HUAWEI

Version 2024

ICT Services and Software 2030

Building a Fully Connected, Intelligent World

Overview

As the communications industry evolves from 2G to 5G, the ICT service and software industry has also undergone an inter-generational upgrade toward standardization, digitalization, and being tool-based. With the rise of new technologies such as Generative AI (Gen AI) and digital twins, the transition from digitalization to intelligence has become immensely popularized. By 2030, AI will be ubiquitous. We will experience more changes: the intelligent and ubiquitous sensing of infrastructure will become a necessity, large models will make progress toward AGI and think like people do, service models will pivot from majority manual labor to majority machine, and enterprise marketing and enablement methods will become increasingly real-time and agile.

Over the next decade, an intelligent transformation will sweep through various industries, echoing the industrial revolutions in the 20th century. Gen AI will empower machines to learn and think with human-like intelligence, catalyzing a pivotal change of production and redefining the work and life for every enterprise, family, and individual, just as the steam engine and light bulbs did back in their times.

Contents

01

Macro Trends and Prospects

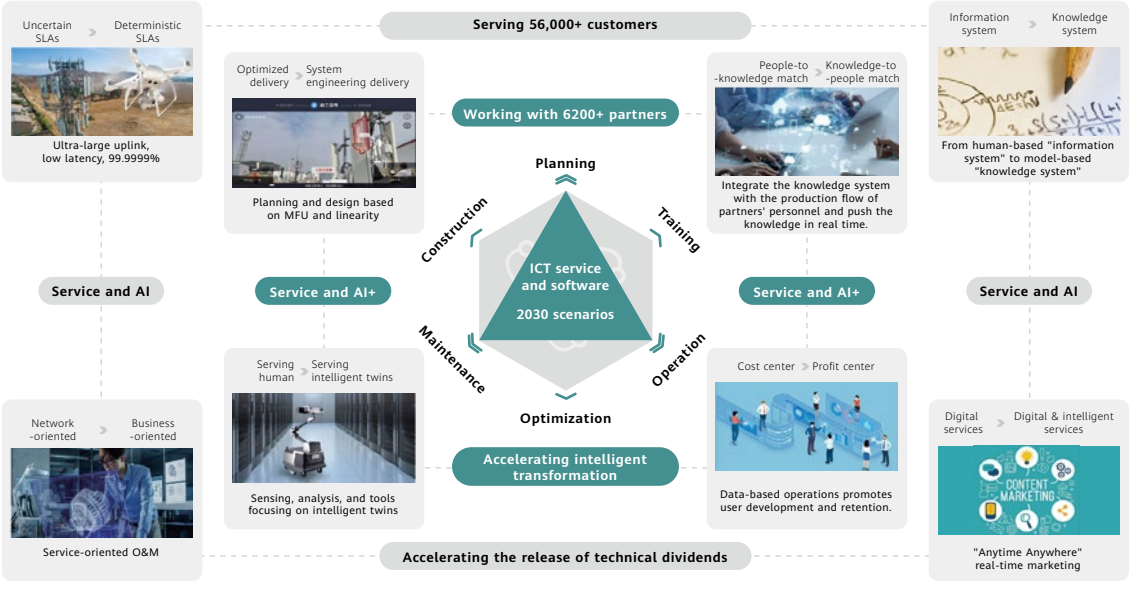
New Technologies, Business, and Models Bring Infinite New Possibilities and Uncertainties Alike

As we rapidly approach an intelligent world, a distinct development trend towards digitalization, intelligence, and low carbonization has emerged. Thanks to certain technologies like naked-eye 3D, AI backpack, autonomous driving oriented to the B2C business, unattended manufacturing factories oriented to B2B business, and mechanical hands and arms for smart mining and ports, forward-thinking businesses are the ones making this happen. With Gen AI as a representative, new technologies such as large models, AI, 5G-A, ultra-large computing clusters, liquid-cooling data centers, digital twins, and agents are making rapid progress. New models of knowledge and data management, AIOps, collaboration between small and large models, as well as AI for Network, and Network for AI have all recently emerged. All these elements work together to drive innovation and the emergence of new, intelligent businesses, sparking new visions and opportunities.

To introduce each new generation of technology and model into the production environment and unleash new productivity, it is essential to achieve continuous evolution alongside the existing business production environment. We must effectively manage complexities and uncertainties, as well as enable orderly evolution of ICT infrastructure throughout its entire lifecycle. It is important to promptly satisfy the new demands for ICT infrastructure driven by new businesses and experiences, foster innovation, and upgrade experiences continuously. The ultimate goal is to maximize investment returns and accelerate the industry's digital transformation. As AI pioneer Fei-Fei Li puts it: "AI is a pervasive technology that will permeate every aspect of our lives and industries like water." As new services, technologies, and models rapidly evolve, the ICT service and software industry will be confronted with unprecedented uncertainties. To unlock business value and

technical dividends, we must harness the power of these innovations while leveraging new technologies to sustain a competitive edge throughout the entire lifecycle. Here are two possible approaches:

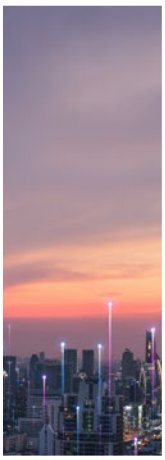
ICT Service and Software 2030 – Future Scenarios: AI+ Changes Service Modes, and + AI Brings New Scenarios

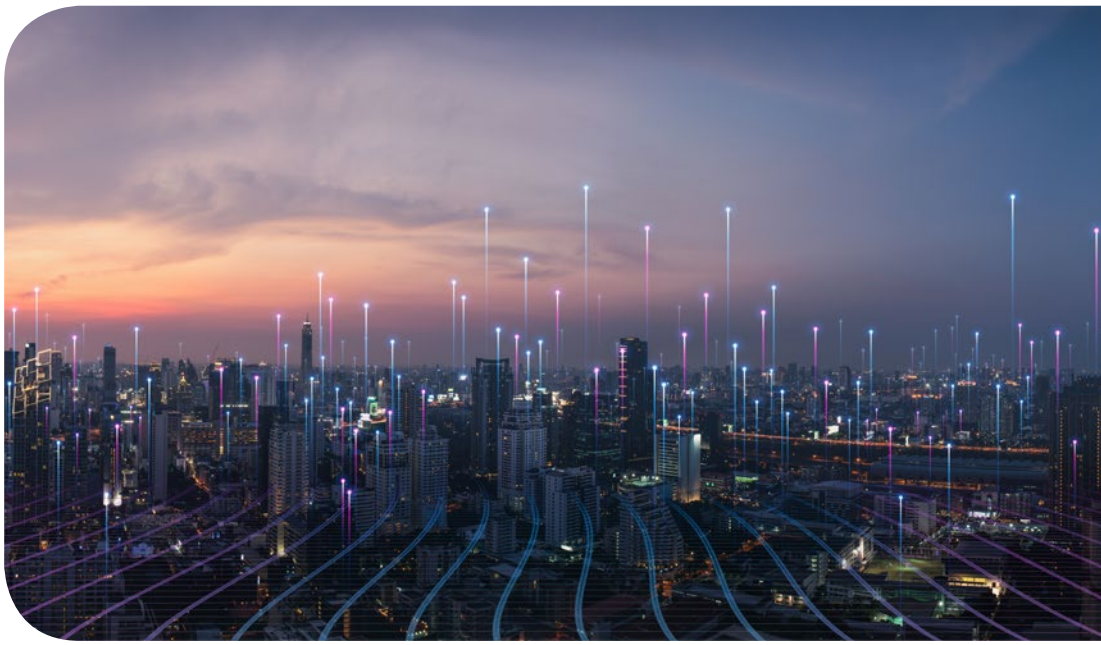


1. Services and AI: Consider what services need to do for better AI development. For example, with the rise of intelligent connectivity of everything, new service level agreements (SLAs) have introduced uncertainties, while network faults are affecting a wider range of services. In this complex landscape, how can we transition from network-centric to business-centric O&M? Data and knowledge management, as a crucial capability for General AI and large models, will redefine the way people learn and are empowered. How do we navigate the future development of ICT talent and services that AI needs?

2. Services and AI+: Consider what AI needs to do for better service development. For example, large models, robots, and embodied AI agents have become an integral part of the future service mode. How do we change the traditional mode of people and platforms through the agent, tool and person mode to improve the efficiency of planning, construction, maintenance, optimization, marketing, and training, as well as reduce costs?

A new horizon is unfolding as we approach 2030. How to utilize definite service capabilities to address numerous uncertain needs of AI and AI+ is a key question that every ICT professional must ponder.





02

Future Scenarios

Planning, Construction, and AI: From Uncertain SLAs to Certain

By 2030, we will have evolved from "connectivity of everything" to "intelligent connectivity of everything". In the digital age, the objects of connectivity are more often people and things in the traditional concept of IoT. According to Gartner and IMT and other organizations' predictions for 2030, AR/VR/MR terminals are expected to occupy 30% of the terminal market. Autonomous driving and unattended industrial manufacturing will also become a reality by 2030. With the help of mechanical arms and hands, unattended factories and mines will become essential for enterprises. On top of that, agents and robots will gradually replace a significant portion of repetitive work currently performed by humans. Huawei predicts that the number of active users for wireless AI agents will reach 6 billion by 2030, By 2030, Huawei predicts that 45% of ICT scenarios will be supported by intelligent agents, and every role will have a dedicated copilot.encompassing not only digital twins in virtual worlds but also embodied AI in the physical world, such as industrial robots, service robots, companion robots, autonomous drones, and self-driving cars. These new service entities will introduce significant uncertainty into future network planning.

We need a network plan based on the new propagation model. Traditional networks, modeled after human interaction, prioritize SLAs that focus on call connectivity rates, call drop rates, mean opinion score (MOS), and call setup delays. The objective is to meet experiential expectations within the bounds of human subjective tolerance, where minor call drops and delays are typically forgivable. However, the network propagation model for intelligent connectivity of things requires providing optimal sensing for machines. As agents increasingly integrate into more B2C lifestyle scenarios and B2B production follows, there is a greater need for deterministic SLAs to ensure ultimate experience in daily life and uninterrupted production. In high-stakes environments like autonomous driving, low-altitude economy, and smart ports, minor setbacks can rapidly snowball into catastrophic failures affecting entire industries,

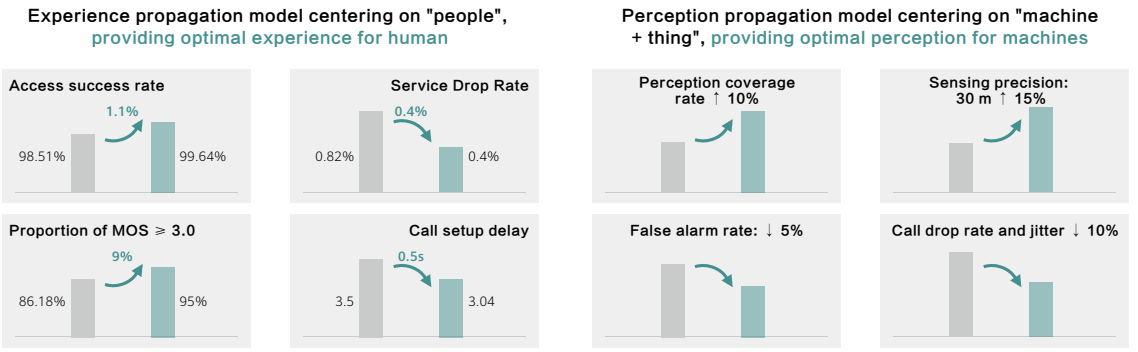
Industry	Business Type	Requirements of Business for the Network														
		Number of Connections per Enterprise Connections per Enterprise	Service Availability (Single User or Single Service)										Security		Trustworthiness		
			Bandwidth Requirement/Single User (Mbit/s)					Service Latency Requirement (ms)									
			B1	B2	B3	B4	B5	T1	T2	T3	T4	T5	S1	S2	M1	M2	M3
			1~10	10~20	20~50	50~100	>100	50~100	20~50	10~20	5~10	<5	Logical Isolation	Physical Isolation	Visibility	Manageability	Operability
Smart Healthcare	16K remote diagnosis and treatment	10					1G										
	Monitoring and Nursing	2,000															
	Holographic Remote Surgery	5					10G										
Smart Grid	Video-based Inspection	-															
	Power Grid Control	-															
	Wireless Monitoring	-															
Smart Manufacturing	Factory Environment	100															
	Information Collection	10,000															
	Operation Control	1,000															

For details, see CAICT 5G E2E Slicing Industry SLA Requirement Research Report.

cities, and global economies. To ensure the high reliability of SLAs, network planning must be compatible with both people-led and machine-led propagation models.

