



Striding Towards the Intelligent World White Paper 2024

Data Communication

Networks Accelerate AI and
AI Redefines Networks



Building a Fully Connected,
Intelligent World

► Contents

01

| Network for AI

Intelligent Productivity Is Ready, with Ongoing Consolidation of AI Infrastructure

Trend 1: Intelligent Computing Clusters Embrace the Era of 100,000+ GPUs/NPUs

Scenario 1: Ultra-Large Single Cluster: High-Quality Intra-DC Network Is Essential for Unleashing Computing Efficiency

Scenario 2: Cross-DC Collaborative Training: Long-Distance Lossless Inter-DC Network, Aggregating Distributed Computing Power

Trend 2: Super-Connectivity AIDC-Access Network Construction Accelerates, Enabling Business Monetization of Intelligent Computing Cloud Services

Scenario 1: Remote Storage-Compute Collaborative Training, Advancing the AIDC-Access Network to a Lossless State

Scenario 2: The Need for Ultra-Fast Delivery of Large Volumes of Samples Highlights the Importance of Building an Elastic AIDC-Access Network

Summary: Collaborative Construction of Intelligent Computing Networks and Power, Enabling On-Demand Intelligence for Enterprises

Recommendations for Action

02

| AI for Network

AI Injects New Innovation Vitality into Networks

Trend 3: The Integrated Development of Digital Twins and AI Accelerates the Pace Towards Level 4 Autonomous Driving Networks

Scenario 1: AI Agents Collaborate with RAG/Small Models to Improve Domain-Specific Q&A and Decision-Making

Scenario 2: Network Change Agents Enable Precise Simulation and Verification, as Well as Error-Free Network Configuration

Scenario 3: Network Fault Agents Enable Intelligent Inspection and Recovery, Efficiently Resolving Silent Faults

Trend 4: Network Security Enters the Era of AI-Based Attack and Defense

Scenario 1: Lightweight Graph AI Detection Models Are Used to Prevent Ransomware Variants

Scenario 2: Self-Learning AI Models Enable Efficient Detection of Encrypted Attacks

Scenario 3: Collaboration Between Small and Large Models Achieves Security Event Noise Reduction and Intelligent Assisted Handling

Summary: Three-Layer Intelligent Architecture, Accelerating Network Intelligence with Network-Security Integration

Recommendations for Action

01

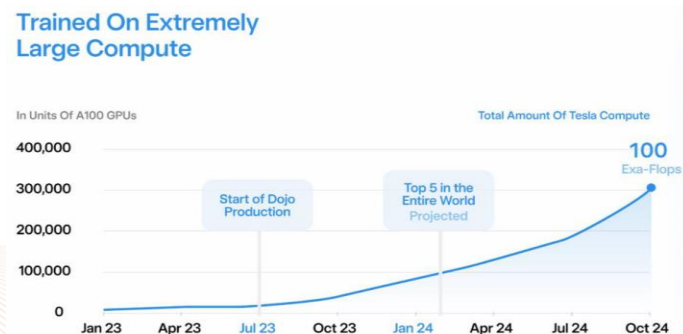
| Network for AI



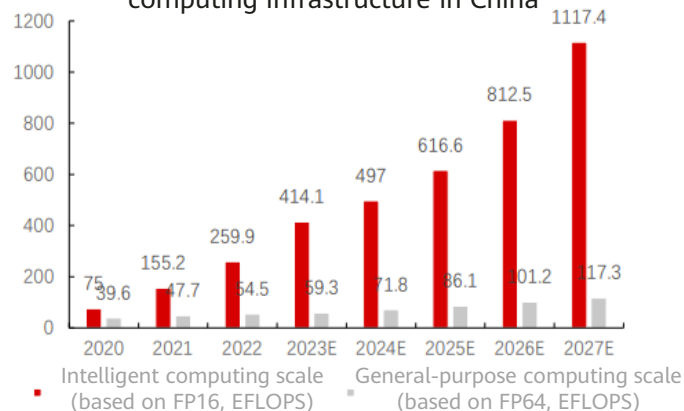
▶ Intelligent Productivity Is Ready, with Ongoing Consolidation of AI Infrastructure

- Foundation model training is continuously accelerating, and intelligent productivity is about to take off.** Foundation model applications have progressed from **phenomenal B2C applications** to **general B2B applications** and are now evolving into **scenario-based B2B applications**. These scenario-based applications represent the core production scenarios for enterprises, making the rapid iteration capability of foundation models essential. For instance, Tesla needs to reduce the workload for self-driving training from one month to just one week, with over-the-air (OTA) updates occurring every two to three weeks to meet safety and competitiveness demands. It is anticipated that the emerging requirements in the intelligentization processes across numerous industries will shorten the training duration of foundation models to days or even hours.
- Investment in AI infrastructure is on the rise, with computing services gaining popularity.** The investment in intelligent computing is accelerating. For instance, in China, intelligent computing power is projected to reach 1117.4 EFLOPS by 2027, reflecting a compound annual growth rate (CAGR) of 33.9% from 2022 to 2027. In contrast, general-purpose computing power is expected to grow at a CAGR of 16.6% during the same period. China's three major tier-1 carriers have unveiled strategies focused on intelligent computing cloud services.
- The collaborative development of networks and computing power establishes a solid foundation for business value realization.** In intelligent computing cloud services, computing power is crucial, while the computing network serves as the backbone. For example, China Mobile is constructing a "4+N+31+X" multi-level intelligent computing center and a computing network called MATRIXES, connecting to third-party computing resources through the Baichuan platform to achieve ubiquitous networking, computing, and intelligence.

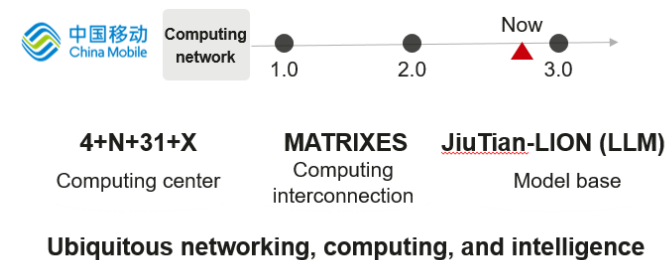
Tesla: Autopilot training continues to improve efficiency



Investment trend of intelligent computing infrastructure in China



China Mobile's intelligent computing strategy



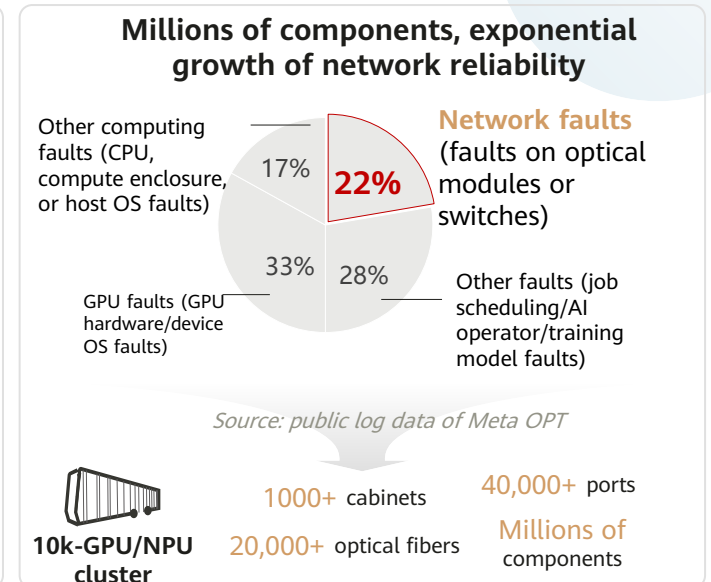
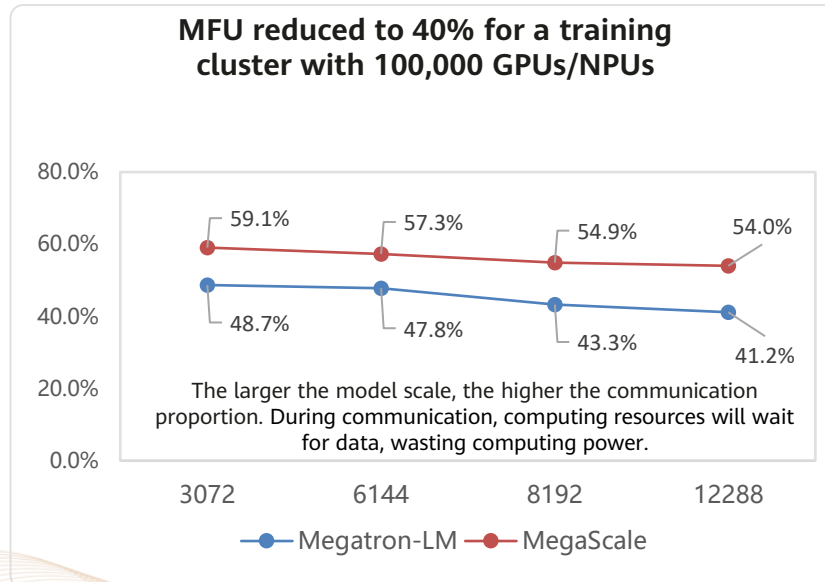
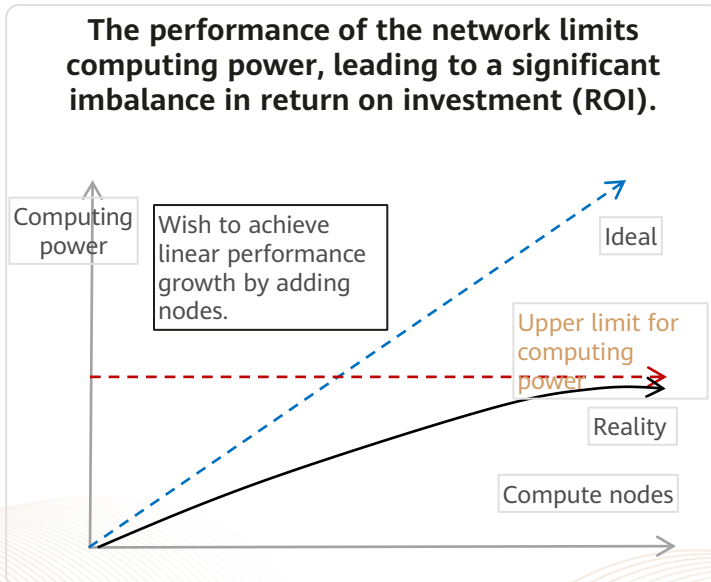
► Trend 1: Intelligent Computing Clusters Embrace the Era of 100,000+ GPUs/NPUs

- **The cluster scale has rapidly increased from 10,000+ GPUs/NPUs to 100,000+ GPUs/NPUs.** The rise of foundation models has led to a continuous increase in their parameters, outpacing Moore's Law regarding the improvement of single-GPU/NPU computing power. As a result, cluster sizes are growing and have now entered the era of 100,000+ GPUs/NPUs.
 - ✓ At the beginning of 2024, Meta released two clusters with 24,576 Nvidia H100 GPUs for training next-generation generative AI models.
 - ✓ In July 2024, Elon Musk announced that the xAI team began training the new chatbot GROK 3 on the Memphis supercluster equipped with 100,000 H100 GPUs.
 - ✓ ByteDance builds an Ampere training cluster with 12,288 GPUs and develops the MegaScale production system for training large language models (LLMs).
 - ✓ In 2023, iFLYTEK built **FlyingStar One**, the first **10,000+ GPU/NPU cluster computing platform** that supports foundation model training.
 - ✓ On February 4, 2024, the Shenzhen Smart City Computing Power Coordination and Scheduling Platform **built the strongest computing cluster with 100,000 GPUs/NPUs** (Hetao-Xilihu-Guangming Science City).
 - ✓ The construction of the new intelligent computing center with **130,000 GPUs/NPUs** in area B of the Runze (Langfang) International Information Port has started and is expected to be delivered by the end of **2025**.



