Network Innovation and Development Alliance
全球固定网络创新联盟

中国电信
CHINA TELECOM

ИНТЕРНЕТ АССОЦИАЦИЯ КАЗАХСТАНА
INTERNET ASSOCIATION OF KAZAKHSTAN
ҚАЗАҚСТАННЫҢ ИНТЕРНЕТ АССОЦИАЦИЯСЫ

Metropolitan Area Network for the AI Era

Network Innovation and Development Alliance

September 2025

Preface

The rapid development of artificial intelligence industry drives the explosive growth of various AI applications. As the critical infrastructure bridging end users and computing resources, metropolitan area networks (MANs) are now facing transformative requirements in network architecture, functional capabilities, and service paradigms.

In 2024, China Telecom pioneered the industry-first ‘computing service-oriented metropolitan area network’ concept and released the ‘computing service-oriented metropolitan area network White paper’, generating significant industry-wide attention and discourse. As a continuation, this white paper provides in-depth analysis of metropolitan area network evolution in the AI era. This white paper first analyzes the development landscape of artificial intelligence from the perspectives of industry advancement and macro policies. Subsequently, it conducts an in-depth analysis of AI application requirements to define the essential network capabilities that metropolitan area networks must possess. This white paper then examines the design objectives, elaborating on the overall architecture and key technologies of metropolitan area networks for the AI era. Finally, it provides technical solutions tailored for typical scenarios.

The following organizations and principal members contributed to the preparation of this whitepaper:

- China Telecom Research Institute: Yongqing Zhu, Zehua Hu, Xia Gong, Shizhang Yuan
- Zhongguancun Ultra Cross Connection New Infrastructure Industry Innovation Alliance: Bo Yuan
- Huawei Technologies Co. Ltd.: Haobin Zhao, Jie Dong, Li Zhang
- ZTE Corporation: Wenqiang Tao, Haidong Zhu, Xiaowei Ji

Directory

Chapter I

Development Trends of Artificial Intelligence

1.1 AI Industry enters a phase of accelerated growth

As the core driving force leading the Fourth Industrial Revolution, the artificial intelligence (AI) industry is experiencing unprecedented rapid development, demonstrating enormous market potential. According to Grand View Research, the global AI market size reached 196.63 billion in 2023 and is projected to increase to 1,811.75 billion by 2030, with a compound annual growth rate (CAGR) of 37.3% from 2024 to 2030. In China, research reports indicate that the scale of the AI industry is expected to expand from 398.5 billion yuan in 2025 to 1,729.5 billion yuan in 2035, with an estimated CAGR of 15.6%. Artificial intelligence has undoubtedly become a powerful engine for global economic growth.

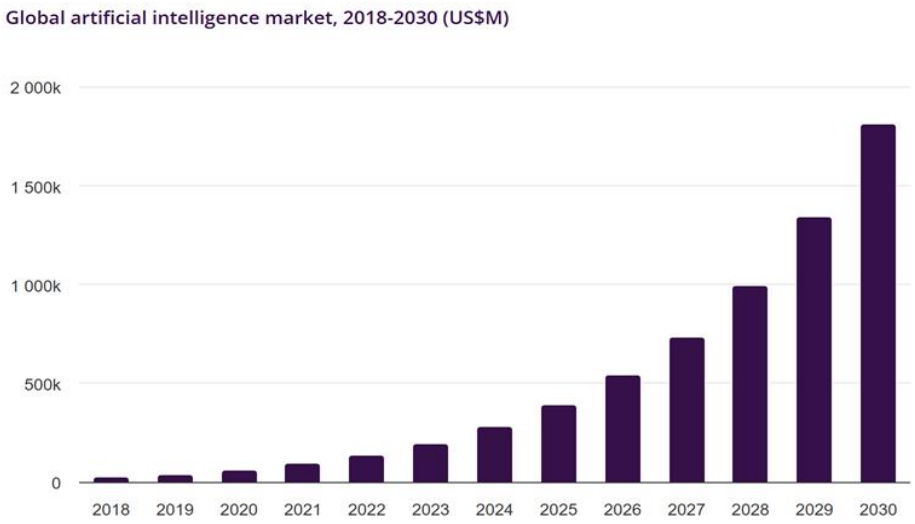

Figure 1-1: Global artificial intelligent market

The global AI industry demonstrates a trend for development of ‘dual-track advancement and diversified flourishing’. On the one hand, global technology giants continue to intensify their AI investments: companies like Google and Microsoft are deepening research and development (R&D) in core AI technologies; Amazon and Apple persist in innovating intelligent cloud services and end-device smart applications, while China's major tech firms such as Baidu, Alibaba, Tencent, and Huawei (BATH) are also making rapid progress in key areas such as AI chip development, large AI model construction, computer vision, and embodied

intelligence. On the other hand, the explosive breakthroughs in generative AI technology have spurred a wave of innovative enterprises worldwide: OpenAI pioneered the commercialization of generative AI with ChatGPT; Anthropic and Cohere specialize in vertical-oriented development; and in 2025, China's DeepSeek significantly accelerated the commercial application of large AI models in inference scenarios. Numerous emerging AI supply chain companies have become investment hotspots, collaborating with industry leaders to form a synergistic innovation ecosystem. This dynamic development pattern that features competition and symbiosis among diverse players not only accelerates the commercial deployment of large language models in finance, healthcare, and manufacturing, but also provides robust momentum for the high-quality development of the digital economy.

Benefiting from the rapid development of AI industry, AI technologies are becoming powerful engines for urban development, injecting unprecedented vitality into various sectors of cities: In transportation field, leveraging the precise predictive capabilities of large AI models optimizes traffic flow and enhances travel efficiency. In healthcare industry, AI-assisted diagnostic technologies enable the rapid and accurate analysis of medical images, helping doctors to formulate treatment plans. In education, customized teaching content is provided based on students' learning progress and characteristics, stimulating their interest and potential. The financial sector utilizes large AI models for risk assessment and investment decision-making, improving the precision and security of financial services. Furthermore, numerous fields such as intelligent manufacturing, intelligent government services, and environmental monitoring have become more efficient, intelligent, and sustainable through the empowerment of AI. The application of AI technologies provides residents with more convenient, comfortable and secure living experiences, leading cities to an intelligent and digital future.

1.2 AI is focal point of global industrial policies

AI has become one of the core driving forces for urban and social development, forming a global consensus：

- The United States launched the ‘White House Smart Cities Initiative’ in 2015, leveraging AI, big data, and the Internet of Things (IoT) technologies to help cities address challenges such as traffic congestion, energy management, and public safety. By 2025, it would further strengthen AI infrastructure through the ‘Stargate Program’.
- The European Union proposed the ‘European Data Union Strategy’ in 2025 to promote AI and big data applications in healthcare, education, and urban governance, supported by the ‘Digital Europe Programme’ to implement AI in critical social and livelihood sectors.
- Japan introduced the ‘Super City’ vision, integrating AI and IoT to create data-driven ‘smart cities’.
- Singapore implemented its 'National AI Strategy 2.0', which combines talent attraction, industrial applications, R&D innovation, and infrastructure to build an AI ecosystem that improves public services and industrial competitiveness.
- The Chinese government also prioritizes AI-driven urban development. In 2024, China’s National Data Administration issued guidelines to deepen smart city initiatives, encouraging AI-powered solutions, such as intelligent analysis, scheduling, regulation and decision making, to comprehensively empower urban digital transformation.

Networks have become critical infrastructure supporting global AI industry development and are receiving high priority from nations worldwide：

- In China, ‘empowering computing through networks’ has been established as a fundamental principle for building smart cities. In October 2023, China's Ministry of Industry and Information Technology (MIIT) introduced the High-Quality Development Action Plan for Computing Infrastructure, which aims to create a group of computing power network city benchmarks in key regions.

- In June 2023, the Singapore government launched its Digital Connectivity Blueprint, proposing the construction of seamless end-to-end 10Gbps domestic connectivity within five years to ensure Singapore's digital infrastructure remains world-class and sets the direction for its digital future.
- In April 2024, Saudi Arabia's Ministry of Communications released the Saudi Arabia's 10Gbps Society White Paper, becoming the first globally to propose an end-to-end high-speed, high-quality Net5.5G network architecture to support the country's intelligent transformation.
- In 2025, the European Commission Digital Europe Programme (DIGITAL) 2025-2027 also emphasized the need to enhance network resilience in various AI scenarios.
- In Kazakhstan, the government is vigorously advancing digital transformation and the development of artificial intelligence (AI).

  In February 2024, Kazakhstan’ s Ministry of Digital Development released the "AI Development Concept for 2024-2029". This document plans to establish a National AI Center, integrate resources across various fields, and promote the extensive application of AI in economic and social sectors. To cultivate AI talents, Kazakhstan has reformed its education system: starting from 2026, AI tools will be incorporated into the curricula of primary and secondary schools; meanwhile, many universities have launched AI-related educational programs, and over 2,000 postgraduate students are currently engaged in AI research.

  Kazakhstan has set up a $1 billion venture capital fund at the Astana International Financial Center to support promising AI startups. In July 2025, with the commissioning of its first supercomputer, the country’ s computing power has been significantly enhanced, laying a solid foundation for AI development.

  In September 2025, President Tokayev announced in his annual State of the Nation Address that the government would establish the Ministry of Artificial Intelligence and Digital Development, aiming to build a comprehensive digital asset ecosystem. At the same time, he proposed the establishment of a National

Digital Asset Fund and the acceleration of the implementation of the Digital Code. These measures are intended to comprehensively drive the country’ s digital development, with the goal of transforming Kazakhstan into a fully digitalized nation within three years.

With the widespread adoption of large AI models and growing demand for applications such as distributed inference, the role of networks in AI development is becoming increasingly prominent. Building a second "information superhighway" dedicated to AI has emerged as a global priority.

1.3 AI technology is developing explosively

1.3.1 AI technology is advancing comprehensively

The development of AI technology demonstrates notable trends of diversified collaboration, high-efficiency evolution, and multi-ecosystem integration：

- At the hardware level, the significant increase in inference scenarios has driven rapid advancements in dedicated AI chips such as TPUs and LPUs, while general-purpose GPUs, combined with cutting-edge technologies like chiplet, 3D stacking, and quantum computing, provide enhanced capabilities for ultra-large-scale AI model training.
- In storage technology, protocols such as HBM3 and CXL have achieved leaps in memory bandwidth and capacity, while architectures such as storage-compute disaggregation meet the demand for building private knowledge bases based on large AI models.
- High-speed interconnect technologies such as UEC, NVLink, UCIe and Falcon break down data transmission barriers, enabling efficient collaboration between distributed computing and heterogeneous architectures.
- On the software ecosystem, open-source frameworks such as PyTorch and TensorFlow are deeply integrating with automated toolchains, combined with cloud-edge-device unified deployment, to achieve end-to-end optimization from training to inference.

- In addition, green computing technologies, including liquid cooling and dynamic power management, contribute to the sustainable development of AI.

1.3.2 Large AI model technology enters rapid development phase

Large AI models have become one of the most widely applied key AI technologies today. From the launch of ChatGPT in 2022 to the rise of DeepSeek in 2025, the field of large AI models has experienced explosive growth. The development of large models exhibits multi-dimensional trends: on one hand, model scale continues to expand with increasing parameter counts, enabling the capture of more complex patterns and relationships to enhance performance across various tasks; on the other hand, multi-modal fusion has become an important development direction, as large models combine text, image, speech and other multi-modal data to achieve more comprehensive understanding and generation of information, expanding their application scenarios. Additionally, greater attention is being paid to model safety, reliability and interpretability, with researchers committed to developing more robust model architectures and training methods to ensure stable operation and trustworthy application of large AI models in complex environments. These trends collectively drive the continuous advancement of large AI model technology, laying a solid foundation for the widespread application of AI. Currently, large AI models are evolving in the following technical directions:

Direction 1: As the parameters and training data scale of large AI models continue to increase, the demand for computing power is also growing rapidly. Single 1K+ GPUs or 10K+ GPUs AI Data Centers (AIDC) can hardly meet the requirements of ultra-large-scale training. Taking Llama 3.1 released in 2024 as an example, its largest model has 405B parameters and requires approximately 15 trillion Tokens for pre-training, with the entire training process demanding 39.3 million GPU/hours (H100) of computing power. **Therefore, adopting distributed training methods and utilizing high-performance networks to enhance the collaborative training**

efficiency across multiple AIDCs has become a necessity for AI development. Currently, multiple operators have completed the commercial deployment of distributed training, achieving the distributed training for 10K+ GPUs, 100B parameters large AI models across AIDCs over distances 100+ kilometers. Among them, China Telecom and Huawei jointly deployed the distributed training service supporting 120KM wide-area RDMA lossless transmission, with training efficiency reaching over 95%.