We need a network plan based on the new propagation model. Traditional networks, modeled after human interaction, prioritize SLAs that focus on call connectivity rates, call drop rates, mean opinion score (MOS), and call setup delays. The objective is to meet experiential expectations within the bounds of human subjective tolerance, where minor call drops and delays are typically forgivable. However, the network propagation model for intelligent connectivity of things requires providing optimal sensing for machines. As agents increasingly integrate into more B2C lifestyle scenarios and B2B production follows, there is a greater need for deterministic SLAs to ensure ultimate experience in daily life and uninterrupted production. In high-stakes environments like autonomous driving, low-altitude economy, and smart ports, minor setbacks can rapidly snowball into catastrophic failures affecting entire industries, cities, and global economies. To ensure the high reliability of SLAs, network planning must be compatible with both people-led and machine-led propagation models.

From a business perspective, it is difficult to calculate the return on investment (ROI) for a network that features intelligent connectivity of things, as traditional indicators like packages, DOU, and penetration rates were designed to gauge people-led networks. Each scenario demands a tailored approach, balancing business models and networking needs, necessitating a flexible planning and GTM pace taking into consideration the intelligence level of the corresponding region and city. As a result, refined investment in system integration will have a higher TTM requirement. To gain a market advantage in complex business

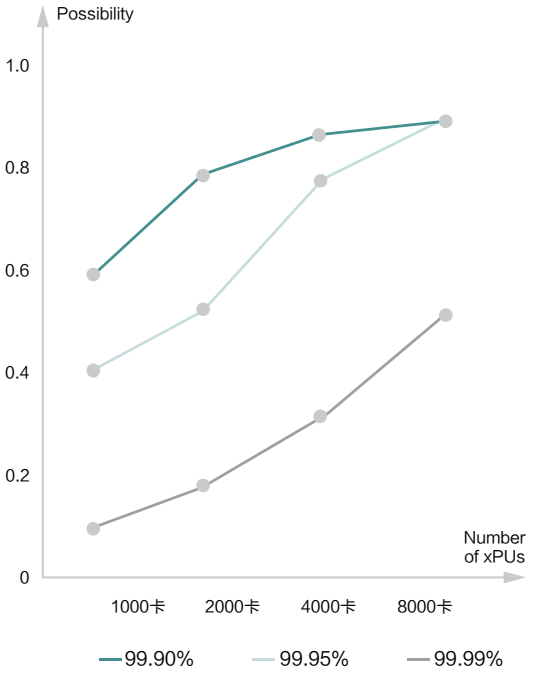


scenarios, it is essential to combine deterministic business scenario SLAs for rapid network upgrade and better ROI. Network planning needs to build real-time network simulation through digital twins, and rapidly define the target network based on the requirements specific to a forward-thinking network construction business scenario. Network planning and design need to simulate the business and network changes in the physical world using digital twins, so as to build human- and machine-led propagation models and network performance simulation and prediction. In addition, planning also needs to be iterated at a small cycle. In this way, we can ensure an accuracy of 99.9% and a 50% shorter TTM.

Planning, Construction, and AI+: From Digital Integration to System Engineering Integration

We are now in the phase of intelligent IT integration. Clusters are scaling up based on scaling laws. Grok 2 and Stargate are now talking about 100,000- and 1,000,000-GPU/NPU clusters. In February 2024, ByteDance unveiled MegaScale, a system featuring 12,288 GPUs, which trained 1,75B models, matching the industry's current peak Google in terms of 10,000-GPU/NPU cluster. ByteDance tried up to 9 optimization methods. In spite of this, ByteDance achieved a Model FLOPs Utilization (MFU) of only 55%, leaving a significant gap to the maximum of 95%. A 1% improvement in MFU will result in over $1 million in cost savings, significantly enhanced performance, and reduced TTM training time. Mason predicts that AI OPEX will rise by 35% compared to traditional computing OPEX, and is expected to increase by over 50% by 2050, driven mainly by water and electricity costs. Intelligent data centers will require AI-driven energy saving based on full-stack DC L1&L2 linkage, as well as high MFU planning.

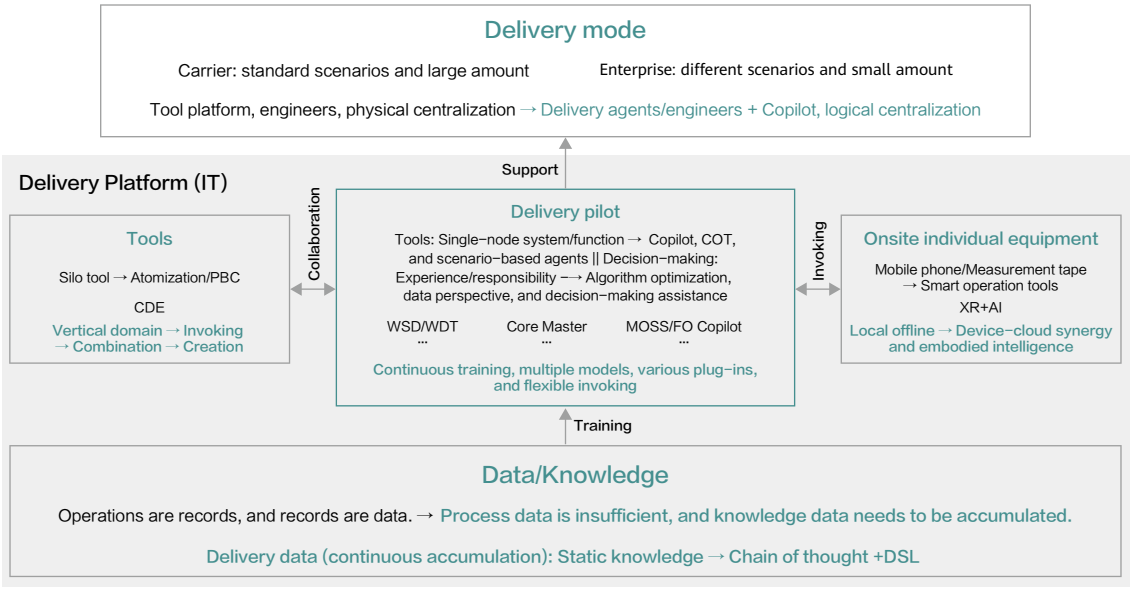
Representative AI Model	Number of Training xPUs
ChatGPT	1000
GPT-4	10,000
Gemini	54,000 TPUs
Grok 2	100,000
AGI	1,000,000 (Stargate project)



AI Business	Computing Power	Network Bandwidth	Network Latency	Bandwidth	Memory Capacity
LLM training	★★★★	★★★★☆	★★	★★☆	★★☆
LLM inference (prefill)	★★★★☆	★	★	★☆	★★★☆
LLM inference (decode)	★☆	★☆	★★★★★	★★★★★	★★★★
Recommendation system training	★★☆	★★★★★	★★★	★★★	★★☆
Recommendation system inference	★	★★☆	★★☆	★★☆	★★★★★

Different AI businesses have unique requirements for building competitive intelligent computing networks. Computing power, network bandwidth, latency, memory bandwidth, and memory capacity needs differ across training and inference scenarios of large language models (LLMs) and small models. The system architecture must be flexible to meet diverse needs. The network provides the foundation for connectivity, allowing for flexible combinations of computing and storage to meet different needs. However, not all five capabilities are needed in a single business scenario at the same time. The network-storage-computing collaborative design, with its characteristics of fully utilizing network to gain extra computing power, utilizing computing to gain more storage power, and utilizing storage to replace computing, is the primary planning direction for future intelligent computing cluster systems. PwC predicts that, by 2030, the potential optimization space for collaborative planning of computing, storage, and networking under the same computational requirements is three times that of the current one. This will make integrated services based on "system engineering" a necessity for meeting future demands for high-complexity MFU and linear scalability.

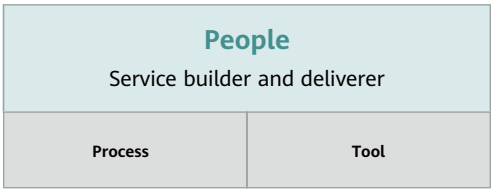
From the delivery mode perspective, future integration services will shift from two-dimensional



to three-dimensional, adding spatial calculation to traditional time (phase and phase effects) and task (outcome and outputs) dimensions to describe services, which involves tracking changes in the system's space over time and tasks using digital twins.

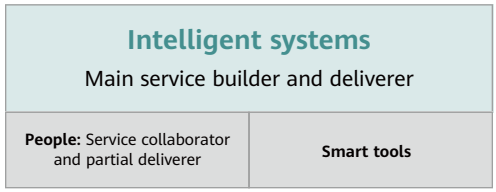
The future delivery model will shift from a traditional, person-plus-tool, physically centralized approach to an agent-plus-copilot, logically centralized model. In the past, delivery was based on the number of sites, (for example, to deliver 3,000 sites, we will need 100 employees and 90 days), and resources were allocated according to the delivery location. The future delivery will be more centralized. Delivery centers will be constructed at the group and province level. Agents and copilots will be used for site surveys, method of procedure (MOP) designs, and original factory configurations. Field work will become more focused and process-oriented. Delivery project managers and technical project managers were crucial in the past, as they possessed project management, key technologies, tool platform capabilities. In the future, models and applications will become the primary builders and deliverers of integrated services. Human-machine collaboration and data-driven approaches will further flatten delivery organizations, resulting in a delivery efficiency improvement of over 50%.

Services focusing on people



- Linear increase of headcount and revenue
- Long service period and small concurrent workload
- High service price, tool-assisted

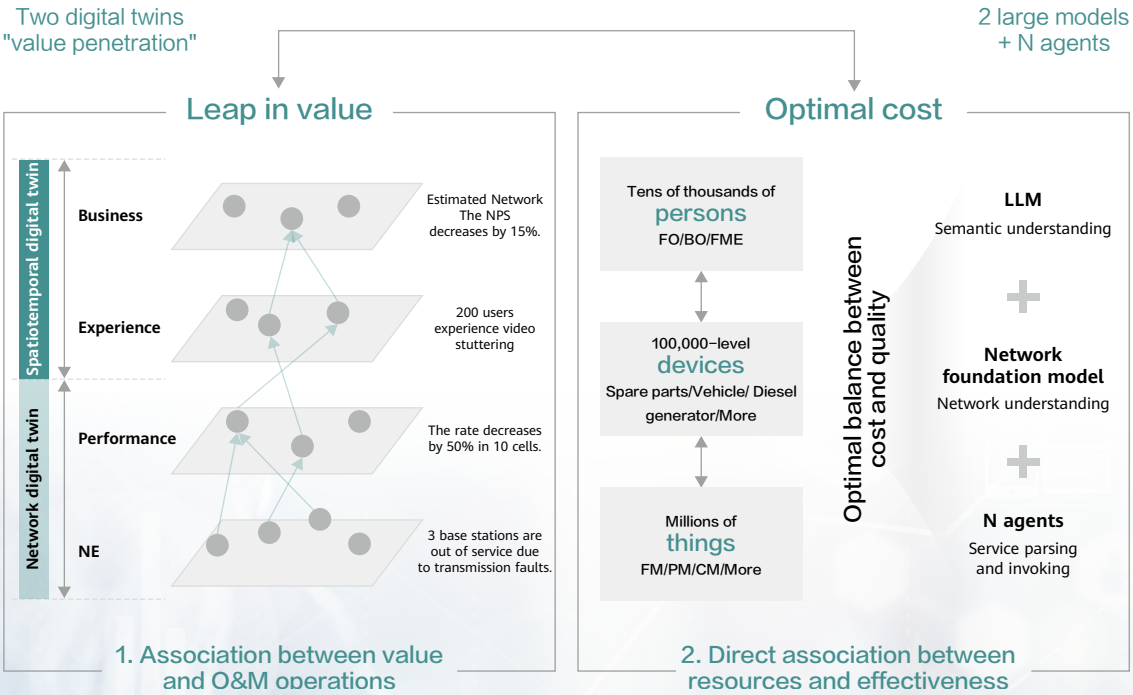
Services focusing on intelligent systems



