JUN 2024 | VOL. 26 • NO. 3 • ISSUE 212

ZTE TECHNOLOGIES

VIP Voice

DITO: Driving Digital Transformation in the Philippines

Expert Views

Development Trends of AI Computing Power

Special Topic

Intelligent Computing

Scan for mobile reading

Cover Figure | *Eric Alberto, CEO of DITO in the Philippines*

ZTE TECHNOLOGIES

JUN 2024 | VOL. 26 • NO. 3 • ISSUE 212

Advisory Committee

Director: Liu Jian
Deputy Directors: Sun Fangping, Yu Yifang, Zhang Wanchun, Zhu Yongxing
Advisors: Bai Gang, Fang Hui, Hu Junjie, Hua Xinhai, Kan Jie, Li Weizheng, Liu Mingming, Lu Ping, Tang Xue, Wang Quan, Zhang Weiqing, Zheng Peng

Editorial Board

Director: Lin Xiaodong
Deputy Director: Huang Xinming
Members: Deng Zhifeng, Dai Yanbin, Huang Xinming, Jiang Yonghu, Ke Wen, Kong Jianhua, Liang Dapeng, Liu Shuang, Lin Xiaodong, Ma Xiaosong, Shi Jun, Xia Zejin, Yang Zhaojiang, Zhu Jianjun

Sponsor: ZTE Corporation

Edited By Shenzhen Editorial Office
General Editor: Lin Xiaodong
Deputy General Editor: Huang Xinming
Editor-in-Chief: Liu Yang
Executive Editor-in-Chief: Yue Lihua
Special Topic Editors: Wang Weibin, Shi Yuqing, Feng Kejun
Circulation Manager: Wang Pingping

Editorial Office

Address: NO. 55, Hi-tech Road South, Shenzhen, P.R. China
Postcode: 518075
Website: wwwen.zte.com.cn/en/about/publications
Email: yue.lihua@zte.com.cn

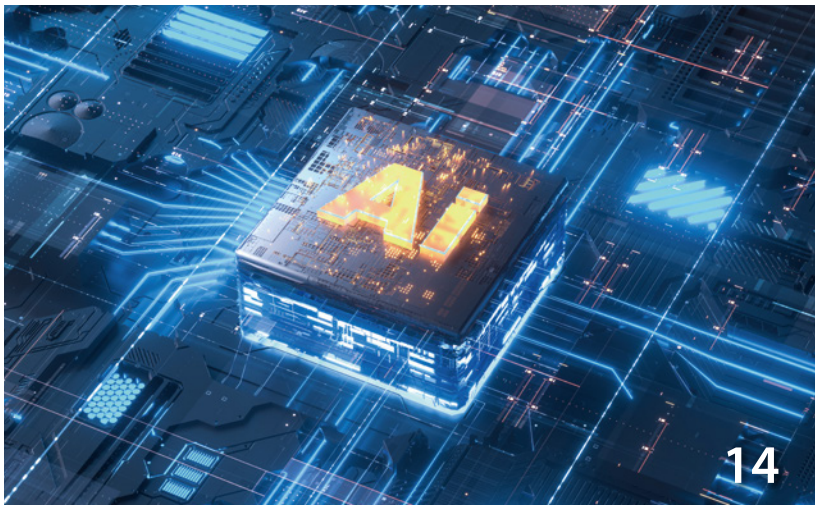
Statement: This magazine is a free publication for you. If you do not want to receive it in the future, you can send the "TD unsubscribe" mail to magazine@zte.com.cn. We will not send you this magazine again after receiving your email. Thank you for your support.

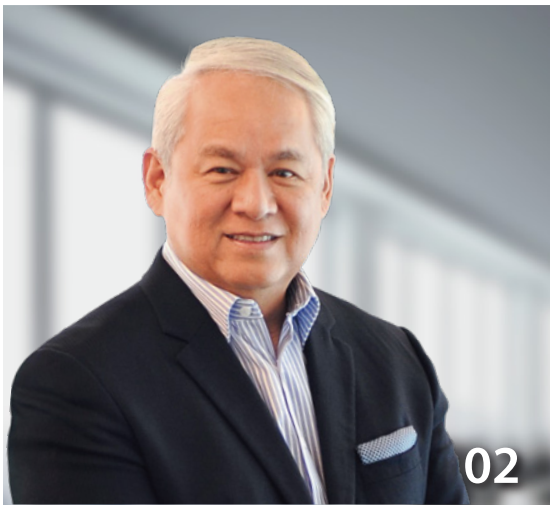
CONTENTS

VIP Voice

Expert Views

Special Topic: Intelligent Computing





Press Clipping

Success Stories

DITO

Driving Digital Transformation in the Philippines

Reporter: Zhao Xi

With the ongoing shift to digital lifestyles in the Philippines, DITO Telecommunity (DITO) is strategically positioned to capture opportunities in the country's digital transformation. Eric Alberto, CEO of DITO, outlines the country's evolving digital landscape, the company's strategies in driving digital transformation, and its future key goals, while emphasizing the importance of developing strong vendor partnerships.