Direction 2: **Software engineering optimization has become the key pathway to break through AI hardware bottlenecks, driving large AI models toward cost-effective development, and accelerating the adoption of AI across industries.** The open-source DeepSeek-V3 in 2025 completed pre-training in just two months using only 2,048 GPUs through algorithmic optimization, while the DeepSeek-R1 model further compressed the training cycle to 2-3 weeks. This 'low-cost & open-source' solution significantly lowered the technical threshold for large AI models, directly leading to two notable changes: First, the relatively low usage costs triggered explosive growth in large AI model-based applications, resulting in surging AI traffic within cities that requires network to ensure efficient AI traffic steer; Second, through full-stack software engineering optimization spanning 'algorithm & hardware & system', AI inference latency was reduced by over 60%, driving exponential growth in AI inference demand.

Direction 3: **The intelligent interaction of multi-edge agents reflects AI technology's transformation from centralized to distributed systems and from single intelligence to collective intelligence, driving breakthroughs in real-time performance, autonomy, and collaboration of AI.** Large AI models can achieve lightweight deployment through techniques like model distillation, making them compatible with resource-constrained scenarios such as consumer-grade GPUs, mobile devices, and IoT equipment, thereby promoting the development of edge-based small intelligent devices. At the software level, the widespread adoption of Multi-Agent technology enables multiple terminals to collaboratively complete

complex tasks, further advancing large-scale interactive applications of edge agents. Google's introduction of the A2A and MCP protocols for agent interaction in 2025 signals AI's impending transition from the ‘cloud computing’ architecture of B2B, B2C, and C2C to the ‘granular computing’ architecture of A2A, M2M, and X2X, with the frequent interactions between intelligent computing particles place higher demands on the reliability, and bandwidth capacity of network.

Direction 4: In September 2024, OpenAI launched the o1 model with Chain-of-Thought (CoT) mechanism, which achieves higher accuracy by extending thinking time during inference, marking a paradigm shift from pursuing response speed to emphasizing deep reasoning. This transformation has driven the shift of computing power demand from pre-training to inference, breaking through the limitations of Scaling Law: **while pre-training relies on 10K+ GPUs Scaling-up clusters, inference can be implemented through Scaling-out architectures composed of a small number of GPUs, promoting the evolution of AI infrastructure toward distributed and flexibly scheduled systems.** Additionally, the significantly increased deployment demands on AI inference have raised requirements for large-scale inference performance improvements. Network-based distributed inference has become a key direction for future urban AI applications, necessitating networks to support distributed AI inference deployment. In response, NVIDIA introduced the Dynamo framework, adopting a PD-separated architecture to optimize resource scheduling and computing efficiency in large-scale AI inference.

1.4 Challenges to MAN from large-scale AI commercialization

Building a comprehensive AI urban ecosystem has become the core pathway for upgrading urban systems to advanced intelligence. In this process, the concept of ‘City as a Computer’ has gradually gained global consensus: by deeply integrating computing power, storage, and terminals through metropolitan area networks (MANs), cities are transformed into distributed ultra-large-scale computing systems, enabling

citywide intelligent management through millisecond-level data flow and real-time decision-making. Existing broadband networks, mobile networks, dedicated government and enterprise networks, and cloud networks within cities connected various users. However, traditional MANs struggle to meet the requirements for carrying urban AI services, whether in terms of network architecture or core technologies.

1.4.1 Challenges in data circulation

The training of large AI models and the construction of knowledge bases typically require data volumes at the TB/PB scale, which imposes higher throughput requirements on data transmission networks. Simultaneously, the computing traffic of large models exhibits significant elastic characteristics, demanding extremely high network reliability. Substandard and non-deterministic networks may result in insufficient data transmission bandwidth, excessive latency, or frequent packet loss, thereby compromising the availability of computing resources. Furthermore, version iterations of large models and knowledge base upgrades in AI systems also depend on stable network support. Poor network quality can constrain the implementation of these functions, ultimately reducing the overall operational efficiency of AI infrastructure.

The rapid development of large-scale inference applications and A2A computing paradigms has introduced new challenges to urban AI data circulation: on one hand, MANs need to meet the efficient data communication and interaction requirements between distributed inference nodes; on the other hand, the A2A mode has led to exponential growth in high-frequency interaction traffic across intelligent agents, which not only significantly increases the bandwidth requirements of edge networks but also requires MANs to ensure the reliability of information interaction between intelligent agents. Therefore, to realize the vision of ‘City as a Computer’, it is urgent to build a new ultra-interconnected network different from traditional MANs to meet the transmission requirements of AI computing data flows and enable MANs to

effectively support efficient computational data circulation.

1.4.2 Challenges in O&M

When MANs carry AI services, network management and maintenance (O&M) face greater challenges. From service model perspective, AI has transformed network traffic patterns: large AI model training can cause sudden traffic surges, while frequent interactions between intelligent agents also generate bursty communication, requiring networks to possess predictive planning and maintenance capabilities. AI services also demand higher network reliability, even minor faults during model training may lead to complete task resets. When large-scale inference services replace manual services in cities, networks must ensure service experience.

Consequently, traditional management models that rely on manual intervention and route convergence to ensure basic network availability can no longer meet the performance demands of AI services. AI services require higher fault self-healing rates and lower latency in network optimization decisions, pushing network operations toward high autonomy to fulfill needs like predictive maintenance, service awareness, and elastic optimization. The question of how to equip networks with highly intelligent management and operational capabilities, namely automating the adjustment of network resources and configurations based on the intentions and states of computing services, has become a key focus for AI-oriented MANs.

1.4.3 Challenges in security and trustworthiness

With the rapid adoption of large models, vast amounts of urban data are being utilized for analysis, computation, and processing. Data from enterprises, households, and individuals constitute private domain traffic, posing significant security risks: For households and individuals, private domain traffic involves sensitive data such as personal information and consumption behaviors, where leaks could lead to privacy violations; for enterprises, private domain traffic encompasses R&D data, production data, and operational data, where breaches could undermine competitiveness or even

trigger legal disputes. Since data transmission faces potential threats such as theft, tampering, and loss, MANs must have robust data protection capabilities to ensure data confidentiality, integrity, and availability.

Traditional AAA (Authentication, Authorization and Accounting) systems and data encryption technologies based on traffic flows struggle to meet the security and trust requirements of AI scenarios. However, emerging technologies like blockchain and quantum encryption offer innovative solutions for trustworthy data circulation: blockchain provides immutable, end-to-end traceable trust mechanisms for AI data flows through distributed ledgers and smart contracts; quantum encryption leverages breakthroughs like quantum key distribution to fundamentally enhance anti-eavesdropping capabilities for data transmission. MANs must integrate these innovative mechanisms to establish a trusted foundation for large-scale urban AI deployment, providing critical infrastructure support for the widespread implementation of metropolitan AI services.

Chapter II

AI-Driven Requirements for MAN

2.1 AI applications annovation aontinues to accelerate

In early 2025, DeepSeek spearheaded a transformative wave in generative AI, driven by its exceptional performance and industry-leading cost efficiency in LLM training and inference, accelerating the commercialization of AI technologies. Today, AI applications have entered the stage of scaled deployment, serving diverse scenarios across home(toH), consumer (toC), and business (toB), with penetration into multiple vertical industries including media, legal services, education, and manufacturing.

2.1.1 AItoH scenarios

AI significantly enhances the professionalism, interactivity, and personalization of home services, enriching home scenarios. Currently, the industry is gradually reaching a consensus on building an integrated smart home ecosystem that combines connectivity, computing power, and intelligence. Through cloud-network-edge-device collaboration, providing broadband users with an smart cloud services, supporting various AItoH scenarios including smart home and home assistants:

- Smart home: Smart TV, smart refrigerators and other smart home products, utilizing AI technologies like voice recognition and computer vision, now support intelligent capabilities including natural language interaction, user habit learning and contextual adaptation. These smart home products can dynamically adjust lighting, temperature and humidity based on user preferences, while employing facial recognition and behavior analysis technologies to enhance home security.
- Home assistants: Smart home assistant products, including smart speakers and domestic robots, employ natural language processing and other AI technologies to enable harmonious human-machine dialogue. These products achieve precise intent understanding to execute tasks including schedule reminders and information retrieval, while enabling contextualized services such as appliance control and security monitoring through seamless IoT interoperability.

2.1.2 AItoC scenarios

AI revolutionizes the interactions between consumer and service, driving enhanced user experiences and fostering market innovation. The AI innovation landscape is witnessing rapid proliferation of various vertical applications. Major industry players are actively deploying AItoC solutions across smart terminals, personalized services, and digital lifestyle domains, leveraging metropolitan AI services to enhance user experience and retention. The current AItoC applications primarily encompass the following categories:

- Productivity Enhancement: AI applications such as intelligent search, automated summarization, content generation, and code assistance have significantly improved efficiency for both individuals and organizations. These applications streamline complex workflows, enabling users to focus on higher-value strategic initiatives while fostering innovation and competitive advantage.
- Creative Generation: AI applications including design automation, image generation, video synthesis, and music composition are revolutionizing the content creation industry. These applications augment creative ideas for content creators.
- Entertainment: AI applications such as AI cameras and virtual companions are transforming user experience through novel interaction paradigms and enhanced engagement. These applications leverage user profiling to deliver personalized entertainment services, elevating the enjoyment of digital experiences.

2.1.3 AItoB scenarios

AI demonstrates formidable capabilities in data analytics and decision support, enabling enterprises to achieve significant operational efficiency improvements and substantial cost reductions. Furthermore, AI exhibits exceptional capabilities in data processing and content generation, enabling enterprises to access novel business opportunities. The technological convergence of AI, 5G, and edge computing is accelerating industrial intelligent transformation, establishing a closed-loop value

system of ‘high-speed connectivity+real-time computing+intelligent decision-making’ that is reshaping entire processes from production to maintenance:

- Accelerating Product Development: During the requirements analysis phase, AI leverages natural language processing and sentiment analysis to rapidly mine massive user feedback and market data, enabling precise identification of latent needs and pain points. In the conceptual design phase, AI automatically produces hundreds of viable solutions based on historical data and design specifications for engineers to evaluate, significantly compressing design cycles. For engineering validation, physics-informed AI simulation systems accurately predict product performance parameters, substantially reducing verification costs.
- Enhancing Operational Efficiency: AI empower enterprises to achieve intelligent and high-efficiency operations through automated process enhancement, optimal resource allocation, and strengthened supply chain management. For instance, AI-driven monitoring systems conduct real-time surveillance of supply chain nodes, predicting potential disruptions and demand fluctuations to dynamically optimize inventory levels and logistics planning. Futhermore, AI-driven maintenance systems analyze sensor data and historical maintenance records to accurately forecast failure patterns, enabling proactive maintenance scheduling that significantly reduces unplanned downtime.