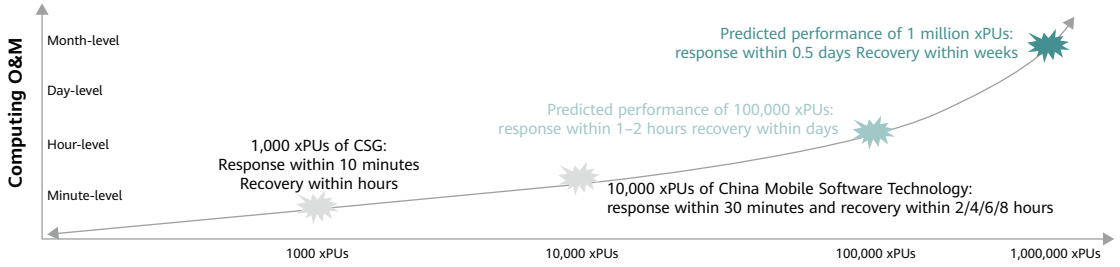
- Use intelligent systems to replace most personnel and upgrade from manual operations to human-machine collaboration.
- Intelligent systems automatically generate processes and data-driven, and multiple concurrent operations are supported.
- Upgrade from functional tools to intelligent tools, improving efficiencyv

Maintenance, Optimization, and AI: From Network-oriented to O&M-oriented

By 2030, as network architecture becomes more complex, O&M objects will include cloud, network, storage, edge, and device. This complexity makes it harder to determine network operations, increasing the need for skilled network change engineers. In recent years, network incidents frequently occur. Compared with five years ago, the proportion of major ICT network faults has increased by 45% in 2024. The primary reason is that as networks become more and more complex, traditional network-based O&M cannot perceive the terminal. It is difficult to connect data and algorithms across the entire stack involving network elements (NEs), performance, experience, and business. With advancements in smart technologies such as Gen AI, large models, and digital twins, the future of O&M will shift from focusing on networks to focusing on business needs.



Computing O&M mode: demarcation efficiency grows exponentially as cluster size increases



Furthermore, as computing power grows according to the scaling law, large-scale model manufacturers now typically use tens of thousands to millions of xPUs. This makes traditional network-based O&M impractical. A report of an Internet vendor on a training cluster involving tens of millions of xPUs shows that over 54 days of training, there were 466 job interruptions, averaging 8 per day. Most interruptions (41%) were caused by software problems, cable issues, and network faults. By 2030, estimated losses from an interruption involving millions of xPUs could exceed CNY100 million under the current maintenance model.

For traditional network O&M, we mainly cope with major and urgent complex network changes from pre-event, in-event, and post-event perspectives.

Pre-event involves maintenance engineers preparing in advance for emergency plans and making corresponding emergency plans to prevent potential impacts on business in the future. In-event involves preparing network change operation scripts in advance and standardizing engineer operation principles. Post-event involves backtracking and summarizing, and iterating on shortcomings in the pre-event and in-event phases, serving as a basis for subsequent case studies. As networks become increasingly complex, this traditional O&M approach finds it difficult to completely avoid major accidents caused by human error. At site X, an engineer mistakenly input an extra '0' in the traffic threshold configuration, leading to a signaling storm that resulted in a two-day disruption of communication services for 30 million users across the province.

Driven by advancements in technologies like Gen AI, digital twins, knowledge graph, and embodied AI, industries are shifting towards replacing traditional network-based O&M with service-based intelligent O&M.By 2030, 30% of leading carriers will deploy digital twin systems using 5G-A as part of their intelligent transformation.

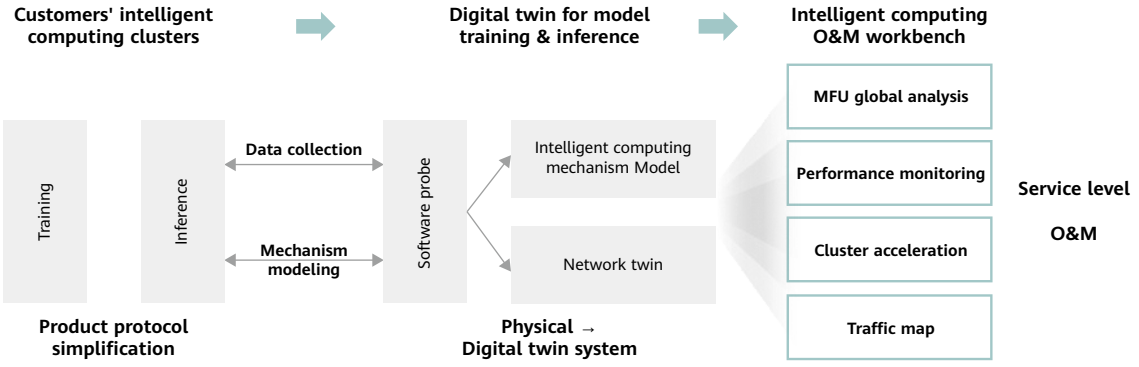
- Start with the end in mind. Simplify computing network protocols based on O&M business needs, providing real-time visibility and access to E2E data.

- Build real-time network digital twins. Create a knowledge graph for O&M to track how network issues affect services. This connects the physical network to the digital world, and links services to sort out key performance indicators (KPIs) and key quality indicators (KQIs), making it possible to visualize and manage the impact of operations and changes on services.

- Develop new applications based on service-level O&M, integrating network and services to enable global visualization and unified management of ICT O&M.

With these capabilities, fault recovery will take hours instead of days, network fault response will take seconds instead of hours, and standby board replacement can be a regular task on a weekly or monthly basis, instead of an emergency operation completed in 4 hours.

Oriented to business O&M:
Streamlining the full stack of product protocols, digital twins, and service transformation



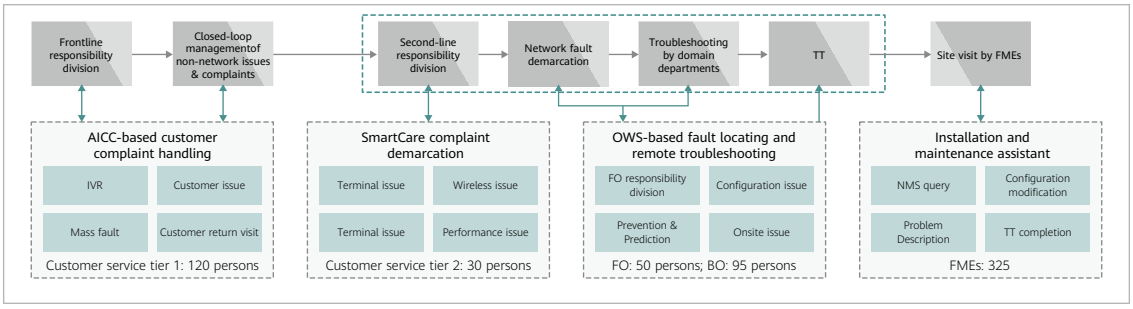
Maintenance, Optimization, and AI+: From Manual to Machine Labor

In the traditional digital landscape, tools and processes were designed with people as the focus. In the intelligent landscape, human-machine collaboration does not require this. There are clear human and machine interfaces. Machines, people, processes, and tools will work together to solve problems.

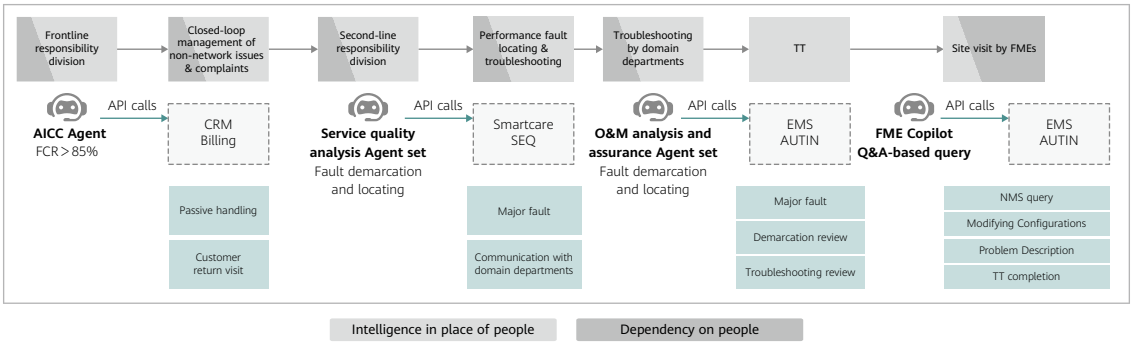
To implement agent-centric O&M, it is generally agreed that three basic technologies need to be built:

1. Computing network O&M large models: Build a computing network O&M large model that understands the O&M mechanism and network protocols based on the basic large model commonly used in the industry, and build role-oriented copilot and scenario-based agents for RE, FO, BO, and FME to reduce repeated manual work. Then, implement real-time dispatch of configuration and commands by incorporating traditional automation large models, and implement unified APIs for the

Digital era: People are the focus. Processes, platforms, and people are interdependent. Processes and small model tools are designed based on the efficiency of manual issue handling.



Intelligent era: Human-machine collaboration is used. People do not need to participate in each phase. Human-machine boundaries are clear. People, processes, and tools are designed by focusing on machine-based issue resolution.



OSS domain, enabling users to use large models easily.

2. Digital twins: The real-time network digital twin system is required to associate services with NEs. Network data in the past was considered black box data, which can be analyzed by layer-by-layer filtering using probes and NMSs. In the future, digital twin systems will use knowledge graphs to collect data quickly at the NE level, aiming to reduce the collection time from 30 minutes to one hour to just a minute.

3. Embodied AI robots: Onsite O&M costs generally account for 60% of the total ICT O&M costs. In the future, embodied AI robots will be deployed in each data center, equipment room, and site. The transformer-based small-sized large model IOS can accurately identify instructions from the NOC/SOC agents, and perform network operations (inspection, live network status awareness, fiber port adjustment, board replacement, etc.) instead of maintenance personnel. This will significantly improve O&M efficiency.