DITO is the fastest-growing telecommunications provider in the Philippines after it was awarded a Certificate of Public Convenience and Necessity by the National Telecommunications Commission in 2019. It aims to improve the Philippines' connectivity with faster and more secure high value 4G and 5G technologies.

As an emerging telecommunications operator in the Philippine market, how does DITO Telecommunity position itself in the rapidly evolving environment? What are the key challenges that you face when planning your strategy?

We are the newest mobile telecom operator in the Philippines. With only two incumbent operators in the country of over 100 million people who are digitally cosmopolitan and savvy, there is a huge opportunity for a third player like us to enter the market, particularly when you look at the evolution of the technology from 3G to 4G and now to 5G. DITO is an exclusively 4G/5G operator, equipped with a brand-new, state-of-the art network. Following the pandemic, a lot of our customers have taken to digital applications and solutions to enable their lifestyles. Therefore, there is much more consumption of mobile data and life-enhancing applications, presenting us with ample opportunities to capture business from our customer constituency.

With the growing digital transformation in the country, what strategies have DITO Telecommunity devised to leverage this trend for a competitive advantage?

DITO has strategically positioned itself at the forefront of the digital transformation wave by implementing several key strategies to gain a competitive edge in the market.

Firstly, DITO's deployment of standalone 4G/5G technology underscores its commitment to providing cutting-edge mobile and fixed solutions. By investing in next-generation infrastructure, DITO ensures high-speed, reliable networks, meeting the evolving demands of consumers and businesses alike. As testament to this, DITO was awarded the Speedtest Award for Top Rated Mobile Network in the Philippines by Ookla at the recent MWC in Barcelona. Prior to this, among the eleven (11) metrics measured by Open Signal, DITO ranked #1 in seven and #2 in three.

Secondly, DITO's decision to forgo legacy network systems and embrace modern, all-digital IT platforms across back-end and front-end operations is instrumental in maintaining agility and scalability. This streamlined approach allows for faster innovation cycles, quicker response times, and a more efficient use of resources, ultimately enhancing DITO's competitiveness in the digital landscape.

Furthermore, DITO leverages Big Data and analytics to guide its network architecture decisions, ensuring optimal performance and personalized service delivery. By harnessing data insights, DITO can proactively anticipate and address customer needs, wherever and whenever they arise. This approach is not only operationally cost-efficient but also customer-centric. It enhances satisfaction and strengthens DITO's market position by staying ahead of emerging trends and evolving customer preferences.

In addition to these strategies, DITO remains committed to ongoing innovation initiatives, fostering partnerships, and cultivating a culture of teamwork.

Source: DITO website

DITO Telecommunity must have a set of standards and expectations when choosing a partner. Could you tell us about what drove your decision to partner with ZTE? What are the main benefits or strengths of this partnership?

Simply put, the most important selection criteria for any vendor are performance and price points. As a startup, we face cost challenges and are prudent in our spending, both in terms of CAPEX and OPEX. We are pleased that from the outset, ZTE understood our strategy for cost management and quality, and supported us with an innovative partnership arrangement. We hope that ZTE will continue this kind of support in the future.

What are the key goals and milestones DITO Telecommunity aims to achieve in the next few years?

DITO Telecommunity has outlined several key goals and milestones to drive its growth and establish itself as a leader in the telecommunications industry over the coming years:

Meet 84% Population Coverage by 2024: DITO aims to fulfill the final network audit requirement mandated by the Philippine government, achieving an impressive 84% population coverage by 2024. This milestone signifies DITO's commitment to providing widespread connectivity and meeting the needs of communities across the country.

Achieve 30% Market Share by 2026: With a strategic focus on market penetration and customer acquisition, DITO aims to secure a rightful market share of 30% by 2026. This ambitious target reflects DITO's determination to compete effectively and gain a significant foothold in the telecommunications market.

Lead in 5G Technology by 2028: As part of its long-term vision, DITO is dedicated to leading the way in 5G technology. By leveraging its infrastructure investments and technological expertise, DITO aims to capitalize on the opportunities presented by 5G networks, paving the way for innovation, and driving new revenue streams by 2025.

Graduate from Start-Up to Financial Stability: In the short and medium term, DITO aims to transition from a start-up organization to achieve financial stability. By focusing on revenue growth, cost optimization, and operational efficiency, DITO seeks to establish a solid financial foundation to support its ambitious growth objectives.