1.1 - Ultra-Large Single Cluster: High-Quality Intra-DC Network Is Essential for Unleashing Computing Efficiency

- **The network is crucial for the training efficiency of foundation models in a cluster.** The communication methods used in AI training differ significantly from those in traditional training, varying based on the architecture of the foundation model. In some cases, communication can account for up to 50% of the training process. As the number of model parameters increases and the cluster size expands, data synchronization times lengthen, making network communication efficiency increasingly impactful on training performance. **Notably, a foundation model utilizing 100,000 GPUs/NPUs demands a high-quality network.**
- **A high-quality network must be non-blocking and exhibit low latency to effectively utilize the computing power of 100,000 GPUs/NPUs.** In large-scale AI scenarios, vast numbers of parameters are distributed across multiple GPUs/NPUs on servers, requiring up to 100,000 GPUs/NPUs to train datasets of tens of terabytes or more. Issues such as uneven network load balancing or high latency can lead to idle computing power, reduced algorithm efficiency, and potential **saturation** during communication among numerous GPUs/NPUs.
- **For long-term training of a large cluster with 100,000 GPUs/NPUs, a high-quality network must be stable and robust.** Training foundation models is a complex system project, and maintaining stable system operations is vital. The network infrastructure is key to ensuring long-term stability during training. For instance, the total training duration for a 100-billion parameter model is 65 days, but frequent faults can cause the model to restart over 50 times, reducing the effective training time to just 33 days. In a cluster with 100,000 GPUs/NPUs, the scale and complexity sharply increase, heightening the risk of faults. Additionally, long fault recovery times can lead to system availability dropping below 60%.

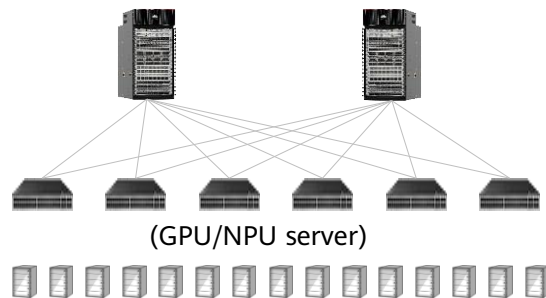


Key Capability 1: Ultra-Large-Scale Networking and 800GE Ultra-High-Speed Interconnection

- **A new networking architecture is needed to support 100,000 GPUs/NPUs.** The two-layer modular-fixed switch networking currently accommodates a maximum of 32,000 GPUs/NPUs (as seen with Meituan), while the three-layer fixed-switch-only networking can support up to 500,000 GPUs/NPUs. This architecture aims to facilitate millions of GPUs/NPUs in the future. However, as we enter an era with over one million GPUs/NPUs, adding more network layers will introduce challenges, including increased hops, higher latency, network complexity, reduced effective load, and elevated interconnection costs. The industry is actively exploring new interconnection solutions based on architectures like Dragonfly and Torus.
- **800GE is now being deployed on a large scale to build ultra-broadband networks.** As 400GE applications become more widespread, data center networks are advancing rapidly towards 800GE. According to a five-year forecast report on the data center switch market released by Dell'Oro Group, the number of 800GE switch ports is expected to surpass that of 400GE by 2025.

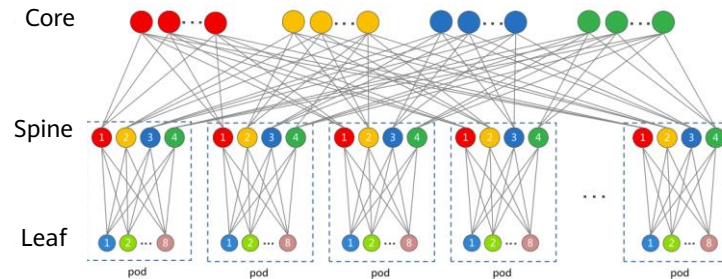
Two-layer Clos architecture

32,000 GPUs/NPUs @ 200GE access
(400GE interconnection)



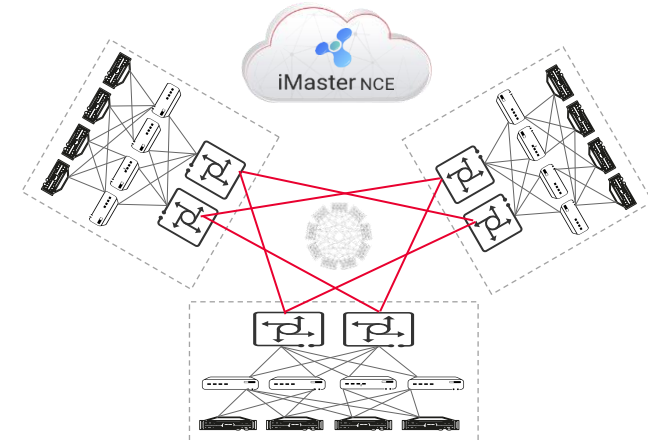
Three-layer Clos architecture

500,000–1 million GPUs/NPUs @ 400GE/800GE access
(800GE interconnection)



Dragonfly-based Groupwise DF+ architecture

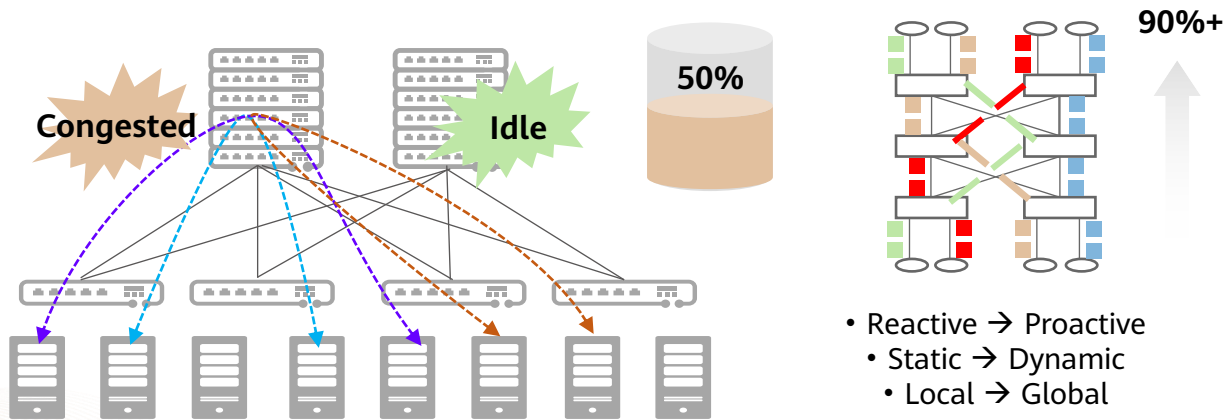
> 1 million GPUs/NPUs @ 800GE access
(800GE/1.6TE interconnection)



Key Capability 2: Network-Level Load Balancing for Zero Congestion and Ultra-High Throughput, Unlocking Computing Power

- **Improving network throughput is the key to improving AI training efficiency:** Currently, leading companies, including top Internet enterprises, large AI R&D firms, and carriers in China, are building or utilizing 10,000-GPU/NPU clusters and actively planning future-oriented 500,000-GPU/NPU clusters. In AI computing scenarios, traffic is characterized by low volume but high single-stream bandwidth. However, the effective throughput of networks is often low due to load imbalances, typically around 50%. When the network architecture shifts from two layers to three layers, the number of network paths increases exponentially, exacerbating load imbalance and potentially reducing overall throughput to as low as 20% (based on lab simulation data with 500,000 GPUs/NPUs).
- **Network-level load balancing for higher network throughput:** To improve network throughput, mainstream industry players are adopting similar optimization strategies. To meet the demands of AI training, there is a deep coordination and adaptation among the peer end, network, and protocol, aiming for network-wide load balancing and achieving throughput rates exceeding 90%, thereby enhancing communication efficiency. Currently, Huawei employs Network Scale Load Balancing (NSLB) technology to achieve 98% effective network throughput in a two-layer Clos architecture (such as Meituan). For three-layer networking, an upgraded intelligent load balancing algorithm has led to further breakthroughs and adaptations.

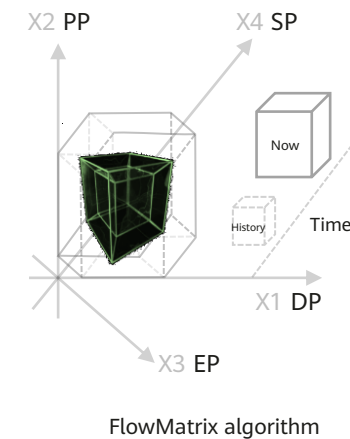
10,000-GPU/NPU model training: Huawei NSLB technology improves the network throughput to over 90%



NSLB-DP: Intelligent load balancing technology for 100,000+ GPUs/NPUs

The three-layer network architecture increases the number of network paths from 4.5 million to 30 billion, increasing the difficulty of load balancing.

Five-dimensional modeling based on the FlowMatrix algorithm, reducing the complexity from M^N to $M*N$, and adjusting network-wide links in seconds



1.2 - Cross-DC Collaborative Training: Long-Distance Lossless Inter-DC Network, Aggregating Distributed Computing Power

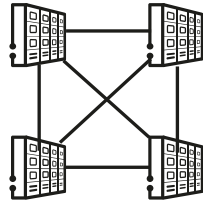
- **The scale of a single intelligent computing center is limited, which has led to the need for multi-DC collaborative training.** This limitation is influenced by factors such as equipment room conditions and power consumption. Cross-DC collaborative training has emerged as an effective method for aggregating computing power. For example, Google Gemini Ultra employs cross-DC collaborative training based on Cloud TPU v4 for model training. Additionally, Microsoft OpenAI Labs has announced that GPT-6, scheduled for release in 2025, will require cross-region training due to power supply challenges. Furthermore, major carrier players in the computing market **have extensive existing CO/DC resources and are looking to maximize their value by leveraging opportunities in intelligent computing construction.** Companies like China Mobile and China Telecom Guangdong are actively exploring solutions to meet the future demands of ultra-large model training and enhance their competitiveness in computing services.

Distributed deployment of intelligent computing centers, breaking power bottlenecks

Single intelligent computing center



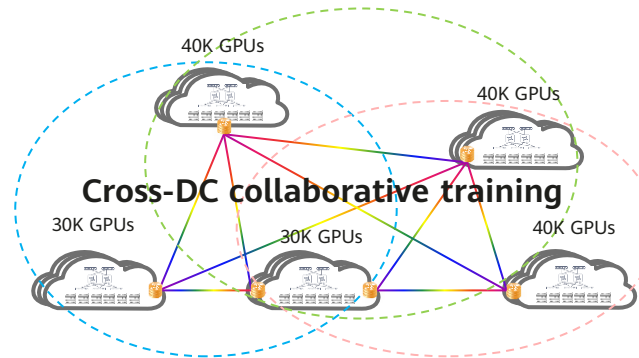
Multiple intelligent computing centers



> 750 mW GTP-6 power consumption

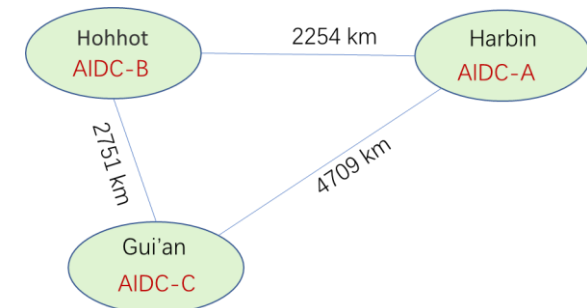
Microsoft predicts that **GPT-6 is limited by grid capabilities** and can only be trained through multi-regional computing centers.