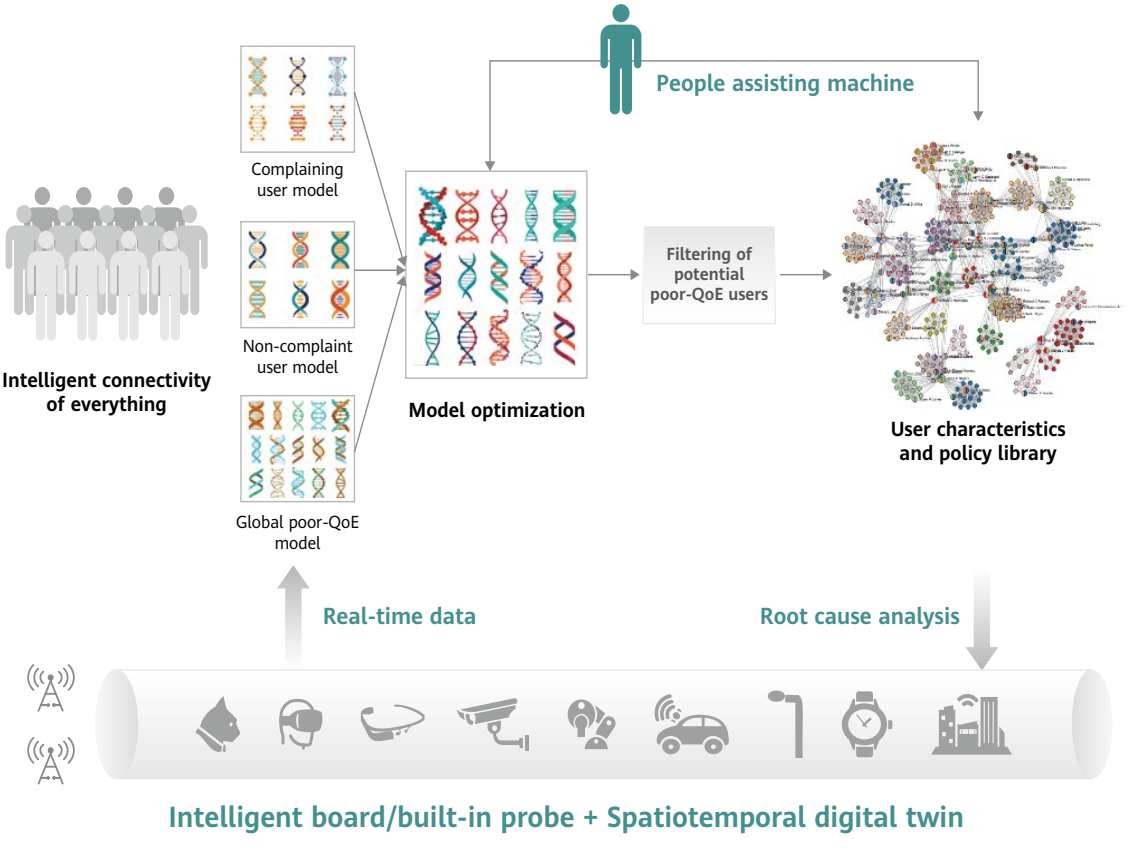
Hardware replacement by a robot hand

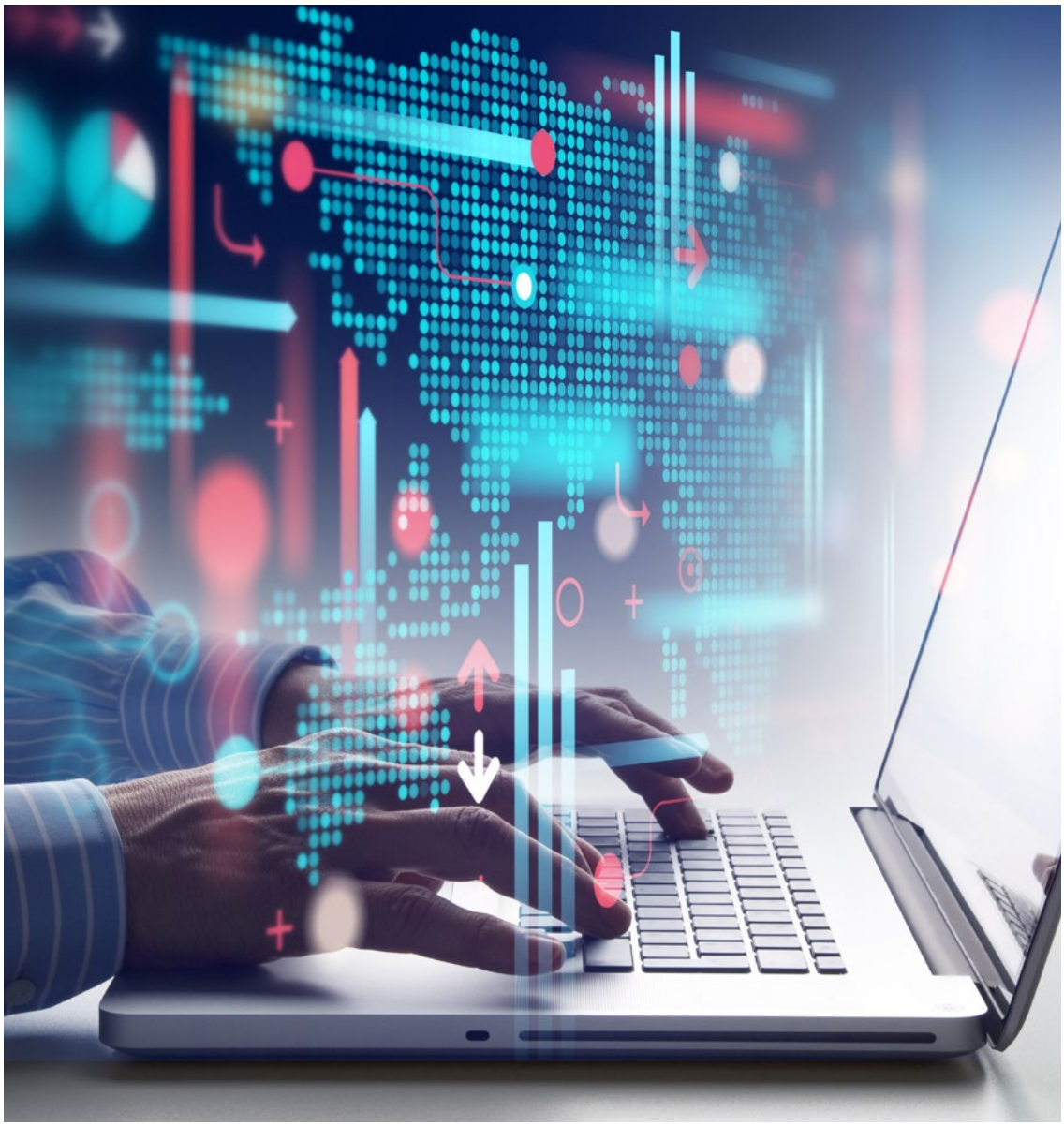
Automatic inspection robot

Optimization and AI: From "People Waiting for Network" to "Network Waiting for People", Fostering the Willingness to Monetize User Experience

By 2030, mobile networks will shift from mainly handling person-to-person and person-to-device communications to mainly handling person-to-agents and agent-to-agent communications. Communications networks will connect not only individuals, but also various sensing, display, and computing resources related to individuals, as well as AI agents. They will connect not only home users, but also related home, car, and content resources. They will also connect machines, edge computing, and cloud resources related to an organization, on top of employees of an organization. In this way, future communications networks can meet various business requirements of an intelligent world.

From an experience assurance perspective, the daily optimization of traditional network is a passive "people-waiting-for-network" approach, where network optimization is a reactive response to customer complaints. It aims to improve product performance by 10-15% based on existing performance, doing the best possible with what's available. In China, an investment of 25,000 person-days is required annually to address a range of routine optimization issues, including collecting complaints about network issues, designing optimization priorities for typical regions/sites, digital road testing, and designing and implementing routine optimization solutions. The "people-waiting-for-network" approach resolves only 35% of network issues, failing to delivery an optimal user experience. A unified large model for network optimization is lacking, causing performance conflicts among multiple optimization solutions. Besides, optimization experience cannot be shared.





By 2030, the network optimization approach will shift to "network-waiting-for-people". We will achieve end-to-end performance sensing using HarmonyOS, smart boards, smart antennas, and fiber iris. We will build a digital twin system (TAZ) based on space-time to precisely predict future traffic, trends, and SLA tendencies for each business category. Using an optimization large model featuring computing & network integration, we will achieve intelligent and autonomous closed-loop of single products in 30% scenarios. For the remaining 70% scenarios, we will adopt a series of methods. First, we will use knowledge graphs and management systems to consolidate core assets, and develop AI driving force based on MR and SEQ performance data generated during daily optimization. We will continuously support model upgrade and iteration, and develop agents and integration twins for VIP assurance and daily optimization. By doing so, we can help network optimization personnel take preventive measures and predict issues based on business changes, quickly generating optimization solutions. Finally, we will invoke traditional small models to perform real-time analysis, decision-making, and closed-loop management of performance experience.

Optimization and AI+: Network Optimization Agent Based on Endogenous Intelligence

According to Joseph Sifakis, a French computer scientist and 2007 Turing Award laureate, the definition of future autonomous network systems in the communication field requires a different approach compared to the general agents based on LLMs. In the communication field, strategies generally need to be formed based on actual network business/network status. Therefore, agents in the field must be able to comprehend network, including its topology, performance, alarms, and events. The high real-time requirements for data in the communication field make it difficult for large models to learn, so digital twins and other technologies are integrated to help. We use real network information to plan and implement solutions, and inject domain expertise into large models through prompts or SFT. This, combined with digital twin models and atomic capabilities for experience, maintenance, and optimization, enables the planning, perception, decision-making, and execution of tasks.

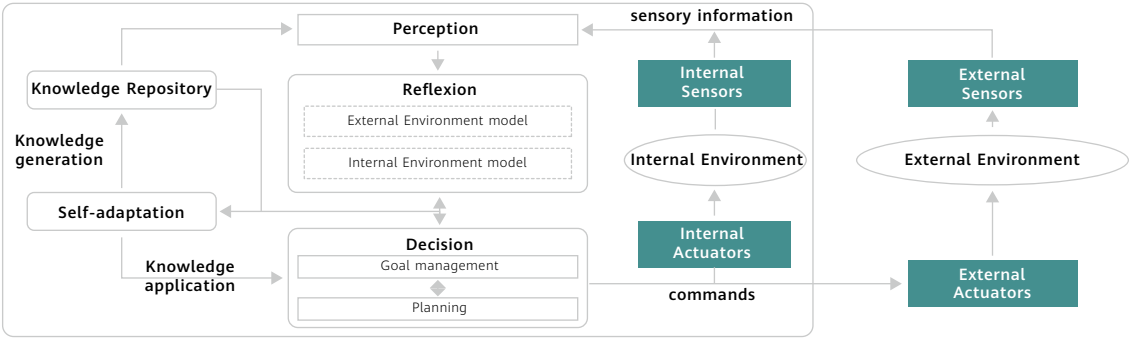
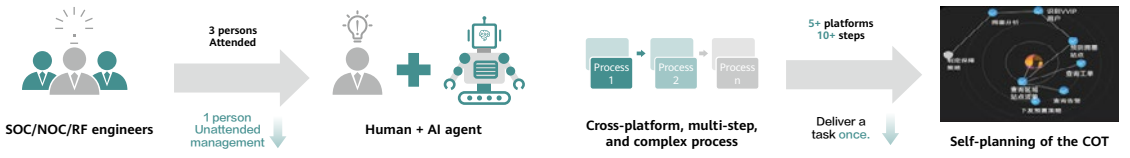


Figure 6: Computational model for cyber physical agent

Take the optimization work pattern in a region as an example. Previously, three network optimization engineers were on standby. Now, leveraging the endogenous intelligence of network optimization agents, unattended network optimization is achieved, requiring only one person to design the agent model.

The network optimization function development process is also simplified. Previously, it involved multiple platforms, over 10 steps, and at least three months. Now, agents design and complete the task in one dispatch with multiple rounds of interactions based on the chain of thought (COT).



Marketing and AI: From Digital Business to Digital and Intelligent Business

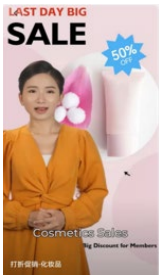
By 2030, the digitalization of marketing businesses transforming into intelligent transformation would have become a consensus. Consulting firms such as Accenture and PwC have identified marketing as the top industry that can be impacted by large models, because there are diverse user demands, numerous human-machine interaction interfaces, and vast opportunities for creative generation and innovation in handling customer needs, from advertising and marketing to billing and sales, as well as customer complaints. This aligns with the three main conditions suitable for current mainstream LLM applications: vast data, creative scenarios, and natural language. The current digital marketing model is no longer sufficient to meet the demands of a future with billions of digital and intelligent humans, autonomous driving cars, and full-line industry intelligence. What do we need in the future?

- Agile innovation: In the digital age, the package design putting people in mind typically takes 3 to 6 months to complete the entire process of market strategy development, resource preparation, package development, and market promotion. By the time a product hits the market, the market has often shifted. In the future, intelligent businesses will need to develop marketing initiatives and package designs based on customer data analysis. Leveraging agile computing and network infrastructure, they can accelerate this process from several months to just a few days. This allows for real-time push of packages at any time, anywhere, in movie theaters, sports stadium VR, and other scenarios.

- Smart creativity: In the digital age, marketing primarily relies on big data user profiling for targeted phone/content pushes. It completely depends on humans for analysis and crafting marketing strategies. By 2030, the efficiency and cost of digital humans in content generation and operations will surpass human capabilities, enabling rapid iteration in response to market changes to meet the latest demands of the masses. Digital humans offer 7x24 online, high-quality interactive solutions to customer issues.

Product overview

Introduction videos for software services, beauty & fashion, 3C electronics, factory workshop, and solutions

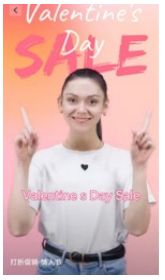


Promotional activities

Voice-over videos on Black Friday, Valentine's Day, Christmas, and other internationalized holiday nodes for voice-over videos.

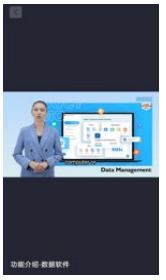
In-feed ads

Google, Facebook advertisements, YouTube, TikTok, Amazon, Instagram platform video



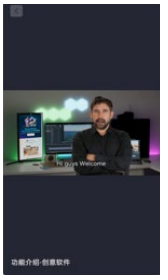
Product recommendation

Video playback for product presentation, details display, comparison and evaluation, purchase guide, etc.



