HUAWEI

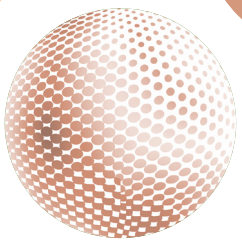
Striding Towards the Intelligent World White Paper 2024

Autonomous Driving Network (ADN)

AI for Network, Ushering in New Era of High Autonomy







Building a Fully Connected, Intelligent World

CONTENTS

01 CSPs in the Intelligent Era / 01

02 Trends Insights / 08

Trend 1 The Vigorous Development of New AI Applications Stimulate Network Traffic Growth and Create New Opportunities for User Experience Monetization 09

Trend 2 Large Models Accelerate the Digital Intelligent Transformation of Enterprises and Redefine Enterprise Network Experience and O&M 15

Trend 3 The Expansion of ICCs Is Promoted by Large Models, and Confronted by Challenges Posed by Intelligent Troubleshooting and Operations Efficiency Improvement in All Domains 20

Trend 4 5G-Advanced Accelerates Integrated Sensing and Communication and Extends Operator Networks to the Sensing Field, Potentially Posing New Challenges to Intelligent Network O&M 26

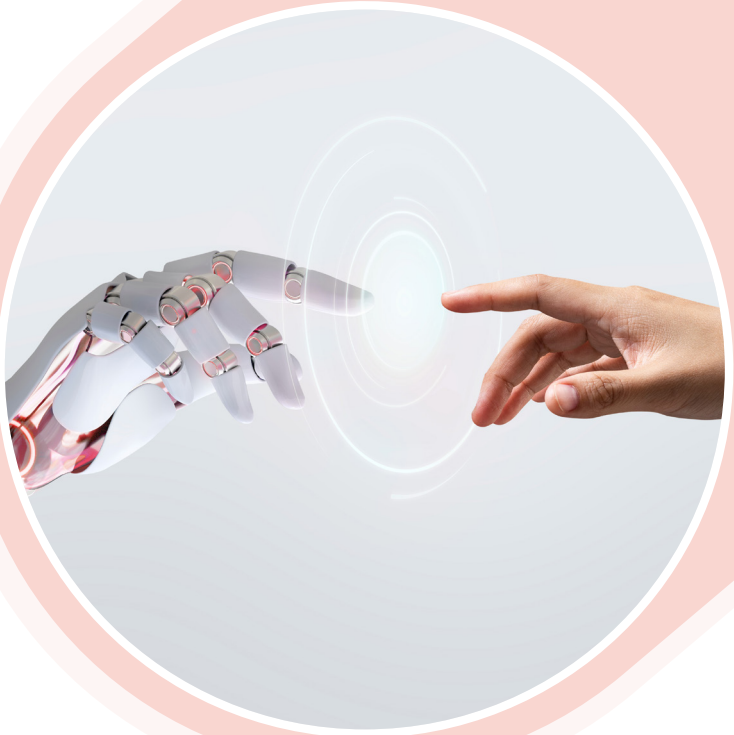
Trend 5 High-Value Scenarios Are Driving the Improvement of the AN Level, and Large Models Are Advancing the Evolution Toward AN Level 4 31

Trend 6 Large Models Work Alongside Other AI Capabilities to Solve Problems in Different Scenarios to Achieve Network Intelligence 38

Trend 7 Multi-Agent Collaboration Will Be a Key Technology for Achieving Highly Autonomous Networks in CSP Networks, and It Is Gaining Attention in Industry Research 45

03 Autonomous Driving Network / 50

Huawei's Autonomous Driving Network (ADN) Enables the Transformation Toward Highly Autonomous Networks and Helps CSPs Evolve Toward Full Intelligence 51



01

CSPs in the Intelligent Era

As the continuous evolution of artificial intelligence (AI) technologies is advancing comprehensive intelligent transformation across all industries, all enterprises are looking to use these technologies to deliver significant value as quickly as possible and become a leading player in the intelligent world. As such, it is critical to envision the future of enterprises in the intelligent era and develop strategies for enterprise development. Enterprises in the intelligent era should feature the 6 A's.



Adaptive User Experience, Auto-Evolving Products, Autonomous Operations, and Augmented Workforce represent the effects of intelligent transformation.

» Adaptive User Experience: Refers to the awareness and understanding of user behavior, requirements, interests, tastes, and environment changes, and proactive adjustment and provisioning of services that best fulfill user requirements. In addition to tailoring current products, special products must be designed from scratch to quickly fulfill personalized user requirements.

» Auto-Evolving Products: Indicates products in the intelligent era capable of self-learning, continuous iteration, and adaptation to changes. These products are self-optimizing and self-evolving.

» Autonomous Operations: Refers to a self-closed loop of awareness, planning, decision-making, and execution, enabling highly autonomous service flows.

» Augmented Workforce: This means that each employee has an intelligent assistant who understands the employee and completes each task efficiently and with high quality.

All Connected Resources and AI Native Infrastructure represent the foundation of intelligent transformation.

» All Connected Resources: Refers to the full connection of enterprise assets, employees, customers, partners, and ecosystems. This integration, along with real-time feedback and digitalization of objects, processes, and rules, improves the amount and quality of information quality. This helps develop a data flywheel and provides significant information-based advantages for enterprises.

» AI Native Infrastructure: Refers to ICT infrastructures systematically constructed to fulfill the requirements of intelligent applications, achieving ICT for Intelligence. Additionally, O&M management and experience assurance of infrastructures must undergo a complete intelligent transformation, achieving Intelligence for ICT.

A vision of communications service providers (CSPs) in the intelligence era is formulated based on the 6 A's.



1. Adaptive User Experience: CSPs deliver an adaptive user experience for mobile broadband (MBB), home broadband (HBB), enterprise private lines, video services, New Calling, smart homes, and other scenarios. Let's consider home products and services as an example.

» Adaptive network: The network is automatically adjusted based on various home application requirements (entertainment, office, and Internet of Things [IoT]) and different room locations to ensure an optimal application experience.

» Adaptive assistant: The assistant functions as a considerate "digital housekeeper" who detects the household situation, understands the needs and emotions of different family members, and offers tailored services.

» Adaptive gym: Based on the workout records of different family members in the past, the system customizes fitness courses for each member by automatically adjusting the themes and difficulty levels to ensure that everyone has the most suitable fitness experience.

» Adaptive entertainment: Based on the preferences of different family members, videos and games are recommended, and the plot and difficulty level of the games are adapted.

» Adaptive home: The lighting, room temperatures, and water temperatures are automatically adjusted based on the user's preferences, time, and weather to provide the most comfortable home environment.

2. Auto-Evolving Products: CSPs offer self-evolving MBB, fixed broadband (FBB), enterprise private lines, video services, New Calling, smart homes, and other services. Let's consider New Calling, which features a more diverse range of advanced functions, as an example.

Personalized images: An increasing number of artificial intelligence-generated content (AIGC) image types enable virtual humans to create real-time expressions that are more lifelike.

» Fun calls: More creative and interesting call backgrounds and features are provided.

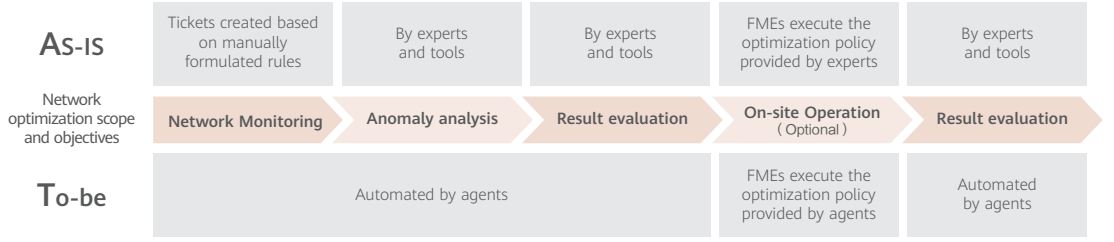
» Intelligent interactions: More intelligent interaction modes and diversified scenarios are offered.

» Real-time translation: More languages are supported, and the translation accuracy is improved.

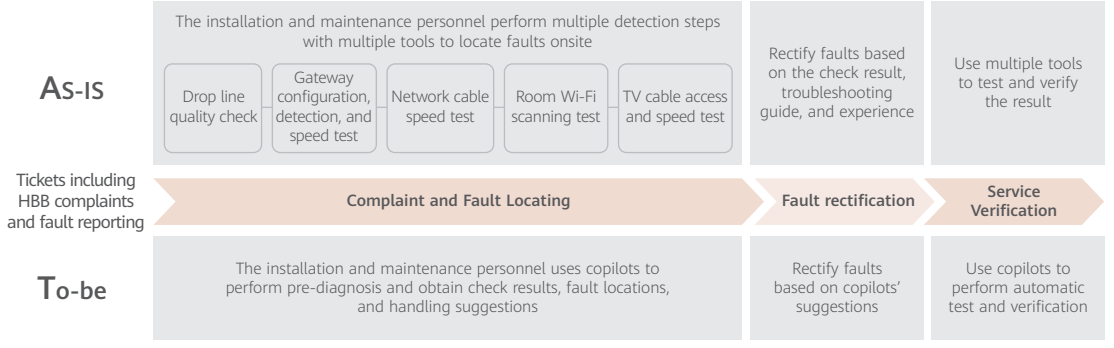
» Intelligent recording: More accurate and comprehensive information is recorded.

» Digital avatars: A more diverse range of scenarios, such as intelligent call answering services and conference speaking, are supported.

3. Autonomous Operations: CSPs deploy the wireless network optimization agent, fault monitoring and handling agent, change monitoring agent, home broadband experience assurance agent, and service provisioning agent to implement autonomous network operations. Let's consider the wireless network optimization agent as an example. It enables autonomous wireless network optimization, reducing the number of low-rate cells by 20% and the network optimization period from one day to one hour.



4. Augmented Workforce: CSPs offer exclusively smart assistants for their employees by deploying the home broadband installation and maintenance copilot, field maintenance copilot, customer service copilot, and development copilot. Let's consider the home broadband installation and maintenance copilot as an example. It reduces the average fault rectification time of installation and maintenance engineers from 60–90 minutes to 30 minutes, and the second site visit rate from 10%–15% to 5%.



5. All-Connected Resources: CSPs develop optical network digital twins, IP network digital twins, wireless network digital twins, and core network digital twins to offer all-connected resources. Let's consider optical network digital twins as an example. The digital twins are deployed based on the fiber/cable layer, optical layer, electrical layer, and service layer.

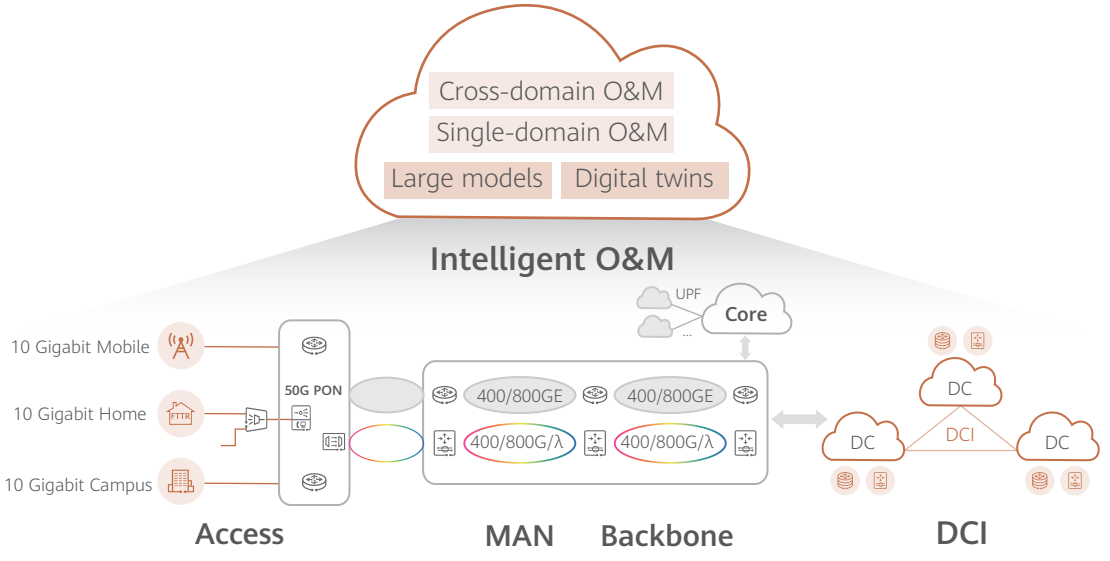
» Service layer: Includes information about bandwidth, latency, availability, bandwidth usage, packet loss rate, bit error rate, and paths.

» Electrical layer: Includes information about the ODU status, route, latency, bit error, alarms, and shared risk link groups (SRLGs).

» Optical layer: Includes information about wavelength resources and status, optical path routes, pre-FEC bit errors, optical power, signal-to-noise ratio (SNR), and alarms.