Collaborative training mode 1: Intra-region computing power aggregation (Guangdong-Hong Kong-Macao Greater Bay Area, Beijing, Tianjin, and Hebei)



GPT-5 (100,000 GPUs): power consumption 380 mW; annual electricity consumption 3.3 billion kWh \approx 1/10 of the annual electricity consumption in Hong Kong

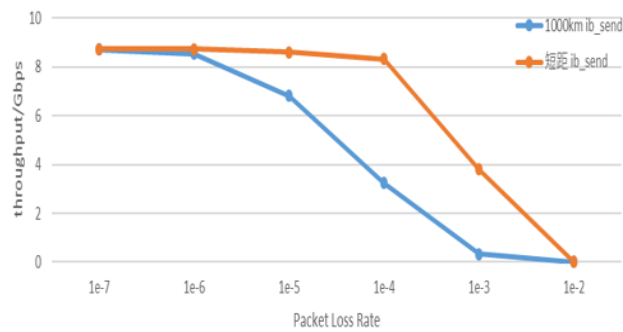
Collaborative training mode 2: Computing power collaboration between hubs, meeting the requirements of ultra-large model training (China Mobile)



Key Capabilities of the Inter-DC Network: Zero Packet Loss, High Utilization, and Effective Response to High Bursts

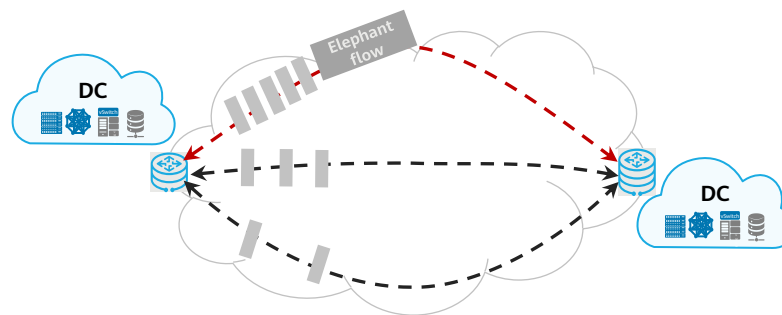
- **Cross-DC collaborative training requires zero packet loss on the DCI network.** Compared to traditional services, AI training data is particularly sensitive to packet loss; even a packet loss rate of 0.1% can reduce training efficiency by 50%, severely impacting collaborative computing effectiveness. Therefore, implementing lossless transmission on the WAN, ensuring efficient operation of collaborative computing, and building a super-connectivity intelligent WAN with zero packet loss are crucial.
- **Cross-DC collaborative training must address the issue of low network utilization.** The traffic characteristics of AI training involve a "small number of service flows and large burst traffic," commonly referred to as elephant flows in the industry. Research indicates that at the 10,000-GPU/NPU level, elephant flows can cause traditional 5-tuple-based load balancing methods to fail, resulting in unbalanced loads and low network utilization.
- **Cross-DC collaborative training needs to tackle the problem of traffic bursts.** In a 10,000-GPU cluster, a single GPU supports 200 Gbps interconnection, and during parameter synchronization, the theoretical traffic rate can reach 51.2 Tbps. However, due to high service bursts and concurrency, the actual instantaneous concurrency can surge to 1600 Tbps. Currently, the interconnection bandwidth between DCs cannot meet these demands, necessitating traffic shaping and convergence on network devices.
- **Leading enterprises have begun implementing cross-DC collaborative training.** For instance, China Telecom Beijing has actively explored remote intelligent computing solutions, conducting tests in three AI training DCs: Yinghai, Wuqing, and Yongfeng. The 100-km collaborative training has proven feasible, with linearity decreasing by **less than 5%** compared to a single DC. However, simulation and evaluation indicate that latency over 1000 km is excessive, limiting support to cross-region **parallel data training**. A DCI interconnection bandwidth of **50 Tbps** or more can accommodate the requirements of **trillion- or 10-trillion-level sparse models**.

The throughput of long-distance RDMA (1000 km) decreases by 60% when the packet loss rate is 0.01%.

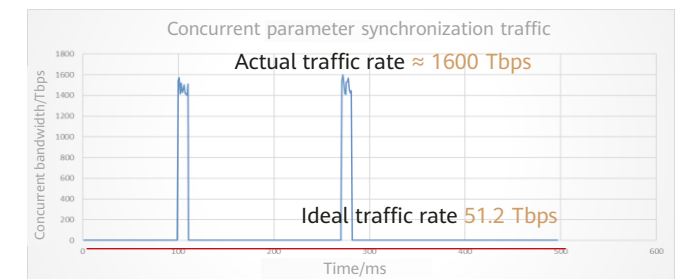


Impact of the packet loss rate on the throughput (1000 km vs. 2 km)

Traditional networks are prone to link load imbalance, resulting in low network utilization.

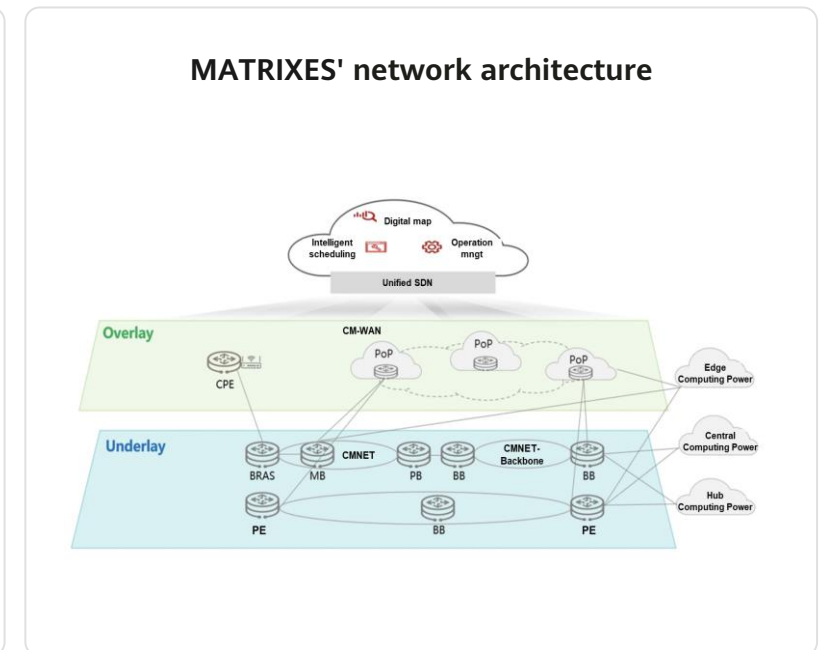
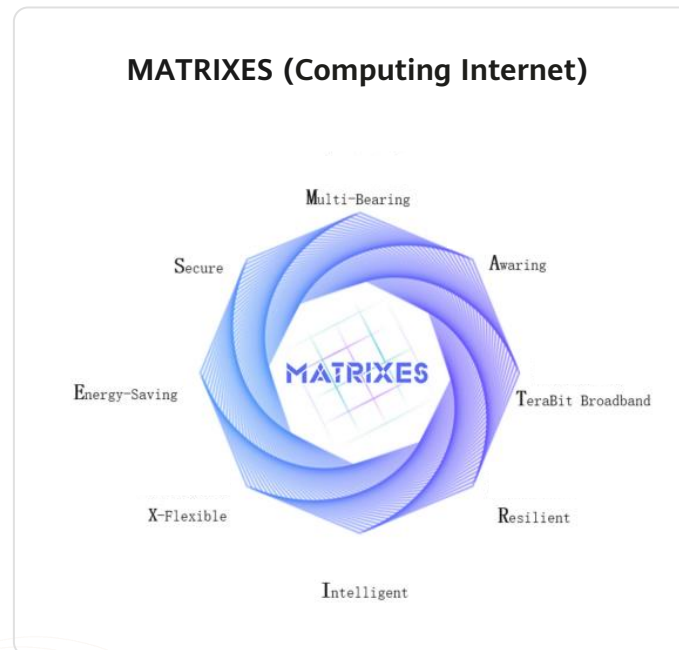
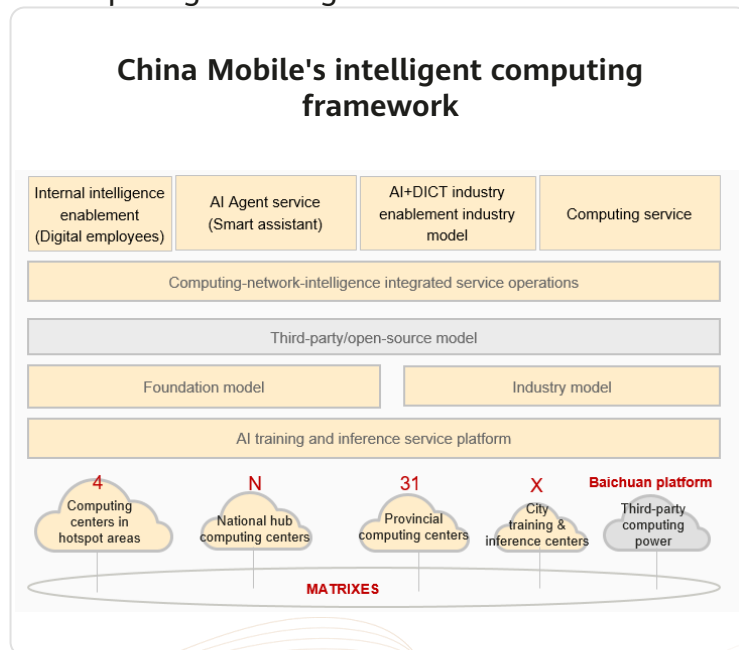


The DCI convergence ratio and training efficiency need to be **optimally balanced**, aiming for convergence at the 100 Tbps level.



Trend 2: Super-Connectivity AIDC-Access Network Construction Accelerates, Enabling Business Monetization of Intelligent Computing Cloud Services

- Super-connectivity AIDC-access network, commercializing "computing power" facilities:** Once the computing power infrastructure is established, the next challenge is how to effectively "serve a large number of customers and optimize computing power" to realize a positive business cycle within intelligent computing DCs. In China, less than 25% of completed intelligent computing DCs currently offer intelligent computing cloud services. The lack of a high-quality network to efficiently connect customers, end users, AI applications, and intelligent computing DCs — facilitating ultra-high-speed data transfer — remains a significant barrier. This super-connectivity AIDC-access network primarily serves two scenarios: foundation model training for industry customers and model inference for a vast number of end users.
- Building a high-quality AIDC-access network has emerged as a new industry trend.** China Mobile has launched MATRIXES, introducing new services such as elastic private lines. China Telecom has upgraded its cloud-network strategies, and China Telecom Shanghai is actively exploring technological innovations and business introductions in new scenarios like "sample data express." Additionally, China Unicom is utilizing CUBE-Net 3.0 as the top-level architecture for network transformation over the next 5 to 10 years, aiming to create a converged service model of "connection + computing + intelligence."



1. Remote Storage-Compute Collaborative Training, Advancing the AIDC-Access Network to a Lossless State

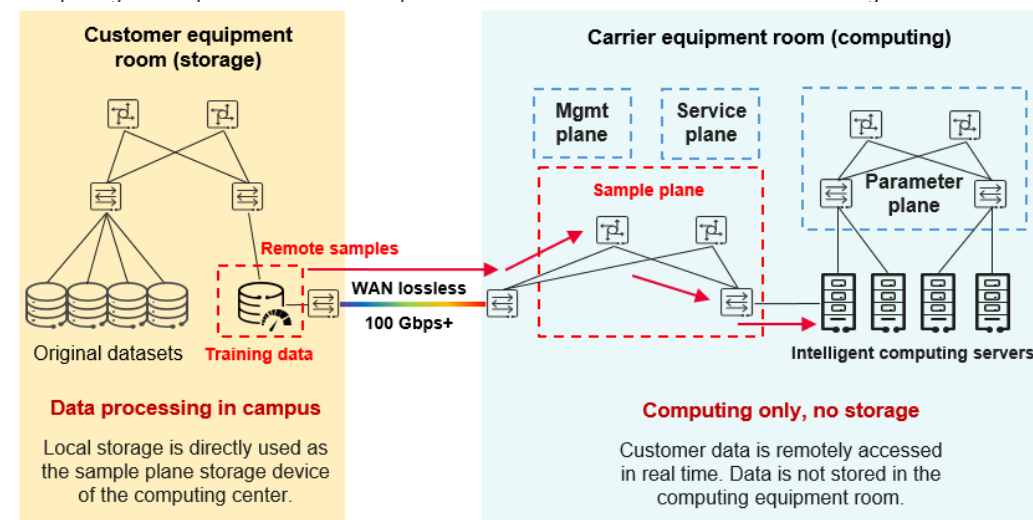
- **Sensitive data must remain on campus, with support for remote storage and compute training.** To ensure data security, some enterprises or industry customers require that data not be transmitted outside the campus during transfers from research institutes to the computing center.
- **Remote storage and compute training necessitates new network capabilities.** On one hand, training services are highly sensitive to packet loss, requiring the AIDC-access network to provide lossless services on demand. On the other hand, the network must maintain end-to-end high effective throughput capabilities.

To enhance data security, sensitive data is kept out of the storage zone in the intelligent computing center.