2.2 AI applications exhibit diverse deployment models

The deployment of AI applications requires meeting differentiated response requirements while considering critical aspects including data security, elastic resource scaling, and system maintenance. Through the coordination of urban AI Data Center (AIDC), Metropolitan Area Networks (MANs), and various deployment models, a hierarchical and collaborative city AI enablement system can be constructed. Common deployment models include: cloud deployment, on-premises deployment, hybrid deployment, federated deployment, and edge deployment.

Cloud Deployment: Internet service providers typically adopt cloud deployment

to enable rapid AI application provisioning and extensive user coverage. Leading enterprises usually build proprietary AIDC to support their own service requirements while offering computing power leasing services. For small and medium enterprises, establishing proprietary AIDC incurs high investment and maintenance costs, making them more inclined to lease computing power for rapid AI application deployment and iteration.

On-Premises Deployment: On-premises deployment is particularly suited for industries such as finance, healthcare, and manufacturing that require stringent data security and compliance. This approach enables enterprises to maintain full control over their data, ensuring all data processing and storage remain within their internal networks while delivering ultra-low application access latency. However, the continuous scaling of AI models results in prohibitively high costs and demanding operational requirements for on-premises deployment.

Hybrid Deployment: Hybrid deployment combines the advantages of cloud and on-premises deployment, enabling enterprises to process sensitive data locally while utilizing cloud resources for non-sensitive data processing. Enterprises can process latency-critical tasks locally while strategically offloading compute-intensive or non-core workloads to cloud, optimizing both on-premises hardware investments and operational expenditures. Hybrid deployment provides enterprises with a balanced solution for performance, security, and cost, making it one of the increasingly preferred approaches for deploying AI applications.

Federated Deployment: Federated deployment leverages distributed computing to enable multiple enterprises to collaboratively train a more effective global model without sharing privacy data. Specifically, each participant trains AI model locally, then transmits the encrypted model parameters to a central server for aggregation, generating an improved global AI model that is subsequently distributed to all participants. Federated deployment facilitates collaborative learning across multiple participants while preserving data privacy, delivering an innovative and practical deployment approaches for AI applications.

Edge Deployment: Edge deployment targets scenarios requiring real-time processing and rapid response, such as autonomous driving, industrial control systems, and smart home. For instance, in autonomous driving, edge-deployed AI applications enable real-time analysis of data from cameras, radars and sensors to facilitate instant decision-making, ensuring rapid response to environmental changes. In industrial control systems, edge-deployed AI applications maintain continuous operation even without stable network connectivity, guaranteeing uninterrupted production.

2.3 AI applications impose new requirements on MANs

MANs interconnects heterogeneous computing resources and diverse user terminals within the region, providing connectivity for various deployment models including cloud deployment and hybrid deployment, and serving as a critical infrastructure for the sustained development of AI. Developing MANs like urban power grids or water grids to enable 'one-point access, on-demand computing' computing power services has progressively become an industry-wide consensus.

The AI models required for different application scenarios exhibit significant variations, which can be categorized by scale into two distinct types: large models and small models. Small AI models typically refer to those with fewer parameters and shallower layers, characterized by their lightweight architecture, computational efficiency, and deployment flexibility. These models are specifically optimized for dedicated tasks and vertical domains, with representative implementations including DistilBERT, TinyBERT, and MobileNet. Large AI models refer to those with massive parameters and sophisticated computational architectures, exhibiting enhanced representational power and superior accuracy to handle more sophisticated tasks, with representative examples including Deepseek, GPT-4, Qwen. Diverse AI model impose significantly differentiated requirements on MANs.

2.3.1 Requirements of large AI model

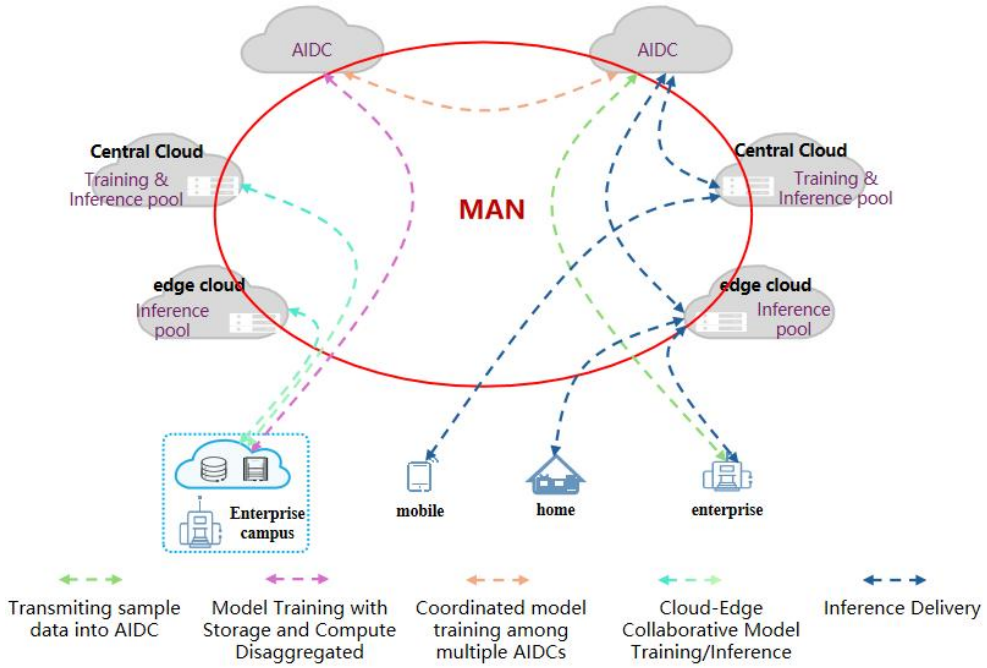

Figure 2-1: Requirements of Large AI Model

The lifecycle of large AI models encompasses multiple stages including sample data transmission, model training and model inference, each presenting distinct data transmission characteristics in terms of volume and patterns, consequently imposing higher requirements on MANs.

1. Transmiting sample data into AIDC

With the rapid advancement of large AI models, data volume is growing at an unprecedented rate. According to the Global DataSphere 2023 report released by IDC (International Data Corporation), China's data volume reached approximately 30 ZB in 2023 and is projected to expand to 76.6 ZB by 2027. Currently, numerous enterprises still rely on shipping physical hard drives to transfer sample data. This 'manual copying + physical delivery' approach is not only inefficient but also carries data loss risks. Existing network-based solutions exhibit significant limitations: traditional dedicated line services adopt fixed-bandwidth monthly/annual subscription models, while enterprises typically require only intermittent sample data transfers, resulting in high costs relative to actual usage. MANs requires capability upgrades to

provide more efficient and cost-optimized sample data transmission services.

MANs should support network scale load balancing to achieve sustained ultra-high throughput exceeding 90% across all links, enabling efficient hourly transmission of terabyte(TB)-scale sample data from enterprise to AIDCs. Simultaneously, MANs should feature highly elastic and agile service capabilities, offering on-demand elastic bandwidth to enterprises through 'just-in-time provisioning' task-based services, while providing multi-level data transmission services (minute-level, hour-level, and day-level) to meet diverse user demands. Furthermore, MANs should possess intelligent computing power orchestration capabilities to dynamically match optimal computing-network resources and transmission paths based on service characteristics including origin, type and coverage area, thereby establishing a more agile and efficient computing power provisioning system.

2. Model Training with Storage and Compute Disaggregated

Numerous industries handle sensitive data with critical security requirements, , such as the experimental and accident data in automotive manufacturing, or consumer transaction records and personally identifiable information in financial. When leasing cloud computing resources, these organizations or enterprises strictly require the localized storage of data and the guaranteed protection against data leakage during model training. To address these data security requirements, model training requires the disaggregated storage and compute architecture (with compute nodes deployed in cloud and storage nodes maintained on-premises), where training data is pulled into memory on-demand without being written to compute node disks.

In this scenario, sample data is directly written from storage nodes to compute node memory across MANs through RDMA technology. Current mainstream RDMA protocols rely on Go-Back-N retransmission mechanisms, making them highly sensitive to latency and packet loss (Even a 0.1% packet loss rate can degrade computational performance by 50%). Therefore, MANs should not only support highly resilient and high-throughput data transmission, but also incorporate precise

flow control to guarantee lossless RDMA transmission, ensuring less than 5% computational efficiency degradation across 100-500 km metropolitan domains. Moreover, MANs should deploy robust data encryption mechanisms to ensure the security of data transmission.

3. Coordinated model training among across AIDCs

The Scaling Law for large AI models persists, with computing power demands having grown by approximately one million-fold over the past decade, and projected to sustain an annual growth rate exceeding 400%. The scalability of individual computing resource pools is constrained by physical infrastructure limitations including space and power supply. Coordinated model training among multiple AIDCs enables the efficient consolidation of geographically dispersed computing power, supporting large AI model training at scales of 100K+ GPUs. The computing power of existing AIDCs is typically small-scale (in China, AIDCs with 100-300 PFLOPS account for over 70% of the total). Therefore, integrating distributed computing power across data centers, research institutions, and cloud service providers will help overcome geographical, facility, and vendor constraints to establish a unified and high-efficiency computing power service platform.

In this scenario, parameter-plane synchronization data is transmitted across MANs while sample-plane data remains stored within enterprise premises, effectively isolating potential data leakage risks. This solution imposes stringent requirements on network bandwidth and latency, mandating MANs to deploy 400G/800G links with RDMA lossless transmission to guarantee zero packet loss during model training. Parameter synchronization between GPUs predominantly relies on AllGather/AllReduce collective communication operations, introducing significant challenges of highly concurrent and burst traffic patterns. Taking the training of a 1000 billion parameters model as an example, a single parameter synchronization cycle in a 16K GPU AI cluster generates over 1.6 PB concurrent traffic. Therefore, MANs require device capability upgrades (including GB-level port buffers and tenant-level queuing) to enable optimized burst traffic processing and collective

communication scheduling, while establishing high convergence ratio network architectures (4:1, 8:1, 16:1) to balance computational efficiency with deployment costs. Furthermore, network failures causing critical issues such as training task interruptions would severely reduce training efficiency. MANs should implement tenant-level network slicing isolation and incorporate network simulation and self-healing technologies to realize Level 4 autonomous network, guaranteeing controllable failure impact scope and rapid service recovery.

4. Cloud-Edge Collaborative Model Training/Inference

The dramatic reduction in large model training and inference costs has enabled enterprises to rapidly adopt AI applications through on-premises deployment of AI Training & Inference server. However, enterprise on-premises computing resource pools encounter significant challenges in capacity expansion and high operational maintenance costs, rendering them inadequate to meet the escalating demands for model fine-tuning and inference. To address this, the cloud-edge collaboration between enterprise on-premises and cloud computing resource pools presents a more efficient, agile, and cost-effective approach to realize elastic computing power scaling. This solution leverages parallel computing techniques including pipeline parallelism and expert parallelism to partition large AI models across on-premise and cloud computing resource pools. By implementing localized deployment of input/output embedding layers, it ensures strict on-premises sample data containment, thereby fulfilling the data security requirements for highly regulated sectors such as financial and healthcare.

In this scenario, MANs should support lossless RDMA transmission to prevent significant computational efficiency degradation caused by packet loss. Simultaneously, MANs requires tenant-level network slicing to ensure effective service isolation, meeting SLA requirements while preventing interference from other service failures. Furthermore, MANs should possess intelligent computing power scheduling capabilities to dynamically select optimal edge resource pools based on user location and service demands, ensuring efficient model fine-tuning/inference

processes.

5. Inference Delivery

AI inference enables large AI models to be applied in real-world scenarios, serving as the critical step for commercialization. By 2027, approximately 70% of new applications are expected to incorporate AI inference models, with concurrent transactions between AI applications and resource pools anticipated to reach the million-scale threshold. Inference delivery comprises two fundamental processes: model delivery, referring to the deployment of AI inference models across multiple edge clouds; and result delivery, denoting the interaction between users and AI inference models to generate required outputs.

