without human intervention.

In 2024, more than 50% of the vehicles sold in China are equipped with L2+ intelligent driving features.

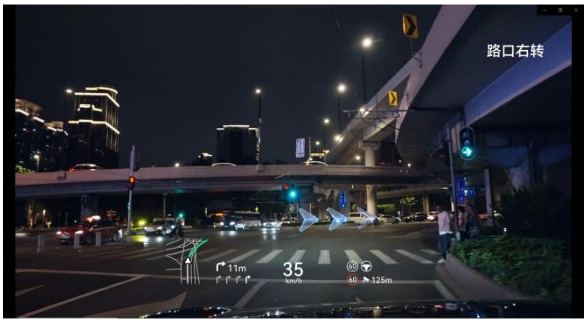


Figure 4-24 L2+ intelligent driving

L4 autonomous driving has also matured, achieving commercialization in various fields such as driverless taxis and unmanned delivery. As of July 2024, the commercial orders for Robotaxi operated by Luobo Fast Run have exceeded 6 million. The 5G-A network, based on key features like low latency and high uplink capacity, can ensure an uplink speed of 20 Mbps and a downlink latency of 20 ms. This supports real-time monitoring of autonomous vehicles, immediate remote control in abnormal scenarios, and real-time uploading of anomaly data to the cloud. These capabilities ensure the continuous operation of vehicles and empower the efficient running of L4 autonomous driving vehicles.

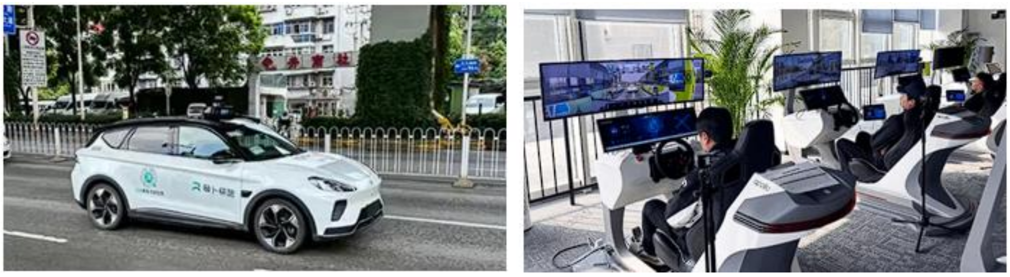


Figure 4-25 Robotaxi application

Smart roads and cloud-based large models will complement individual vehicle intelligence, achieving the ultimate goal of "vehicle-road-cloud integration." Through the stable low-latency 5G-A network, the massive computing power of cloud-based large models will augment the limited computing resources on vehicles. The real-time sensing capabilities of smart roads will also fill in the blind spots of beyond-line-of-sight scenarios for individual vehicles, further enhancing travel safety and efficiency. For example, vehicles can upload real-time data about complex traffic signs or tidal lanes that cannot be parsed locally. Cloud-based AI large models will then parse and recognize these signs and immediately send the results back to the vehicle. This process reduces traffic accidents, allows for real-time selection of the optimal route, and enhances the efficiency of intelligent driving.

4.10 Low-Altitude Intelligent Connectivity

This white paper describes the solutions for communication services and sensing services in low-altitude scenarios using advanced information and communication technologies, artificial intelligence, big data, and other technological means to build a low-altitude intelligent network.

Communication: In low-altitude flying environments, terminal devices such as UAV face challenges like unclear cell coverage boundaries, lack of dominant coverage cells, numerous and disordered neighboring cells, and frequent handovers. These issues can lead to frequent cell switching during drone flights, affecting the continuity and stability of communication, potentially resulting in service interruptions, data loss, and poor communication experiences for low-altitude terminals.

To reduce the number of handovers, improve signal quality, and enhance user experience, this white paper proposes an AI-based continuous flight assurance solution for low altitudes. By leveraging the predictive and decision-making capabilities of AI, considering factors such as the positions of serving and neighboring base stations, current RSRP measurements, drone speed timestamps, etc., intelligent cell switching along the flight path is achieved. Resources are pre-planned based on handover strategies, allowing target cells to optimize resources in advance, ensuring continuous flight assurance for UAV terminals.