» Fiber/cable layer: Includes information about fibers, optical cables, fiber distribution terminals (FDTs), optical distribution frames (ODFs), poles, manholes, optical cable routes, fiber core connections, optical power, optical attenuation, and health status.

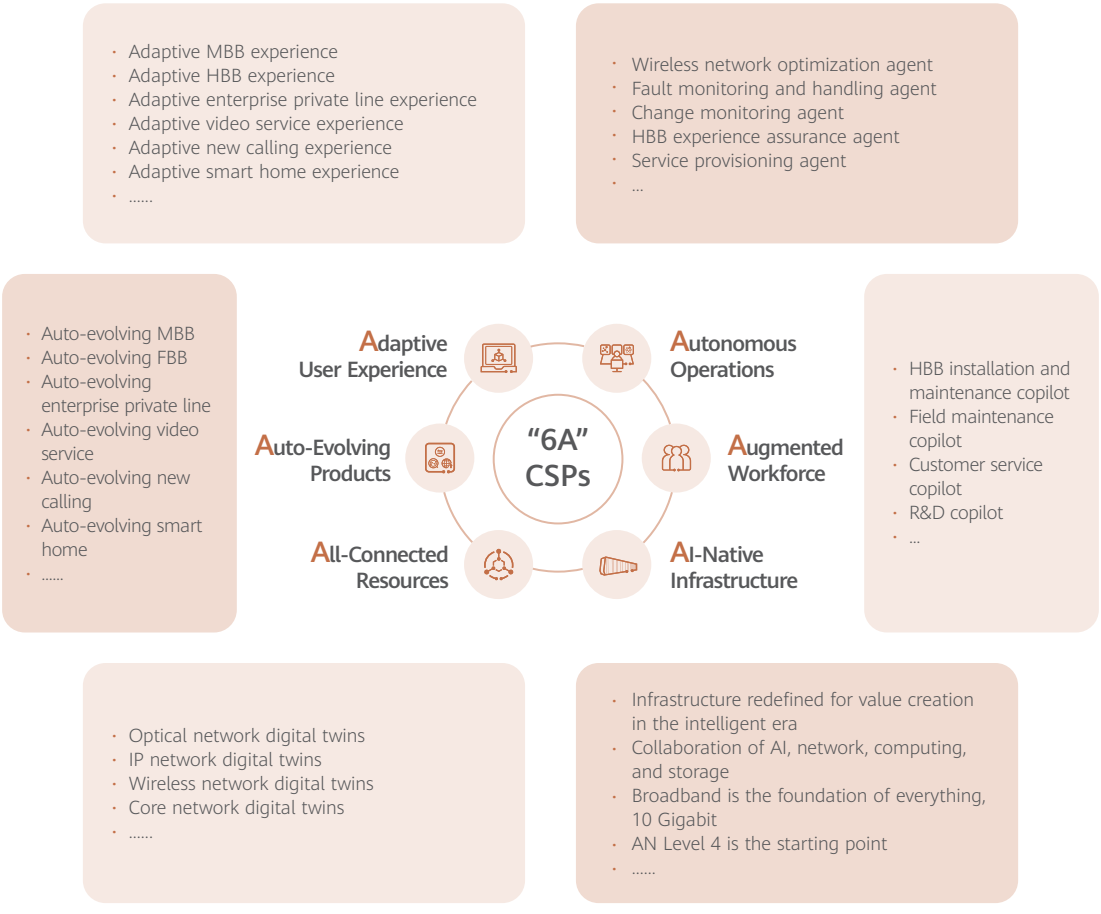
6. AI-Native Infrastructure: CSPs redefine infrastructures to deliver significant value (Adaptive Experience, Auto-evolving Products, Autonomous Operations, and Augmented Workforce) in the intelligent era and implement advanced collaboration of infrastructures and technologies such as AI, networks, computing, and storage.



» Broadband is the foundation of everything, including 10-gigabit access networks (mobile, home, and campus networks), ultra-high speed, low latency, and intelligent and elastic metro and backbone networks.

» With Autonomous Networks (AN) Level 4 as the starting point, scenario-specific agents and role-based copilot applications are developed for high-value scenarios based on large models and digital twins.

As AI technologies and applications are undergoing rapid development, we propose a new vision for CSPs in the intelligent era: "6A" CSPs — a common goal for comprehensive intelligent transformation across the industry. We expect to deliver more value to CSPs by continuously implementing innovative practices and exploration.





02

Trends Insights

Trend 1

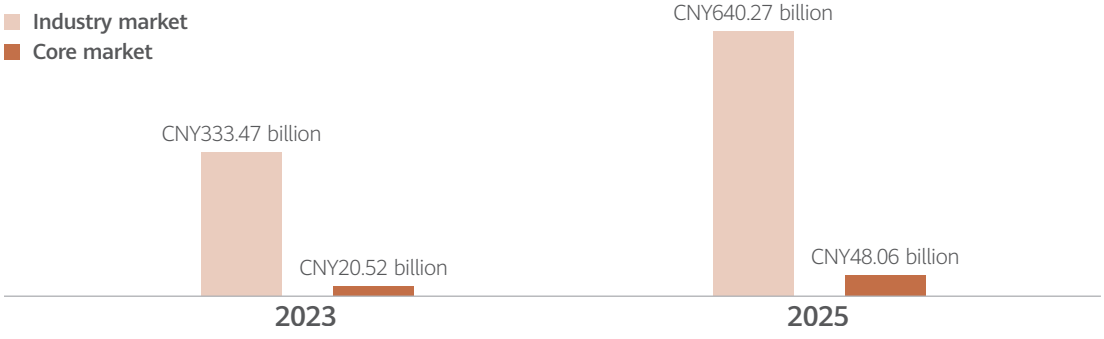
The Vigorous Development of New AI Applications Stimulate Network Traffic Growth and Create New Opportunities for User Experience Monetization



Business Insights

AI technologies are transforming the landscape of the live-streaming industry and related technologies. From e-commerce, gaming, technologies, culture, and tourism to real estate, the virtual human live-streaming service is witnessing rapid growth and extensive application, becoming one of the most trending services. On April 19, 2024, iiMedia Research, a third-party data mining and analysis organization in the new economy industry, released a White Paper on the Development of China's Virtual Digital Human Industry in 2024, announcing that the market scale created by virtual humans in China reached CNY333.47 billion and the core market scale reached CNY20.52 billion in 2023. These numbers are expected to reach CNY640.27 billion and CNY48.06 billion in 2025, respectively.

Scale of China's virtual human industry



The increasing demand for entertainment and continuous iteration of cutting-edge technologies such as AI has significantly stimulated the growth of China's virtual human industry, which in turn is quickly transformed by the thriving Metaverse. Virtual human live streamers have become the most popular service, accounting for 81.40% of all services. According to the "618" shopping festival data released by JD.com in 2024, Yanxi, a virtual human live streamer, was used in over 5,000 live streaming studios for over 400,000 hours, attracting over 100 million views and interacting with audiences 5 million times.



Source: 2024.6.18 Jingdong App

A virtual human can reproduce over 80% of the appearance, movements, and voice of a real person. A virtual human live streamer can help enterprises communicate with customers 24/7, sell products, and increase revenues.

Communications networks provide an increasing number of intelligent connections of people, vehicles, things, and enterprises. This creates huge volumes of new network traffic and leads to higher user requirements on stable network access, interaction experience, network bandwidth, and latency. For instance, an AI assistant requires an end-to-end (E2E) network latency of about 70 ms and an uplink network rate higher than 20 Mbit/s to deliver a "quasi-lifelike" or "lifelike" experience of real-time multi-modal interaction. Using AI to generate 3D object models for display on apps or web pages requires a downlink network bandwidth of 80 Mbit/s to 400 Mbit/s to deliver a click-and-start experience. In robotaxi security monitoring and remote takeover scenarios, videos from six cameras (12 Mbit/s to 24 Mbit/s) need to be uploaded through 5G networks. An E2E network latency longer than 100 ms will significantly affect the control accuracy. Another example is AIGC applications, which are developing rapidly and generating more high-quality videos. AI recommendation is implemented to distribute these videos and attract more viewers, increasing the video definition and viewing duration and resulting in different video traffic usage of different users. This leads to difficulty in network bandwidth convergence and rapid growth of network traffic. According to the 54th China Statistical Report on Internet Development released by the China Internet Network Information Center (CNNIC), the number of live-streaming users in China had reached 777 million as of June 2024, accounting for 70.6% of the total number of netizens. The mobile Internet access traffic had reached 160.4 billion GB, with a year-on-year growth of 12.6% Omdia predicted that the compound annual growth rate (CAGR) of global network traffic will reach 25% under the influence of AI-driven applications, including AI-upgraded applications and net new AI applications.

Impacts

1. The suboptimal experience of live streaming applications causes "frustrations" regarding network speed among users.

The upcoming 5G era invigorates the development of live streaming. The growing popularity of virtual human live streaming heralds an age of universal live streaming on the Internet but also leads to "frustrations" regarding network speed among consumers. According to iiMedia Research, 45.6% of consumers are dissatisfied with the network speed of their traffic packages. Network speed has become a major concern for selecting traffic packages. Live streaming features real-time interaction, frequent communication, and subjective perception, meaning that any blurry images or occasional frame freezing create a significant negative impact or even losses. Real-human and virtual-human live streaming pose new requirements on stable network access, smooth real-time transmission of high-definition images, sufficient network bandwidth, and low latency. Differentiated automatic network quality awareness, analysis, and assurance capabilities have become increasingly vital components in serving high-value network customers.

2. New business opportunities from generative AI (GenAI) applications raise new requirements for network awareness and experience.

Virtual-human live streamers can be used on e-commerce platforms when real humans are absent to increase the live streaming duration, continuously improving brand exposure and increasing the time for user interaction. An E2E network latency of 70 ms is required for virtual-human live streamers to produce real-time videos, for AIGC emotional companion/role-play applications to generate real-time text and images, and for AI assistants to deliver a "quasi-lifelike" or "lifelike" experience. Robotaxi security monitoring and remote takeover require an E2E latency of 100 ms and an immersive video experience. All these scenarios require improved real-time network awareness and interaction experience. Furthermore, AI technologies will gradually evolve toward intelligent robots with advanced intelligent interaction capabilities, delivering a more immersive experience and a stronger sense of companionship and belonging across various fields. A wider range of network applications will demand differentiated real-time awareness and interactive experience, creating new business opportunities and key markets for CSPs in the future.

3. New AI applications transform traffic consumption habits and stimulate CSPs to create innovative business models.

Virtual-human live streaming eliminates the time and space restrictions of consumption and creates new growth engines in the consumption market after midnight. It also leads to a significant increase in network traffic and highlights the value of network traffic during off-peak hours. Due to the limited profit potential of the traffic-based charging mode, CSPs are in urgent need of new business models and innovative business solutions to explore the value of high-value applications and traffic for revenue growth. For instance, a CSP launched the "full-service acceleration" product in Beijing to provide VIP network acceleration channels for users, improving the Internet access rate by over four times on average. A CSP in Zhejiang province deployed the 5G-A differentiated assurance solution to provide key service assurance for users by layer and level. This solution enables seamless network access during concert streaming, live streaming, mobile office services, and other key event scenarios, delivering a next-level communication experience.

Suggestions

1. Provide differentiated automatic network quality awareness, analysis, and assurance capabilities for high-value customers and new AI applications to enhance user experience.

For new AI applications, offer differentiated automatic awareness, analysis, and O&M capabilities for key quality indicators (KQIs) and customer experience indexes (CEIs) to ensure that major user experience issues, including suboptimal network quality, low image definition, and frame freezing of high-value network customers, can be quickly and automatically identified in real-time and automatically resolved through intelligent techniques. Furthermore, congestion points and bottlenecks on networks with suboptimal KQIs and CEIs need to be resolved through periodic and proactive checks and rectification. User experience needs to be improved to eliminate the "frustrations" regarding network speed.

These applications also integrate real-time interaction and awareness technologies, such as sensors, voice, images, videos, and smell, delivering an innovative user experience. CSPs need to consider these new AI applications and traffic as opportunities to pioneer network technology transformation by improving real-time network interaction and awareness capabilities.

2. Design new business models for high-value customers and scenarios to increase CSP revenues.

CSPs should fully leverage the new business opportunities deriving from e-commerce virtual human live streaming and design new business models. For example, CSPs can use innovative business methods such as new charging modes and VIP packages to provide differentiated traffic and deliver a differentiated interaction experience for high-value live-streaming customers and scenarios, while exploring the value of traffic in low-value periods (off-peak hours). These approaches will enable them to increase revenue.



Trend 2

Large Models Accelerate the Digital Intelligent Transformation of Enterprises and Redefine Enterprise Network Experience and O&M



Business Insights

The accuracy of large models continues to increase while the cost of using large models continues to decrease. According to the ten major AIGC application layer trends in 2024 released by IDC, over 53% of enterprises have started innovating AI services, marking the beginning of the digital intelligent transformation era for enterprises. However, the exponential growth of enterprise service complexity and O&M requirements presents numerous O&M challenges.