Industry	Government			Healthcare			Finance		
Intelligent Computing Scenario	Government office automation			Automatic screening of typical cases			Credit risk identification		
Pain points	Confidential protection			Not transmitted out of the hospital			Physical isolation		
	Sensitive information such as official documents, citizens, legal persons, and addresses			Electronic medical records, epidemiological data, and gene data			Bank/insurance/credit investigation/telecom/payment information		

Survey 1: Shanghai Stock Exchange data center

Securities company data is stored in the private domain. When the training frequency is high, training data changes greatly, and a large volume of incremental data is involved, encrypted connections are required to streamline storage and computing and implement remote sample

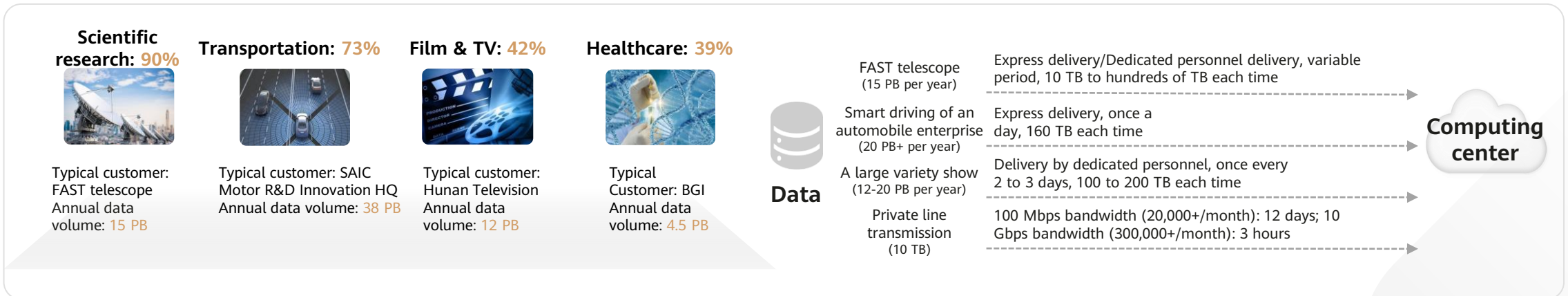


Survey 2: GD Government Data Bureau

Government-facing foundation models are deployed in the SG intelligent computing center, and ZS leases computing power of the intelligent computing center for foundation model training. Due to the sensitivity of the information involved, users prefer to store data locally and leverage remote computing power through the network for foundation model training.

2. The Need for Ultra-Fast Delivery of Large Volumes of Samples Highlights the Importance of Building an Elastic AIDC-Access Network

- **Foundation models are evolving towards trillions of parameters and multimodal, making TB/PB-level sample data transmission a significant challenge.** The traditional hard disk and manual transfer method is inefficient, costly, and prone to data loss. For private line transmission, low bandwidth is insufficient, and high bandwidth is too expensive.
- **Elastic private lines (AIDC-access network) have become a trend.** This approach uses a basic package combined with elastic traffic charging to facilitate task-based large-scale sample transmission. **China Telecom Shanghai** and Huawei are currently conducting scenario-based technical verifications for fast data transfer solutions.

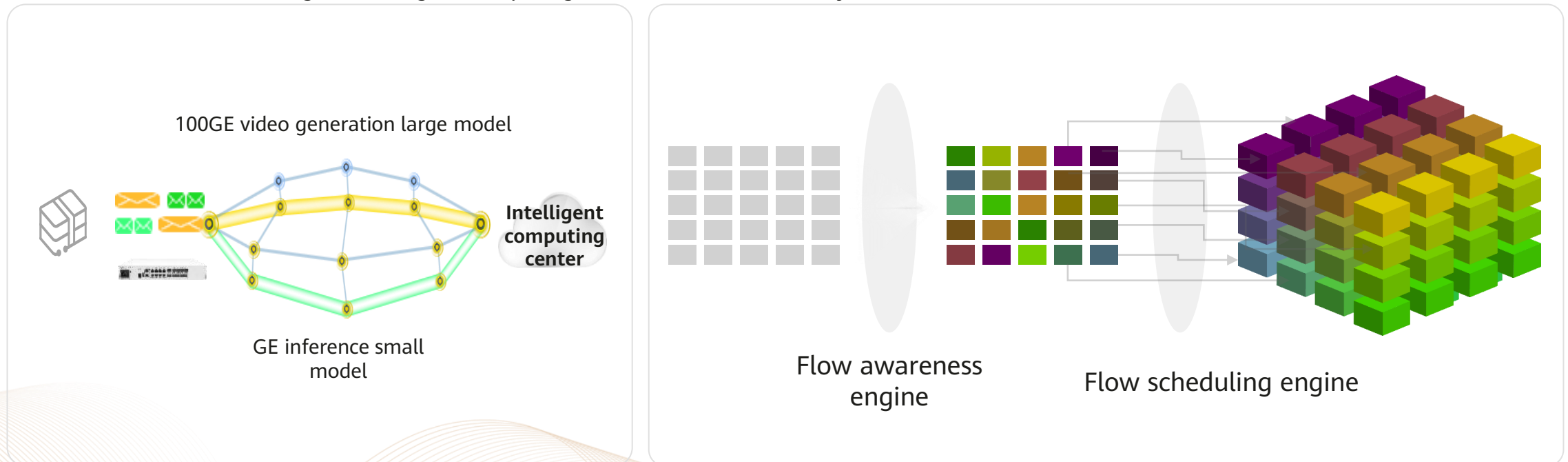


SAIC Motor R&D Innovation HQ - Smart driving case: 20 test vehicles in four cities work for 20 days every month. Each vehicle generates 8 TB data every day. A total of **160 TB** data (40 hard disks) needs to be sent to the Shanghai HQ IDC by express (**three days**). The data is manually copied to the smart driving training center (**three days**). After the data is uploaded, the hard disks are returned. The entire process **takes approximately two weeks**.



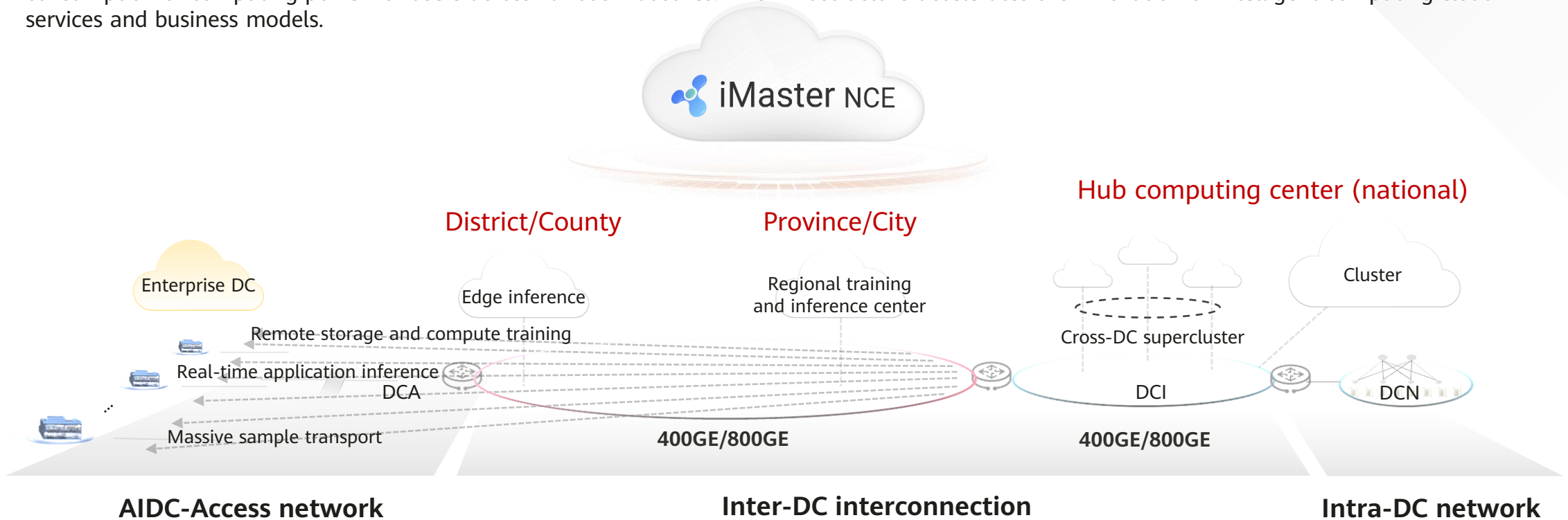
► Key Capability: Highly Elastic, Concurrent Computing

- **By leveraging network capability openness and innovating business models, collaborative elastic scheduling with computing power can support the rapid delivery of massive samples.** The essence of elastic scheduling lies in the new task-based service model. Unlike traditional network services, which allocate fixed bandwidth resources that may not meet the demands of high or low bandwidth scenarios, task-based elastic scheduling enables enterprises to implement on-demand and flexible data transmission. This approach accelerates service processes and enhances market competitiveness. To effectively implement elastic scheduling, network capabilities must be open and work in tandem with computing power scheduling, complemented by innovative business models.
- **Aimed at a vast number of intelligent computing users, boosting overall network throughput and service capabilities through network-wide scheduling is essential for establishing a closed-loop business model for computing interconnection and services.** Typically, the load rate of traditional networks is below 50%. However, intelligent computing services primarily involve elephant flows due to high burst traffic. Therefore, intelligent scheduling is essential to improve the overall load capacity of the network. Even with a load rate of 80%, intelligent computing services remain unaffected, allowing the intelligent computing network to simultaneously serve more users.



Network for AI: Collaborative Construction of Intelligent Computing Networks and Power, Enabling On-Demand Intelligence for Enterprises

- **Intelligent computing infrastructure construction:** Plan the layout and construction timeline of intelligent computing centers based on the service requirements of end users, the actual development needs of customers, and the specifications of equipment rooms and power supply.
- **Intelligent computing network infrastructure construction:** An **elastic ultra-broadband** AIDC-Access network, a **long-distance lossless** inter-DC interconnection network, and a **stable, reliable, ultra-large-scale** intra-DC network with **zero congestion** enable efficient production and on-demand consumption of computing power for users across various industries. This infrastructure accelerates the innovation of intelligent computing cloud services and business models.



► Network for AI: Recommendations for Action

- **Carry out unified planning and collaborative construction for the network and computing center.** First, enhance the collaborative construction of the network and computing center through integrated planning and investment, ensuring that computing power is available on demand for target customers. Second, strengthen the full automation and collaboration of network and computing power at the platform scheduling layer, allowing users to access these resources on demand in a task-based manner. Third, promote the innovation of business models for computing power and networks to make intelligent computing cloud services user-friendly and affordable.

- **Introduce a new DCN networking architecture to accelerate the deployment of 800GE.** Actively implement a new DCN networking architecture along with ultra-high-speed interconnection technologies, such as 800GE, to create ultra-large computing clusters. This approach aims to achieve high linearity while maintaining acceptable networking costs and enhancing overall availability.

- **Accelerate the introduction of innovative technologies and promoting the construction of long-distance lossless inter-DC networks.** Actively collaborate with vendors to integrate innovative technologies for building inter-DC networks. Additionally, partner with leading customers to establish industry benchmarks, fostering industry development, enhancing solution maturity, and incubating new business models through innovation.

02

| AI for Network



▶ AI for Network: AI Injects New Innovation Vitality into Networks

- **The demand for network innovation never stops:** As the cornerstone of the digital world, networks are becoming increasingly important as we stride towards the all-digital, all-intelligence era. Just like water and electricity in the physical world, networks have become an indispensable infrastructure. Today's networks underpin more services and expand the scope of connectivity from people to things and the cloud. In the future, this will even extend to digital humans. Furthermore, the rise of new technologies poses higher requirements on networks, including high network quality, automation, and differentiated services. Networks need to further evolve to meet these needs. As networks connect more entities and support more services, network architecture becomes more complex than ever, not to mention higher requirements for network operations and maintenance (O&M). However, the number of available highly talented network professionals and enterprises' investment in network O&M do not increase accordingly. As such, there is an urgent demand for network autonomy. Beyond this, as cloud computing and AI technology advance, network attacks become more rampant. It is imperative to draw on emerging technologies and techniques to further ensure network security. In a word, the demand for network innovation has never stopped and becomes more urgent, instead.

- **AI injects new vitality into network innovation:** AI technology advances can comprehensively improve network capabilities to new levels. During network planning and deployment, we can leverage AI to forecast requirements and guide through precise capacity expansion based on historical and real-time traffic data. We can also draw on AI to simulate planning for enterprise networks such as wireless networks and provide guidance on installation and commissioning. As for network O&M, we can harness the power of AI to efficiently utilize network resources and automatically adjust resources for optimized network performance and user experience. We can also capitalize on AI to analyze application experience, identify behavior, and ultimately implement proactive experience assurance and precise security protection. Beyond this, we can tap into AI for fault detection, demarcation, locating, and troubleshooting, as well as network energy consumption optimization. One or more of these AI use cases have been proven valuable on networks. **Chain-of-thought (CoT) technology in foundation models** can break down and process complex issues, accelerating experience optimization and troubleshooting across cross-domain networks. We firmly believe that in the future, CoT will drive **systematic mass AI adoption** in networks, injecting new service capabilities into networks.