Sensing: Integrated communication and sensing technology can serve the needs of low-altitude drone detection and sensing by utilizing network-side data to provide target sensing services. This can be divided into two major scenarios: airspace management and route protection. Airspace management is primarily used for monitoring unauthorized drones, while route protection focuses on protecting fixed routes for logistics or drone companies to avoid collisions. Traditional low-speed radar detection technologies have limitations such as difficulty in urban deployment, complex continuous networking, and high deployment costs. Integrated communication and sensing technology, by adding sensing functions to traditional base stations, achieve integrated communication and sensing, offering capabilities such as position and speed detection, trajectory tracking, and target recognition. It has advantages like contiguous networking and cost-effectiveness.

China Mobile, addressing the broad-spectrum sensing needs of low-altitude and aquatic areas, has pioneered medium-to-low frequency integrated communication and sensing technology and millimeter-wave and sub-6GHz high-low frequency coordination technology solutions. To address issues like gaps in sensing architecture and air interface design, insufficient coverage distance and height, low precision, and complex networking and evaluation, they propose innovative technologies and solutions including agile deployment architectures, hybrid waveform new air interfaces, super-angle hardware, high-precision sensing functions, integrated communication and sensing networking, and new sensing evaluation models. As a new paradigm in the evolution of mobile communication networks, integrated communication and sensing possess the following technical capabilities:

- Integrated Communication and Sensing Air Interface Design: By sharing spectrum resources, integrated air interface designs are conducted, achieving bidirectional integration of communication data and sensing information. Key designs such as "OFDM+LFM hybrid wave" sensing waveforms and integrated communication-sensing frame structures enable shared resource utilization for both communication and sensing performance.
- Integrated Communication and Sensing Device Capabilities: Communication and sensing share antennas, RF modules, some signal processing equipment, and other hardware resources, effectively increasing device integration and work efficiency. Based on this, software-coordinated optimization of communication and sensing functions is achieved.
- Data and Service Openness Capabilities: Sensing data can not only be directly presented on sensing platforms but also integrated with other systems according to customer needs, enabling multi-source data AI fusion processing. The sensing results are displayed in various forms to different customers, providing customized service experiences.

Application Case 1: Low-altitude Security in Zhejiang

In April 2024, low-altitude security applications were launched in Hangzhou, Zhejiang Province. Collaborating with Zhejiang Mobile, a low-altitude security demonstration area was created around the Olympic Sports Center in a complex urban environment. This initiative achieved precise tracking and identification of drones within the core area (~1 square kilometer), assisting public security in real-time monitoring and reporting and countermeasure systems. Additionally, it provided drone sensing security and timely detection and alarms for unauthorized flights in an area of nearly 5 square kilometers. A single sector can cover 1.2 km, enabling precise detection of drones with an RCS of 0.01 square meters at altitudes below 300 meters.

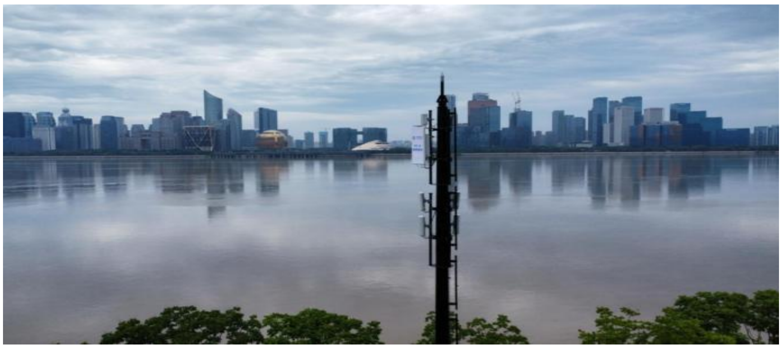


Figure 4-26 Low-altitude security application case

Application Case 2: Marine Wildlife Protection in Fujian

In 2024, applications for marine wildlife protection are being implemented in Xiamen, Fujian Province. Using integrated communication and sensing technology, continuous beyond-visual-range monitoring of the Baiji dolphin's core protection zone is achieved, along with vessel trajectory tracking and early warning functions in designated sea areas. A cloud platform base, acoustic and visual data center for marine life, and customized large model services have been established to create a smart protected area spatiotemporal foundation.