In this scenario, MANs should provide low-latency and high-bandwidth data transmission capabilities with ubiquitous coverage and seamless access to ensure the service quality of AI applications. MANs should also incorporate deterministic service capabilities, enabling precise traffic identification and optimized path selection to enhance transmission determinism and reliability.

2.3.2 Requirements of small AI model

Small AI models feature compact architecture, low computational demands, and rapid response capabilities. These models are typically designed for specialized tasks and demonstrate unique advantages in resource-constrained environments such as smart terminals and IoT devices. With their widespread deployment across smart home systems, industrial IoT applications, and mobile platforms, they are imposing more requirements on MAN.

1. Inference Delivery

In real-time AI inference, small AI models are predominantly deployed on edge devices proximal to data sources, enabling instantaneous processing of input data and generation of predictive outputs to achieve ultra-low latency response. For scenarios requiring greater computational power, edge nodes collaborate with cloud in a hybrid deployment architecture. Edge devices process routine high-frequency inference

requests locally, while computationally intensive or anomalous cases are offloaded through MANs to cloud for deep analysis. MANs should provide guaranteed bandwidth, deterministic low-latency path, and intelligent traffic orchestration capabilities to ensure reliable and real-time AI inference service delivery.

2. Federated learning

Federated learning is a critical training paradigm for small AI model. It adopts Federated deployment that significantly enhances model efficacy while ensuring local data privacy preservation, imposing three critical requirements on MAN: First, real-time parameter synchronization demands guaranteed periodic connectivity for participants to maintain training continuity; Second, data transmission security requires end-to-end encryption for model parameters to prevent any potential model leakage; Third, MANs should incorporate dynamic resource allocation capabilities, allocating greater bandwidth to higher-priority participants based on their differential training progress.

2.3.3 Requirements of hybrid AI model

Hybrid AI model deploys lightweight small AI models at edge while hosting large AI models with advanced comprehension and reasoning capabilities in the cloud. Through efficient collaboration between these models, hybrid AI Model fully leverages the small model's advantages in low-latency response and personalized adaptation while harnessing the large model's capabilities in multi-modal understanding and generalized intelligence.

The coordination between large AI models and small AI models is primarily reflected in two critical aspects: data interaction and model updating. For data interaction, edge-deployed small AI models perform localized data collection and preprocessing before transmitting critical data to cloud-deployed large AI models for analysis, with the calculation results subsequently delivered back to edge devices for execution. This process requires MANs to provide deterministic service capabilities that ensure low-latency, high-bandwidth, and highly stable data transmission. For

model updates, cloud-deployed large AI models can distribute optimized parameters or models to edge devices through techniques such as knowledge distillation, enabling continuous iteration of small AI models. This process relies on the network-level load balancing capability of MANs to realize high throughput, particularly during concurrent updates across massive edge devices.

2.4 AI applications driven MANs toward next-generation evolution

The rapid advancement of AI applications is imposing more stringent demands on MAN: At the architectural level, MANs should support efficient north-south and east-west traffic steering to meet cloud-edge and inter-cloud coordination requirements, while enabling elastic scalability to achieve ubiquitous user access. At the technical level, MANs should incorporate capabilities including network-scale load balancing, flow-level precise flow control, and high oversubscription ratio networking to support ultra-large-scale AI computing traffic, while ensuring real-time interactive experience through deterministic services and tenant-level network slicing. At the operational level, MANs should strengthen service-oriented capabilities to provide flexible and agile computing power services for users, while enhancing intelligent O&M capabilities to ensure high stability and reliability of services. At the scheduling level, MANs should establish an intelligent cross-domain coordination system to achieve global optimal allocation of computing power, storage and network resources. These requirements are accelerating the transformation of MANs into next-generation intelligent, converged, and deterministic AI infrastructure, enabling continuous innovation in cloud-network integrated products and services.

Chapter III

MAN Architecture for the AI Era

3.1 MANs design objectives

The metropolitan area networks (MANs) design for AI era focuses on the construction of the next-generation network infrastructure with deeply integration between computing and network, and is intelligent, efficient, secure, and reliable. The core directions are as follows:

1. Integrated computing and network, converged bearer

Centered on the computing resource pool, MANs can integrate the heterogeneous computing power of general computing, intelligent computing, and supercomputing. SRv6 and other technologies are used to uniformly schedule and intelligently orchestrate network, cloud, and compute resources, breaking physical isolation. Supports lossless transmission of heterogeneous computing power across domains and collaborative training of multiple AIDCs, building a foundation for cloud-network synergy innovation. Unified access of fixed, mobile, and cloud services and converged bearer of multiple services, achieving ubiquitous access of users. By building intelligent, agile, secure, and reliable high-quality network infrastructure, MANs can effectively support efficient collaboration of multi-dimensional services and provides end-to-end all-scenario connection services for digital transformation and smart upgrade of industries.

2. Elasticity, agility, flexibility, and efficiency

Based on the Spine-Leaf modular architecture and IPv6 Enhanced technology foundation, agile network expansion and service provisioning in minutes can be achieved. Intelligent identification of elephant flows and network flow-level scheduling enable network-level load balancing and refined management and control of service flows, ensuring high throughput and low latency transmission performance, implementing quick traffic grooming and high user experience access, and comprehensively improving the overall network transmission efficiency.

3. Precise control and dynamic convergence

Based on intelligent flow identification and precise flow control technology, and the deterministic delay forwarding and network convergence optimization mechanism,

the RDMA high-performance lossless interconnection architecture is constructed. Based on the intelligent flow-level scheduling capability and the flexible computing power-oriented network architecture, dynamic collaboration between enterprise computing power and hub computing power centers is supported. On-demand service flow adaptation and precise resource orchestration effectively support TB-level data throughput requirements for large AI model training and inference, achieving the optimal balance between network construction costs and computing efficiency.

4. Intelligent O&M, security and reliability

The AI-driven intelligent management and control system is deployed to build intelligent O&M capabilities such as flow-level scheduling optimization, fault self-healing, and network simulation. The dual-plane redundancy architecture and cross-domain disaster recovery mechanism ensure high system availability. Network slicing, tenant-level flow control, and standard security interfaces are used to establish a multi-layer security isolation system. Integrated with technologies such as zero-packet-loss transmission assurance and end-to-end QoE degradation detection to ensure reliable transmission over all paths of data pipelines and redundant protection for multi-plane computing boundaries, ensuring secure and reliable service running throughout the service lifecycle.

3.2 Overall MAN architecture

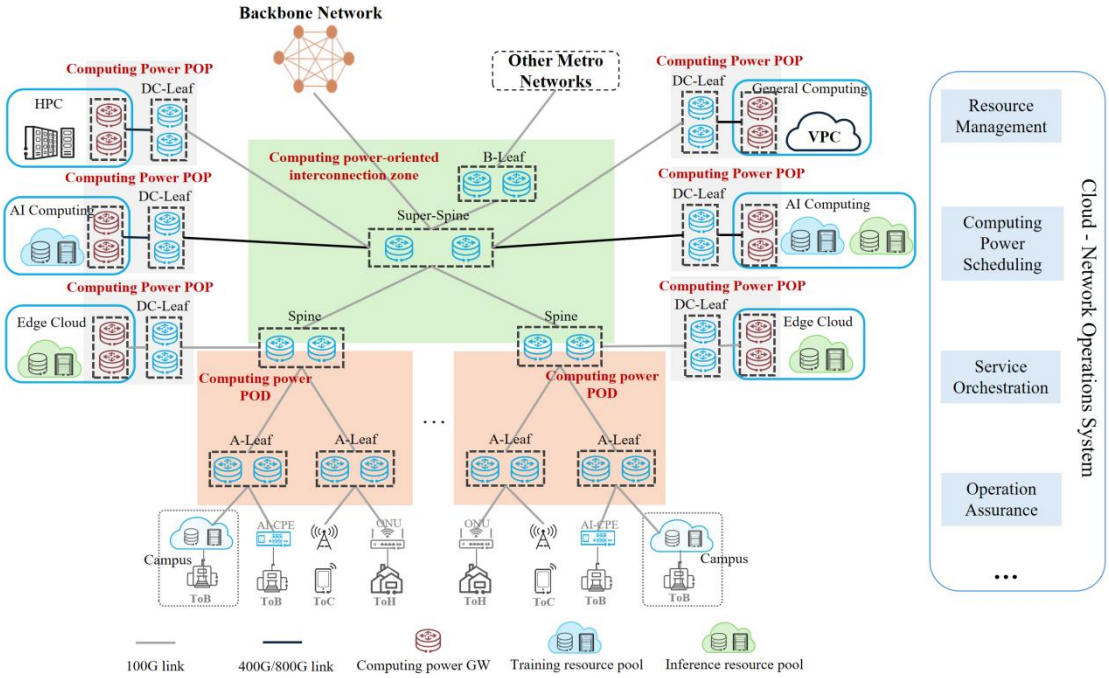

Figure 3-1: MANs Architecture for the AI Era

The MAN architecture for the AI era consists of three core modules: Computing power-oriented Point Of Delivery (POD) zone, Computing power-oriented Point Of Presence (POP) zone, and Computing power-oriented interconnection zone. The three modules seamlessly collaborate with the cloud-network operation system through standard interface and protocols. A tailorable and hot-swappable building-block architecture is used to achieve elastic scaling of computing resources as required.

- Computing power-oriented POD zone: This module introduces PODs in the data center to the MAN. Based on the spine-leaf modular architecture, this module enables efficient access of customer terminals and enterprise branches through optical fibers, PON, and 5G, and supports large-capacity data exchange, fast convergence and traffic diversion of fixed and mobile services and intelligent computing services in a region. SRv6 and EVPN technologies are used to carry multiple services in a unified manner. The cloud & network operation system is used to achieve automatic service provisioning and intelligent O&M. The network slicing technology provides customized bandwidth and security

assurance for intelligent computing, industry, and public services.

- Computing power-oriented POP zone: As the interconnection point between the cloud network and the bearer network, this module interconnects with the computing power resource pool in a standard manner to implement on-demand scheduling and elastic allocation of computing power resources, so as to support integrated computing-network services. Serves as the network anchor of the computing resource pool, it connects to provincial/regional spine nodes, opens up inter-computing channels, and supports cross-domain resource collaboration and disaster recovery. Interworks with the computing power-oriented POD zone to provide end-to-end lossless connection between users in different PODs and computing power resource pools.
- Computing power-oriented interconnection zone: This module serves as the hub between the MAN, backbone network, Internet, and industry private networks. It simplifies the connection between the MAN and external networks and between various computing power resource pools, implements flexible component expansion, and efficiently diverts traffic between components. Uses 400G/800G high-speed links, network-level load balancing, and SRv6/EVPN technologies to achieve efficient inter-domain traffic forwarding and path optimization. The network slicing technology provides differentiated bandwidth and security assurance for intelligent computing services, ensuring stable service interconnection and user experience.

The three module zones together constitute the MAN architecture oriented to the AI era. Each module zone plays a specific role to ensure the efficient bearing of AI services on the MAN. The computing power-oriented POD zone functions as the user access entry and connects to the computing power-oriented POP zone through spine devices to build efficient transmission channels between users and the computing power resource pool. The computing power-oriented interconnection zone and computing power-oriented POP zone are connected to the cross-domain computing power pool collaboration network to implement intelligent scheduling of computing

power resources. The three modules use standard technologies such as SRv6 and EVPN to ensure end-to-end service logic consistency and provide high-quality network bearer capabilities for AI services.

Based on the concept of hierarchical decoupling and collaborative design, the architecture builds an integrated computing service network featuring edge access, core scheduling, and cross-domain collaboration. It uses the cloud-network operations system to implement unified management and control and intelligent scheduling of network-wide resources. The cloud network operation system focuses on four core modules: resource management, computing power scheduling, service orchestration, and operation assurance. Resource management integrates network and computing power resources to achieve global visibility and management, and computing power scheduling dynamically optimizes resource allocation based on service requirements. Service orchestration implements quick service deployment and end-to-end integration through automated processes. Operation assurance uses intelligent monitoring and analysis technologies to ensure stable system running and user experience. The collaborative operation of modules provides a solid foundation for the computing power requirements of high concurrency, low latency, and high reliability in the AI era.