In the journey toward digital transformation, enterprise O&M encounters new challenges due to the advent of all-wireless offices, cloud-based applications, video-based applications, collaboration, and intelligent transformation. These developments catalyze a shift in traditional work models, concurrently giving rise to higher O&M standards and new requirements. Cloud-based, video-based, collaborative, and intelligent applications have higher requirements on network bandwidth, latency, and stability. For instance, in video conferencing and online collaboration scenarios, any network fluctuations may interrupt conferences or collaboration, causing significant negative impacts on the office experience.

Offices are evolving from a single headquarters to multiple branches, which are distributed worldwide. The production mode is evolving into smart manufacturing. Data centers, as the cornerstone of enterprise information, are growing in scale and being migrated to the cloud, forming a hybrid cloud and multi-cloud architecture. These transitions lead to a rapid increase in the enterprise network scale. The increasing device diversity is also a trend that deserves attention. From traditional servers, routers, and switches to Internet of Things (IoT) devices such as cloud cameras, smart sensors, and automated equipment, a diverse range of devices presents a significant O&M challenge, expanding the scope of routine maintenance and increasing the maintenance

complexity. In addition to ensuring the stable running of all devices, the network must quickly respond to various emergencies and continuously optimize network performance to ensure service continuity and growth. Furthermore, the continuous development of enterprise services and further digital transformation extends service deployment from a single network environment or geographical location to multiple network domains, including internal, external, cloud, data center, and IoT networks. Cross-domain deployment significantly increases the complexity and the time required to complete the deployment process, from service design and system integration to verification and provisioning.

Network security, another essential topic of network O&M, becomes more challenging as malware variants can be created more quickly, thanks to recent technological advancements. Conventional detection methods based on feature codes can no longer effectively detect these variants. Virus and malware variants can easily bypass traditional defense measures and penetrate enterprise intranets, causing data leakage, system breakdown, and other negative impacts. As more intelligent tools and technologies become available, attackers can use malicious large models, automated attack tools, and the like to collect and analyze crucial information about enterprises and formulate more accurate and covert attacks. Such attacks bypass traditional security monitoring and defense mechanisms, leading to long-term incubation and continuous penetration, posing significant threats to enterprise information security.

Impacts

1. Suboptimal O&M capabilities may become an obstacle to enterprises' digital transformation and innovation.

From the perspective of operations efficiency, network faults or device performance deterioration issues have a direct impact on services, lower production efficiency and customer satisfaction, and even lead to economic losses. For instance, enterprises may suffer significant losses when production lines stop, data centers break down, or key service systems are unavailable. To fulfill increasingly complex O&M requirements,

enterprises need to invest more labor, material resources, and financial resources, including recruiting professional O&M personnel and purchasing advanced O&M tools and technical services, which increase operational expenditure (OPEX). In a fast-changing market, enterprises need to flexibly adjust their business models and operation strategies — a task that is impossible without advanced O&M capabilities.

2. Accelerating service deployment and improving the efficiency and quality of cross-domain deployment become critical issues for enterprise O&M.

From the perspective of service agility, enterprises cannot quickly respond to market changes when service deployment takes a long time. In a fierce market environment, enterprises often want to launch new products or services to seize opportunities quickly. A long deployment period will undoubtedly impact the competitive edges of enterprises. From the perspective of cost-effectiveness, time-consuming deployment increases OPEX, which includes labor costs, time costs, and potential service losses caused by delayed deployment in addition to hardware and software costs. From the perspective of user experience, inefficient deployment may cause service interruption or delay, affecting user satisfaction and loyalty.



3. Technical skills and security operations policies need to be improved to ensure enterprise information security and service continuity.

From the perspective of service security, these challenges threaten the information security and service continuity of enterprises. Once an enterprise's information system is infected by malware or intelligent attacks, sensitive data may be leaked, and services may be interrupted, impacting the enterprise's reputation, customer relationships, and service competitiveness. From the perspective of operations cost, enterprises need to invest more resources and funds to enhance security protection measures and address these security challenges. Purchasing advanced security devices, software, and services, and improving the technical skills and emergency response capability of the security team significantly increase OPEX. Security challenges may also affect the service quality and user satisfaction rates of enterprises. For example, excessively stringent security control measures may cause inconvenience to user operations and affect user experience. Exposing security vulnerabilities may cause panic and dissatisfaction among users. These factors may harm the brand image and market position of an enterprise.

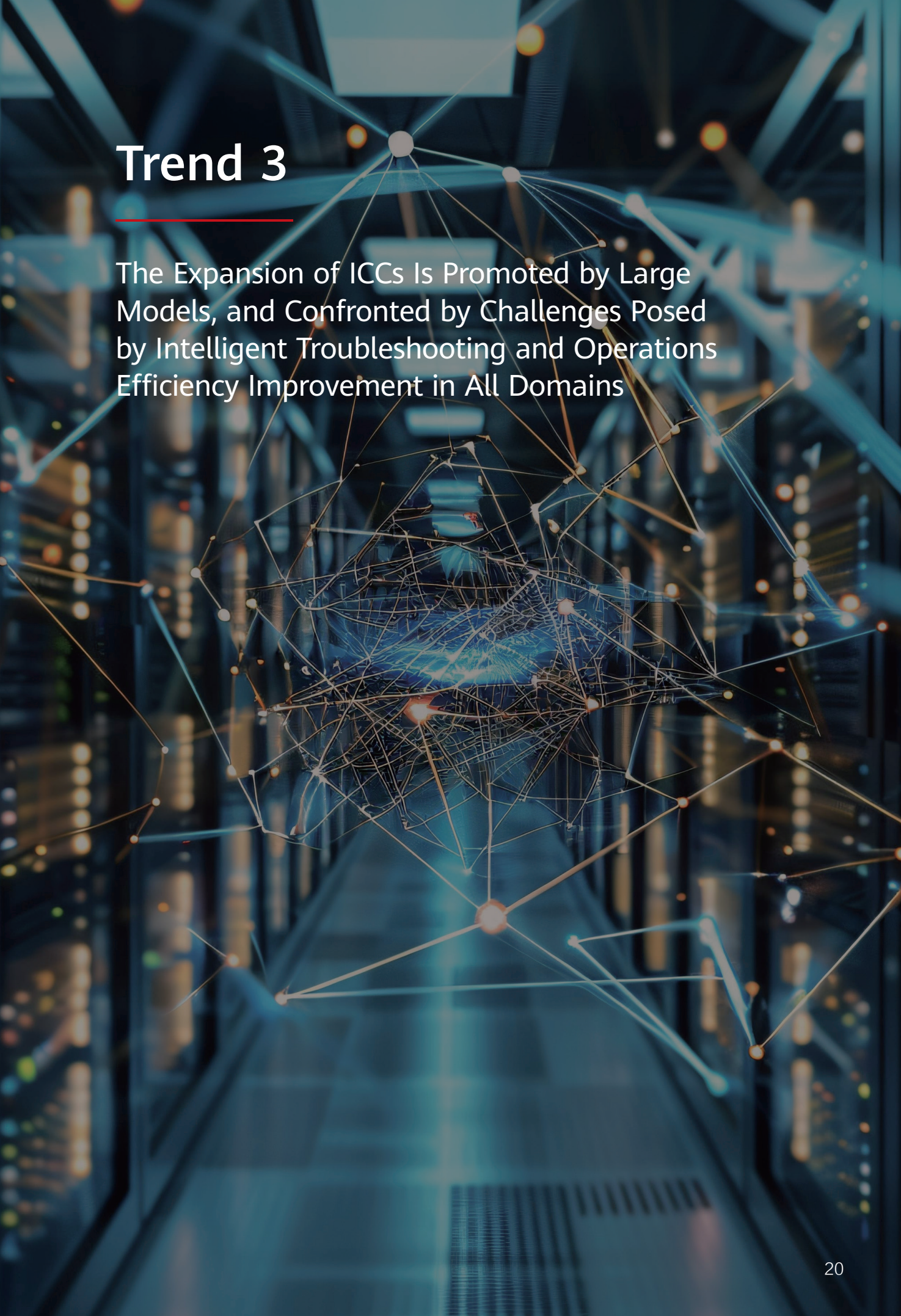
Suggestions

Develop intelligent O&M capabilities for enterprise networks based on digital twin and large model technologies to achieve zero security risks, zero-wait service provisioning, zero network interruption, and zero service congestion.

Enterprises need to focus on full-stack management system interconnection, perform centralized digital twin modeling for infrastructure resources in all domains, and implement full-stack convergence of terminals, CT devices, IT devices, and applications. They also need to develop a unified entry, full-stack topology, and out-of-the-box core capabilities, and implement full-stack visualization, awareness, and simulation. Based on advanced AI technologies such as large models, enterprises need to develop scenario-specific agents for various typical network O&M scenarios to implement self-closed loops of processes and services. They also need to develop copilots as intelligent digital assistants for O&M personnel and users in each phase of the process to improve O&M collaboration efficiency and reduce maintenance costs.

Trend 3

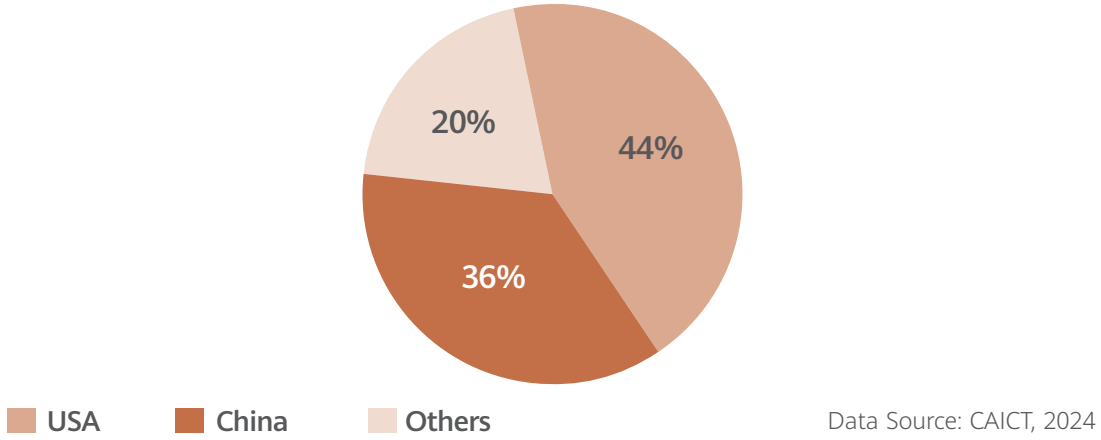
The Expansion of ICCs Is Promoted by Large Models, and Confronted by Challenges Posed by Intelligent Troubleshooting and Operations Efficiency Improvement in All Domains



Business Insights

Large models are developing rapidly across the globe and have numbered 1,328 by the first quarter of 2024, with over 80% of them coming from China and North America.

Global AI Models Distribution



Data Source: CAICT, 2024

Take China as an example. On January 11, 2023, China's State Information Center released the Intelligent Computing Center Innovation and Development Guide. The report estimates the economic benefits generated by investment in intelligent computing centers (ICCs). By 2025, 80% of enterprises will use ICCs, and the investment in ICCs for a city can drive the growth of core AI industries by 2.9 to 3.4 times and related industries by 36 to 42 times. In the past two years, the development of AI large models has placed new requirements on computing power, algorithm platforms, and data. According to the scaling law, traditional CPU-centric cloud computing infrastructures cannot offer orders of magnitude higher computing power required by large models. With the fierce competition among myriad large models, the AI industry is witnessing the professional division of labor, as well as a development trend of intelligent computing centering

on GPUs and NPUs. On October 8, 2023, China's Ministry of Industry and Information Technology (MIIT) and six other authorities issued the Action Plan For The High-quality Development Of Computing Power Infrastructure. According to the plan, by 2025, China aims to achieve over 300 EFLOPS in computing power, increase the percentage of intelligent computing power to 35%, and build 50 ICCs. Over 100 ICC construction and operation projects have been launched in China.

This example proves that exclusive, large-scale, and time-consuming large-model training demands more stable AI server clusters in ICCs. On the one hand, AI cluster servers are much less stable than traditional computing servers. Faults related to optical modules, power supply, heat dissipation, and network connections may impact the availability of AI servers. Additionally, it is more complicated to train large models than traditional distributed AI, sometimes taking weeks or even months. Large-scale AI cluster servers with 1,000 GPUs/NPUs or 10,000 GPUs/NPUs in ICCs may result in frequent AI server faults during AI training. When training is performed on clusters with 1,000 GPUs/NPUs, faults occur every one to three days, resulting in two times longer training duration than compared to when no fault occurs. On the other hand, during the training of large models, the model and optimizer statuses are stored in the device memory. If a single AI server is faulty, all AI server processes in the cluster will be blocked, resulting in the loss of model and optimizer statuses. This, in turn, will interrupt the entire training, and prevent fault cause and location from being quickly identified. Furthermore, the success rate of large model training will decrease and so training costs will be high if AI cluster servers have low computing efficiency, faults occur frequently, troubleshooting takes a long time, or training cannot be resumed quickly after it is interrupted. In a nutshell, AI large model training has higher requirements on the stability, fault detection, and training fault tolerance capabilities of AI cluster servers in ICCs. It is imperative that ICCs have capabilities such as real-time fault monitoring, resumable training, automatic isolation of faulty nodes, and quick fault locating and rectification.