Figure 4-27 Marine wildlife protection application case

Application Case 3: Vessel Monitoring in Jiangsu Yangtze River Waterway

In July 2024, vessel monitoring applications were initiated in Nanjing, Jiangsu Province. Collaborating with the Jiangsu Maritime Bureau, three 4.9G integrated communication and sensing base stations were deployed along the Yangtze River section in Nanjing, providing continuous coverage from Zhongshan Dock to the Nanjing Yangtze River Bridge and Bagua Island. This enabled precise positioning and comprehensive monitoring of illegal vessels, unregistered boats, and fishing vessels within the waterway, effectively protecting water economy and enhancing water security.

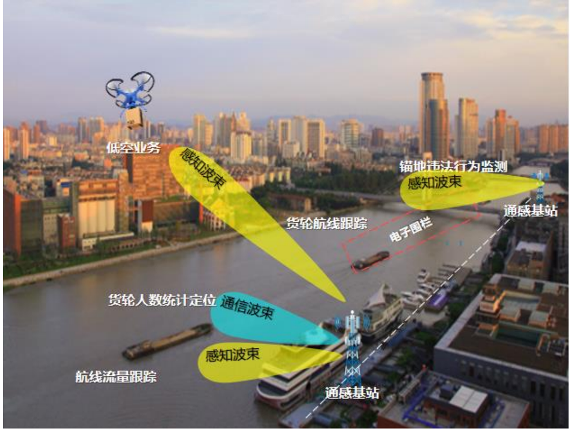


Figure 4-28 Vessel monitoring in Jiangsu Yangtze river waterway application case

Application Case 4: Airport Low-altitude Security in Yunnan

In July 2024, airport low-altitude security applications were launched in Baoshan City, Yunnan Province. A large-scale low-altitude surveillance network and system platform were established around the airport, implementing functions such as drone bird dispersal, drone inspections, and clear airspace patrols. Effective monitoring and early warning for low-altitude drones, runway personnel and vehicles, and park safety were realized. Through the application of integrated communication and sensing technology, the airport's low-altitude security capabilities were comprehensively upgraded, creating a smart service brand experience demonstration area. Yunnan Mobile signed the first 5G-A commercial cooperation agreement with Baoshan Airport, exploring the first commercial cooperation model, and rapidly replicating it in other airports like Wenshan, Lijiang, and Dali.

Figure 4-29 Airport low-altitude security application case

Application Case 5: Bridge Micro-deformation Monitoring in Jiangsu

In 2024, bridge micro-deformation applications are underway in Nanjing, Jiangsu Province. Collaborating with Sujiaoke, the global first feasibility verification pilot for millimeter-level micro-deformation monitoring on the Qixia Bridge was completed. Under light wind and no rain conditions, deformation monitoring accuracy reaches the millimeter level, allowing continuous dynamic monitoring of bridge structural conditions. The pilot results have been recognized by the provincial transportation department. Future research will focus on improving monitoring capabilities under adverse weather conditions to provide efficient lightweight solutions for bridge health monitoring and enhance infrastructure safety management.

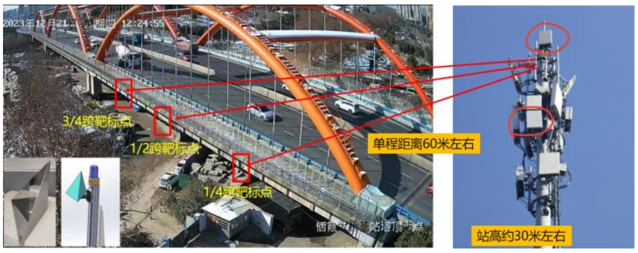


Figure 4-30 Bridge micro-deformation monitoring application case

Application Case 6: Low-altitude Air Routes in Shanghai

In May 2024, low-altitude air route applications were launched in Jinshan District, Shanghai. Collaborating with Zhejiang Mobile, Shanghai Mobile leveraged multi-frequency coordination advantages to optimize the network innovatively for low-altitude scenarios. This resulted in global 5G-A low-altitude network coverage over a 100-kilometer cross-sea air route, including full coverage of 5G signals in coastal, near-sea, and far-sea regions. It provided robust network support for low-altitude fresh produce transport routes spanning 100 kilometers horizontally and 300 meters vertically, significantly enhancing low-altitude network levels and boosting low-altitude economic development.