3.3 Key modules of MAN

3.3.1 Computing power-oriented POD zone

The computing power-oriented POD zone is the edge access layer of the MAN and provides converged access for customer terminals (2C), enterprise branches (2B), and home users (2H). Aggregates traffic level by level through base stations, CPEs, leaf nodes, and spine nodes to form a wide-coverage and flexible computing service entry. In addition, deep and shallow edge computing can be mounted on demand, providing customers with low-latency and high-experience computing services. Its core functions include:

- Converged access: supports multiple access modes, such as optical fiber, PON,

and 5G, implementing "one-line for multi-computing". A single line can meet the access requirements of Internet, cloud services, and multi-computing power pools.

- Elastic bandwidth: Provides elastic access capabilities from 0 to 100 Gbit/s, adapting to changes in customers' computing power requirements.
- Pooling scheduling: Supports deep and shallow edge computing power pooling and cross-POD scheduling, flexible coverage based on the service scale or service scope, achieving efficient computing resource transmission.

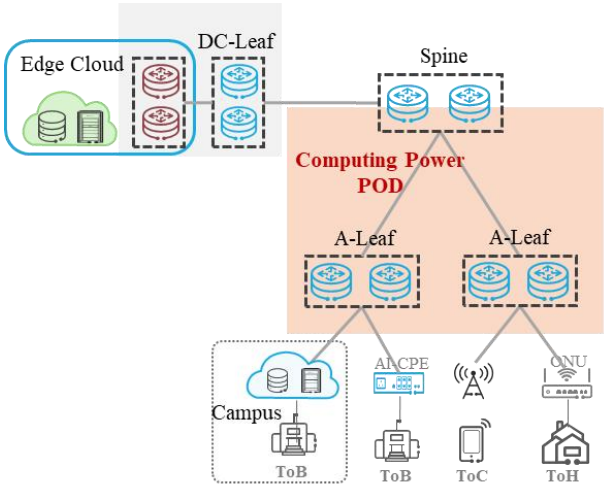

Figure 3-2 Computing power-oriented POD Zone

The computing power-oriented POD zone uses a wide-coverage and level-by-level convergence networking architecture to dynamically balance the computing power resource utilization while reducing network coverage costs. Precise scheduling and control at the flow level ensure data transmission quality based on the RDMA protocol, and effectively support long-distance lossless transmission in scenarios where massive samples are quickly processed and storage and computing are separated. In addition, the operation system has the flexible and agile resource configuration capability. By dynamically adjusting network paths and bandwidth resources, the operation system effectively copes with service traffic fluctuation and ensures service continuity and stability.

3.3.2 Computing power-oriented POP zone

The computing power-oriented POP zone connects the MAN and computing power resource pool through the computing power gateway, implementing standardized and fast interconnection between the sample plane network and service plane network of the MAN and computing power resource pool. The computing POP zone provides standardized functional zone interconnection policies and deployment guidance, supporting integrated bearing and resource scheduling of multiple services. Its core functions include:

- Modular networking: Standard modules connect to heterogeneous computing resource pools (self-owned or third-party) to implement resource pooling and unified management.
- End-to-end lossless connection: Connects to provincial/regional spine nodes and associates with multiple computing power PODs to provide low-latency and high-reliability connections between users in different PODs and computing power pools.
- Intelligent computing service support: Stream-level precise flow control is used to meet service requirements such as sample input calculation, model training with storage and compute disaggregated across AIDCs, and cross-cluster collaborative training.

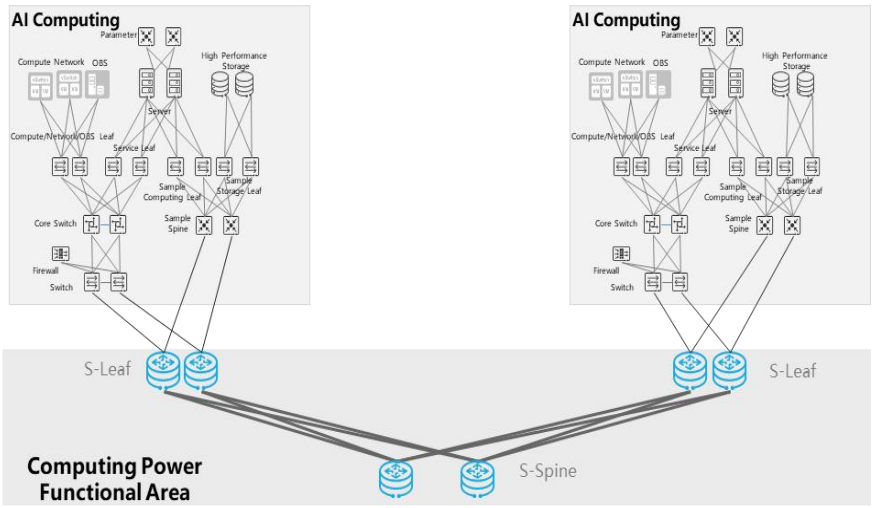

Figure 3-3 Computing power-oriented POP Zone

The computing power-oriented POP zone supports multi-service convergence and supports various services, such as general computing and intelligent computing. As the north-south traffic scheduling hub, the computing gateway intelligently interconnects with the sample plane and service plane networks of the computing resource pool through standard interconnection policies. The computing power POP establishes high-speed links with provincial or regional spine nodes and multiple PODs to form an end-to-end connection between users, computing PODs, computing POPs, and computing resource pools.

3.3.3 Computing power-oriented interconnection zone

As the core hub node of the MAN, the computing power-oriented interconnection zone connects the computing power-oriented POD zone and computing power-oriented POP zone to the backbone network and Internet egress through 400G/800G high-speed links. The network slicing technology provides differentiated bandwidth and security assurance for intelligent computing services, ensuring stable service interconnection and user experience. Its core functions include:

- Differentiated services: Based on technologies such as precise traffic identification to classify and mark different service flows, so as to provide differentiated service quality assurance for different types of services, and meet different requirements for latency, bandwidth, and packet loss rate of various services. Ensure that mission-critical services and high-value services can obtain priority processing and better network resources.
- Traffic scheduling and steering: Schedules and manages the traffic in each functional zone of the MAN in a unified manner, and steers the traffic to different links and paths based on the network load, service requirements, and predefined policies. In this way, the traffic is evenly distributed and the network resource utilization can be improved.
- High-speed network interconnection: As the hub for connecting MANs to backbone networks, other MANs, computing power-oriented POPs, and

computing power-oriented PODs, the uses 400G/800G links to implement high-speed interconnection between different networks. Exchanges routing information with external networks, ensures that data packets can be correctly forwarded between the MAN and external networks, and ensures smooth transmission of various services between different network domains.

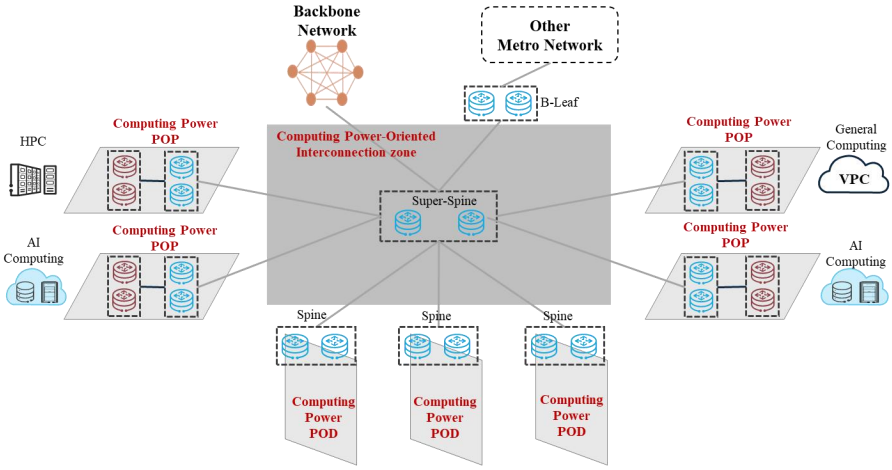

Figure 3-4 Computing power-oriented interconnection zone

The computing power-oriented interconnection zone builds an integrated MANs base of "high-speed interconnection and intelligent scheduling". Support service innovation through differentiated services, release computing resource efficiency through intelligent scheduling, break down cross-domain collaboration barriers, and enable computing power circulation and converged applications of new network infrastructure.

Chapter IV

MAN Key Technologies for the AI Era

4.1 Integrated computing and network, converged bearer network

4.1.1 Unified service bearer network

In the AI era, MANs are facing new challenges brought by the sharp increase in computing power collaboration and data transmission requirements. Therefore, the unified protocol stack is urgently required to carry multiple fixed, mobile, cloud, and computing services simultaneously, reducing network complexity. Service deployment and O&M efficiency is significantly improved. The SRv6 and EVPN-based converged architecture provides an ideal solution for unified service bearer network. It implements logical isolation and flexible scheduling of services on a single network, avoiding architecture redundancy caused by multiple traditional networks and greatly improving network resource utilization. Its core strengths are as follows:

- Unified user access: SRv6 supports cross-domain end-to-end connections based on IPv6 native protocols. Enterprise users can meet multiple service requirements with only one access, significantly reducing access complexity.
- Unified service bearing: EVPN provides flexible Layer 2/Layer 3 VPN services and SRv6 source routing capabilities to dynamically adapt to various service SLA requirements, implementing intelligent traffic scheduling and resource optimization.
- Convenient service provisioning: Intelligent O&M technologies such as autonomous driving network enable automatic service orchestration and minute-level service provisioning, significantly improving network agility. In addition, SRv6's network programmability lays a foundation for AI-driven network optimization, further improving network resource utilization and intelligence.

4.1.2 Intelligent scheduling of computing power

In ubiquitous computing power scenarios, MANs face the core challenge of dynamic matching of computing power supply and demand. Therefore, key problems such as resource dispersion, requirement diversity, and task real-timeness need to be resolved. MANs needs to build an intelligent scheduling mechanism for computing power. The mechanism implements dynamic pricing and task allocation by real-time sensing of supply and demand status and algorithm optimization, ensuring efficient utilization of computing power resources and meeting users' core requirements for low latency, high reliability, and low cost.

The core objective of intelligent scheduling of computing power is to achieve dynamic matching between supply and demand and improve the collaboration efficiency of computing power resources. Based on the geographical location, resource type, and real-time load of the supplier, as well as the service SLA requirements and task characteristics of the demander, the global computing power awareness and unified measurement system are constructed. In this process, SRv6 uses flexible and programmable features to deeply bind computing power scheduling and service requirements through network path optimization. Intelligent scheduling of computing power builds a closed-loop system featuring dynamic resource awareness, SRv6 path optimization, and intelligent decision-making to implement precise scheduling of heterogeneous resources across domains and provide low-latency and high-elastic computing power assurance for computing scenarios such as general computing, intelligent computing, and supercomputing.

4.2 Elasticity, agility, flexibility and efficiency

4.2.1 Task-based scheduling

The task-based scheduling technology facilitates the off-peak transmission of non-real-time tasks (such as data backup tasks) and improves the utilization of idle network resources. This technology is based on the intelligent closed-loop mechanism

of "requirement awareness-resource prediction-dynamic fulfillment", which improves resource utilization efficiency and user experience. First, the operation system receives user transmission requirements through standardized interfaces and performs multi-dimensional feasibility evaluation based on historical bandwidth data. and feed back a committed transmission time window to the user. Second, the network slicing technology is used to dynamically allocate physical port resources, and dedicated transmission channels are provided for users. Finally, the transmission quality is monitored in real time during the task execution and bandwidth resources are automatically released after the task is complete.