Trend 3: The Integrated Development of Digital Twins and AI Accelerates the Pace Towards Level 4 Autonomous Driving Networks

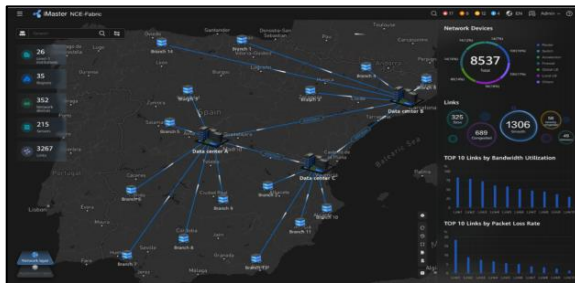
1. Network Digital Twins (NDTs) are a solid foundation for network intelligence.

- ✓ NDTs revolutionize traditional network management approaches because they provide trial and verification for change operations such as adjustment, maintenance, and optimization, slashing trial-and-error costs. That's why standards organizations like ITU, IETF, ETSI, TMF, and CCSA are actively defining NDT architectures, technologies, and standards.
- ✓ Likewise, leading industry players like China Telecom, China Mobile, Cisco, H3C, and Juniper Networks have successively unveiled their NDT architectures and products. According to Gartner's prediction, 50% of network vendors will provide NDT functions in their solutions by 2026.

2. Network foundation models have moved from hype to reality.

- ✓ Carriers and equipment vendors actively engage in network foundation models through in-house development and external procurement. Gartner predicts that by 2027, 40% of CSPs worldwide will adopt AI and ML and more than 90% of enterprises will integrate AI into network management.
- ✓ At MWC Barcelona 2024, Huawei launched NetMaster, the industry's first network foundation model. NetMaster builds on Huawei's more than 30 years of ICT network prowess, and is trained with an impressive arsenal of over 50-billion-level corpus. With these traits, NetMaster covers 44 scenarios in 6 categories in the data communication field.

In 2022, Huawei released the industry's first digital twin base.



1600+ practices @ AIS in Thailand, XL in Indonesia, Orange in Spain, China Zhesang Bank, etc.

- Holographic network visibility: network-wide SLA visualization
- Optimal network performance: 30% lower network latency according to the P3 test
- Optimized network autonomy: path optimization time shortened from 3 months to 3 minutes

ITU launched the intelligent O&M architecture.

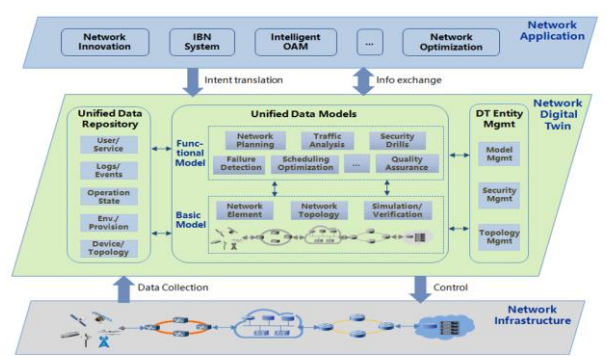


Figure 2: A reference architecture of Digital Twin Network

In December 2023, ITU officially released the international standard ITU-T Y.3550 for intelligent O&M.

Typical network foundation model products

	Developed an in-house Jiutian AI model to provide network intelligence.
	Set up GTAA to develop LLMs tailored to the telecom industry.
	Connected to Microsoft's foundation models to provide Q&A capabilities.
	Developed NetMaster based on Pangu models.
	Connected to OpenAI to provide Q&A capabilities.

Enterprises with a wealth of network O&M data will gain advantages in building network foundation models.

Network Intelligence Drives O&M Transformation from Simple Q&A to Complex Scenarios

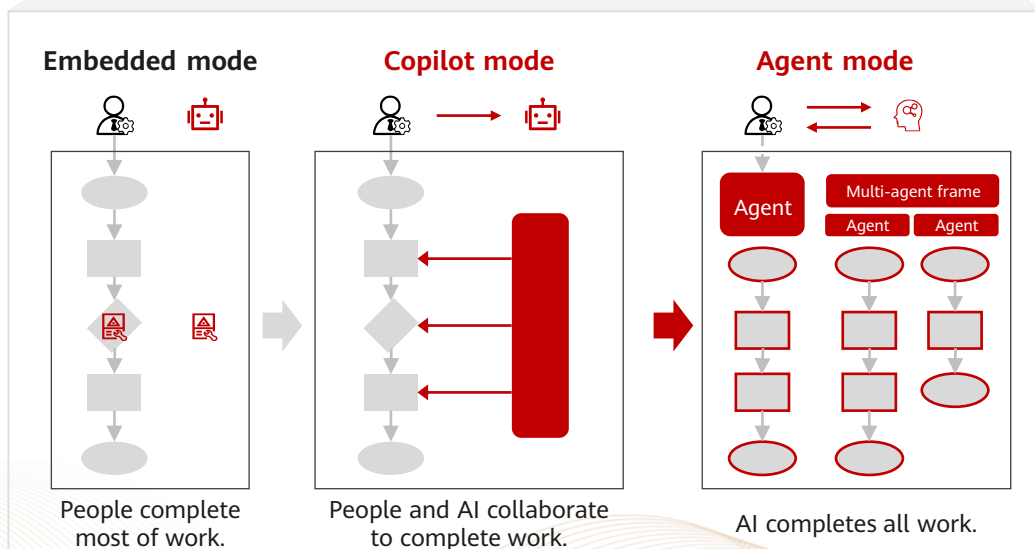
1. Human-machine collaboration transforms from copilot to agent modes, driving network intelligence to evolve in phases.

- ✓ The rapid maturity of the foundation model industry drives network intelligence to evolve from embedded intelligence to copilot- and AI agent-centric intelligence. This further supplements the capabilities of network foundation models, ultimately reshaping network O&M paradigms.
- ✓ **Role-based copilots reshape O&M experience:** By enabling intelligent language interactions, copilots can understand complex technical problems and provide accurate solutions for O&M personnel, helping them quickly resolve network issues.
- ✓ **Scenario-based agents enable scenario-specific autonomy:** AI agents provide self-closed-loop capabilities for different O&M scenarios to implement intelligent O&M processes. LLM-based agents will be the megatrend for AI innovation in the next 5 to 10 years.

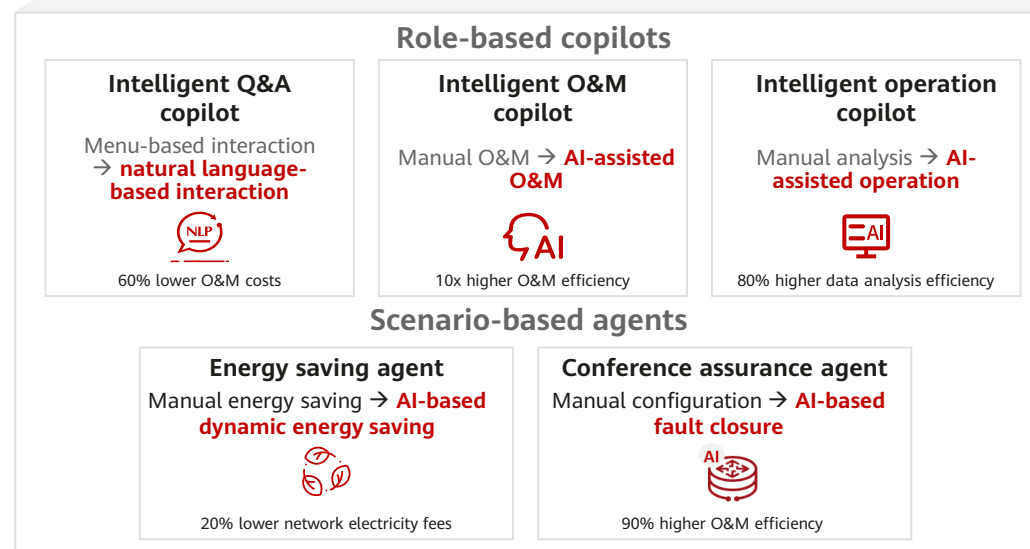
2. Mainstream vendors are actively developing copilot and AI agent products.

- ✓ Copilot and agent capabilities depend on network foundation models. Well-established equipment vendors boast a wealth of high-quality training corpora and have innate advantages in developing competitive copilot and agent products. They can flexibly develop copilot and agent products throughout the network lifecycle from planning and construction to maintenance and optimization.
- ✓ Regarding network foundation models, Huawei has released a range of products based on roles and service scenarios. By leveraging atomic capabilities that can be intelligently combined and generalized, these purpose-built products help quickly build up scenario-specific intelligence for improved O&M efficiency. A typical example is agents for energy saving on a smart green campus. Such agents can detect campus-wide service loads in real time and predict tidal enterprise traffic. This translates into over 20% network energy savings and 30% building energy savings.

Copilots and agents are two core phases of network intelligence.



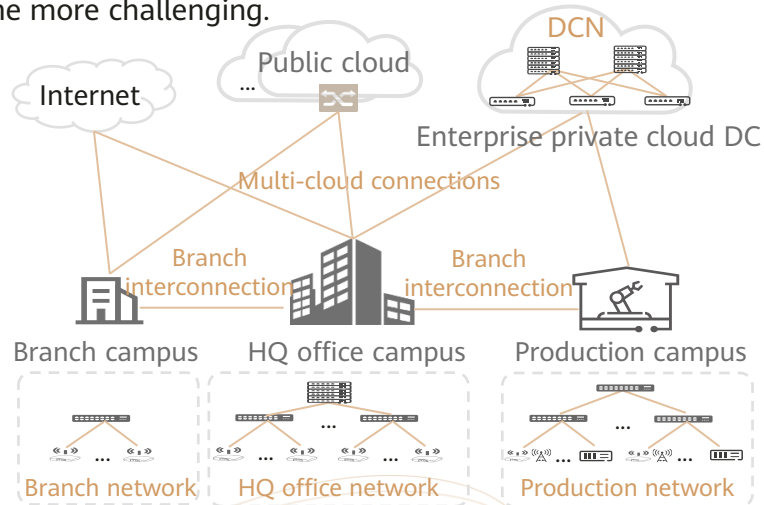
Huawei released copilot and agent products for campus networks.



► Autonomous Driving Networks Reshape Enterprise Network Experience and O&M

Challenges for enterprise networks

- The advent of fully-wireless offices and the rise of cloud, video, collaboration, and intelligent applications make it difficult to ensure employees' office experience.
- The growing demand for enterprise office, production, data center, branch, and multi-cloud connections leads to larger network scale and more device types, as well as larger scope and higher complexity of routine maintenance.
- Service deployment involves multiple network domains, and it's a time-consuming process from design and integration to verification and provisioning.
- The number of virus variants increases exponentially, attacks become more intelligent, and security operation risks and unknown threats become more challenging.



Leverage telecom foundation models and network digital twins to develop enterprise scenario-based agents and role-based copilots. The resulting benefits include but are not limited to the following:

Zero service suspensions • Ensures deterministic experience for any user, on any device, in any application.

Zero network interruptions • Gains real-time visibility into network-wide status, proactively detects and eliminates network risks, and automatically recovers from non-hardware faults.

Zero-wait provisioning • Automatically generates network configurations, performs simulation and verification, and provisions new services in real time.

Zero security risks • Continuously improves the capability of detecting virus variants and unknown threats, and automatically handles security threats.

Network optimization agent

Network fault agent

Intelligent customer service copilot

Configuration generation agent

Security assistance copilot

...