Figure 4-31 Low-altitude air routes application case

5 Industry Revolution of New Models

The integration of 5G-A and AI can be applied to various scenarios for individuals and industries, providing users with new capability services, new application services, and new terminal services. Based on these innovative services, telecommunications operators can collaborate with the upstream and downstream of the 5G industrial chain to jointly research and explore innovative business models, so as to fully unleash the new value of 5G-A.

5.1 Three Innovative Business Models for ToC/BtoC

For individual customers, on the one hand, in addition to charging for data traffic, telecommunications operators can, based on new network and AI capabilities, charge individual customers for function fees, service fees, network differentiation package fees, etc. On the other hand, application service providers or terminal manufacturers can cooperate with telecommunications operators to provide network guarantee services for their high-end applications/terminals to obtain additional revenue.

5.1.1 Capability Services

5G new calls can provide individual customers with intelligent functions such as real-time translation and interesting calls. In terms of Internet access, telecommunications operators can provide differentiated uplink and downlink rates, access priorities, and QoS guarantees. In terms of calls, when individual customers use the AI functions of 5G new calls, telecommunications operators can charge function fees, service fees, etc. In terms of Internet access, telecommunications operators provide different levels of network capabilities according to different package levels, thus realizing the monetization of network capabilities.

Example: Differentiated Network Capability Packages for Individual Customers

A certain telecommunications operator launched two tiers of 5G packages. The low-tier package has a data traffic of 50GB, with a maximum downlink rate of 500Mbps and a maximum uplink rate of 50Mbps; the high-tier package has a data traffic of 150GB, with a maximum downlink rate of 3Gbps and a maximum uplink rate of 200Mbps, and the high-tier package can also provide priority access, QoS guarantee and other services for applications such as games, videos, and instant messaging.

5.1.2 Application Integration Services

The 5G-A network can provide different levels of network guarantee services based on business types, locations, times, etc. Application service providers such as video, live streaming, cloud gaming, and AR/VR cooperate with telecommunications operators to design application/member content that includes 5G-A network rights. Telecommunications operators provide network guarantee services for customers who subscribe to these applications/members. Application service providers can obtain higher revenue based on differentiated network services, and telecommunications operators can also charge application revenue sharing from application service providers according to the number of subscriptions to applications/members and network usage.

Example: Membership Charging of a Cloud Gaming Service Provider

A certain cloud gaming service provider launched two tiers of paid membership services, namely VIP membership and Super VIP membership. For Super VIP members, the cloud gaming service provider cooperates with the telecommunications operator to provide them with ultra-high-definition images and high-level 5G-A network guarantee services. The Super VIP membership fee is 20 yuan more per month than the VIP membership fee. For each newly developed Super VIP member, the telecommunications operator charges the cloud gaming service provider 10 yuan per month as the network service fee.

5.1.3 Terminal Integration Services

The 5G-A network can provide different levels of network guarantee services based on terminal types. Terminal manufacturers such as drones, smart cars, and AI computers cooperate with telecommunications operators to provide 5G-A network guarantee services for their high-end or designated terminals. Terminal manufacturers can charge individual users more value-added service fees through network services, and telecommunications operators can charge terminal sales revenue sharing from terminal manufacturers according to the sales volume of guaranteed terminals and network usage.

Example: Charging for High-End Smart Drones

A certain drone manufacturer provides two tiers of products for live streaming drones. The high-end live streaming drone product has a built-in SIM card, which can provide remote ultra-high-definition and low-latency real-time video transmission services. When customers purchase high-end live streaming drones, in addition to the one-time purchase cost, they also need to pay the terminal manufacturer 600 yuan per year as the real-time video transmission service fee, and the telecommunications operator charges the terminal manufacturer 300 yuan per year as the network usage fee.