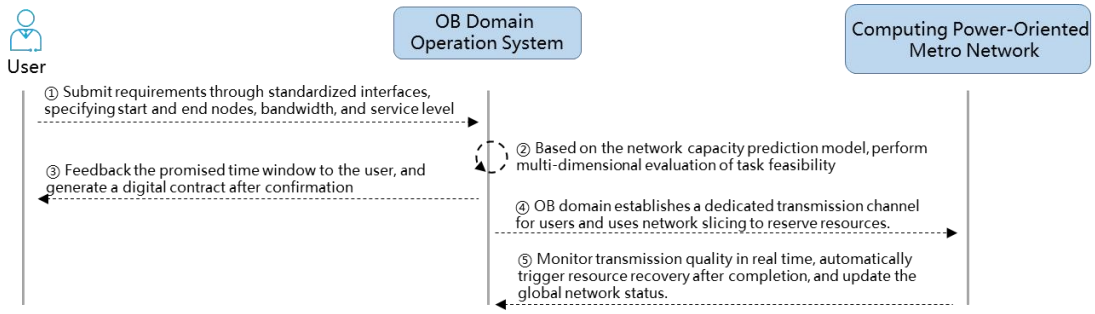

Figure 4-1: Task-based scheduling process

The task-based scheduling technology implements end-to-end automated service processes, greatly optimizes the response time from requirement submission to resource readiness, and significantly improves network-wide resource utilization. This technology establishes a precise time-effective guarantee mechanism, and relies on path pre-computation and dynamic optimization algorithms to ensure the deterministic commitment of transmission time-effectiveness. The digital twin network is used to simulate complex tasks, build intelligent resource scheduling capabilities, and implement conflict avoidance and global optimal orchestration in multi-task concurrency scenarios. Finally, the dynamic and accurate matching between network resource supply and user demand fluctuation in seconds is achieved.

4.2.2 Elastic bandwidth

In typical application scenarios, such as massive sample data storage, enterprises face bandwidth configuration problems caused by periodic data transmission peak

hours. Long-term use of high-bandwidth private lines will waste resources during idle periods, while low-bandwidth private lines will idle computing resources due to long transmission delay. The elastic bandwidth technology enables dynamic on-demand expansion of bandwidth resources and relies on the service agility of the management and control domain, effectively solving the dilemma of "high bandwidth cannot be used and low bandwidth cannot be used".

The elastic bandwidth technology implements dynamic scheduling of bandwidth resources by constructing in-depth collaboration between the network and the operation system. This technology receives bandwidth adjustment instructions from users based on standard service interfaces, builds automatic service orchestration capabilities based on the operation system, and supports minute-level synchronous adjustment of port rates, QoS policies, and routing entries. The entire process forms a closed-loop control of requirement awareness, policy generation, and resource reconstruction, implementing elastic scaling of the bandwidth of a single private line within the range of 100 Mbit/s to 10 Gbit/s/100 Gbit/s. This technology not only provides enterprises with minute-level online scale-out and scale-in agile response capabilities, effectively copes with the instantaneous requirements of network resources caused by burst services, but also supports precise charging based on duration and usage, significantly improving the network service capability.

4.2.3 High-bandwidth links

The rapid development of intelligent computing services imposes higher requirements on link bandwidth. To meet the requirements for uploading TB/PB-level enterprise sample data in minutes or hours at a high speed, MANs are accelerating link upgrade. Edge nodes use 100G high-speed access, and 400G high-bandwidth links are deployed at the aggregation layer to carry aggregated traffic. In addition, to cope with the insufficient computing power of a single intelligent computing center, resources of multiple intelligent computing centers need to be integrated to support large model training. In this context, 400G high-speed links have been widely used on

data center parameter plane networks, and MANs need to be upgraded to 400G architecture to improve bandwidth utilization and dynamic scheduling capabilities and build high-capacity network infrastructure.

Increasing the rate of a single port is a key technology for efficient and low-cost transmission of ultra-large-scale traffic. It has become the core evolution direction of the intelligent computing Internet. Currently, the 400GE port technology for metro interconnection has become mature. Large-scale deployment of 400G interconnection links can effectively reduce the single-bit transmission cost of intelligent computing interconnection, lay a foundation for future evolution to 800G technology, and continuously optimize the transmission cost per bit.

4.2.4 Network-level load balancing

large AI model implements distributed training based on aggregate communication. Traffic has the characteristics of high synchronization, large traffic, and periodic transmission. In this service mode, each equal-cost path in the network carries a large number of data flows at the same time. As a result, the traditional hash-based load balancing technology cannot achieve complete balance between paths. Network-level load balancing is used to solve the problem of packet loss caused by congestion on a non-faulty homogeneous network in the cross-AIDC collaborative training scenario. In the non-fault scenario, the network device does not have faults such as optical module damage and intermittent link disconnection. In the homogeneous scenario, the bandwidth and delay of the network device are symmetrical and synchronized. This technology effectively improves network transmission efficiency in intelligent collaborative training scenarios by optimizing the traffic allocation mechanism.

Network-level load balancing implements conflict-free and balanced scheduling between paths through unified network-wide traffic planning. In this mechanism, the network device first collects the traffic information of the service in real time and reports the information to the network controller. The network controller runs the

global route selection algorithm based on the topology status and traffic characteristics, and intelligently allocates the optimal transmission path to each flow. Finally, the controller delivers the path decision to the network device to perform path adjustment. This dynamic traffic scheduling mechanism based on the global perspective implements efficient and even load distribution, achieves the end-to-end flow transmission efficiency of over 95%, and effectively ensures the efficient and stable running of the training process.

4.3 Precise control and dynamic convergence

4.3.1 Intelligent identification and scheduling of elephant flows

In the AI era, the traffic characteristics of MANs are undergoing a remarkable transformation. The traditional service mode based on massive small and micro flows is gradually evolving to new service forms such as AI training and distributed computing, which are characterized by high bandwidth and long-term elephant flows. Such heavy-traffic services are prone to network congestion and cause overall throughput performance deterioration. Therefore, an intelligent traffic identification and scheduling system is required to improve network resource utilization and ensure efficient transmission of key AI services and overall network performance.

The intelligent identification and scheduling technology of elephant flow traffic builds a closed-loop optimization system of "perception - decision making - execution" to maximize the global network capacity. This technology detects elephant flows in real time through in-depth traffic feature analysis, and reports fine-grained data such as flow features and throughput to the controller in real time by using the Telemetry technology. Based on the SRv6 programmable feature and real-time network situation (such as topology status and link load), the controller establishes an accurate matching model between traffic requirements and resource provisioning and dynamically generates an optimal SRv6 scheduling policy. By intelligently guiding

elephant flows to the optimal path, this technology not only ensures that the throughput of AI services approaches the physical bandwidth limit, but also significantly reduces the link congestion probability through flow-level precise scheduling, building a transmission environment with high throughput, low latency, and low congestion, and providing reliable assurance for large-scale data exchange. For RDMA service, it can also split the elephant flow based on the information in the inner headers of the packet, so as to implement fine-grained traffic identification and management.

4.3.2 Precise flow control

With the rapid development of intelligent computing services, such as cross-AIDC collaborative training and cloud-edge collaboration training/inference, the wide application of the RDMA transmission protocol poses higher requirements on the flow control mechanism of the MAN. Currently, PFC mechanism is widely used in data centers to ensure lossless transmission. However, the coarse-grained control at the port queue level is prone to head block and false damage problems. In contrast, the data flow-based precise flow control technology implements flow-level precise backpressure control through fine-grained identification based on IP 5-tuple. This technology not only effectively resolves the inherent defects of traditional PFC, but also dynamically optimizes flow control policies based on real-time network status. This feature ensures efficient and stable data transmission in complex WAN multi-tenant scenarios and provides a better bearer environment for RDMA services.

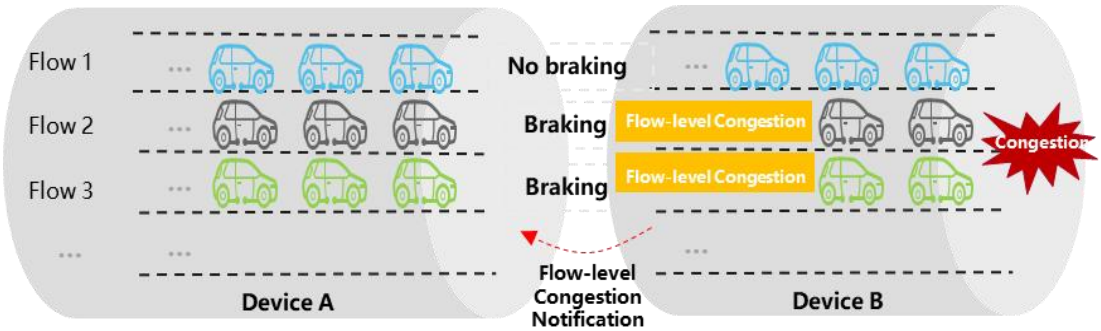

Figure 4-2: Flow-level precise flow control technology

To meet the lossless transmission requirements of the RDMA protocol, the flow-level precise flow control technology builds a fine-grained control system. This technology breaks through the limitations of traditional PFC physical port-level control. It allocates an independent queue buffer to each RDMA service flow and monitors the buffer water level in real time, implementing finer traffic management. When congestion occurs on a specific service flow, the system isolates and stores the packet in the dedicated buffer. When the queue depth exceeds the preset threshold, the system sends flow-level backpressure signals to the upstream device in hop-by-hop backtracking mode. This technology only limits the rate of the congested flow, effectively avoiding the problems caused by the traditional PFC technology, such as queue head congestion. The practical test shows that the proposed technology can control the end-to-end packet loss rate of wide area RDMA transmission below 0.001%, and keep the stable throughput rate above 95%. Then the risk of network congestion spreading is eliminated completely by the fault domain isolation mechanism between service flows.

4.3.3 High convergence ratio network

In the cross-AIDC collaborative training scenario, MANs needs to carry large-scale parameter plane data synchronization between multiple data centers. For example, the 200 Gbit/s transmission rate of a single NIC leads to the peak burst traffic on the parameter plane as high as 2000 Tbit/s. If the non-convergence networking solution is used, the construction cost is high. Therefore, the high convergence ratio networking technology implements efficient convergence of cross-DC collaborative training traffic by using an in-depth collaborative optimization of a set communication algorithm and a network architecture, thereby significantly reducing deployment costs of network infrastructure between multiple AIDCs.

This technology innovatively adopts the three-in-one collaborative mechanism of "algorithm-based peak reduction, cache peak clipping, and scheduling acceleration", which can maintain over 95% end-to-end computing efficiency under the network

architecture with high convergence ratios, such as 32:1 and 64:1. A new network paradigm adapted to cross-AIDC collaborative training is constructed. Its core lies in the reconstruction of the aggregate communication process by using the hierarchical gradient aggregation algorithm, effectively reducing the number of computing cards for cross-DC communication, and realizing the initial convergence of network bandwidth. In addition, smart routers with large-capacity buffers are deployed on MANs. The dual mechanism of "burst buffering + queue scheduling" is used to split training tasks into controllable microburst flows. The cache is used to absorb traffic impact and priority scheduling is used to ensure the timely transmission of GPU control signaling. Avoids idle waiting of computing resources, thereby significantly reducing bandwidth requirements across data center network while ensuring training efficiency.

4.3.4 deterministic service network

In recent years, with the booming rise of intelligent computing service, the application scenarios of inference service are becoming more and more extensive. The transition of inference service from single mode to multi-mode and the continuous evolution towards real-time interaction imposes more stringent requirements on the network. On the one hand, the network needs to be characterized by deterministic low latency to ensure real-time, smooth presentation of inference result, avoiding frame freezing and latency. On the other hand, deterministic bandwidth services are also indispensable. They can ensure stable network bandwidth and uninterrupted data transmission during a large amount of data transmission without congestion. Therefore, the network needs to provide deterministic network services.