Additionally, the exponential growth of AI applications may contribute to AI reference, which requires 100 times higher computing power than AI training. IDC has predicted that, by 2026, cloud-based inference will account for 62.2% of computing power demand, and AI training for 37.8%. It means that AI inference will become a major market in the AI industry. In terms of the application of large models, the core challenge facing AI inference is cost. To address this challenge, medium-sized AI large models that deliver considerable benefits are widely used in various vertical industries to meet computing power requirements and reduce the costs of AI inference.

Above all, improving the efficiency of cluster O&M (converged and full-stack O&M for computing power, networks, and storage devices), as well as prediction, minute-level demarcation and locating, and closed-loop capabilities for faults, are the key to maximizing the value of ICCs' computing power.

Impacts

1. Intelligent fault detection, resumable training, and automatic isolation of faulty nodes for AI cluster servers are key to efficiently utilizing the resources of ICCs.

First, it is necessary to develop real-time automatic detection and root cause analysis capabilities for faulty nodes in AI cluster servers. With these capabilities, O&M personnel can collect the status and incident logs of AI cluster servers in real-time to identify whether the servers and networks are normal. By doing so, O&M personnel can quickly identify faulty processes, locate fault causes, and rectify faults. A prime example is the complex process of locating optical link faults on cluster servers, which involves computing power, networks, and storage devices. O&M personnel need to physically visit an equipment room with professional tools for detecting optical link faults, and detecting a single fault takes one hour. For tricky AI software faults like "Notify Wait" timeout faults that occur frequently, the average time required for manual analysis is greater than three days. Second, AI cluster servers must be able to resume training in minutes, which is the key to ensuring that AI training interruption does not lead to suboptimal availability of computing power. Generally, data is backed up before training can be resumed, and checkpoints are set to check and update the backup data during the training. It usually takes several hours to diagnose, isolate, and rectify a fault, wasting a huge volume of computing power. To address this challenge, the AI model and checkpoints of the optimizer need to be periodically persisted to reduce the checkpoint time and minimize the extra AI training time caused by faults. Third, capabilities such as intelligent fault prediction and fast automatic isolation of fault nodes must be developed for AI cluster servers to reduce the probability of faults and their impact on AI training.



2. Precise operations of computing power, networks, and storage devices in all domains for AI cluster servers are the key to increasing the investment efficiency of ICCs.

Using O&M operations on AI clusters is not a well-trodden path. In the industry, there are no widely used methods for checking large-scale AI cluster risks through NPU chip health checks, HCCL bandwidth tests, intermittent internet connection checks, optical component fault prediction, and frequent component fault prediction. Additionally, ICCs involve full-stack integration of software and hardware about computing power, networks, and storage devices in all domains. The traditional methods are inefficient for checking risks in clusters with 10,000 GPUs/NPUs and over 100,000 computing power, network, and storage components. Full software and hardware checks take more than three days.

Moreover, inefficient nodes and networks cause a loss of over 50% of computing power. Due to the strong synchronization feature of large model training, a single point of failure (SPOF) may cause significant performance deterioration in jobs and difficulty in fault locating due to fault diffusion, leading to tens of thousands of checkpoints. As

a result, manual demarcation and locating can take hours or even days, presenting a major challenge for ICCs.

Suggestions

1. Develop intelligent fault handling capabilities to improve O&M efficiency.

CSPs need to implement automatic and intelligent fault detection, root cause identification, and quick rectification for AI cluster servers to minimize the training interruption time caused by faults. Meanwhile, the minute-level resumable training capability must be developed for AI cluster servers to optimize the checkpoint process, reduce the checkpoint time, and minimize the extra AI training time. Additionally, CSPs need to implement intelligent prediction and fast automatic isolation of fault nodes to reduce the probability and impact of faults and quickly resume the training process.

2. Develop fault prediction and prevention capabilities to improve full-stack operations efficiency in all domains.

Before AI training is performed, a full-stack risk check and precise component-level fault prediction must be implemented for ICCs to ensure good AI training performance in AI cluster servers, reduce the job failure probability, and prevent computing power loss. Additionally, the risk check period and prediction time need to be reduced from days to minutes. When AI training or AI inference is performed, it is critical to quickly detect and analyze inefficient nodes or networks and improve the computing power efficiency of AI clusters. To this end, the deterioration detection capability needs to be transformed from passive mode (reporting in hours through manual tickets) to proactive mode (intelligent detection in minutes), and the demarcation and locating time needs to be reduced from dozens of hours (manual handling) to less than 30 minutes (automatic handling).

Trend 4

5G-Advanced Accelerates Integrated Sensing and Communication and Extends Operator Networks to the Sensing Field, Potentially Posing New Challenges to Intelligent Network O&M



Business Insights

Wireless air interface sensing represented by the low-altitude economy is growing fast, with favorable policies being issued. According to the National Airspace Basic Classification Method released by the Civil Aviation Authority of China (CAAC) in December 2023, the country organizes its airspace into 7 classes, of which classes G and W are unregulated. This defines the low-altitude airspace where electric vertical take-off and landing (eVTOL) aircraft, light and small drones, and common aircraft can legally fly. On February 27, 2024, Shenzhen showcased the world's first eVTOL flight on a cross-sea, inter-city route. The flight from Shekou Ferry Port in Shenzhen to Jiuzhou Ferry Port in Zhuhai took just 20 minutes, a journey that would take 2.5 to 3 hours by car. On June 28, the "low-altitude + rail" air-rail intermodal transport project was officially launched in the East Square of Shenzhen North Railway Station. It formed a low-altitude transportation network with Shenzhen North Railway Station as the center, enabling travel to more than 90% of the Guangdong-Hong Kong-Macao Greater Bay Area within an hour. In the logistics field, Meituan launched 22 drone delivery routes in 8 business districts in Shanghai, Shenzhen, and other cities, providing delivery services for office buildings, scenic spots, and hospitals. It had completed a total of more than 210,000 orders by the end of November 2023. The average time of drone deliveries in 2022 was about 12 minutes, improving efficiency by nearly 150% when compared to traditional deliveries and saving nearly 30,000 hours for customers.

The emerging low-altitude economy can open up new business opportunities for CSPs. (1) Massive new connections will be added. For instance, drones mentioned earlier

fly beyond line of sight (LOS), which means that video stream backhaul is required to control their operation remotely. This requires ultra-large uplink video connections, providing new opportunities for CSPs. According to China's National Three-Dimensional Transportation Network Planning Outline, the market size of China's low-altitude economy is expected to exceed CNY6 trillion by 2035. Furthermore, the number of commercial and industrial drones is expected to reach 26 million, and the number of drone pilots will grow to 630,000. Consequently, the number of connections will expand to tens of or even hundreds of millions. (2) Low-altitude surveillance teams will have more demanding sensing requirements, as they must be able to detect abnormal intruders, such as unauthorized flights, in real-time to ensure safe passage on low-altitude routes. However, traditional radar-based low-altitude detection solutions face multiple challenges, including site re-selection and blocking during deployment, high costs, and poor feasibility. To address these issues, CSPs can innovate sensing services by using deployed wireless infrastructure.

In addition to providing communication connections, integrated sensing and communications (ISAC) can expand the business scope for CSPs in the security field. In indoor scenarios involving Wi-Fi, innovative services such as smart home guard and smart healthcare can be provided based on wide-coverage home broadband networks and Wi-Fi sensing. This eliminates privacy concerns inherent in camera-based camera solutions. Additionally, outdoor innovative services empowered by 5G-Advanced wireless networks include offshore/river channel intrusion management based on wireless sensing. With the use of higher frequency bands, groundbreaking services such as bridge/road micro-deformation detection based on millimeter wave (mmWave) technology can be introduced. For the industry sensing field characterized by optical fiber sensing, a range of novel services can be offered. According to the research on optical network integrated sensing and communication architecture and key technical solutions for multi-scenario applications released by IMT-2020 (5G) Promotion Group in September 2024, optical fiber sensing allows digital management of optical cable resources while also advancing innovative environment sensing and security monitoring services. These include detecting urban road safety risks, geological and hydrological data, and intrusions in oil and gas pipelines.

Impacts

1. Novel services in the low-altitude economy necessitate ISAC network construction.

The low-altitude economy is driving a transformation from traditional ground coverage to ground and low-altitude 3D coverage and sensing. Basic networks (including base stations) need to be upgraded to offer ISAC service functionalities. Additionally, a sensing-oriented network management system needs to be built to efficiently and quickly process flight data of drones flying at high speeds. This covers the overall sensing services that involve multiple base stations, with capabilities in obtaining and analyzing sensing data and achieving accuracy in sensing results.

2. ISAC networks require new operation modes and O&M capabilities.

In the ISAC indicator design of 3GPP, communication and sensing performance is measured by two sets of indicators. Communication indicators focus on capabilities such as channel capacity, spectral efficiency, and SNR. In contrast, key sensing indicators include sensing accuracy, speed, detection rate, and false detection rate. With the development of services, various indicator systems pose higher requirements on network O&M and quick response. For instance, operation KPIs and technologies need to be developed for connection services throughout the lifecycle of network planning, construction, maintenance, and optimization. Additionally, sensing-oriented innovative services require full-lifecycle operation capabilities.

In terms of specific O&M capabilities, wireless networks are expected to evolve from providing traditional ground coverage to enabling ground and low-altitude 3D coverage. Another major challenge is the increasing complexity of wireless environments. Low-altitude networks are affected by multiple factors, including low-altitude flight, small objects, complicated electromagnetic environments, and obstacles. Consequently, there is a need for more network planning and maintenance, simulation technologies with higher accuracy, and network optimization technologies with higher levels of intelligence and automation.

Suggestions

1. In the wireless network field, industry stakeholders must seize the opportunities offered by the low-altitude economy to innovate low-altitude ISAC services, expand new connection services, and provide sensing data services for low-altitude flight safety.

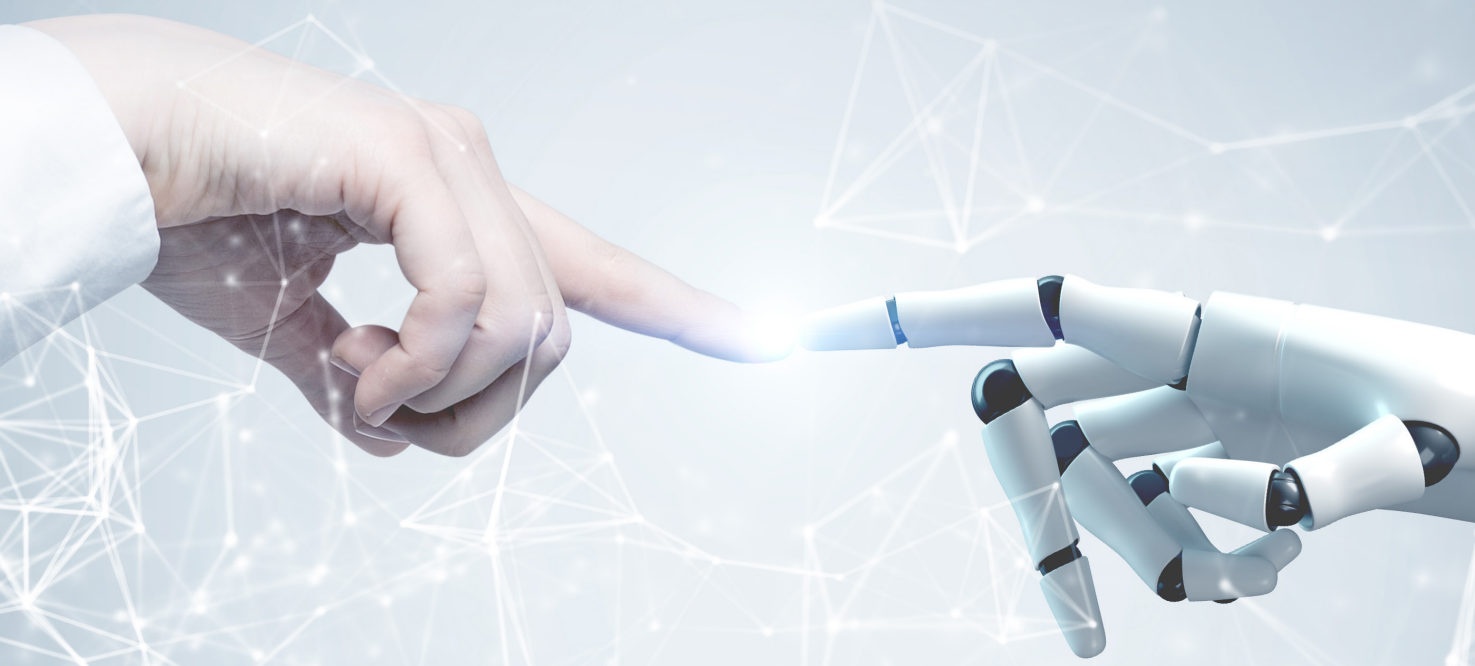
2. In the security field, stakeholders must participate in industry research and collaboratively advance service innovation in home networks, industry sensing, and other directions.