Content marketing

Google SEO, Facebook YouTube video and e-commerce ads



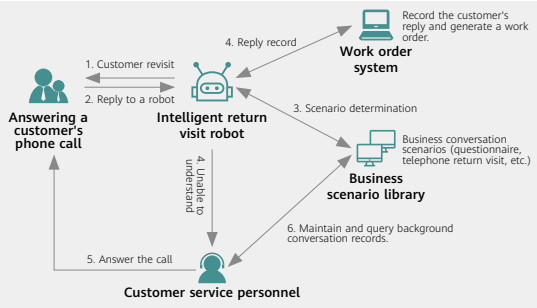
Video tutorial

Customer service, support team content, product operations, Q&A, and explanation videos

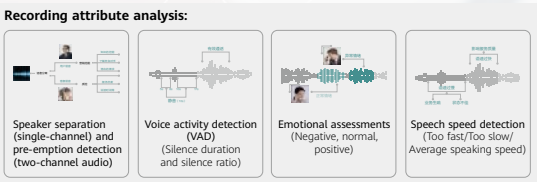
Marketing and AI+: From a Cost Center to a Profit Center

Enterprises have traditionally viewed customer service centers as cost centers. However, with digitalization and intelligence, they are discovering that engaging with complaining customers can attract new users and revenue, making customer service a key revenue driver in the digital landscape.

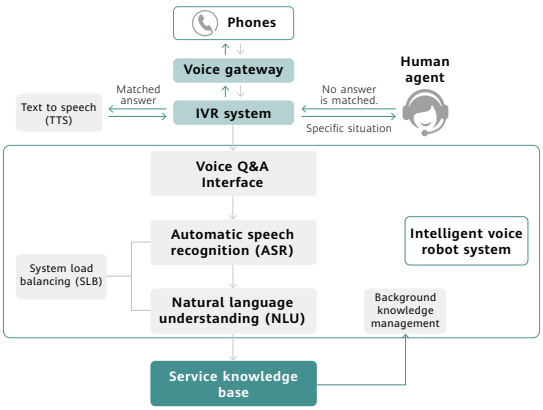
- In the customer cultivation and acquisition phase, digital human intelligent outbound calls are used to replace traditional manual outbound calls. Underpinned by AI and data-based personalized recommendations, the outbound call success rate is improved by three times, and the outbound call efficiency is improved by 300%.



- In the activation and retention phase, digital humans are used to recommend and interpret products in dialog mode, and accurately identify customers' willingness and emotions based on customers' tone, reducing the AHT by 50%.



- In the complaint handling phase, digital humans can handle more than 90% of problems based on the service knowledge base and voice recognition capability. Only 10% of complaints are transferred to manual personnel. This improves the first call right (FCR) from 75% to 90%.

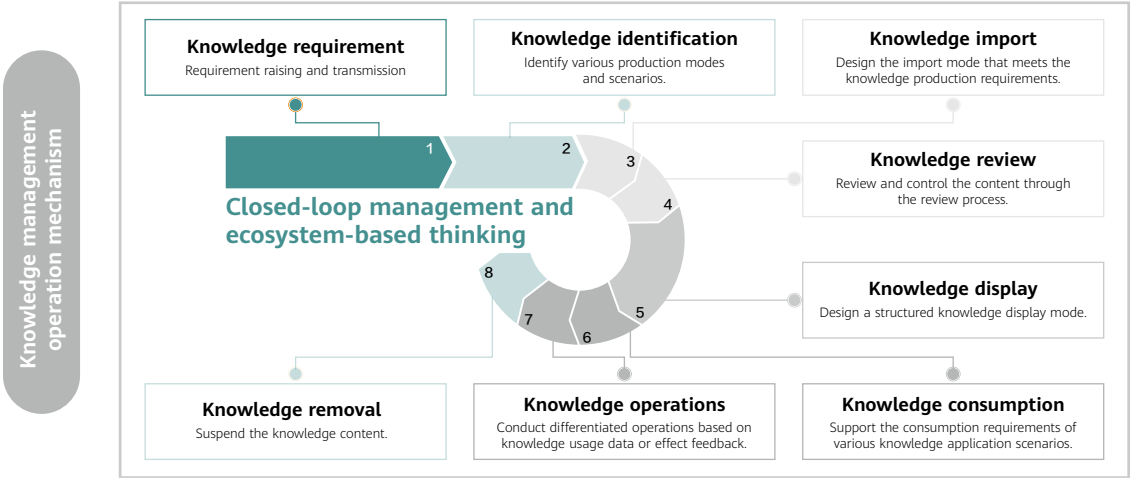


After a customer accesses the intelligent response service, the system implements real-time voice interaction with the customer based on the settings. The background automatically matches related knowledge points through the knowledge base system.

- **The knowledge content can be matched.**
  The system automatically selects voice broadcast or SMS push based on the complexity of the knowledge content.
- **The knowledge content cannot be matched.**
  The system sends messages to the customer, indicating that the related knowledge point cannot be matched and the call is transferred to the agent.

By 2030, digital marketing will keep evolving towards intelligence featuring digital humans powered by large models. Customer service centers will shift from being cost centers to profit centers. Large models will drive a new paradigm for digital business, allowing industries to create more agilely and flexibly while generating ongoing economic benefits.

Enablement and AI: From an Information System to a Knowledge System



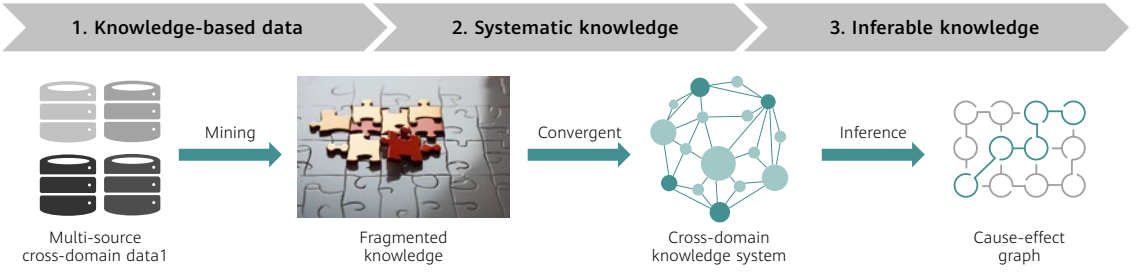
Previous knowledge systems were designed from a human perspective, focusing on learning materials like Word documents, slides, case libraries, and FAQs. Seasoned engineers compile this knowledge through extensive practical experience, and pass it down through generations. Unstructured knowledge makes up over 50% of existing knowledge (Accenture 2024 Insight). Only 5% of unstructured data is effectively managed by enterprises. In the landscape of large models, we need to convert the unstructured knowledge and experience into a format that large models can use, for example, token (for LLMs) and pitch (for videos and images). Besides, professional teams are often lacked during the need identification, import, review, display, operations, and consumption management of traditional knowledge and experience. To equip large models with human-like intelligence, their training data must be up-to-date and highly accurate.

In addition to the mechanism of knowledge operations, a Gen AI knowledge management platform is essential for quick knowledge mining, convergence, and inference. By 2030, Accenture and PwC predict that 55% of enterprises will have deployed knowledge management systems.

1. Knowledge mining: Quickly convert daily expert knowledge into standard data through the portal website and platform, and sort out fragmented knowledge.

2. Knowledge convergence: Associate multi-dimensional tags based on different roles, support keyword-based, tag-based, and association search, and integrate them into the enterprise production flow.

3. Knowledge inference: Perform inference on the knowledge required by various roles, push the knowledge based on scenarios based on the copilot, and automatically recall and generate the knowledge based on the role feedback.



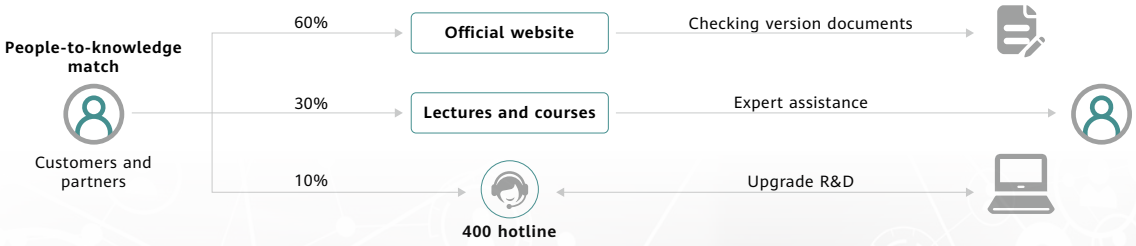
Enablement and AI+: From People-to-Knowledge Match to Knowledge-to-People Match

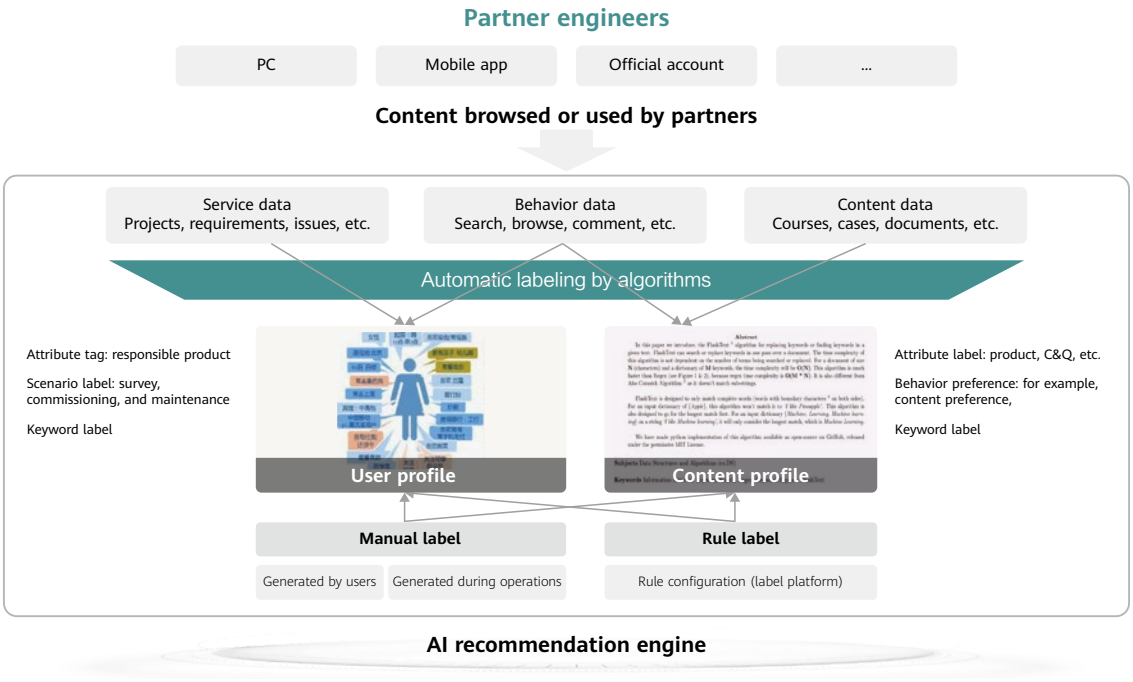
The traditional enablement system is designed based on the fragmented time of human. From the perspective of knowledge acquisition of industry users and partners, we can learn the following information:

- 60% of knowledge is obtained through online searches. General knowledge is typically acquired through industry cases, product manuals, and vendor websites. However, this process is time-consuming and often yields unsatisfactory results.

- 30% of knowledge comes from expert lectures and online and offline courses. Advanced knowledge, such as strategies, new platforms, and new technologies, is acquired through expert lectures and communications. The effectiveness of this enablement depends on the experts' level of knowledge. Furthermore, enablement in this form is not frequent enough, with an average of less than 10 enablement sessions per person per year.

- 10% of knowledge involve private technologies in the industry. Common practitioners do not have the enablement mechanism and can only seek help from vendors' R&D personnel.