5.2 Three Innovative Business Models for ToB

For industrial customers, in addition to personalized and customized services such as 5G+DICT solutions, it is the future development trend for telecom operators to provide lightweight, standardized and closely coordinated product services based on network capabilities and AI applications. These services include capability invocation services with the capability platform as the main carrier, task-based services based on cloud services, and business collaboration services that coordinate network capabilities with customers' businesses.

5.2.1 Capability Invocation Services

The 5G-A network can provide customers with capability invocation services such as slicing and edge computing through API invocation. Solution providers, system integrators, industry customers, etc. can independently build personalized network solutions based on the operator's network capability open platform to meet customers' network requirements for bandwidth, latency, reliability, etc. Moreover, the network capabilities can be invoked immediately when needed, which has the characteristics of a short deployment cycle and quick business launch. Telecommunications operators can charge according to the type of invoked capabilities, the number of invocations, etc.

Example: 5G Network API Service Proposed by Ooredoo

Qatari telecommunications operator Ooredoo has proposed four types of API invocation services based on the 5G network, including SIM Swap, Device Status, Device Location, and Quality on Demand. For example, when a bank application wants to check whether the current user is in a specific area, it can use the API to obtain the location information of that mobile phone number.

5.2.2 Task-Based Services

Telecommunications operators integrate computing power, network, AI, big data and other capabilities and resources, and provide users with relatively standardized task-based services (such as edge cloud rendering, live-streaming digital human,face recognition, etc.) based on cloud services. When customers use task-based services, they do not need to pay attention to complex underlying business logic, and can build their own digital and intelligent applications, effectively reducing the application threshold. Task-based services are paid according to the number of task invocations, duration, or data volume, and are suitable for application scenarios that do not require too much personalized development.

Example: Mobile Cloud Platform Provides 5G Task-Based Services

China Mobile's Mobile Cloud platform provides a series of cloud products such as elastic computing, edge cloud, video services, and artificial intelligence, including task-based services such as video live streaming, cloud rendering, speech recognition, and image recognition. For example, for speech recognition services, Mobile Cloud provides standard API interfaces and commonly used types of SDKs, and with access guidance, enterprises can flexibly customize and conveniently invoke according to their business needs.

5.2.3 Business Collaboration Services

Based on the close coupling of 5G network capabilities and customers' businesses, telecom operators and customers complement each other with their respective advantageous resources. Telecom operators mainly focus on the construction of the 5G-A network, and jointly build intelligent applications with customers. They may even cooperate in business operation to ensure the effectiveness of customers' business operation while sharing the revenues. Telecom operators charge the cost for the network construction part, and at the same time, they also obtain revenues according to the actual operation effects of customers' businesses.

Example: Business Collaboration in Stadium VR Service

A certain stadium enhances the value for its members and increases the ticket prices of sports events through the VR service. The telecom operator provides it with a high-quality 5G-A network featuring large bandwidth and low latency, and they jointly build a VR application to improve users' VR experience. The telecom operator first charges part of the network construction cost, and then obtains a share of the revenue according to the income situation of the VR service.

6 Global Industry Collaboration Proposal

The integration of 5G-A and AI is progressively demonstrating its transformative potential. Looking ahead to the future, the continuous evolution of 5G-AxAI presents exciting opportunities for innovation. The development of AI agents and large models in 5G-A network infrastructure is expected to make AI applications more agile and efficient. This, in turn, will enhance the performance and reliability of 5G-A networks. The combination of high-speed connectivity, real-time data processing, and intelligent decision-making is unlocking new potential across multiple industries. To accelerate this process, GTI proposes three cooperation initiatives.

Joint Research on Advanced Technology. GTI suggests establishing open laboratories to complete the optimization cycle from early-stage technology innovation to business validation. This initiative aims to enhance technical reserves and foster industry development by focusing on cutting-edge research.

Sharing of Innovative Resources. GTI proposes the construction of an open and collaborative innovation community. This community will optimize the investment in innovation and maximize the efficiency of resource utilization. By sharing resources, we can drive down costs, increase efficiency, and accelerate the development of new technologies and applications, thereby empowering industries to upgrade and stay competitive in the rapidly evolving digital landscape.