3. In fields such as network planning, simulation, and optimization, stakeholders must collaboratively develop innovative, intelligent O&M technologies for ISAC networks to empower intelligent O&M of 5G-Advanced ISAC networks.



Trend 5

High-Value Scenarios Are Driving the Improvement of the AN Level, and Large Models Are Advancing the Evolution Toward AN Level 4



Industry Insights

AN strives to provide fully automated networks and intelligent infrastructures, agile operations, and all-scenario services of ICT. It is not just a trend but a game-changer for the industry. It helps CSPs seize new business opportunities, improve customer experience, reduce costs, and increase revenue. In recent years, the rapid evolution of AN has provided a fertile ground for the development of automatic and intelligent O&M for new services such as AIGC services, large model applications, and low-altitude economy. As of August 2024, 61 industry partners have signed the AN Manifesto, a sign of the promising future of AN. 14 leading CSPs have set their strategic goal of achieving AN Level 4 by 2025 to 2027, including China Mobile, China Telecom, China Unicom, MTN, AIS Thailand, Deutsche Telekom, and Vodafone, paving the way for a brighter future in the telecommunications sector.

The commercial practices on AN have allowed leading CSPs to gain significant benefits, reinforcing their strategies to advance to AN Level 4. At the Digital Transformation World (DTW) held in June 2024, China Mobile announced a significant milestone — it reached AN Level 3.2 in 2023 and L3.5 in 2024, with a 95% automation rate of high-value scenarios. This remarkable achievement, coupled with the development of 3,000 AI baseline capabilities, allowing for minute-level network configuration, fault rectification, and network optimization, is a testament to the potential of AN. This is expected to save 5,000 labor costs annually. China Mobile's future goals, including reaching AN Level 4 in 2025 and focusing on automating all high-value scenarios, inspire the company to strive for cost-effective and efficient O&M featuring highly automatic and intelligent network operations capabilities with E2E automation as the basis and AI+ as its core characteristics.

The improvement of AN capabilities relies heavily on the rapid development and convergence of AI technologies. These technologies provide strong support for the

communications industry by optimizing network performance, reducing operation costs, and promoting innovation and commercial deployment. One ground-breaking technology is GenAI, which can generate multi-modal content, including text, image, sound, video, and code. Consequently, it can be applied to network O&M, including automatic configuration and code generation, analysis, and reasoning of complex problems in chains of thought. At the 2024 Mobile World Congress (MWC) in Barcelona and Shanghai, the telecom foundation model with GenAI as its core has become the hottest topic in the communications field.

According to Omdia, GenAI core capabilities are typically used in the communications industry to generate more accurate responses for chatbots and optimized scripts for customer service agents, as well as provide effective reasoning for network maintenance. This helps CSPs reduce manual intervention or even achieve complete automation in multiple service processes, significantly improving O&M efficiency. Consequently, the industry proposes that the GenAI-based telecom foundation model is revolutionary and holds the key to achieving Highly Autonomous Networks. This proposal has been widely discussed at global industry conferences, including DTW and MWC. In response to this proposal, TM Forum established a GenAI AN standard working group in December 2023. Additionally, the China Communications Standards Association (CCSA) initiated the large model pioneer plan to promote the implementation of AN Level 4 enabling technologies such as large model/GenAI and agents.

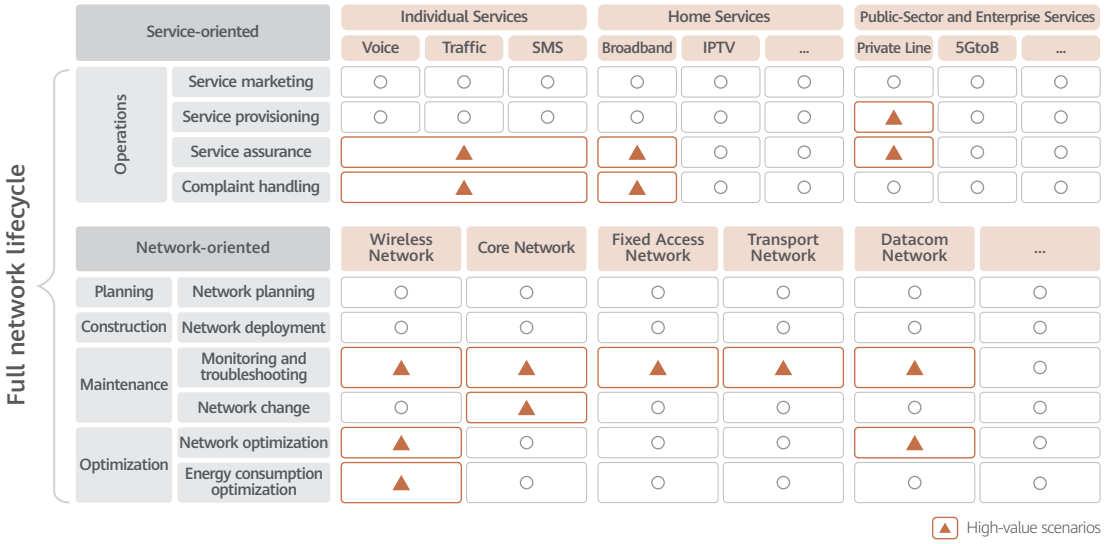
TM Forum has incorporated AN as one of the three core missions, according to its statement. At the DTW conference in June 2024, TM Forum, in collaboration with industry partners, including China Mobile, Vodafone, Telefónica, Huawei, Ericsson, and AsiaInfo, along with Professor Joseph Sifakis, the 2007 Turing Award winner, unveiled the Autonomous Networks Level 4 industry blueprint: high-value scenarios. This report introduces the Level 4 industry blueprint, which includes the vision and objectives, high-value scenarios, architecture, and evolution path. It serves as a valuable reference for CSPs looking to plan and deploy AN Level 4.

1. High-value scenarios should be sorted based on business value to achieve Level 4 phase by phase.

In the face of numerous scenarios divided by communications services and networks, CSPs should prioritize investing resources and energy in scenarios with high business

value to maximize the return on investment (ROI). The Level 4 industry blueprint categorizes implementation scenarios into two phases. Phase 1 (2025–2027) emphasizes single-domain maintenance/optimization scenarios, while phase 2 (2028–2030) centers on multi-domain E2E complex scenarios. TM Forum, on the other hand, adopts a different standard to classify CSP scenarios into three value ranges: high, medium, and low. It identifies 15 high-value scenarios in operations, maintenance, and optimization, based on operations value and technology maturity (as shown in the following figure).

AN Level 4 High-Value Scenarios in 2025-2027



2. AI/GenAI should be introduced to each network layer to build various copilot and agent applications.

In the past few years, AN exploration has focused on point-level use cases, contributing to the identification of hundreds of capabilities through dozens of O&M tasks. However, the fragmented implementation process and outcomes cannot be used for unified deployment. To solve this problem, technical breakthroughs of the telecom foundation model with GenAI as its core have improved the core capabilities such as content generation, chain-of-thought analysis, and multi-modal simulation. These capabilities have enabled the development of new processes and capabilities that are interdependent and feature large granularity, offering application-level solutions for AN practices. The Level 4 industry blueprint proposes an architecture with three layers (business, service, and resource operations) and four closed loops. This architecture introduces full-stack

AI/Gen AI capabilities to build two types of application-level capabilities, including role-based copilots and scenario-specific agents. This helps to enable autonomy in single-domain high-value scenarios and establish a foundation for cross-domain collaboration.

Impacts

The evolution towards AN Level 4 marks a new stage in the development of AN, where the "machines assisting humans" approach has transformed into "humans assisting machines". The Autonomous Networks Level 4 industry blueprint: high-value scenarios paper serves as a crucial reference for industry development. As AN Level 4 continues to expand, cutting-edge technologies such as telecom foundation models and digital twins will be developed to overcome existing limitations, offering the following benefits for the communications networks:

1. Reshaping the O&M mode: Traditional CLI- and GUI-based O&M has low efficiency and requires personnel to have high-level skills. For instance, network maintenance staff need to locate specific operation interfaces across multiple systems and perform operations on each interface to analyze and summarize the data for the desired information. To address these issues, the traditional O&M mode has been upgraded to the intelligent O&M mode. Role-specific copilots provide intelligent O&M methods such as intelligent Q&A, intelligent query, and intelligent report generation based on natural language interaction. Additionally, scenario-specific agents can implement automatic closed-loop management based on the preset targets with minimal or no interference.

2. Reshaping system capabilities: The automation and intelligence capabilities in awareness, analysis, decision, and execution of the system will be improved to unleash the full potential of the telecom foundation model. **Awareness:** Active and passive resources, as well as the network performance and application experience, can be detected in milliseconds. **Analysis:** Intelligent reasoning now relies on chains of thought

instead of preset rules, improving the capabilities to solve complex non-deterministic problems. **Decision:** AI simulation has replaced traditional mechanism simulation, resulting in higher efficiency and precision as well as the ability to solve generalization difficulties. **Execution:** Manual assurance has been upgraded to automatic verification and correction, ensuring reliable and efficient execution.

3. Reshaping service processes: The Level 4 target state involves building an automatic Level 4 process for services and O&M driven by intents rather than personnel. This involves eliminating service breakpoints, simplifying O&M nodes, and minimizing manual intervention by streamlining processes in the high-value scenarios of the AN map based on business value, enablement, and breakpoints.

4. Reshaping the integration mode: Traditionally, the customized development of NBIs requires several months for development, testing, and version release. With the new integration mode, however, scenario-specific APIs can be generated online using the generative capabilities of large models, allowing for loose coupling and simplified integration of interfaces between systems. This accelerates service rollouts and business monetization.



Suggestions

1. Implement commercial practices of Level 4 centered on high-value scenarios. Copilot and agent applications based on high-value scenarios of single-domain maintenance, optimization, and operations have been released and delivered significant value in live-network practices. However, these practices are currently only implemented in a few advanced provinces in the Chinese mainland. Level 4 marks a fresh new start for all CSPs. It is recommended that more CSPs actively implement Level 4 practices to experience the business value of large-granularity Level 4 solutions.

2. Transform and reserve talent to maximize large model capabilities. The AN Level 4 industry blueprint outlines that copilots and agents of large models will facilitate the integration and simplification of E2E O&M processes as well as the breaking of organizational boundaries and streamlining service processes. This involves a shift in the O&M approach from human-to-human collaboration to human-machine or even machine-machine collaboration, requiring that O&M personnel should be capable of utilizing large model applications. In recent years, the communications industry has seen an increasing need for "prompter engineering", which emerged after the launch of ChatGPT, to take full advantage of the telecom foundation model.

3. Collaborate with the industry to improve the execution standards of the Level 4 industry blueprint. Based on the framework in Autonomous Networks Level 4 industry blueprint: high-value scenarios, CSPs should implement the five principles of Level 4 development and reach a broader industry consensus. Furthermore, the industry should refine the objectives, key effectiveness indicators (KEIs), and scenario design for high-value scenarios. It should also develop and enhance Level 4 standards, including evaluation standards, CSP requirements, KEIs, key enabling technologies, and interface definitions based on high-value scenarios. Additionally, a more comprehensive evaluation mechanism and tools should be provided.

Trend 6

Large Models Work Alongside Other AI Capabilities to Solve Problems in Different Scenarios to Achieve Network Intelligence



Technology Insights

1. Multiple AI capabilities can be leveraged to meet the service requirements of communication networks for accuracy, timeliness, and security.

GenAI has become the driving force for the digital transformation of communications networks. GenAI redefines network management through automatic event response, network fault and risk identification, and network event handling solution optimization, changing the network O&M management mode and improving the automatic O&M management level. TM Forum estimates that GenAI can save US$20 billion a year for CSPs in terms of network interruption and service degradation.

After innovative exploration, GenAI now has been applied to the communications industry. For example, Ask AT&T, a ChatGPT-based GenAI platform launched by AT&T, is used to empower employees. Ask AT&T is helping employees finish routine work, such as Q&A and code conversion, allowing employees to focus on more complex and valuable tasks. It is worth noting that AI has a vast space for development. GenAI is one of the ways to solve real-world problems, but there are also other AI technologies, such as machine learning, graph computing, predictive AI, optimal decision-making, and heuristic rules, that are crucial for solving scenario-specific problems. GenAI needs to be combined with other AI and non-AI technologies to achieve the accuracy, timeliness, and explainability requirements of intelligent communication networks.

Gartner's AI prisms provide CSPs with a comprehensive assessment of GenAI use case examples to aid in use case selection. Gartner's report points out that GenAI models are more suitable for use cases, such as content generation, conversational interaction, and knowledge discovery, and are not well-positioned to solve problems independently in use cases, such as prediction, planning, and decision-making.