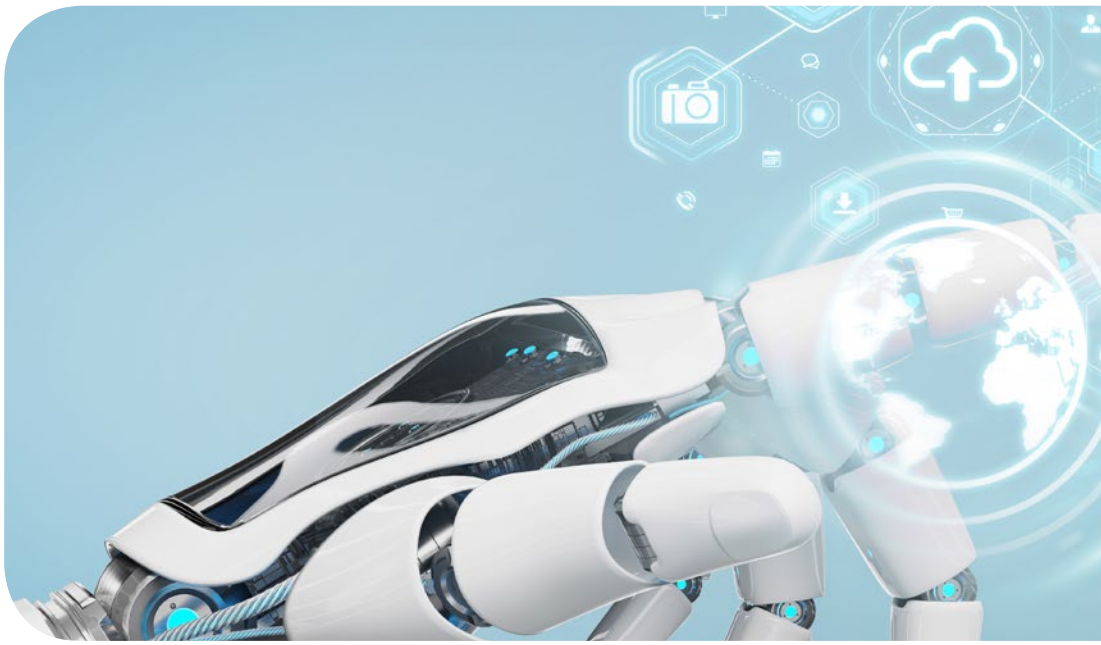
As Gen AI technology gradually penetrates into various industries, knowledge management is becoming an essential element for utilizing large models. From the production of enterprise knowledge to its storage and application, the process will become more standardized and intelligent. Various knowledge application assistants will be integrated into the production flow, providing real-time push notifications to employees with different roles at each stage, thereby significantly enhancing their work efficiency. At the same time, the knowledge of the enterprise can be continuously and rapidly converted into tokens, providing precise knowledge materials for feeding into large models, making the models smarter and more intelligent. With the aid of knowledge management systems and knowledge assistants, employee enablement in enterprises will shift from the traditional "people-to-knowledge match" to "knowledge-to-people match", and from previous fragmented enablement to lifelong enablement based on real-time push notifications aligned with production workflows. The platform identifies business, behavioral, and content data for each role based on employee/partner access records, creating user and content profiles for each role. The system will integrate with employees' daily ERP systems, including OA, OSS, and BSS, to proactively push relevant knowledge to those tho need it most at the right time.

By 2030, Accenture predicts that 80% of enterprise knowledge acquisition will move online through automated and interactive methods. Offline learning will focus on hands-on training, experience, debates, and other practical courses. This integration will help industries and companies continuously promote TECH4ALL and adopt intelligent technologies.

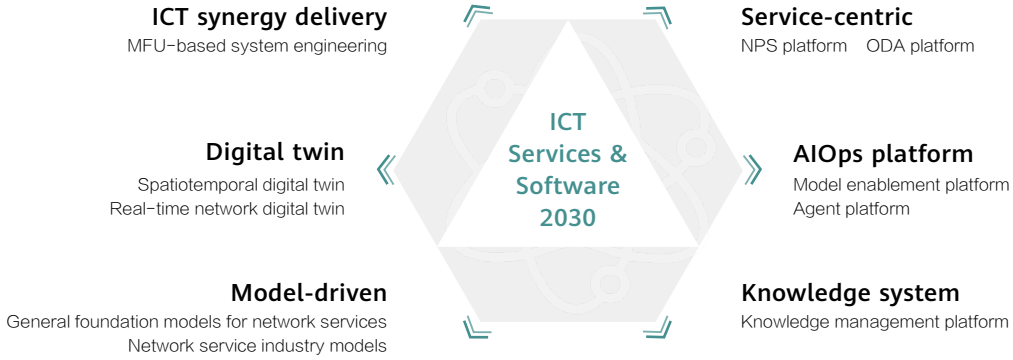
The next-generation enablement platform/ community product will feature online, open, and orchestration as its three key characteristics, aiming to facilitate rapid acquisition, categorization, sharing, and distillation of knowledge experiences across global domains.

- Online real-time learning will reduce the time it takes for humans to acquire knowledge by 90%.
- Cross-regional knowledge sharing and collaboration efficiency will increase by 200%.
- The training period for mid-level engineers will be reduced from 3 years to 6 months.

03



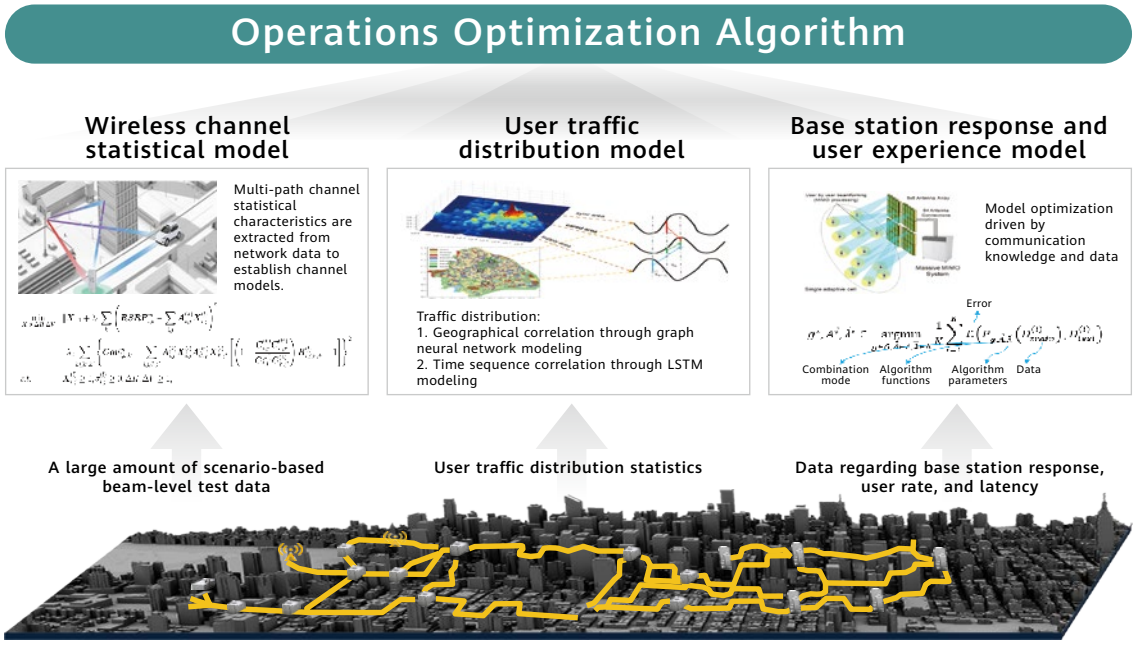
Vision and Core Technologies of ICT Services & Software 2030



Digital Twin

1. Spatiotemporal digital twin: The network optimization of mobile communication systems involves a category of technical issues that are difficult to be modeled in a unified manner. Parameters are mutually dependent or contradictory, making it difficult to establish global optimization models. TAZs identify the distribution differences of service types based on users, services, and networks. Differentiated rate assurance objectives are configured across clusters to maximum the potential of each cluster, improve carriers' revenues, and provide real-time data for network service models. The collection time required is reduced from hours to minutes. The following describes related models:

- Wireless channel statistical model: A large amount of scenario-based beam-level test data is collected across networks. Then, multi-path channel statistical characteristics are extracted from the collected network data to establish channel models.
- User traffic distribution model: User traffic distribution statistics are collected to establish geographical correlation through graph neural network modeling, and time sequence correlation through LSTM modeling.
- User experience model: Data regarding base station response, user rate, and latency is collected to enable model optimization driven by communication knowledge and data.

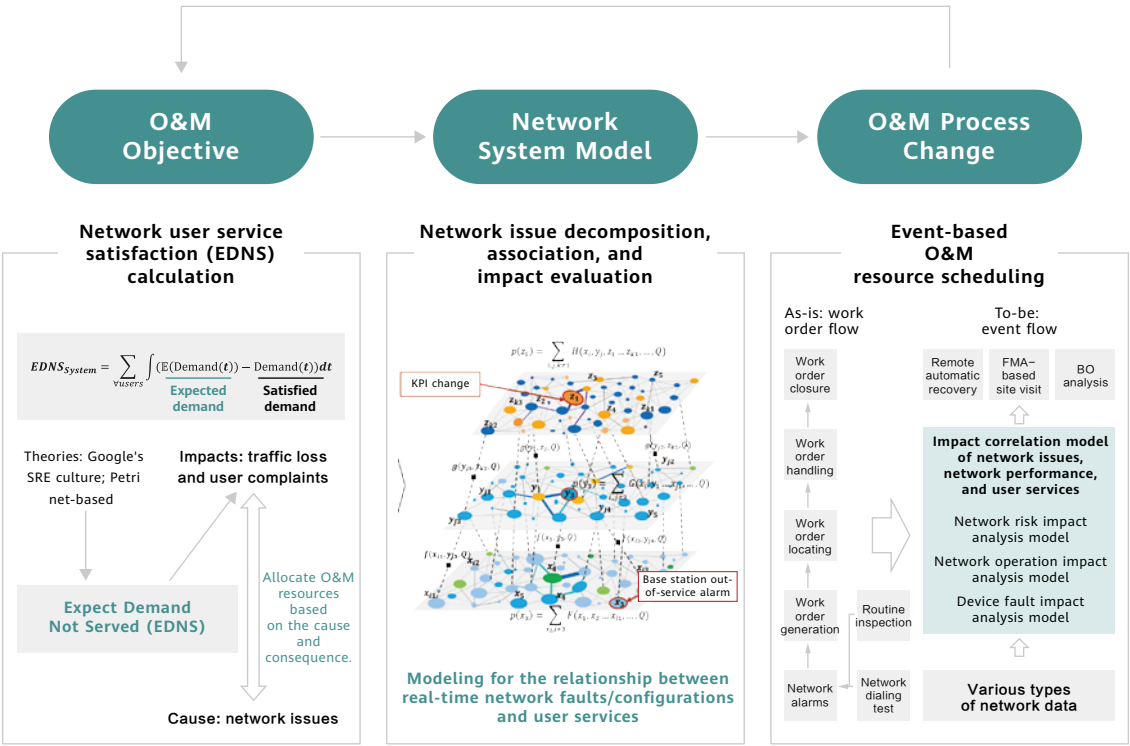


2. Real-time network awareness twin: It measures the impact of network issues on services, and associates the physical world with the digital world for different services. It can also review KQIs and KPIs, as well as visualize and manage the direct impact of each operation and change on services.

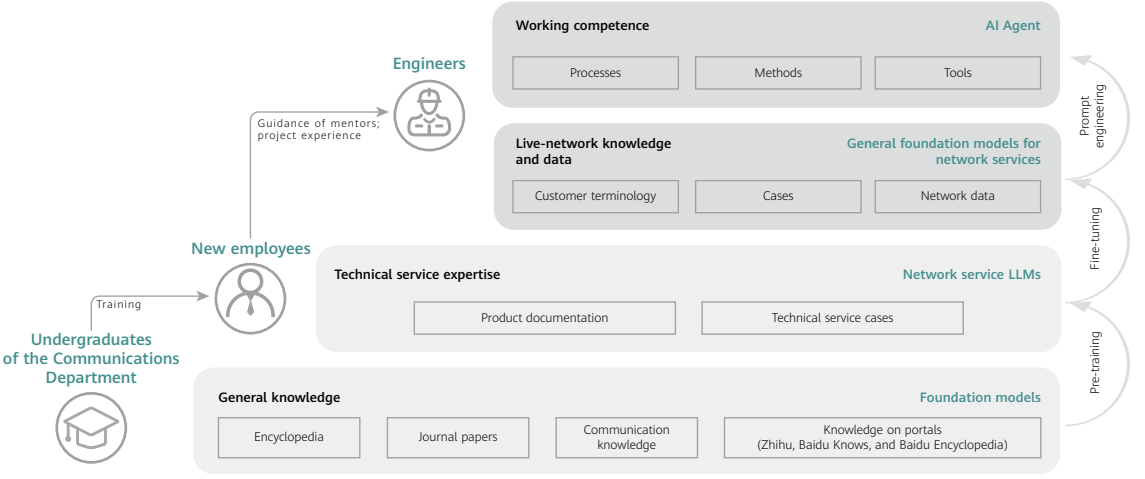
- Algorithm selection based on service objectives: Based on service requirements and the SRE reliability theory, the EDNS calculation logic is established to analyze the logic of service loss reduction.