Scenario 1: AI Agents Collaborate with RAG/Small Models to Improve Domain-Specific Q&A and Decision-Making

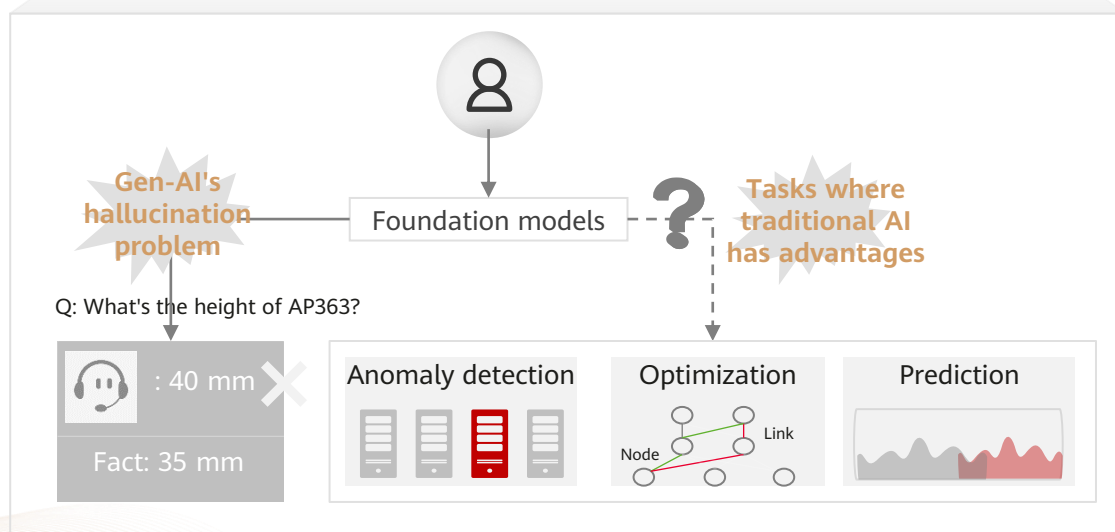
1. Occasional hallucinations in foundation models cannot be eliminated, and Retrieval Augmented Generation (RAG) is the main workaround.

- ✓ **The hallucination issue in foundation models cannot be eliminated.** This is because the training data of any foundation model cannot cover all knowledge and scenarios.
- ✓ **RAG is a widely accepted method of hallucination avoidance.** RAG dynamically retrieves information from external knowledge sources and uses the retrieved data as a reference to organize answers. This greatly improves the accuracy and relevance of responses.

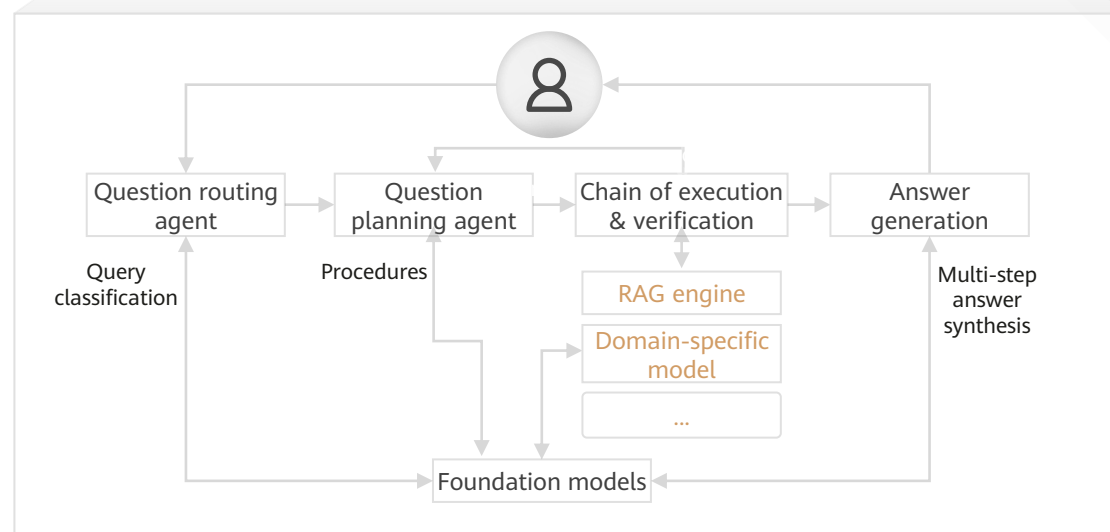
2. AI agents collaborate with RAG/small models to improve Q&A accuracy.

- ✓ Agents are used to plan questions and implement multi-turn conversations to supplement missing information. Plus, step-by-step RAG and multi-step answer synthesis are combined to handle complex questions.
- ✓ The paradigm of collaboration between small and large models can inherit domain-specific complex service logic. More importantly, it can inject new vitality into traditional AI by leveraging the understanding, expression, and generalization capabilities of foundation models.

As-Is: Foundation models have hallucination issues.



To-Be: AI agents and RAG/small models collaborate for faster and more accurate responses.



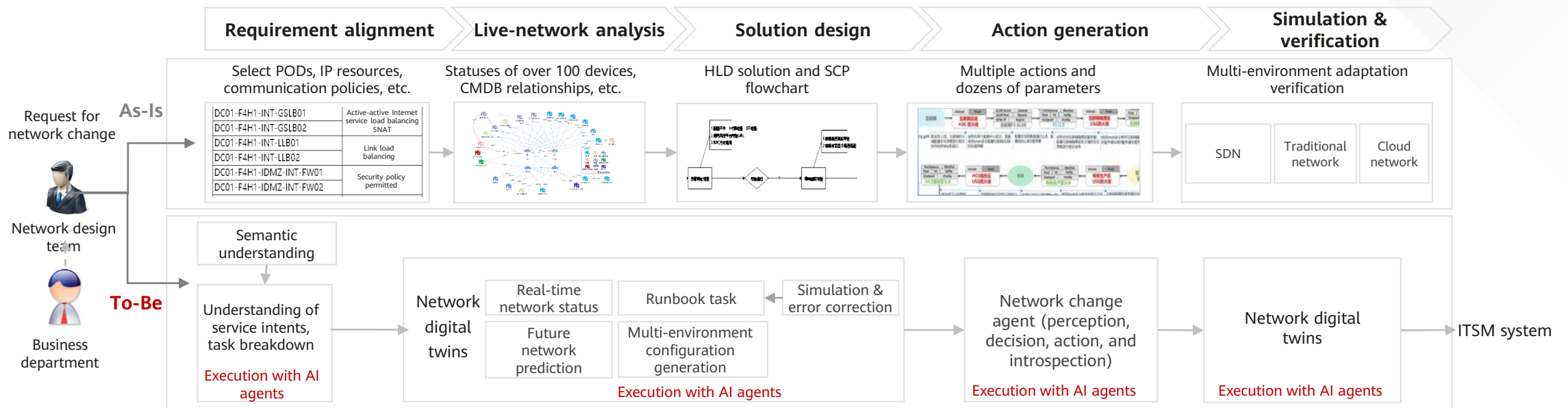
Scenario 2: Network Change Agents Enable Precise Simulation and Verification, as Well as Error-Free Network Configuration

1. Network change is time-consuming and error-prone. More than 50% of change failures are caused by manual errors.

- ✓ In the past two years, more than 10 carriers around the world have encountered major network accidents, affecting tens of millions of end users and causing immeasurable losses.
- ✓ 50% of such network accidents are caused by manual errors, which cannot be radically eliminated.

2. Network change agents understand user intents and enable online simulation, ensuring error-free network configuration.

- ✓ Network digital twins provide full-stack online simulation and verification. Specifically, based on network element (NE) configuration data, the system simulates device routing protocol behavior, accurately generates NE protocol routing tables and global routing tables, and performs analysis based on routing entries to verify network impact.
- ✓ Network foundation models upgrade interaction modes. Users just need to specify the intent for network change. Network change agents then automatically generate network change configurations based on actual service scenarios, automatically verify these configurations, intelligently identify configuration risks, and automatically deploy configurations after successful verification.



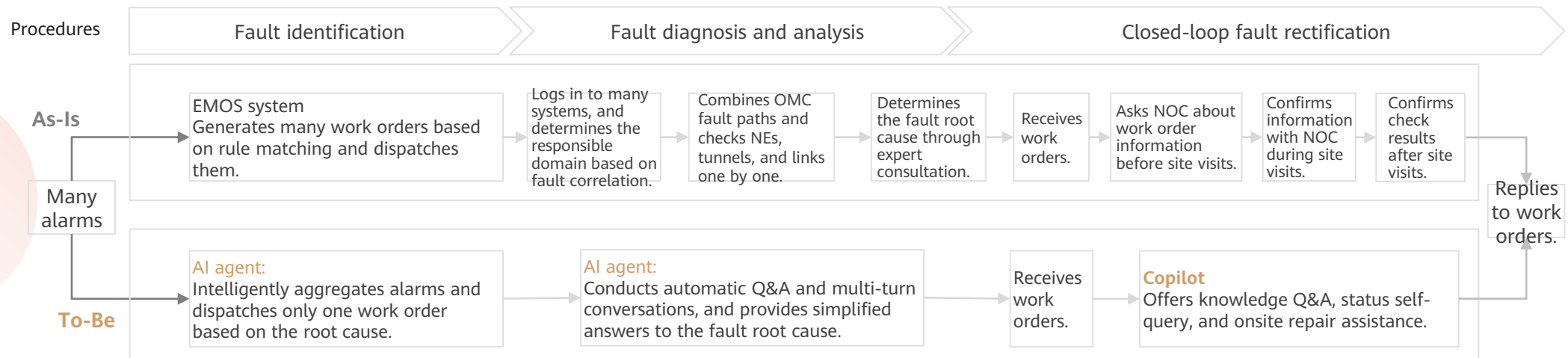
Scenario 3: Network Fault Agents Enable Intelligent Inspection and Recovery, Efficiently Resolving Silent Faults

1. Troubleshooting is driven by trouble tickets, manual fault locating is time-consuming, and proactive prevention is impossible.

- ✓ IP network protocols are complex and involve many devices. There are various causes for network faults, such as faults in components and forwarding mechanisms.
- ✓ No alarm is generated for 90% of packet loss events. There is no choice but to increase manpower to locate faults. According to statistics, 15% of silent faults consume 80% of manpower, and the mean time to repair (MTTR) is about 10 hours.

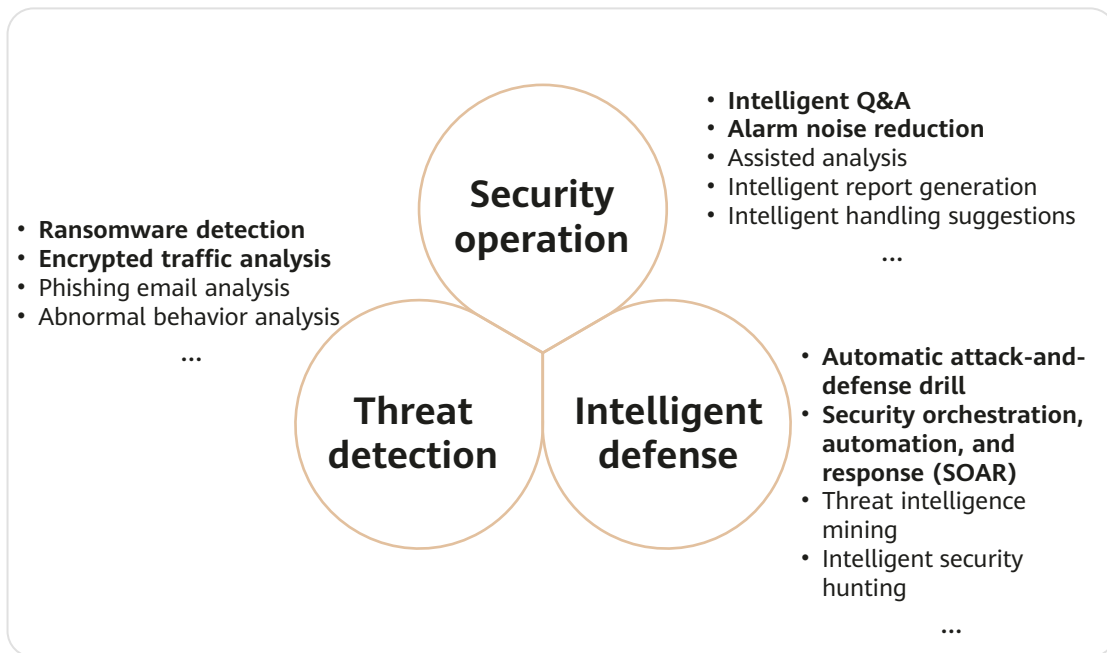
2. Network fault agents enable network self-inspection, self-analysis, and self-troubleshooting, efficiently resolving silent faults.

- ✓ Network fault agents proactively identify silent faults through preventive maintenance inspection and multi-dimensional data analysis.
- ✓ Network fault agents intelligently aggregate alarms and precisely dispatch work orders based on root causes.
- ✓ By deploying Huawei's network foundation models, China Mobile Guangdong Branch improves the automated fault diagnosis rate from 60% to 90%.