The deterministic network technology can partition network resources into different logical networks and provide independent logical networks for different services to implement differentiated services. With SRv6 and Flex-E technology, it can plan data transmission paths more flexibly and optimize network traffic distribution. using SRv6 with the network controller, low-latency path computation is

implemented, effectively ensuring service latency. The network controller collects network topology and link status information in real time. Based on the collected information and the path programmable feature of SRv6, the network controller computes the optimal path that meets the latency requirement for services. When service data enters the network, the SRv6 path is forwarded along the planned SRv6 path to avoid congested nodes and links, reducing the transmission delay. In addition, through Flex-E bandwidth reservation mechanism allocates dedicated bandwidth resources to services, ensures that the bandwidth requirements of specific services can be met even when the network is congested or busy, preventing service performance deterioration caused by bandwidth contention, ensuring service bandwidth and service experience even when the network load is heavy.

4.4 Intelligent O&M, security and reliability

4.4.1 Intelligent O&M capability

With the rapid development of technologies such as 5G, Internet of Things (IoT), and edge computing, MANs are facing challenges such as traffic surge, service diversification, and strict service quality requirements. The traditional O&M mode based on manual rules and static policies cannot meet the real-time, reliable, and flexible network requirements in the AI era. Therefore, an intelligent O&M system is urgently required for MAN networks. TMF defines network autonomy as six levels (L0 to L5). The core of the intelligent O&M system for MANs in the AI era is to build a new network AI brain that features self-sensing, self-analysis, self-decision-making, and self-deployment, helps the network autonomy level to evolve from L3 conditional autonomy to L4 advanced autonomy, and finally achieves the goal of L5 complete intelligent autonomy.

The MAN intelligent O&M system is constructed based on multiple key technologies. First, distributed probes and embedded AI chips are deployed to realize multi-dimensional real-time network status awareness. Second, the intelligent analysis engine built based on deep learning processes massive O&M data in real time. Finally,

the SDN controller and automatic orchestration system are used to deliver and adjust policies in seconds. The following are the key intelligent O&M capabilities:

- High-precision simulation: Real-time online digital mirroring network is constructed to implement multi-level visualized simulation of physical topology, routes, tunnels, VPNs, and flows. The system automatically synchronizes live network configurations, BGP routes, and traffic characteristics, establishes a benchmark mirroring network model, and constructs a pre-evaluation system for configuration changes based on the digital twin technology. When the network configuration changes, the system automatically generates a new mirroring network. Compare and analyze the topology status, traffic distribution, and route convergence efficiency before and after the change, and provide an impact assessment report to effectively identify potential high-risk configuration errors. In addition, with the dynamic traffic modeling technology, the system can simulate routing policies and traffic changes in milliseconds, accurately predict the evolution trend of key performance indicators such as delay fluctuation and packet loss rate threshold, providing data support for network optimization decision-making.
- AI diagnosis: A multi-dimensional fault self-diagnosis model is established based on the second-level fault feature extraction on the device side, knowledge graph inference, and time series pattern mining. The system adopts the big model thinking chain technology to realize intelligent alarm aggregation and status trend prediction, and supports dynamic fault root cause reasoning and potential risk identification. The online knowledge injection mechanism enables the system to perform guided diagnosis on unknown faults and generate closed-loop handling suggestions, forming a complete intelligent O&M solution.
- Self-healing network: A complete closed-loop fault handling mechanism is built based on the O&M knowledge database and dynamic orchestration capabilities of large models, implementing automatic processing in the entire process of "sensing-diagnosis-decision-making-execution". For non-hardware faults, the

system automatically implements recovery policies such as redundant path switchover. For hardware faults, the system generates precise maintenance orders based on digital twin simulation. In addition, based on the correlation analysis of network topology status, device health indicators, and traffic patterns, the system can predict potential faults in advance and implement intelligent DR with minute-level self-healing.

4.4.2 Tenant-level network slice isolation

MANs needs to carry traditional services and intelligent computing services in a unified manner and meet differentiated SLA requirements in different service scenarios. The dual isolation mechanism between logical and physical resources effectively prevents resource preemption and ensures the deterministic service capability of key indicators such as bandwidth and latency for training and inference services. As a new IPv6-based network solution, the tenant-level slice isolation technology makes full use of SRv6 programmability and IPv6 address space advantages to provide multiple tenants with network slice services that share physical resources but are logically isolated. A core mechanism of this technology is as follows: A source node encapsulates a unique slice identifier according to a tenant requirement, and nodes along the path implement slice identification by parsing a packet, and execute a predefined forwarding policy.

Tenant-level network slicing technology has three core advantages: First, slice identifiers are used to represent fine-grained resources, ensuring that indicators such as bandwidth and latency between slices do not interfere with each other. Second, SRv6 network programmability supports flexible service orchestration, meeting the requirements for fast service rollout. Third, it provides a highly reliable network slicing solution to achieve optimal resource utilization and accurate service assurance in a multi-tenant environment.

4.4.3 End-to-End security assurance

A multi-level defense-in-depth system is required for MANs in the AI era to cope with data leakage and horizontal penetration risks in multi-tenant environments. Its core is to implement E2E tenant data isolation and encrypted transmission, especially in computing power scheduling and cross-domain communication scenarios to ensure data confidentiality and integrity. Based on the SRv6 VPN and network slicing technology, a three-level isolation mechanism of "access device-network slice-VPN" can be constructed to effectively block security threats by decoupling the physical layer, protocol layer, and service layer from all dimensions. In addition, security group policies and cross-domain traffic trustlist management and control are used to achieve zero cross-penetration of tenant data.

The security architecture uses the dual protection mechanisms of slice isolation and VPN encryption to upgrade security capabilities from passive defense to active immunity, achieving the security goal of "no data is sliced, no risks are crossed, and no plaintext traffic is left". In the future, cutting-edge technologies such as quantum encryption (including post-quantum cryptography and quantum key distribution) and trusted execution environment will further enhance the security protection capability of the network. Convergence of these technologies will promote the evolution of intelligent computing networks to a zero-trust architecture featuring "active immunity, dynamic awareness, and full-chain trustworthiness", providing a solid security foundation for service innovation in the AI era.

4.4.4 Green and low carbon network

As the key infrastructure that supports the development of artificial intelligence and digital economy, MANs are facing the challenges of high energy consumption and low efficiency. As computing power requirements grow exponentially and network bandwidth pressure continues to rise, problems such as high single-bit power consumption and sharp increase in heat dissipation costs of the existing 100G platform become increasingly prominent. As a result, the OPEX and carbon emissions

increase at the same time. In addition, the multi-layer network architecture brings device redundancy and protocol conversion loss, which further aggravates the energy efficiency bottleneck. Under the background of the "double carbon" strategy, it is urgent to use technological innovation to revolutionaryly improve network energy efficiency and build MANs with high energy efficiency, large bandwidth bearing capability, and intelligent scheduling features.

The green and low-carbon transformation of MANs focuses on three technical paths: In terms of ultra-high-speed platform upgrade, 400G/800G networks can significantly reduce single-bit energy consumption and support long-distance lossless transmission, meeting the requirements of high-bandwidth scenarios such as large AI model training. In terms of intelligent energy saving system, AI-based real-time load prediction and multi-factor decision-making algorithms implement dynamic optimization and adjustment of device power, heat dissipation policies, and optical module status. In terms of architecture reconstruction, SRv6 and EVPN is used to simplify network layers, promote flattened architecture, and use SDN to implement precise resource scheduling. Through the collaborative innovation of ultra-high-speed platforms, intelligent control, and simplified architecture, we will build a new computing network with high energy efficiency and low emissions, providing green infrastructure support for high-quality digital economy development.

Chapter V

Typical Deployment Scenarios

5.1 Scenario 1: Transmitting massive sample data to AIDC

Scenario Characteristics: All three stages of large AI model development—pre-training, post-training, and fine-tuning—require transmitting massive volumes of sample data to AIDC. During the pre-training phase, data volumes have reached the PB scale. While the sample data volume per user in the post-training and fine-tuning phases is relatively smaller (typically at GB/TB levels), the aggregate data volume surges significantly as the number of users increases. Hence, MANs must meet the ultra-high-throughput demands of training data delivery scenarios and possess tenant-level slicing capabilities to ensure secure isolation between different tenants.

Solution:

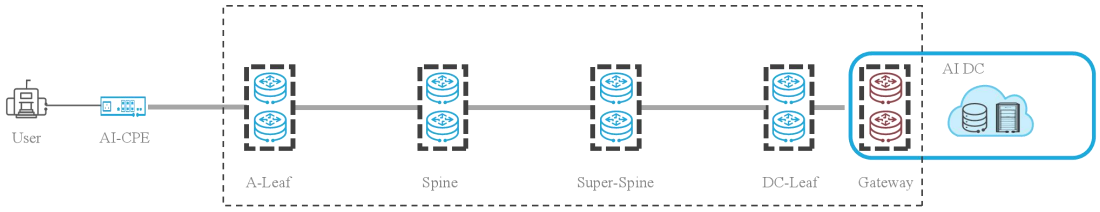

Figure 5-1 : Transmitting massive sample data to AIDC solution

MANs require key technical capabilities like tenant-level slicing isolation and network-level load balancing to support massive sample data delivery scenarios:

- Tenant-level slicing isolation: Isolates training data traffic from regular service traffic using hierarchical slicing technology, effectively preventing resource contention between tenants;
- Network-level load balancing: Implements conflict-free balanced scheduling across all network paths through unified traffic planning, significantly improving network resource utilization.

5.2 Scenario 2: Model training with storage and compute disaggregated

Scenario characteristics: Industries such as finance and healthcare impose extremely high security requirements on private sensitive data. When leasing third-party AIDC for large model training, they demand that private data can not be stored on third-party AIDC. Therefore, in the remote training scenario, sample data storage nodes and AIDC are deployed across wide-area networks. Sample data is pulled on demand for training and immediately discarded after computation, effectively meeting the data security needs of sensitive-data customers. MANs must meet the RDMA lossless transmission requirements of this remote training scenario and possess capabilities such as tenant-level network slicing and data encryption to ensure sample data is not compromised during transmission.

Solution:

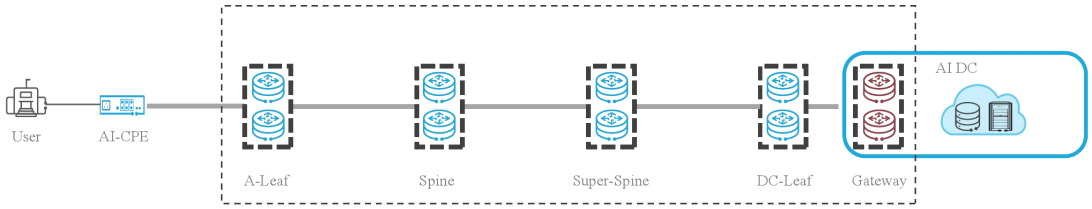

Figure 5-2: model training with storage and compute disaggregated solution

MANs needs to possess key technical capabilities such as tenant-level slice isolation and RDMA wide-area lossless transmission to support the remote training scenario:

- Tenant-level slice isolation: Supports isolating remote training traffic from ordinary service traffic, avoiding traffic throttling during congestion control affecting other services;
- RDMA wide-area lossless: Through flow-level precise flow control, avoids packet loss during the sample pulling synchronized with training, ensuring computational efficiency does not degrade during remote training;
- Elephant flow load balancing: Based on transport layer information, splits traffic, load balances multiple sub-flows of one elephant flow onto different slice paths,

achieving high-throughput transmission;

- Data encryption: Supports end-to-end encrypted data transmission, guaranteeing the security of sample data during transmission.