- Network system model: The impacts on networks, resources, and services are analyzed through modeling. Models and algorithms are designed to analyze the association between each network fault, operation, and service.

- Accurate fault handling based on event flows: The proportion of false alarms is reduced, and precise service-centric O&M is implemented.

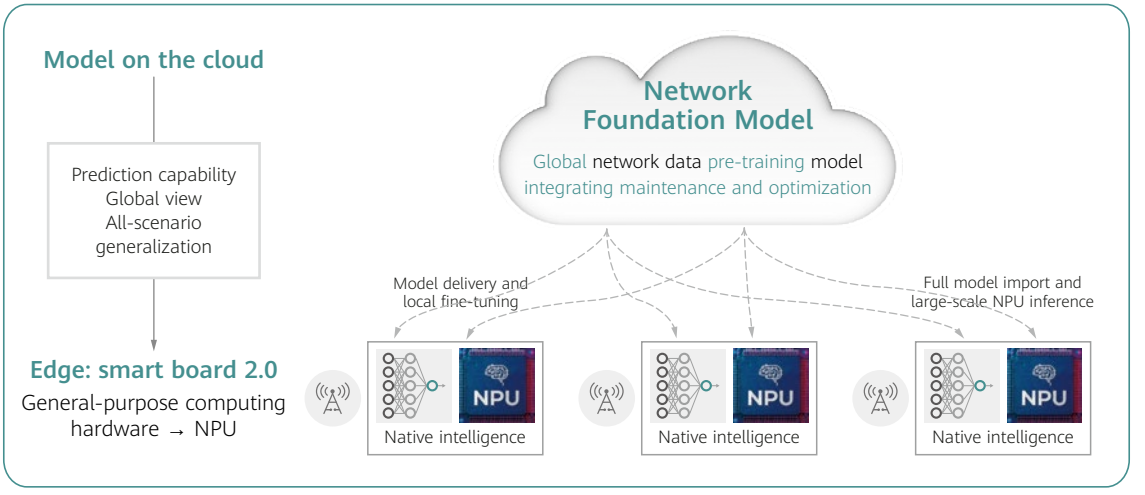


Model-driven



1. Network service LLMs: Global mainstream foundation models, such as Llama 3, Mistral, ERNIE Bot, and Zhipu, are evaluated, adapted, and pre-trained in advance. Common industry knowledge, including information on the C114 website and Wanfang Data, as well as product guides, case libraries, and exam libraries, is aggregated. By doing this, network service LLMs are equipped with basic ICT service knowledge and industry knowledge mastered by high school students. This lays a solid foundation for industry upgrade, and enables models to process and train unstructured knowledge.

2. General foundation models for network services: A general industry model that understands ICT protocols, signaling languages, planning, construction, maintenance, optimization, operations, and OSS/BSS is needed to solve increasingly complex service issues. The model should be able to understand the CoT of complex tasks and achieve an accuracy of over 99%. It needs to provide a knowledge system that consists of structured data and is equivalent to the industry knowledge level of undergraduates. The number of training parameters required for the general industry model is much less than that for LLMs. Small models with 7–10 billion parameters can be fine-tuned and RAG can be quickly deployed to support the production of carriers and industry customers.



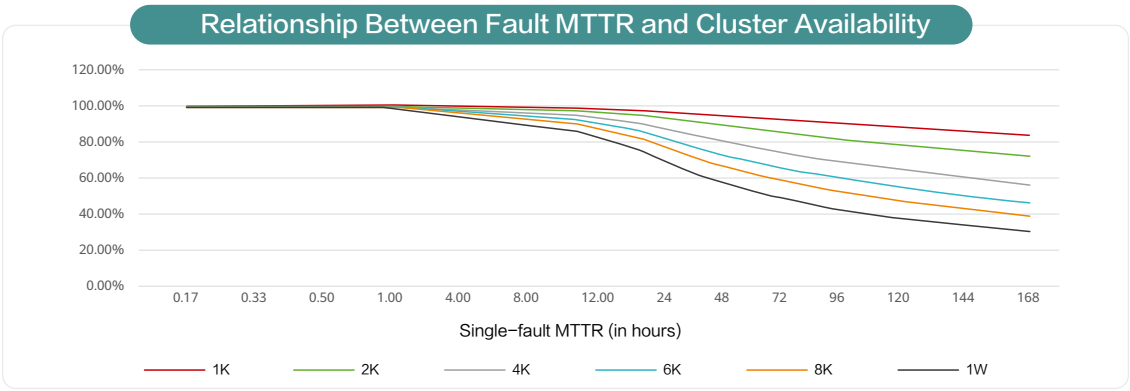
ICT Synergy Delivery

A larger cluster scale (n) indicates a greater impact of single-fault indicators (such as MTBF and MTTR) on cluster availability. According to a report of Meta Llama 3's training clusters with hundreds of thousands of cards in 2024, 466 interruptions occurred during the 54-day training, with an average of 8 interruptions per day. 41% of the interruptions were caused by software exceptions, cable issues, and network faults. OpenAI's Stargate Project will need 1 million cards, which will pose higher requirements on cluster performance (MFU). Each day of interruption will cause an economic loss of tens of millions of dollars. To solve this issue, we need to continuously improve the following two indicators regarding cluster integration and O&M:

1. The cluster linearity is related to network link stability (latency, performance, and configuration consistency) and NPU subhealth. It is strongly dependent on deployment models, configuration optimization, as well as routine subhealth governance and maintenance.
2. The cluster availability is measured by four indicators: single-fault MTTR (h), single-node exception rate (f), number of nodes occupied by jobs (n), and job duration (x).

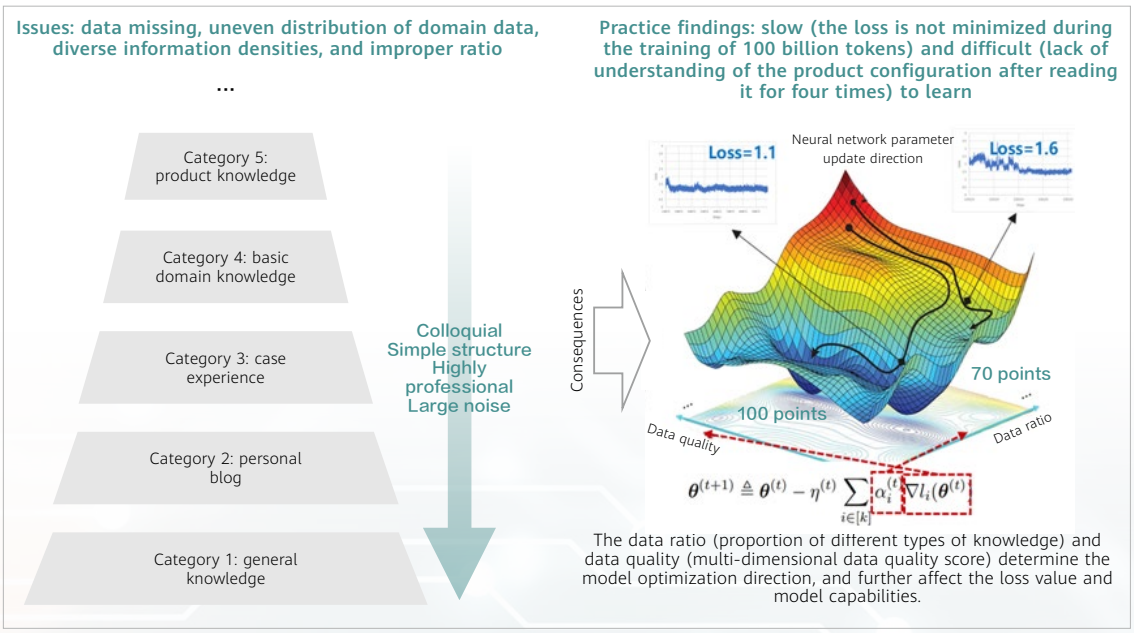
To achieve high availability of training and inference as well as the optimal availability and MFU of computing clusters, three integration capabilities need to be built:

1. Load scheduling at the global, region, and node levels
2. Determined and competitive cluster availability: Faults and subhealth can be detected in advance for proactive maintenance. More than 4000 clusters can run stably for over 30 days. Resumable training after breakpoints is supported, and single-card exceptions have no impact on jobs.
3. Cost-effective linear network, AI computing, and storage performance optimization; hardware pipeline for detecting, isolating, and rectifying subhealth and faults.



Data Engineering

Data is the foundation for the training and inference of large models. In existing practices of large models, 85% of enterprises encounter difficulties in learning. The core reason is that the data engineering required by large models needs to build core advantages of domain-specific models in terms of data scope, data quality, and training efficiency. To achieve this, the following nine core data engineering technologies are required:



The data ratio (proportion of different types of knowledge) and data quality (multi-dimensional data quality score) determine the model optimization direction, and further affect the loss value and model capabilities.

1. Parsing of domain-specific complex process data: Extracts and outputs complex text information of diverse types (signaling protocol interaction, alarm cause-and-effect diagram, and process), with an extraction and parsing accuracy of over 80%.

2. Multi-modal complex information tokenizer: Extracts texts based on the layout analysis of visual models, as well as modules like text matching and table parsing; achieves a PDF content extraction accuracy of over 95%.

3. Efficient domain data synthesis: Achieves self-growth of high-quality domain data — from unlabeled to labeled and from no CoT to CoT — and expands data in typical domain scenarios by 10 times.

4. Automatic data quality evaluation: Automatically evaluates the integrity, accuracy, consistency, timeliness, and quality of pre-trained data, as well as improves efficiency.

5. Domain data augmentation: Augments samples to expand the training dataset, enhances the accuracy and completeness of content semantics as well as the text diversity, improves the model training efficiency, and increases the data diversity by 10 times.

6. Knowledge locating and sourcing: Establishes the association among model capabilities, parameters, and data for overall improvement, locates capability weaknesses, and augments weakness-related data to improve the efficiency of locating data bad cases by 10 times.

7. Optimal data ratio: Breakthroughs have been made in model scaling technologies that obtain the optimal ratio of general data to domain-specific data as well as the optimal ratio for data of different domains in incremental training scenarios. The loss is minimized, and the model training efficiency is improved by 50%.

8. Data curriculum learning: The learning sequence affects the final model effect. The scaling law is used to find the optimal learning sequence and improve the domain knowledge answer accuracy by more than 30%.

9. Data annealing training: Builds the optimal domain-specific data subset, greatly improves model capabilities through multi-stage training and optimal data annealing at the end of training, and achieves a five-fold increase in the annealing rate.