► Trend 4: Network Security Enters the Era of AI-Based Attack and Defense

- **The use of Generative AI (GenAI) simplifies the launch of security attacks, posing new challenges to network security.**
 - ✓ Artificial General Intelligence (AGI) tools reduce the time for hackers to generate new threats from several months to several hours or even minutes.
 - ✓ Foundation models help attackers quickly detect vulnerabilities in software and services. For example, **FraudGPT can be used to write malicious codes** to automatically and intelligently create malware; **WormGPT allows attackers to easily create phishing and email attacks.**
- **Vendors actively introduce AI to redefine network security protection capabilities.**
 - ✓ The current global shortage of network security workforce has exceeded over 4 million, with the demand growing twice as fast as supply.
 - ✓ GenAI offers significant advantages in threat detection and response, security operations assistance, and other areas, making it a crucial choice for security vendors.
 - ✓ Microsoft officially put Microsoft Copilot for Security into commercial use in May this year, aiming to help users defend against attacks at machine speed; Google released a dedicated model for network security last year, which has been applied to the cloud security competence center. Similarly, global network security giants Palo Alto and CrowdStrike have integrated the security operation capabilities of foundation models into their security operation platforms.



Foundation models applied in security

Spatiotemporal feature analysis model

Threat analysis technology built on the threat occurrence window, location characteristics, alarm sequence vector space, dictionary parameter space, and asset configuration space

Threat intelligence model

High-quality threat intelligence center built using multiple analysis methods, based on intelligence sources including but not limited to the SuMap global asset radar, distributed honeypot system, and more than 200 threat intelligence exchange data sources in and outside China

Intelligent in-depth awareness engine

Asset security status rating model built on multiple factors, such as attack intents, policies, times, and results to display the most critical alarm information and asset threats to users

Machine learning engine

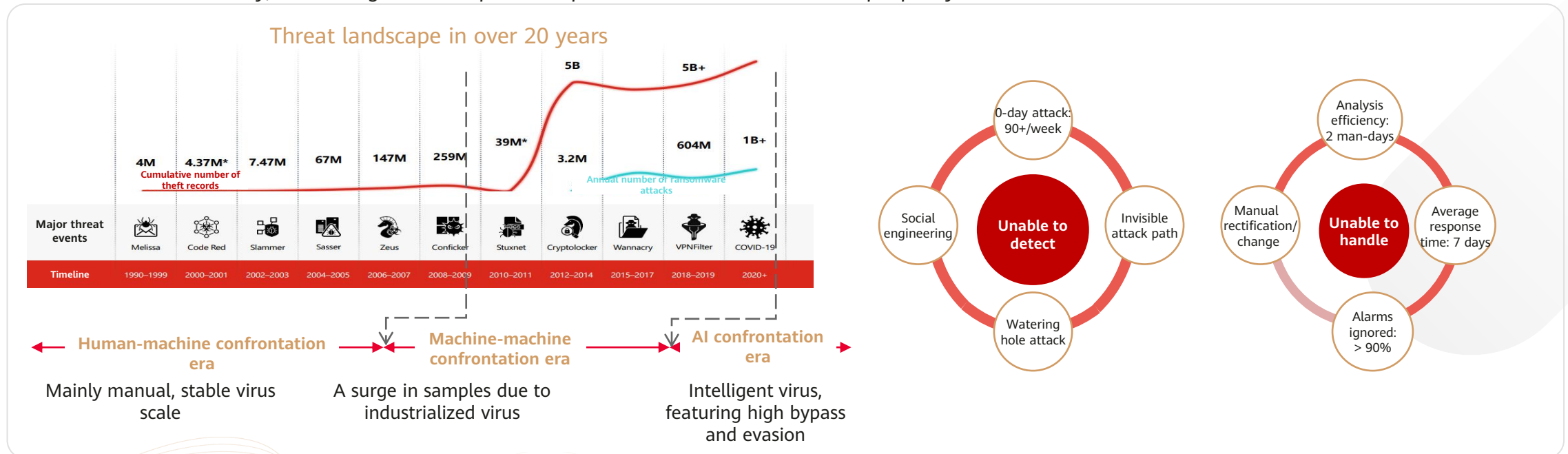
Foundation model used to construct optimization policies for adaptive detection of anomalies and deviations, as well as abnormal attack sources and methods, based on massive sample data on the cloud and specific user characteristics

Risk confidence evaluation model

Risk confidence evaluation model for continuous generation of risk confidence, risk level confidence, and threat availability confidence, focusing on advanced threats, based on massive samples on the cloud and personnel analysis and evaluation

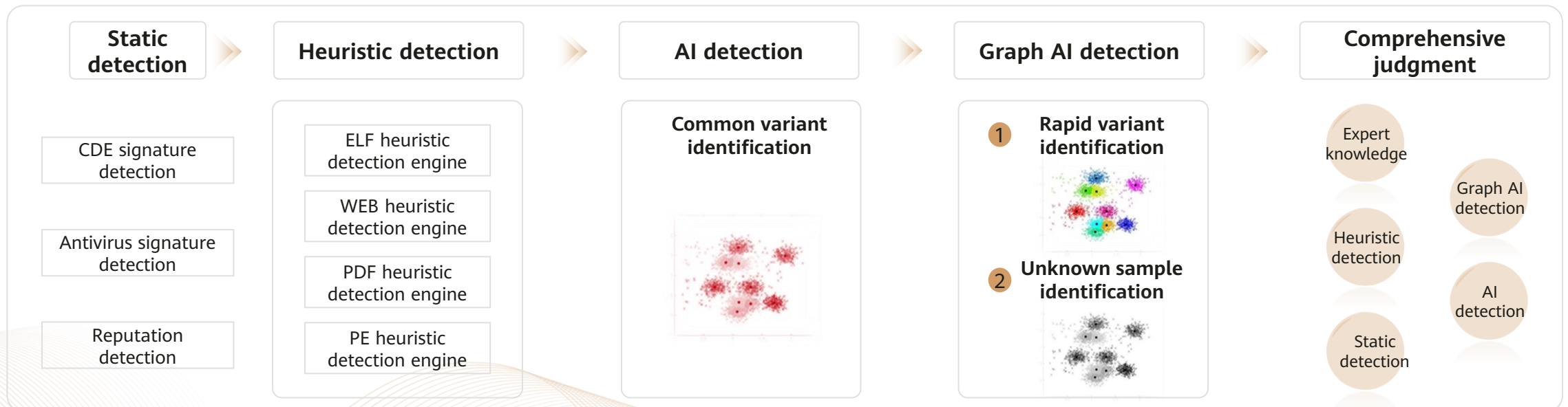
High Threat Variability and Evasion Puts Huge Strain on Unknown Threat Detection and Security Operation in the AI Era

- **Virus threats increase exponentially, making detection of unknown viruses and variants a challenge for the industry.**
 - ✓ Difficult to detect and defend against unknown viruses and variants: According to AV-TEST, an authoritative evaluation organization, there are 330,000 new viruses generated every day, meaning hundreds of millions of new viruses every year. As a result, traditional detection methods relying on signature databases fail to cope with these viruses.
 - ✓ Unable to use traditional methods to defend against encrypted traffic: Zscaler reports that 95% of traffic is encrypted and 86% of attacks are initiated through encrypted channels. This makes traditional content-based signature detection ineffective, highlighting the urgent need for new technologies.
- **A myriad of threat logs and alerts are overwhelming for manual handling, with 90% of alarms being ignored.**
 - ✓ Difficult to manually handle a lot of false positives in logs and alarms: A network security center receives millions of alarms every month, making it impossible for manual handling. Worse yet, the majority of these alarms are false positives. Therefore, there is an urgent need for an effective noise reduction method. According to a survey, on average, a Security Operations Center (SOC) team can handle only 9% of alarms, with the majority being ignored.
 - ✓ Slow manual response and handling: After a security incident is detected, the recovery period is long and collaborative handling is unavailable. According to the *SOC Economics* survey, 30% of organizations plan to expand their SOC teams to 10–15 people by 2024.



Scenario 1: Lightweight Graph AI Detection Models Are Used to Defend Against Ransomware Variants

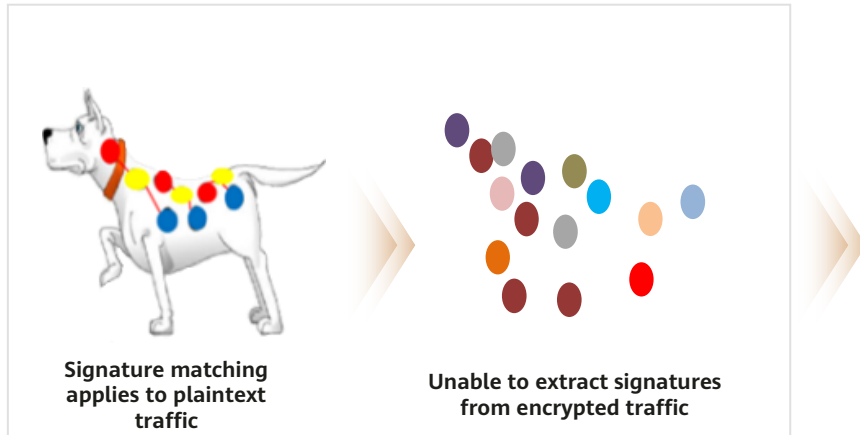
- **Ransomware attacks have become professional and organized. Currently, AI technologies have become a major tool of attackers to generate ransomware variants to break through network defense.** Ransomware attacks have gone through three phases: (1) Ransomware attacks developed slowly from 1989 to 2009. During that phase, the number of ransomware attacks increased slowly, and the attack intensity and severity were low. (2) Since 2010, ransomware developed rapidly, with its variants emerging every year. In addition, the attack scope expanded and the methods continuously updated. (3) Ransomware attacks frequently occurred since 2015. In 2017 alone, the WannaCry ransomware attack broke out worldwide, affecting 300,000 users in at least 150 countries and causing a total loss of over US\$8 billion. **Nowadays, ransomware attacks have been industrialized, scaled, and organized.** With the rapid adoption of new technologies, such as cloud computing and AI, the ransomware attack exposure increases, resulting in unprecedentedly frequent and various attacks. For example, the Matrix virus alone has hundreds of variants.
- **AI technologies have been introduced to improve the accuracy of ransomware variant detection.** For ransomware detection, the industry has gone through four phases: static detection, heuristic detection, AI detection, and graph AI detection. In terms of the latest graph AI detection technology, the AI algorithm for extracting malicious code DNA is used to extract core segments of malicious code for fast variant detection. Moreover, this technology can cope with unknown threat attacks. It uses the CPU command flow and graph AI algorithm to adaptively learn the CPU command sequence baseline. This reduces the false positive rate and invalid alarms, and achieves a 99.9% detection rate of unknown malware.