Use-case family	Generative models' current usefulness	Use-case examples
Prediction/Forecasting	LOW	Risk prediction, customer churn prediction,sales/demand forecasting
Planning	LOW	Operation research,optimization,routeplanning
Decision Intelligence	LOW	Decision support, augmentation,automation
Autonomous Systems	LOW	Self-driving cars, advanced robotics,drones
Segmentation/Classification	Medium	Clustering, customer segmentation,object classification
Recommendation Systems	Medium	Recommendation engine, personalizedadvice, next best action
Perception	Medium	Object detection,recognition, analysis
Intelligent Automation	Medium	Intelligent document processing,objectcharacter recognition,robotic processautomation, hyperautomation
Anomaly Detection/ Monitoring	Medium	Abnormaltransaction detection, outlierdetection, monitoring
Prediction/Forecasting	High	Text generation, image and videogeneration, synthetic data
Content Generation	High	Virtual assistant, chatbot, digital worker
Conversational User Interfaces	High	Knowledge store, search, mining

Source:Gartner (March 2024)

A standalone GenAI model for prediction may not be the optimal choice. Some have tried using GenAI models for prediction tasks, such as trend prediction, traffic prediction, rate prediction, and network status time series prediction, but such models are not designed for statistics or autoregressive modeling based on given network status data. Predictive AI technology is more suitable for these prediction tasks.

Current GenAI models also lack the network optimization capability. GenAI is not effective in high-value network scenarios, such as energy saving optimization, path optimization, multi-objective combinatorial optimization, and configuration policy optimization — because the optimal solution can only be found through iterative optimization by delivering optimization policies to the live or simulated network environment. As such, it is essential to design a dedicated network optimization algorithm based on reinforcement learning and automatic control technologies. In practice, network optimization typically requires various technical approaches. By integrating GenAI models into their network optimization process, CSPs can enhance the planning and scheduling capabilities of their network optimization system.

Network decision-making is a complex, systematic project. It requires case-by-case analysis to develop appropriate solutions and actions and fulfill the expected objectives.

However, the output of current GenAI models still contains hallucinations and lacks explainability. Using such output, CSPs' decision-making may pose critical technical risks. In a complex network environment, GenAI is not able to achieve autonomous decision-making. Key policies, such as awareness, analysis, decision, and execution, still rely on human experience, control, and guidance.

2. A model of a proper scale can be tailored for a specific scenario.

The size of a model must be assessed based on the specific requirements of a service scenario. A large model trained with a large amount of text might be very effective. For example, OpenAI's GPT-4 exhibits great performance on assessment benchmarks through natural language, but it does not mean that it is suitable for all types of network service requirements. A larger model usually implies higher training and running costs. It is estimated that training an LLM with hundreds of billions of parameters may cost hundreds of millions of US dollars. The performance of a large model depends to a large extent on the quality of training data. Therefore, to be used in use cases, large models need to be trained with high-quality domain-specific data.

However, it is impractical for a large model to work with resource constraints due to its high computing and storage requirements. A lightweight model that can balance between performance and resources may stand out. For example, for a customer service system that requires real-time feedback, a large model that responds quickly and only requires a moderate amount of resources may be preferred to a large model with slightly higher accuracy but slow response. Selecting a GenAI large model requires the consideration of multiple factors, including service requirements, cost-effectiveness, data quality, deployment environment, and security compliance. CSPs should conduct multi-dimensional analysis to ensure that the selected large model meets the needs of their service scenarios.

Impacts

To meet the service requirements for higher efficiency and lower costs in network maintenance, optimization, and operations, the use of GenAI will help to enable more convenient, flexible, and diversified solutions.

1. GenAI makes it possible to break down network intents into sub-intents and consolidates the paradigms of human-machine intent interaction and machine-machine intent understanding.

LLMs are very effective at understanding user intents, accurately generating content, and distributing tasks downstream. Their natural language interfaces have become de-facto standards for human-machine interaction in this intelligence era. Large models are trained through tool learning and provide APIs without training and accurate calling service, driving a paradigm shift in intelligent system integration. Specifically, large models can call APIs based on user intents and orchestrate APIs based on user objectives. They also ensure intent understanding, objective breakdown, task orchestration, execution, and reflection in the network domain, as well as unprecedentedly smooth interaction between humans and systems.

2. GenAI works with other AI technologies to improve the efficiency and effect of task execution in network maintenance and optimization.

In network maintenance scenarios, such as fault diagnosis and risk identification, GenAI and specialized network AI algorithm models automatically analyze alarms, indicators, and logs to detect network fault events, diagnose the root causes of faults, generate fault analysis reports, quickly match historical cases of the same type, and accurately recommend fault rectification solutions. In this way, the models help achieve self-diagnosis of network faults and potential risks, as well as automatic closure of network O&M tickets, improving network O&M efficiency.

To achieve online network self-optimization, GenAI can work with traditional AI algorithm models to improve work efficiency by replacing manual 24/7 monitoring, automatically detecting deterioration, generating optimization policies, generating and checking network configurations, and simulating and verifying the optimization policies. In the process, engineers can predict network status changes through network digital twins, compare network load and traffic changes before and after the optimization policies are delivered to ensure the multi-objective network optimization effect, network bandwidth, latency, and rate, and maximize network resource efficiency.

3. GenAI poses new requirements for AI computing on CSPs' network devices.

Some AI applications highly rely on service response of seconds or even milliseconds. To help obtain precise network status data and analyze the massive amount of data in real-time, AI computing should be at work on both the NE side and the management and control unit. Through intelligent analysis and near-data computing, AI computing can give these AI applications a competitive edge with new capabilities, such as fault self-diagnosis, network self-optimization, link adaptation, spectrum self-awareness, user experience assurance, and real-time service simulation.

Suggestions

1. Build industry and domain corpora and provide high-quality data for the applications of GenAI large models.

Domain-specific data governance: Identify and collect available data sources of the industry or domain, including internal databases, public datasets, and experience and knowledge documents. Clean data by deleting redundant data and filtering out invalid, outdated, and inaccurate data to ensure data quality.

Domain-specific data labeling: Determine a proper amount of data, classify and label the data based on expert experience and knowledge in the industry to improve data explainability and accuracy, and integrate the domain-specific data into GenAI models to facilitate model pre-training, post-training, or retrieval augmented generation (RAG).

Data security compliance: Ensure that all collected and used data comply with local privacy and data protection regulations and laws as well as AI acts. Where necessary, take security measures, such as data encryption, access control, and auditing, to protect data from unauthorized access or disclosure.

2. Select AI technologies based on specific scenarios for a systematic solution.

For specific use cases, such as network prediction, network maintenance, network optimization, and autonomous decision-making, select appropriate GenAI models and classic AI solutions to solve problems systematically. For these use cases, classic AI algorithm models are more reliable, controllable, easy to understand, and resource-saving than a solution that depends only on a GenAI model. For example, a machine learning model can be used for deterministic classification; a reinforcement learning model for deterministic optimization; and an inference system based on a logical rule and a knowledge graph for deterministic demarcation and locating.

GenAI integrates open network APIs and domain-specific AI algorithm models through language user interfaces (LUI) to build more intelligent and professional network agents and make up for its limitations, as well as improve the accuracy, real-time performance, explainability, and security controllability of all AI solutions.

3. Deploy dedicated AI computing power close to the service running side to back up the performance of models of proper scales.

Upgrade AI servers or AI intelligent boards for network management and control units and NEs to realize efficient and real-time AI computing capabilities for network services, such as converged awareness of network status, root cause analysis (RCA), and fault recovery assurance. Provide professional and 24/7 online network O&M assistants for O&M personnel and FMEs as well as more intelligent and high-quality zero-fault experience for users.

During the deployment of GenAI-based network assistants and communications agents on the live network, GenAI large models must meet use case requirements. A model of a large scale occupies more computing resources and consumes more energy. A model of a small scale may not have sufficient resources for complex planning and inference. Considering this, it is recommended that the number of parameters of a GenAI large model be somewhere between 7 and 38 billion.

Trend 7

Multi-Agent Collaboration Will Be a Key Technology for Achieving Highly Autonomous Networks in CSP Networks, and It Is Gaining Attention in Industry Research



Technology Insights

Multi-agent technology has sparked significant research interest from both academia and industry, particularly as networks continue to evolve in complexity. This technology is key to achieving layered network autonomy and cross-domain E2E closed-loop management, a trend that is rapidly gaining momentum.

Agent and multi-agent topics are dominating major AI summits such as AAAI, ICML, ICRL, and IJCAI in 2024. The number of papers on multi-agent collaboration is also on the rise. The research covers various directions, including multi-agent reinforcement learning, collaboration mechanisms, communication interaction, competition and confrontation, and cross-domain application. Researchers are exploring the use of LLM-based multi-agent systems, which enable multiple agents to collaborate and leverage their strengths to tackle complex issues. A 2024 paper titled "Large Language Model-based Multi-Agents: A Survey of Progress and Challenges", based on the review and analysis of 81 papers in the multi-agent field, defines LLM-based multi-agent systems. Open-source multi-agent frameworks, such as Microsoft's AutoGen, MetaGPT ("Meta Programming for A Multi-Agent Collaborative Framework," accepted for oral presentation at ICLR 2024), Agents, Camel, and ChatDev, are rapidly evolving, significantly enhancing our understanding and exploration of multi-agent communication and collaboration.

Multi-agent collaboration technology offers a promising approach to enhance cross-domain fault demarcation and locating. CSPs' cross-domain fault agents and vendors' single-domain fault agents have unique strengths. By establishing effective southbound/northbound collaboration between them (including objective breakdown, negotiation, and interaction interfaces), we can combine these strengths to achieve cross-domain fault closure. For instance, it was initially difficult to determine if power outages or

endpoint transmission issues caused BS out-of-service faults. This difficulty was further complicated by the inaccurate association of wireless and transmission resources, leading to a low success rate of ticket dispatch based on root causes and repeated ticket dispatch on both the wireless and transmission sides. On average, over 50 tickets were dispatched daily for BS out-of-service faults caused by transmission issues at the aggregation layer or higher. The wireless and transmission O&M teams had to collaborate extensively and manually transfer the list of faulty BSs. Therefore, CSPs are in urgent need of improved cross-domain fault demarcation and location capabilities and enhanced efficiency.

The TM Forum released the Autonomous Networks Level 4 industry blueprint – high-value scenarios report in June 2024. This report explains how multi-agent collaboration enables E2E closed-loop autonomy in complex scenarios, including E2E customer complaint handling, cross-domain fault demarcation, E2E service assurance, and wireless network collaborative optimization. China Unicom Research Institute, China Mobile Information Technology Center, China Academy of Information and Communications Technology (CAICT), and China Telecom Research Institute collaboratively launched the Catalyst project "GenAI empowers computing force network." The project aims to efficiently schedule cloud-edge-device computing network resources based on customers' personalized service requirements, providing one-stop intelligent service support. E2E computing network service provisioning involves multi-agent collaboration. To facilitate the E2E orchestration of computing network services, it is essential to explore objective breakdown and negotiation between the master and other agents and to combine the strengths of all the agents involved.



Impacts

1. Multi-agent collaboration enables E2E closed-loop management in complex scenarios.

By leveraging the strengths of multiple agents, this approach can overcome individual limitations, accomplish complex tasks in CSP networks, and deliver excellent flexibility and scalability. It provides key enabling technologies for E2E closed-loop management

of complex telecom networks. The trend is to evolve from a single-agent to a multi-agent collaboration, where multiple agents collaborate and communicate to streamline various CSP processes and cover high-value scenarios. Multi-agent collaboration also redefines the mode of system integration. Traditional system integration relies on atomic APIs and requires a long process to launch a new function: multi-layer and multi-system joint design, API development, onsite integration, and joint commissioning, which can take months to complete. Multi-agent collaboration upgrades structured APIs to natural language interfaces through which OSS agents negotiate intents with single-domain agents for objective breakdown and collaborative closed-loop management, eliminating the need for system integration.

2. Multi-agent systems are expected to become new communication objects and drive innovation in communication network services.

AI agents are evolving rapidly, and we can expect to see many more agents in the future, including intelligent terminals, embodied intelligence (such as robot dogs and intelligent service robots), virtual intelligent assistants, and digital persons. These agents will introduce new communication objects and service scenarios into future networks. Future networks must offer new connection services for agents with varying forms and capabilities, including digital identity authentication, interconnection and interworking, and task collaboration.

3. Multi-agent collaboration still faces technical challenges.

(1) Currently, there are two technical approaches in the communication industry: multi-agent collaboration based on reinforcement learning and multi-agent collaboration based on LLMs. The industry is still exploring the future evolution and potential technical complementarity of these two approaches.

(2) Multi-agent systems face unique capability requirements and key technical challenges in communication and collaboration. The industry is still exploring solutions, and no consistent methodology or framework is currently available. From a collaboration perspective, the industry must implement objective breakdown, coordination, and conflict prevention between different agents. Multi-layer collaboration and message transfer between agents may magnify hallucination layer by layer. From

a communication perspective, the existing communication mechanisms (such as decentralized communication, centralized communication, and hybrid communication) have their advantages and disadvantages. CSPs need to carefully select a communication mechanism based on the application scenario characteristics of agents. The natural language–based communication message model is yet to mature.