Co-Creation of Application Ecosystems. GTI encourages collaborative exploration of high-impact use cases and refinement of scalable business models to empower industrial upgrading. By focusing on practical use cases and aligning technical innovation with market demands, this initiative seeks to identify the most promising areas for the application of 5G-A and AI technologies.

This collaborative framework is designed to harness the synergies between 5G-A and AI, ensuring that technological advancements translate into tangible benefits across various sectors and contribute to a more connected and intelligent future.

7 Glossary

5G-A	5G-Advanced

5QI	5G QoS Identifier
ADC	Application Data Channel
AI	Artificial Intelligence
AIGC	Artificial Intelligence Generated Content
AGI	Artificial General Intelligence
AGV	Automated Guided Vehicle
AR	Augmented Reality
ASR	Automatic Speech Recognition
BBU	Base-band Unit
BDC	Bootstrap Data Channel
CC	Component Carriers
CNN	Convolution Neural Network
CV	Computer Vision
DBSCN	Density-Based Spatial Clustering of Applications with Noise
DC	Data Center
DCSF	Data Channel Signaling Function
DL	Deep Learning
DNN	Deep Neural Networks
DTMF	Dual Tone Multi Frequency
DTN	Digital Twin Network
DPD	Digital Pre-Distortion
ELAA	Extremely Large Antenna Array
eMBB	Enhanced Mobile Broadband
XR	Extended Reality
FNN	Feedforward Neural Network
FTTR	Fiber to The Room
FSA	Flexible Spectrum Access
HOF	Handover Failures
GAN	Generative Adversarial Network
GBR	Guaranteed Bit Rate
IMS	IP Multimedia Subsystem
IMS DC	IP Multimedia Subsystem Data Channel
IoT	Internet of Things
IOV	Internet of Vehicles
LLM	Large Language Model
MAML	Meta Action Markup Language
MF	Media Function
MIMO	Multiple Input Multiple Output
MLP	Multilayer Perceptron
MOS	Mean Opinion Score
MR	Measurement Report
MRI	Magnetic Resonance Imaging

NDT	Network Digital Twin
NF	Network Function
NFV	Network Function Virtualization
NLP	Natural Language Processing
NWDAF	Network Data Analytics Function
O&M	Operations&Maintenance
PA	Power Amplifier
PCF	Policy Control Function
PEFT	Parameter-Efficient Fine-Tuning
PLC	Programmable Logic Controller
PM	Performance Measurement
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RLF	Radio Link Failures
RNN	Recurrent Neural Network
RRM	Radio Resource Management
SA	Service Awareness
SDN	Software-Defined Networking)
SLAs	Service Level Agreements
SUL	Supplementary Uplink
TTS	Text-to-Speech
UCBC	Uplink Centric Broadband Communication
UE	User Equipment
URLLC	Ultra-Reliable Low-Latency Communications
UPF	User Plane Function
V2X	Vehicle-to-Everything
XR	Extended Reality

8 References

[1] GTI. “5G-A x AI: New Era, New Opportunities, New Value.“

[2] GSMA. “The State of 5G 2024. Introducing the GSMA Intelligence 5G Connectivity Index”

[3] telecoms. “Maximising the Impact of 5G: 2024 Survey Report”

[4] OMDIA. “5G-A Industry Intelligence Maturity Index”

[5] ZTE. 5G+AI for Integrated Communication and Computing. ZTE News.

[6] R2-2408326 Discussions and evaluations on RRM measurement prediction, NTT DOCOMO, INC.

[7] R2-2405004 Discussions on other aspects related to RLF/HOF prediction, NTT DOCOMO, INC.

[8] R2-2408327 Discussions on measurement event prediction, NTT DOCOMO, INC.

[9] M. Li et al., "Automatic Data Generation and Optimization for Digital Twin Network," in IEEE Transactions on Services Computing, doi: 10.1109/TSC.2024.3522504.

[10] Q. Li et al., "Native Network Digital Twin Architecture for 6G: From Design to Practice," in IEEE

Communications Magazine, doi: 10.1109/MCOM.001.2400216.