5.3 Scenario 3: Collaborative model training across multiple AIDCs

Scenario Characteristics: During collaborative large model training across multiple geographically dispersed AIDCs, intermediate data generated in each training iteration (optimizer parameters, gradients, etc.) must be synchronized among all AIDCs before proceeding to the next iteration. This cycle repeats until training completion. Parameter-plane data synchronization relies on the RDMA which is highly sensitive to packet loss, with concurrent data volumes reaching terabytes. Consequently, MANs must deliver ultra-high throughput and lossless transmission capabilities, while incorporating high-convergence networking to balance bandwidth costs with training computational efficiency.

Solution:

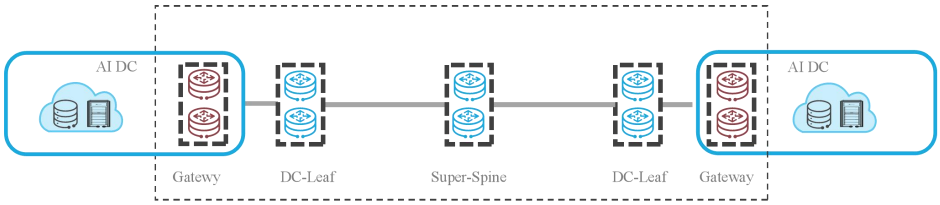

Figure 5-3：Collaborative model training across multiple AIDCs solution

MANs solutions must support the following key technologies:

- Network-level load balancing: Achieves conflict-free inter-path flow scheduling across the entire network through unified traffic planning;
- Lossless RDMA over wide area networks: Prevents packet loss during distributed training via per-flow precise flow control, ensuring no degradation in computational efficiency;
- High-convergence networking: Implements efficient convergence of collaborative training traffic through collective communication algorithms and network optimization, reducing network infrastructure deployment costs.

5.4 Scenario 4: Cloud-Edge collaborative model training/inference

Scenario Characteristics: The local deployment approach of training-inference integrated machines in corporate park struggles to meet enterprises' rapidly growing demands for model fine-tuning and inference. Thus, cloud-edge collaborative training/inference between integrated machines and computing resource pools has emerged as a critical direction for enabling elastic scaling of enterprise computing resources, thereby supporting large model application deployment. Cloud-edge collaboration relies on model partitioning, requiring MANs to support inter-layer parameter plane data synchronization. This necessitates lossless RDMA transmission capabilities with ultra-high throughput.

Solution:

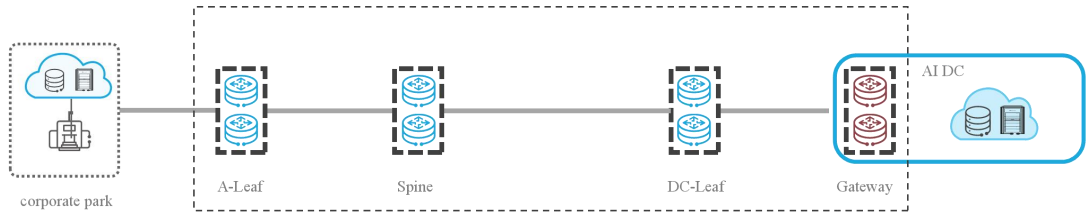

Figure 5-4：Cloud-Edge collaborative model training/inference solution

MANs solutions must support the following key technologies:

- Network-level Load Balancing: Achieves conflict-free balanced scheduling across all paths in the entire network through unified network-wide traffic planning;
- RDMA Wide-Area Lossless Networking: Prevents packet loss during model training via flow-level precise traffic control, ensuring computational efficiency remains undiminished throughout collaborative training and inference processes.

5.5 Scenario 5: Inference delivery

Scenario Characteristics: Pre-trained large models are typically at the gigabyte (GB) scale. During deployment, they need to be distributed from training clusters to

multiple inference clusters. MANs must provide ultra-high throughput to ensure transmission efficiency during distribution, alongside robust security mechanisms to safeguard model integrity. Furthermore, after inference models are deployed to edge nodes, they must rapidly respond to users' high-concurrency, real-time inference requests, necessitating deterministic service capabilities in MANs.

Solution:

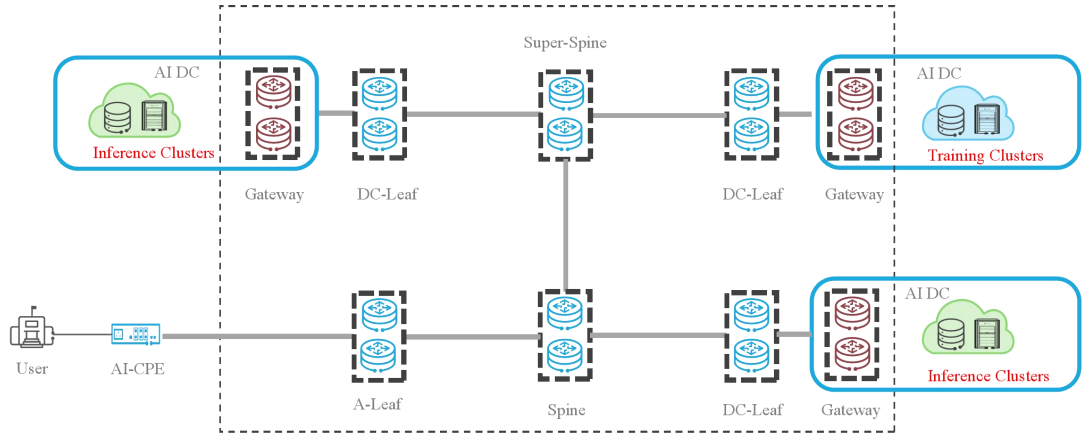

Figure 5-5: Inference delivery solution

MANs solutions must support the following key technologies:

- Network-level Load Balancing: Through network-wide traffic orchestration, it achieves conflict-free balanced scheduling across all paths during inference model deployment;
- Security Encryption: Combined multi-tiered encryption technologies safeguard data security during transmission;
- Deterministic Low Latency: Ensuring optimal user experience during real-time interaction with inference applications.

5.6 Scenario 6:Federated learning

Scenario Characteristics: During the multiple AIDCs' federated learning process, each participant trains models locally using private domain data. Model parameter gradients are exchanged to achieve model parameter aggregation. MANs must provide stable connectivity for devices participating in federated learning while safeguarding data transmission security.

Solution:

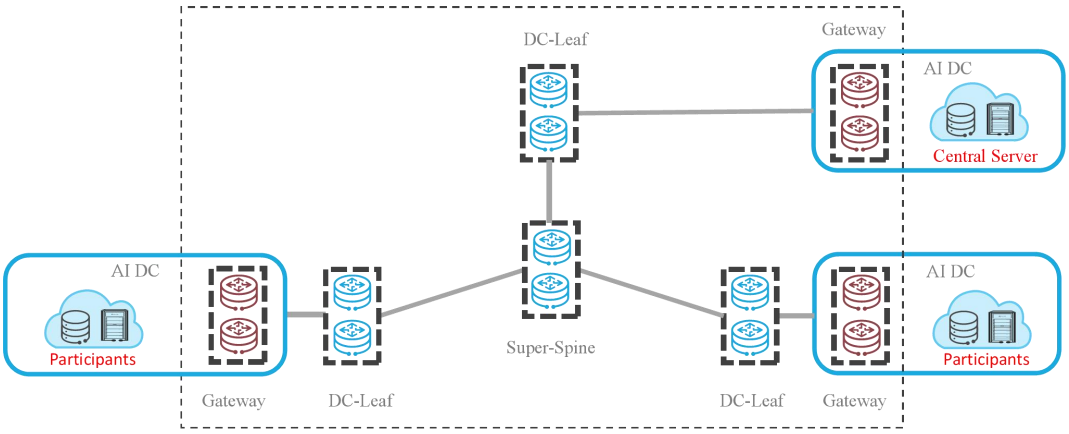

Figure 6-6: Federated learning solution

MANs solutions must support the following key technologies:

- Secure encryption: Through multi-level encryption technology combination, guarantee the security of the data transmission process;
- RDMA Wide-Area Lossless Networking: Prevents packet loss during federated learning via flow-level precise traffic control, ensuring computational efficiency remains undiminished throughout federated learning.

5.7 Scenario 7: Multi-agent system / A2A

Scenario characteristics: Multi-Agent System (MAS) implements real-time interoperability, dynamic task collaboration and secure communication between agents through A2A (Agent-to-Agent) protocol, imposing explicit and strict requirements on network: dynamic task delegation between A2A agents demands low network latency to prevent task chain blocking; A2A needs to handle long-duration tasks (e.g. in-depth research analysis lasting hours to days), requiring maintenance of stable persistent connections; permission isolation (e.g. Agent A can only invoke specific interfaces of Agent B) requires network support for fine-grained access control.

Solution:

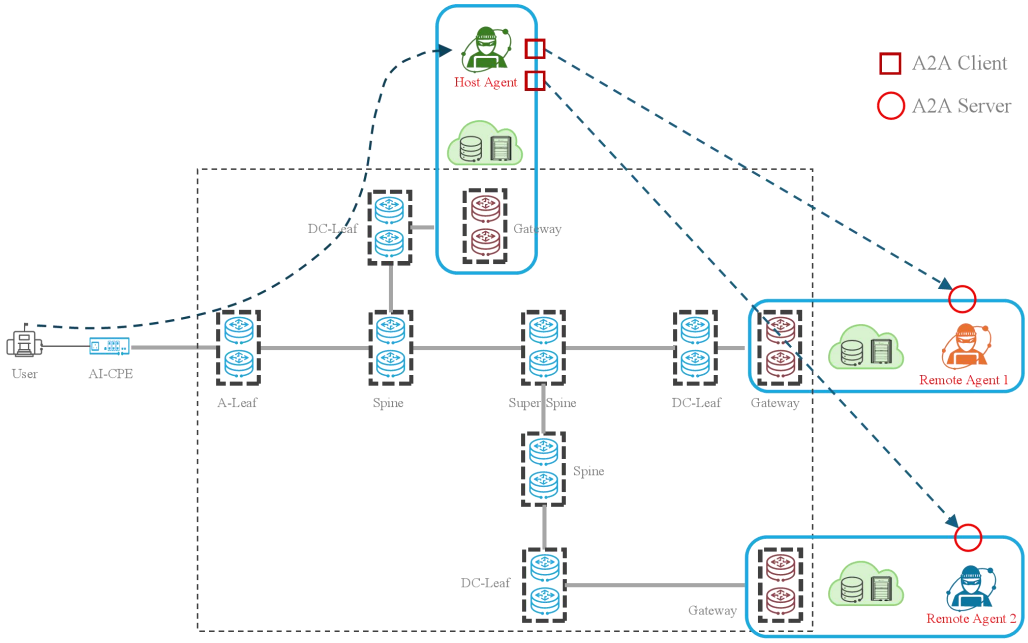

Figure 5-7: Multi-agent system / A2A solution

MANs solutions must support the following key technologies:

- Low-latency path: Metropolitan Area Network constructs millisecond-level low-latency guarantee for task delegation among agents;
- Network reliability: Metropolitan Area Network provides stable and reliable network paths, ensuring long-duration tasks among agents.

Chapter VI

Conclusions and Future Perspectives

This white paper examines the development trends of artificial intelligence and corresponding service requirements, conducting comprehensive research on application scenarios, network architectures, key technologies, and deployment solutions for metropolitan area networks. It actively promotes the evolution of conventional metropolitan area networks into next-generation computing service-oriented metropolitan area network, thereby facilitating technological innovation and practical deployment.

The planning and construction of metropolitan area networks should be driven by both user demands and advancements in computing-network convergence technologies. Through the research and analysis presented in this white paper, we seek to stimulate broader industry participation and discussions. We look forward to collaborating with partners across the ecosystem to develop next-generation metropolitan area networks featuring comprehensive coverage, elastic scalability, lossless wide-area connectivity, ultra-high reliability, and intelligent automation.

媒体合作：contact@nida-alliance.org
工作机会：contact@nida-alliance.org
业务合作：contact@nida-alliance.org
官方网站：www.nida-alliance.org