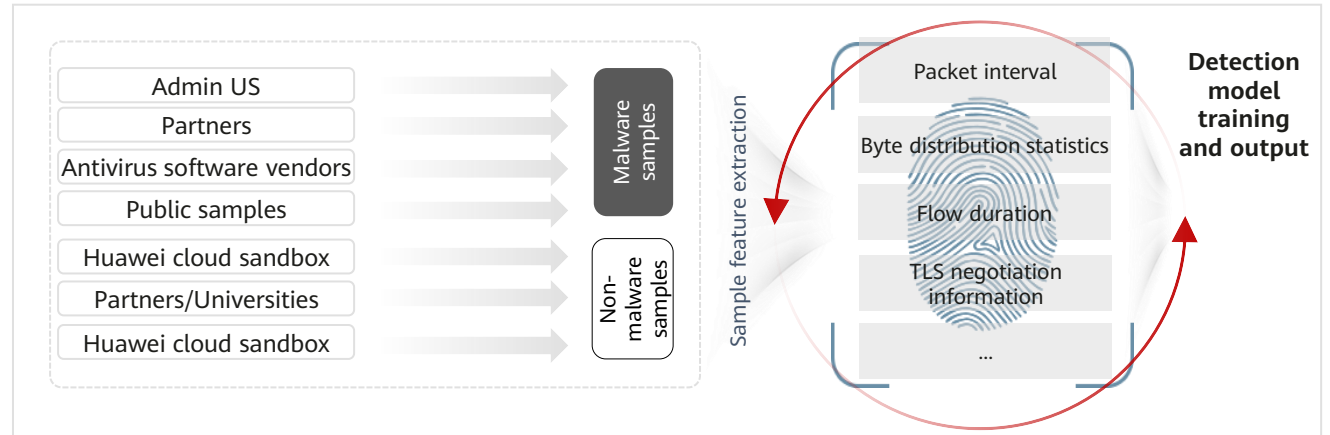
Scenario 2: Self-Learning AI Models Enable Efficient Detection of Encrypted Attacks

- **Encrypted malicious traffic is hidden, anonymous, and difficult to detect, posing a great challenge to security protection.**
 - ✓ Encryption technology, a powerful tool for protecting information security, has been widely used in various network communication scenarios. According to Google's report, over 90% of Internet traffic is now encrypted. A third-party report found that 80% of enterprise intranet traffic is also encrypted.
 - ✓ However, it is difficult to directly parse and monitor encrypted traffic. This enables malicious users and hackers to evade security checks and launch various network attacks, posing great challenges to network security monitoring. From 2020 to 2022, the proportion of attacks using encrypted channels increased from 57% to over 85%. According to a survey, more than 95% of enterprises have encountered encrypted traffic attacks.
- **Machine learning and deep learning facilitate efficient detection of encrypted attacks.**
 - ✓ Vendors in the industry are proactively introducing AI technologies to detect threats in encrypted traffic without decryption. For example, Huawei develops a multi-stream Encrypted Communication Analytics (ECA) algorithm and AI model self-learning engine to accurately detect threats and attacks hidden in encrypted traffic. With this technology, Huawei can detect 100% of encrypted threats within 30 minutes, a six-fold improvement over the industry average of 3 hours.

Many encryption-based attacks (86%), making packet content inspection ineffective

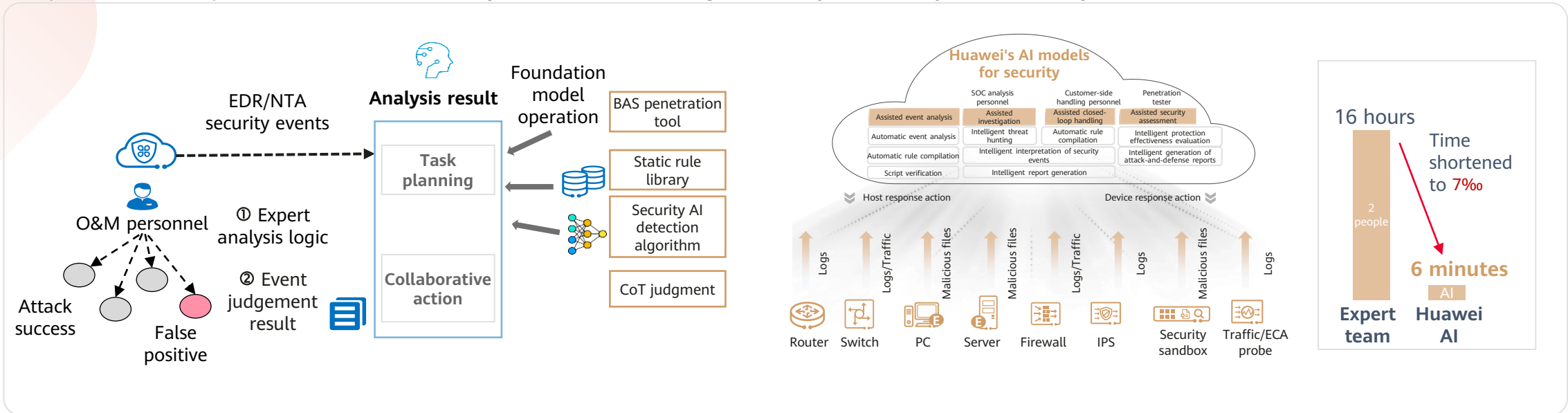


Huawei-developed multi-stream ECA algorithm + AI model self-learning engine



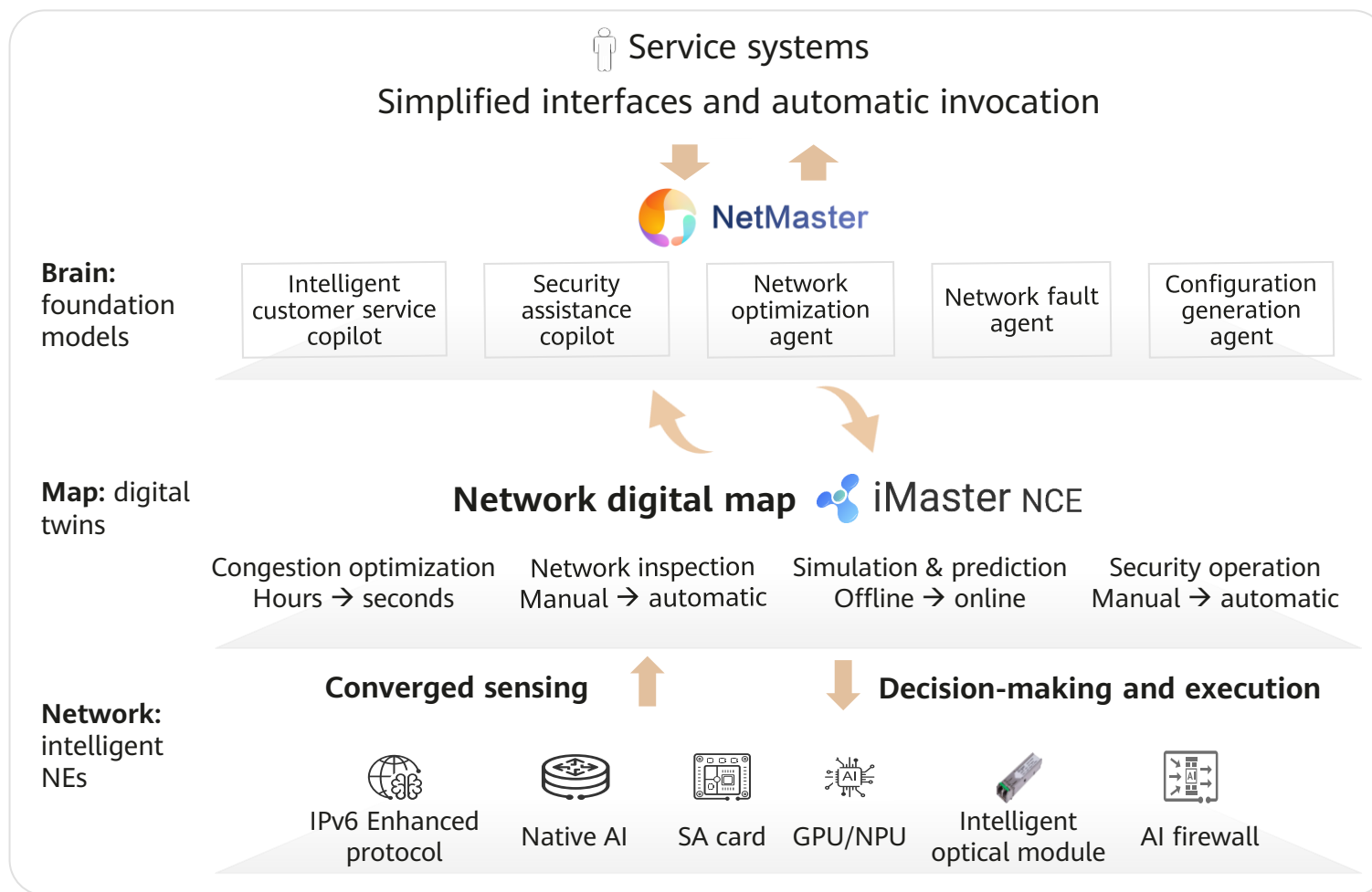
Scenario 3: Collaboration Between Small and Large Models Achieves Security Event Noise Reduction and Intelligent Assisted Handling

- A myriad of security alarms require intelligent analysis and assisted decision-making.** As cyber-attacks increase (for example, a 104% year-on-year increase in global cyber attacks in 2023), security alarms are also on the rise. This brings huge pressure to the security team, and puts huge strain on manual analysis approaches. As a result, 90% of alarms are ignored. When it comes to Huawei's business process & IT, for example, approximately 1 billion logs and 100,000 alarms are generated every day. Assuming that one security analysis expert can process a maximum of 500 alarms per day, this means more than 200 such experts are required to process all alarms. Worse yet, security talent is insufficient, with the global shortage of as high as 4 million highly skilled security professionals. Against this backdrop, a common need in the industry is to leverage technological innovations such as AI to identify far-reaching security events and intelligently assist security handling.
- Collaboration between small and large models achieves efficient security event noise reduction and intelligent assisted handling. Network foundation models** intelligently invoke traditional machine learning algorithms and operational analysis rules to implement collaboration between large models, small models (17 small AI models), and expert rules (8000+ Qiankun CloudService rules). As a result, the mean time to detect (MTTD) for security events is shortened from hours to minutes. Plus, AI agents identify the **semantic context of security events, intelligently schedule and orchestrate** various security tools, small AI models, and expert rules, and implement **autonomous security defense and handling**. In this way, the security event handling time is slashed from weeks to minutes.



AI for Network: Three-Layer Intelligent Architecture, Accelerating Network Intelligence with Network-Security Integration

Three-layer intelligent architecture



Leverage three-layer collaboration (namely, brain, map, and network) to enable E2E converged sensing, intelligent inference & analysis, AI simulation & decision-making, and reliable execution, reshaping system capabilities and accelerating network intelligence with network-security integration.

- **Brain:** Network foundation models are used to reshape service processes and O&M paradigms into intent-driven, automated ones based on natural languages. In particular, copilots bring brand-new O&M experience typified by natural language-based interactions. Agents build up scenario-specific autonomy, fault closure, and automatic threat event analysis and handling.
- **Map:** Network digital twins are built based on intelligent analysis of full data to achieve a shift from static visualization to multi-dimensional sensing and visualization, from offline network simulation to real-time network simulation, and from single-point optimization to network-wide collaborative optimization. In this way, a high-definition (HD) network digital map takes shape, offering precise navigation for autonomous driving of networks.
- **Intelligent NEs:** A cornerstone for network intelligence is equipment intelligence, that is, changing from passively accepting and executing intents to proactively generating intents and handling them in a closed-loop manner. Beyond this, millisecond-level micro insight into network and service status, accurate fault source tracing and locating, edge inference, and real-time decision-making ensure zero service interruptions and effectively cope with the exponential growth of virus variants and hidden penetration attacks in the AI era.

▶ AI for Network: Recommendations for Action

- **Leverage digital twin technologies to achieve comprehensive network visualization.** Network visualization is the first step towards intelligence. Network digital twins can simulate and verify all operations that need to be performed on the live network, and continuously evaluate, modify, and optimize operational procedures based on feedback, ultimately minimizing the impact on the real-world network. Furthermore, network digital twins record network status and behavior in real time and support historical data tracing and playback. In this way, pre-verification can be completed without affecting network operations, greatly reducing trial-and-error costs.
- **Tap into network foundation models to build a cornerstone for network intelligence.** Network foundation models are the bedrock of network intelligence. In this regard, enterprises need to specify their strategic directions and start deploying network foundation models as soon as possible, no matter whether to interconnect with existing foundation model platforms or to build network foundation models through enterprises' own capabilities.
- **Develop tailor-made intelligent applications for real-world service scenarios.** Copilots and agents are similar to mobile apps on smartphones. Enterprises need to build up their copilot and agent capabilities in typical service scenarios to reduce risks while improving efficiency.
- **Actively introduce AI security protection technologies to enhance cyber security capabilities.** Enterprises need to gradually develop and build all-round security protection capabilities through diverse means. In doing so, they can effectively tackle security threats, improve security detection capabilities — especially in detecting unknown threats and threats in encrypted traffic such as ransomware variants — and continuously increase security operation efficiency.
- **Embrace AI from now on.** As AI technologies advance, the deployment paths and benefits for network intelligence become more prominent, and we are at a milestone moment for mass deployment of network intelligence. AI itself does not completely replace humans, but better assists humans, instead. AI networks will comprehensively improve network availability, optimize efficiency, improve performance, and do more with the same or fewer resources.



Building a Fully Connected, Intelligent World

Net5.5G

Network Foundation for the Intelligent World

AirEngine Wi-Fi 7

| **NetEngine** Routers

CloudEngine Switches

| **HiSecEngine** Security Gateways



Scan for more information