Service-centric

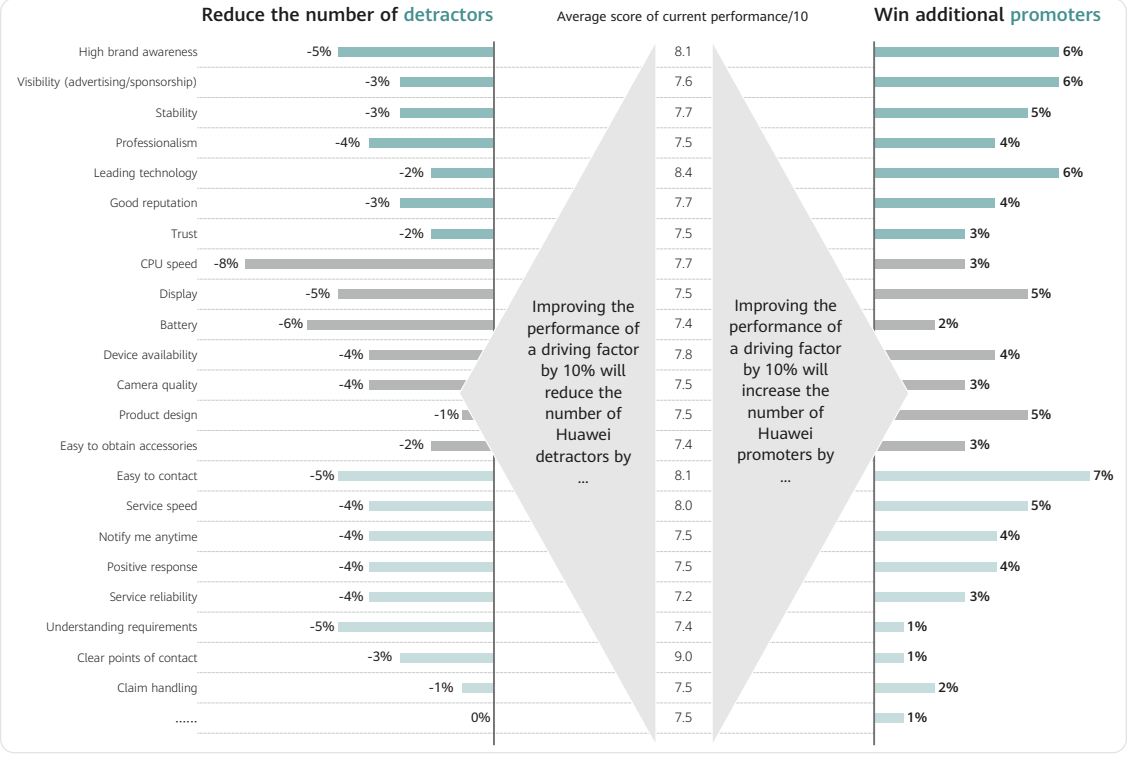
NPS digital analysis platform: To implement service-centric experience management, the major challenge lies in how to identify issues, determine the impact scope, and develop high-quality simulators through digital methods.

1. Promoters and detractors are identified based on the Kano theory. With promoters and detractors as the sampling frame, simple linear regression analysis is performed to identify the reward and penalty driving forces of different factors.

2. Based on the driving forces of depreciation and recommendation, the overall impact on the NPS is deduced. That is, the NPS increase achieved by improving each experience indicator by 10%.

3. Simulators are required to help evaluate the NPS changes caused by the improvement or deterioration of related indicators, in order to develop targeted policies.

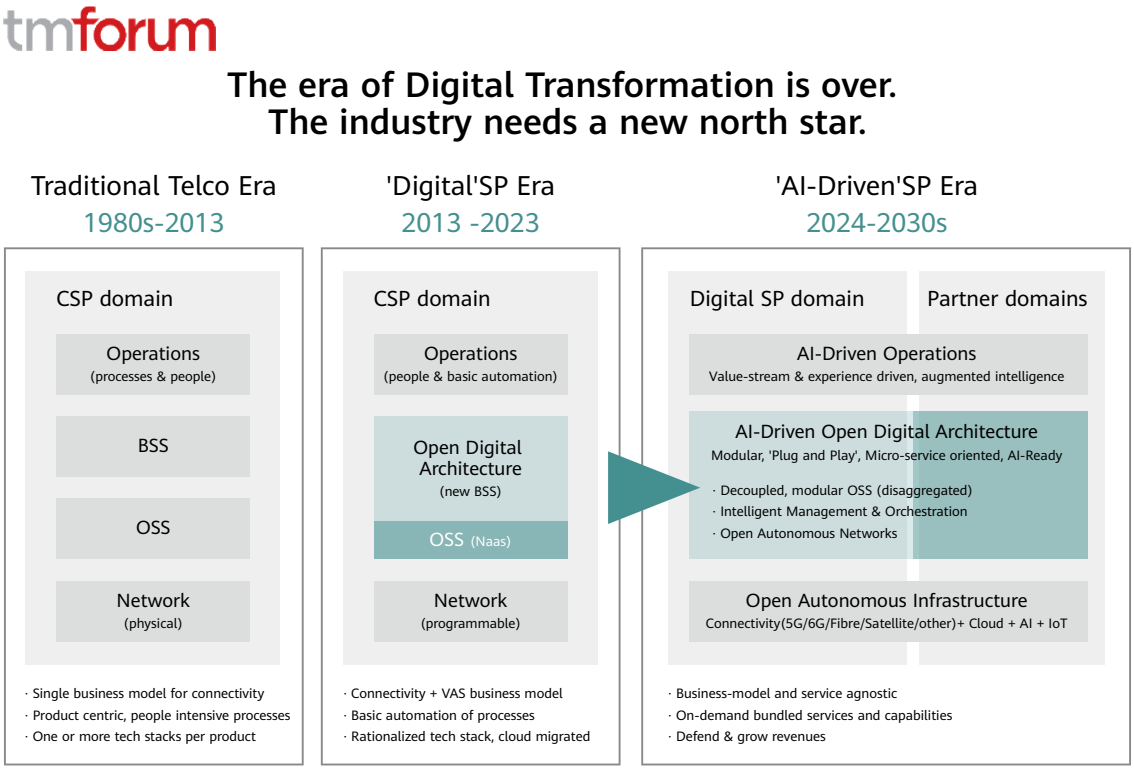
The common practice in the target industry is to specify the driving forces of reward and penalty for each indicator through modeling and analysis. Through cross-analysis with the satisfaction matrix, the reward factors that need to be guaranteed and the penalty factors that need to be preferentially improved are identified. The improvement priorities and measures of specific indicators are analyzed based on the driving forces of satisfaction, pain point occurrence rate, and reasons for customer recommendation and non-recommendation.



ODA O&M platform: Although CSPs have built IT systems that can integrate components from multiple vendors, these components are generally provided by communications and software vendors. As CSPs now compete on a larger stage, the ability to integrate components of vendors from industries other than communication becomes essential. When a new way (for example, Agents) of interacting with customers is available, it is unrealistic to wait for its version dedicated to the communication industry. In addition, as open source projects like open network automation platform (ONAP), Open Source MANO (OSM), and Open Baton emerge, any multi-vendor definition must include open source software. Many CSPs also begin to embrace open source in their future-oriented IT systems. TM Forum's next-generation OSS architecture ODA is also designed in attempts to solve such technical issues.

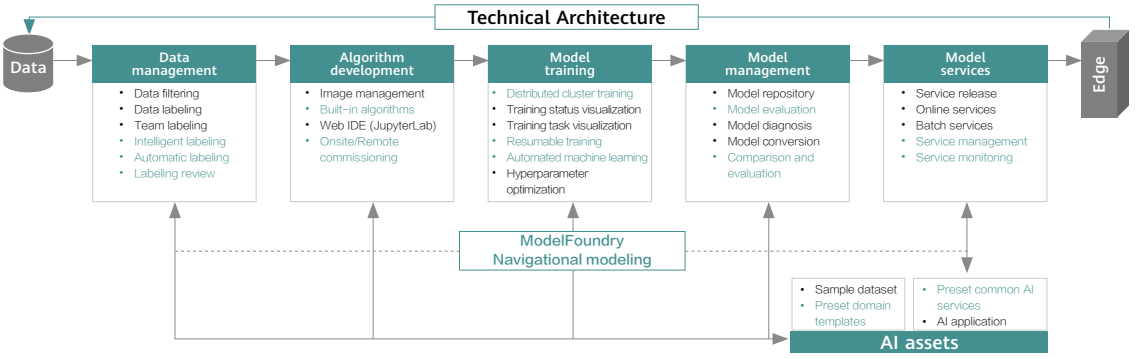
- Cloud components: Newly developed applications are managed in the common repository as cloud components that comply with ODA standard interfaces and specifications.
- DevSecOps: automatic software version verification and release platform.
- The ODA component repository consists of six subdomains: Party, Core Commerce, and Production that carry core processes, as well as Decoupling & Integration, Engagement, and Intelligence that support service processes. The component repository presents a panorama of BSS & OSS components, including 28 components whose functions have been preliminarily defined (that is, numbered components). Most of them are distributed in the Production and Party subdomains.

These new features make it possible for carriers to manage the digital ecosystem on a large scale across borders. For example, globalized automakers can reach an agreement with multiple CSPs on autonomy and IoV agreements. ODA also considers future-oriented key requirements and concepts that will not be considered in independent projects, such as AI, unified BSS/OSS architecture, and unified data-centric approaches.



Source: TMF Strategic Review 2023,12

AIOps Platforms



To prepare for numerous agents in the future, we need to develop model enablement platforms and agent platforms oriented to ICT services and software. With these platforms, customers can quickly develop and operate their own industry-specific models and agents based on the framework. In addition, model training assets can be quickly replicated across scenarios.

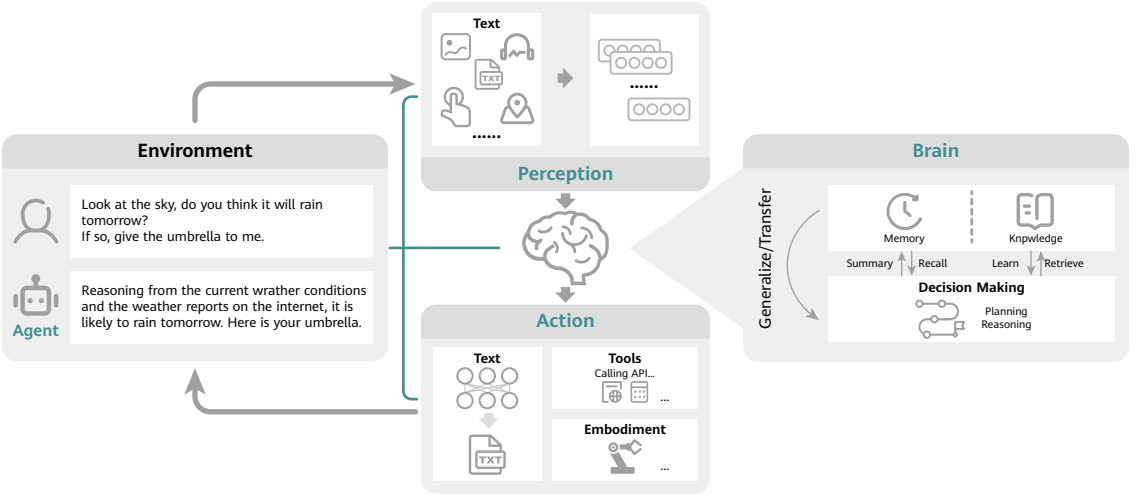
A model enablement platform provides five major services: model selection, knowledge management, model training/fine-tuning, model evaluation, as well as model compression and inference. A comprehensive LLMOps system and open-source tool chain have been established outside China. In China, Baidu and Zhipu are building their own tools based on open-source systems, and developing model enablement service packages and AIOps tool chains based on industries like banking and Internet.

An AI agent platform provides three major capabilities: awareness, thinking, and execution.

1. Awareness: Includes user task acquisition, environment status awareness, and feedback detection. These functions enable agents to obtain required information regarding digital twins, smart boards, and IoT.

2. Thinking: Includes planning, inference, knowledge learning, and memory storage. These functions enable agents to make analysis and decisions by referring to the CoT of people.

3. Execution: Answers texts, uses programming tools (such as APIs) and entity tools, and invokes tools like traditional small models in real time.



04

ICT Services & Software 2030 Initiative

The key to intelligence is the collaboration between traditional business talent and new ICT talent. To enable machines to think like humans and drive intelligent transformation, business experts in each domain must actively learn and gain an in-depth understanding of AI.

By 2030, AI is expected to promote equality, openness, and security. It will enable every individual and organization to benefit from technologies and quickly upgrade infrastructure, businesses, and personnel.

Let's work together to usher in a new era of intelligence.

Glossary (Acronyms and Abbreviations)

Acronym	Full name
5G	5th generation mobile communication
AGI	artificial general intelligence
AHT	average handle time
AIOps	artificial intelligence for IT operations
COT	chain of thought
EDNS	Expected Demand Not Served
ERP	enterprise resource planning
FCR	first call resolution
GenAI	generative AI
IMT	International Mobile Telecommunications
IoT	Internet of Things
KQI	key quality indicator
LLM	large language model
MFU	model FLOPs utilization
MOP	method of procedure
MOS	mean opinion score
OA	office automation
RAG	retrieval-augmented generation
SLA	Service Level Agreement
SRE	system reliability engineer
TAZ	traffic autonomous zone
TECH4ALL	TECH4ALL initiative
TTM	TTM

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