(3) Research on multi-agent standards is still in its early stages, and no unified standards or specifications are available yet. Researchers are still working on the requirements for interoperability, semantic transfer accuracy, and compatibility between different systems.

Suggestions

The evolution from a single agent to multi-agent collaboration aims to enable E2E closed-loop management in complex scenarios of carrier networks and accelerate the transformation toward high-level autonomy. This is the overwhelming trend and the ultimate goal. However, this transformation faces unique challenges due to unclear high-value scenarios, primitive industry standards, and immature key technologies. In the first stage toward AN Level 4, all industry stakeholders must collaborate and accelerate the implementation of agent-based single-domain closed-loop management in high-value scenarios. As a key technology in the second stage toward AN Level 4, multi-agent collaboration lags in commercial implementation in the telecom field. The telecom industry and academia should collaborate to overcome the challenges of multi-agent collaboration technology and advance its development and application in the telecom field. Specific suggestions are as follows:

Suggestion 1: CSPs should establish business scenarios and service requirements for multi-agent collaboration and conduct technical research on high-value scenarios.

Suggestion 2: The telecom industry should start standard planning and research justification to advance the standardized development of multi-agent collaboration technologies. This is crucial for the application of multi-agent collaboration technology.

Suggestion 3: The academic community, including universities, should research key technical directions and challenges of multi-agent collaboration, including coordination mechanisms, communication protocols, and interaction interfaces.

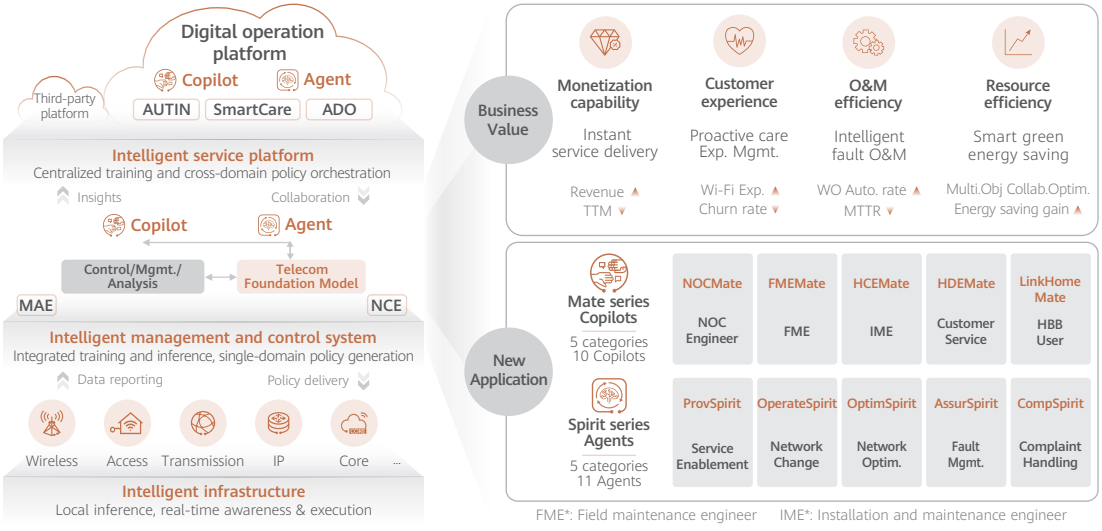


03

Autonomous Driving Network

Huawei's Autonomous Driving Network (ADN) Enables the Transformation Toward Highly Autonomous Networks and Helps CSPs Evolve Toward Full Intelligence

AN has become an industry consensus in recent years and is the optimal choice for CSPs to achieve comprehensive intelligent transformation. ADN is one of the four core strategies of Huawei's Communications Network 2030 and is also Huawei's solution for the global AN industry. ADN aims to develop self-fulfilling, self-healing, self-optimizing, and autonomous networks based on connectivity and intelligence. Based on the principles of single-domain autonomy, cross-domain collaboration, value-driven approaches, visualization, and productization, Huawei collaborates with CSPs and enterprises to develop self-configuration, self-healing, and self-optimizing capabilities,



ADN solution overview

offering a superior experience of zero-wait, zero-touch, and zero-trouble to consumers and public-sector and enterprise customers.

ADN delivers critical value to CSPs from the aspects of business, experience, efficiency, and energy efficiency.

1. Monetization capability+: Network capabilities can be provided as services to help customers improve monetization capabilities, enable network as a service (NaaS), implement zero-wait service provisioning, shorten product time to market (TTM), and facilitate agile rollout in various industries. For instance, fast service provisioning can be implemented without visiting enterprise campuses, reducing the provisioning duration by more than 75% and helping enterprise users quickly, accurately, and stably provision services.

2. Customer experience+: Key indicators such as the service quality fulfillment rate and complaint handling timeliness need to be improved to advance customer experience and satisfaction and deliver a superior experience across the entire lifecycle. For instance, in-home broadband experience assurance scenarios, the home visit rate is reduced from 65% to 10%, user complaints are reduced by 60%, and the average revenue per user (ARPU) is improved by differentiated broadband.

3. O&M efficiency+: In-depth AI applications, including predictive maintenance and dialog-based O&M, are utilized to significantly save manual workload, reduce the operation time per unit, and enhance O&M efficiency. For example, in IP network fault management scenarios, the root cause of a fault can be quickly diagnosed, reducing 90% of alarms. The ticket processing duration is shortened from 3.5 hours to 30 minutes.

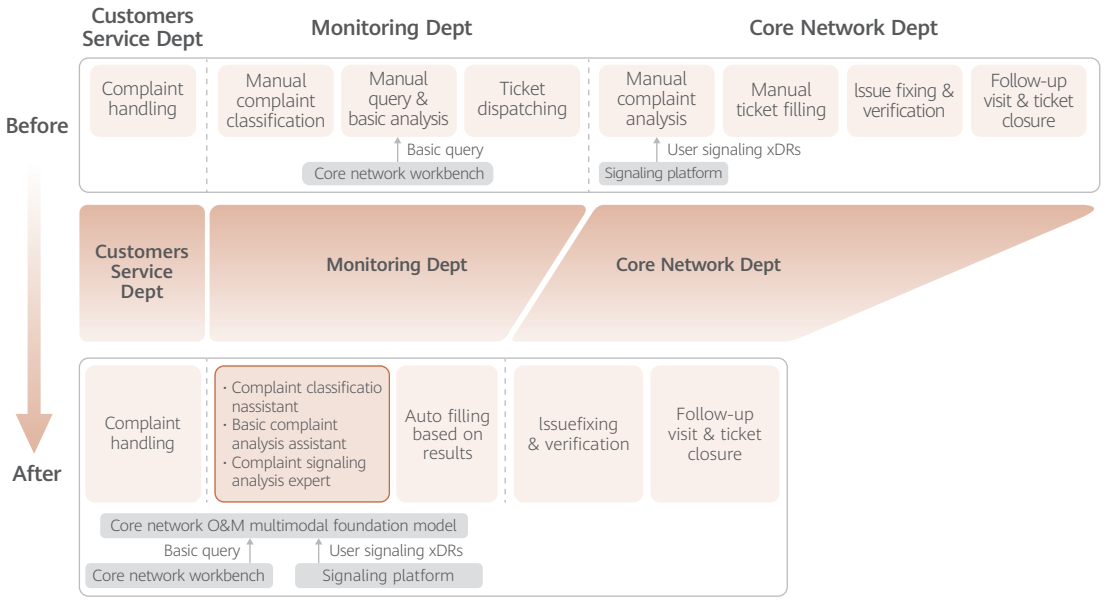
4. Resource efficiency+: Energy consumption of network devices needs to be reduced through approaches such as multi-dimensional collaboration. In addition, dumb network resources need to be visualized to improve data accuracy, thereby guaranteeing precise allocation of network resources. Additionally, network paths need to be optimized to prevent network congestion and improve network resource utilization. For example, wireless network optimization has achieved a 35% increase in network energy savings and a 20% increase in the cell rate.

Huawei has developed the ADN Level 4 solution for high-value scenarios based on various key technologies, including the telecom foundation model, converged awareness, and digital twin. The solution offers crucial application capabilities through role-based copilots and scenario-specific agents to help CSPs and enterprises improve employee abilities, enhance user experience, and deliver more significant value through digital intelligent network productivity. The solution enables the evolution toward Highly Autonomous Networks in numerous typical scenarios.

Core Network Complaint Handling: CompSpirit Redefines the Complaint Handling Process and Enables Efficient O&M for CSPs

As the core network continuously evolves to cloud-native and full convergence, it is becoming even more complex. In particular, the number of managed objects and associated network risks are increasing exponentially, while the SLA requirements for problem handling are rising year by year. Traditional core network O&M methods cannot properly support service development. In complaint handling, for example, complaint analysis may involve a range of different combinations of NEs, interfaces, protocol types, and signaling messages. Among these, signaling analysis is typically the most challenging. A single signaling message can involve the exchange of more than 100 machine language messages, which need to be interpreted by experts with years of experience in the core network field. This is a skilled task that can be especially time-consuming when performed manually.

To handle complex complaints, Huawei has launched CompSpirit, which is based on the multimodal large model for core network O&M. CompSpirit can quickly comprehend complaint intents, demarcate complex processes, simplify operation processes, and expedite the closure of the entire complaint handling process. The proportion of complaint tickets with the handling process moved forward has increased to more than 20% on average, and the E2E complaint ticket handling duration has decreased from 14.6 hours to 5 hours, improving the efficiency by 64%.



» **Accurate complaint classification:** The data preprocessing and parameter extraction capabilities of the large model are used to accurately extract keywords from complaint tickets, and the BERT model is used to classify complaints. Based on seven types of high-frequency complaint scenarios, more accurate complaint classification is implemented, improving the complaint classification accuracy from 40% to over 90%. This has successfully eliminated the breakpoints between complaint classification and demarcation. The complaint demarcation API is automatically invoked to automatically backfill trouble tickets. In this way, the analysis of complaints is shifted forward from the core network department to the monitoring department.

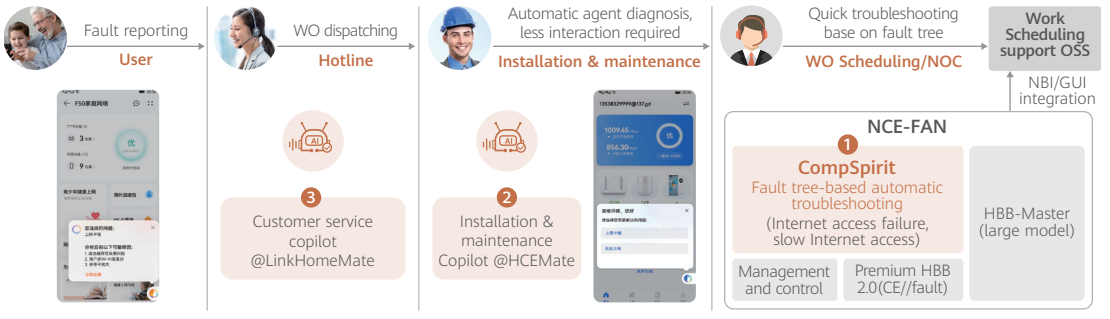
» **Automatic signaling parsing:** The unique signaling large model adopts a three-layer modeling approach to signaling behavior, allowing for a better understanding of service logic. By merging signaling and semantic convergent coding, the large model enables natural language-based Q&A signaling analysis, which simplifies the process of signaling analysis. The model is capable of analyzing signaling issues layer by layer, just like human experts. Through dialog-based Q&A interactions, even non-experts can complete signaling analysis within 5 minutes and offer recommended root causes and related cases. Such results are comparable to having core network O&M experts with more than five years of experience. What's more, the time required for a single signaling analysis phase has been reduced from 4 hours to just 5 minutes.

The solution has now been implemented in the production process of China Mobile Zhejiang, effectively adding more than 30 experienced digital employees to the team. This has resulted in a redefined complaint-handling process and a positive cycle of core network O&M transformation.

Optical Access Poor-QoE Identification: CompSpirit for Intelligent Poor-QoE Identification and HCEMate for Accurate Onsite Maintenance, Greatly Improving User Experience

As the digital economy rapidly develops, premium service experience has become a core demand of home broadband users, with more users willing to pay higher prices for a better broadband experience. CSPs are also keen to enhance user stickiness, increase revenue, and explore new marketing opportunities by providing premium experience-based operations. These usher in a new era of experience-based operations for home broadband services. As home network users are increasingly demanding better broadband access experience, their various STAs, and apps demand differentiated user-specific services from optical access networks. Traditional manual O&M is no longer up to the task.

Huawei's IntelligentFAN solution offers HCEMate for field engineers, LinkHomeMate for home broadband users, and CompSpirit for experience assurance. This solution helps automatically identify, intelligently diagnose, and quickly rectify user experience issues, ensuring a premium home broadband service experience and significantly improving operations and O&M efficiency.



Solution overview

» **CompSpirit for experience assurance:** automatically demarcates and locates faults based on the fault tree.

In the case of poor home broadband experience, CompSpirit automatically extracts the network KPI and KQI data at the moment when the faults occurred with the help of time-space correlative analysis, identifies the fault root causes within 30 seconds based on the fault tree-based diagnosis algorithm, and generates feasible solutions. About 30% of these faults can be automatically rectified and closed remotely, while the rest are handled by the CSP's NOC onsite. This solution helps rectify faults before users file complaints, reduces onsite visits by 30% and fault diagnosis time by 80%, and increases the E2E troubleshooting efficiency by 50%.

» **HCEMate + CompSpirit for automatic troubleshooting:** provide intelligent Q&A based on the large model and troubleshooting assistance.

HCEMate can automatically provide installation and maintenance knowledge and guidance for field engineers when they are working onsite. Additionally, CompSpirit can be automatically invoked to quickly locate faults and generate solutions based on the fault tree. The efficient collaboration of agents greatly improves the operation efficiency of field engineers and reduces the time required for handling home network issues by 80%.

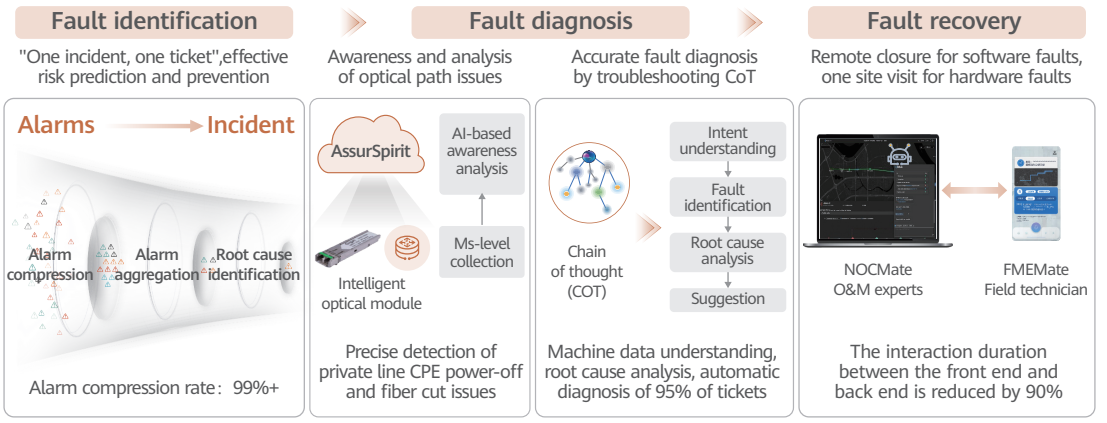
» **LinkHomeMate + CompSpirit for self-healing:** enables users to rectify faults by themselves.

When a fault affecting the broadband network experience is detected, LinkHomeMate can quickly diagnose the fault. CompSpirit can be automatically invoked to quickly demarcate and locate the fault and determine whether the fault can be rectified by the user on their own. If it can, CompSpirit proposes a troubleshooting solution and instructs the user to quickly rectify the fault. In this way, 30% of faults can be resolved before users file complaints. If the fault cannot be rectified by the user, the user can report the fault to the CSP, improving the onsite troubleshooting efficiency by 50%.

IP Network Troubleshooting: FMEMate and AssurSpirit Help CSPs Improve Quality and Efficiency

IP networks are complex, with multiple layers, sophisticated routing protocols, and dynamically changing routes. A single network failure can generate multiple alarms. Traditionally, identifying faults from multiple alarms relied on predefined rules, leading to repeated ticket dispatch. O&M personnel had to troubleshoot faults based on their experience and system functions. For faults requiring field work, field maintenance engineers (FMEs) had to collaborate with NOC personnel remotely, but the entire troubleshooting process lacked automation, reducing O&M efficiency.

To address this, the ADN solution, based on key technologies such as digital twin and telecom foundation model, offers a scenario-specific troubleshooting agent called AssurSpirit, as well as two digital assistants, NOCMate and FMEMate, for NOC personnel and FMEs. These tools enable automatic fault analysis, significantly improving O&M personnel's troubleshooting efficiency. AssurSpirit automatically identifies root alarms by filtering, aggregating, and intelligently correlating all alarms. It uses intelligent optical modules to collect ms-level optical power data and large models to analyze chains of thought, helping break down complex issues. By calling NMS functions, AssurSpirit locates root causes and offers troubleshooting suggestions. By calling tools or interfaces, it automatically resolves software commissioning issues. NOCMate and FMEMate facilitate onsite fault rectification. Their intelligent Q&A and auxiliary troubleshooting capabilities help FMEs quickly rectify faults without the need for collaboration.

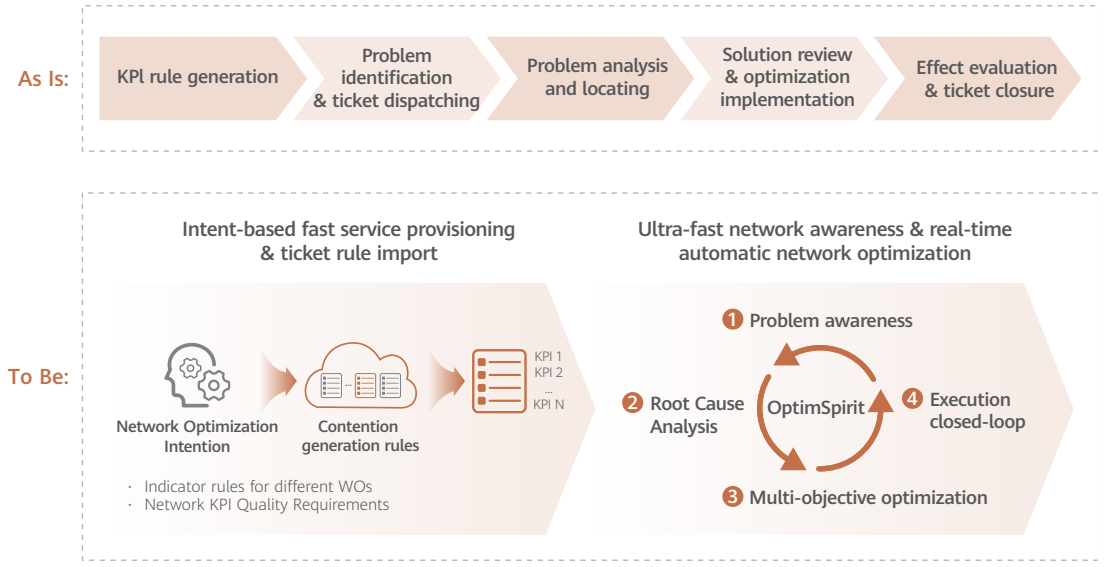


This ADN solution has been successfully implemented at China Mobile Guangdong. It interworks with upper-layer troubleshooting and ticket systems, automating the entire process from fault identification, locating, and rectification to verification. This has resulted in a 15% reduction in trouble tickets, an 89% decrease in fault locating time, and the creation of over 100 digital employees.

Wireless Network Optimization: OptimSpirit Redefines the Routine Network Optimization Process and Implements Real-Time Automatic Optimization

Nowadays, CSPs place increasingly high requirements on wireless network performance, and the traditional network optimization process often faces many challenges. These include high labor costs, low efficiency, dependence on expert experience, delayed responses, and support for only single-objective optimization. As the network scale expands, the network architecture has become increasingly complex, driving up O&M costs. CSPs are in urgent need of more intelligent approaches to reduce costs and boost efficiency.

To enhance routine network optimization scenarios, Huawei has launched OptimSpirit, which is based on the wireless intelligent agent. It integrates functions such as intent-based service provisioning, ultra-fast network awareness, intelligent analysis, and automatic optimization. The aim is to redefine the routine network optimization process by replacing the traditional problem-driven and human-centered handling process with the self-discovery, self-analysis, and self-closed-loop handling process. With the implementation of real-time dynamic network optimization, the solution has increased the problem self-handling rate to 20%, resulting in improved efficiency, cost reduction, and network stability for CSPs.



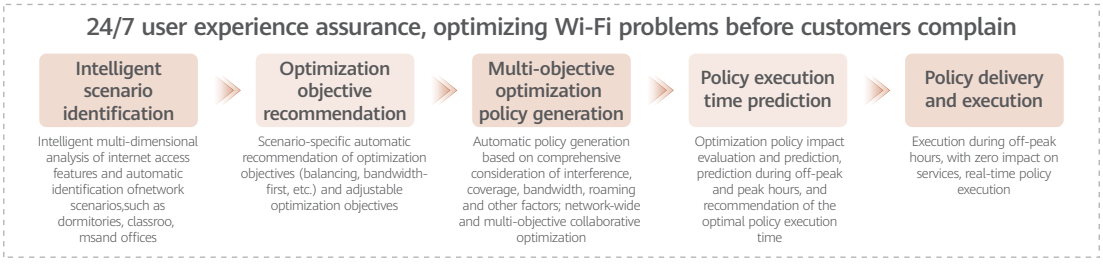
» **Dynamic real-time network awareness:** Based on the time sequence model of network KPI changes and network computing power, the system can identify any unusual bursts of traffic or performance deterioration across the network. The identification results are then promptly reported to the upper-level management and control node. The new intelligent architecture allows network layer and NE layer resources to converge and coordinate in mere seconds, marking a significant advancement over traditional network monitoring capabilities. Furthermore, by analyzing short-term trends, it is possible to predict network KPIs in future periods and proactively mitigate potential network or user experience risks.

» **Real-time multi-objective optimization:** L2 parameter decision-making and optimization based on online iterative reinforcement learning can be carried out by either the environment agent model or online decision-making model. The environment agent model uses target KPIs as the training input after live network data preprocessing and feature processing. The online decision-making model generates recommended configurations as output after training on the live network data and the augmented dataset obtained through simulation. OptimSpirit is capable of automatically recommending optimization configuration combinations based on user intents and can deliver the optimal solution. This enables single-objective to multi-objective collaborative optimization. As a result, weak coverage is improved by 10%, and the proportion of low-speed cells is reduced by 20%, ultimately achieving automatic network optimization and an enhanced service experience.

Campus Network Fault Closed Loop: OptimSpirit Enables 24/7 Network Experience Assurance and Resolves Problems Before Users Complain

The construction of campus networks is a major driver for quality education. Wi-Fi access is used on most campus networks, especially in student dormitories. On these networks, many services, such as online courses and videos, are running concurrently, necessitating the need for Wi-Fi experience assurance. However, the short distance between dormitories, numerous obstacles, dense population, and unauthorized interference sources often cause problems such as Wi-Fi interference, unstable signals, and slow network speed. These problems are often passively identified and located based on complaints, resulting in low handling efficiency.

To address the challenge of difficult Wi-Fi troubleshooting in university dormitories, Huawei launched the campus network optimization agent based on the telecom foundation model and leading algorithms, including the multi-objective Wi-Fi optimization algorithm. The agent integrates campus scenario identification, optimization objective recommendation, multi-objective optimization policy generation, policy execution time prediction, and policy delivery and execution. It provides 24/7 online protection for Wi-Fi experience on campus networks, identifies network risks in advance, and implements proactive optimization, transforming passive troubleshooting into proactive assurance and significantly improving service continuity and customer satisfaction.



This solution has been piloted in a top university in China. It enables the dormitory Wi-Fi network to automatically resolve non-hardware issues, improving the fault interception rate to over 80%, significantly reducing the number of fault tickets, and enhancing network experience and customer satisfaction.

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P. R. China
Tel: +86-755-28780808
www.huawei.com

Trademark Notice
HUAWEI, HUAWEI , are trademarks or registered trademarks of Huawei Technologies Co.,Ltd
Other Trademarks,product,service and company names mentioned are the property of thier respective owners

GENERAL DISCLAIMER
THE INFORMATION IN THIS DOCUMENT MAY CONTAIN PREDICTIVE STATEMENT INCLUDING, WITHOUT LIMITATION , STATEMENTS REGARDING THE FUTURE FINANCIAL AND OPERATING RESULTS, FUTURE PRODUCT PORTFOLIOS, NEW TECHNOLOGIES,ETC. THERE ARE A NUMBER OF FACTORS THAT COULD CAUSE ACTUAL RESULTS AND DEVELOPMENTS TO DIFFER MATERIALLY FROM THOSE EXPRESSED OR IMPLIED IN THE PREDICTIVE STATEMENTS. THEREFORE, SUCH INFORMATION IS PROVIDED FOR REFERENCE PURPOSE ONLY AND CONSTITUTES NEITHER AN OFFER NOR AN ACCEPTANCE. HUAWEI MAY CHANGE THE INFORMATION AT ANY TIME WITHOUT NOTICE.
Copyright © 2024 HUAWEI TECHNOLOGIES CO., LTD. All Rights Reserved.
No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co.,Ltd.