Overall, looking ahead to the medium to long term, DITO aspires to achieve leadership in the digital space. This involves not only maintaining technological excellence but also fostering innovation, cultivating strategic partnerships, and delivering exceptional customer experiences. By becoming a trusted leader in the digital ecosystem, DITO aims to shape the future of telecommunications and drive sustainable growth for years to come. ZTE TECHNOLOGIES

Development Trends of AI Computing Power

Zhu Kun
Chief Engineer of ZTE Cloud Computing Planning

With the advent of ChatGPT, artificial intelligence (AI) has rapidly become a key force for social progress. The extensive application of AI technologies has brought great changes to our lives and work, relying heavily on robust computing infrastructure. AI training tasks and inference applications demand high-performance, large-scale parallelism, and low-latency interconnections, necessitating diverse requirements for computing, storage, and network interconnections. In addition, the demand for AI power aggregation also triggers innovation in the infrastructure management platform (Fig. 1).

AI Chips

In addition to the high-performance matrix operations required for AI model training and inference, larger parameter values in AI models necessitate greater memory capacity. Additionally, extensive data exchange among multiple AI chips demands high bandwidth and low latency in interconnection buses. Therefore, AI chips must meet three major requirements: computing power, memory, and interconnection buses.

In terms of computing power, AI uses machine learning technology based on multi-layer neural networks, requiring extensive matrix operations such as multiplication, convolution, and activation functions. Traditional CPUs have lengthy data flows, allocating more space to control and cache units, with computing units occupying only 25% of the space. Generally, there are just a few dozen arithmetic logic units (ALUs), with efficiency not being high in processing these parallelized and vectorized operations. Graphics processing units (GPUs), however, allocate 90% of the space to computing units, enabling parallel processing of dense data with thousands of ALUs. Since 2017, mainstream AI chip manufacturers have released AI GPUs dedicated to matrix computing acceleration, enhancing computing performance for large-model training. In addition to hardware, GPU manufacturers usually provide the corresponding development platforms like NVIDIA CUDA, allowing developers to directly program and optimize GPUs to fully utilize their computing capabilities.

In terms of memory, transformer model parameters increase by an average factor of 240 every two years, while AI memory capacity only doubles every two years, failing to keep pace with model growth. To address this, a feasible solution is the use of super nodes with unified memory addressing. For example, by customizing an AI server

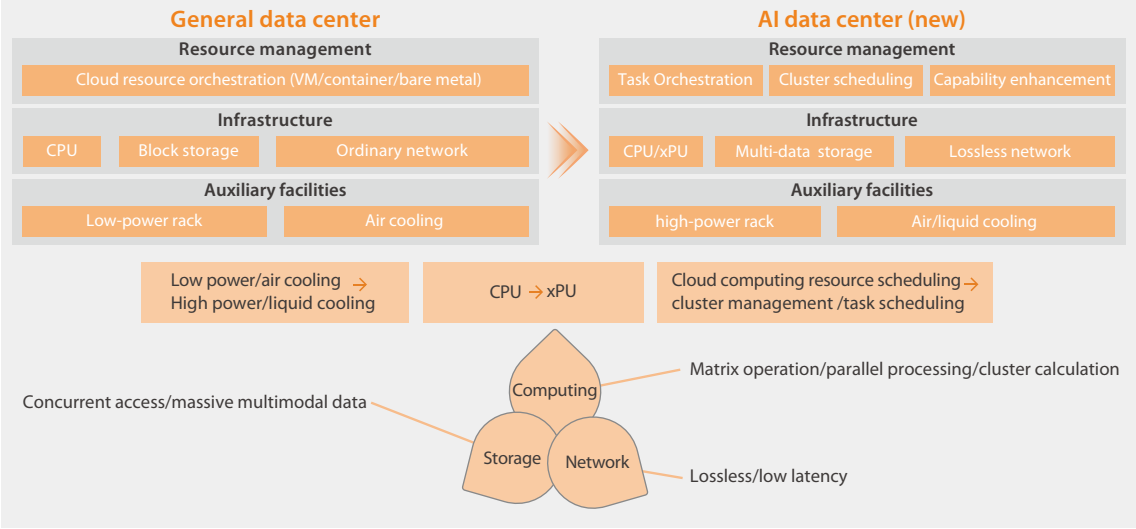


Fig. 1. Innovative architecture of the AI data center.

and forming a super node (including 256 GPUs and 256 CPUs) through high-speed interconnection technology, memory capacity can increase by 230 times. In addition, AI chips use the von-Neumann architecture where computing and storage are separated, leading to significant energy consumption (60% to 90%) during data migration. Estimating based on 60% of the maximum power consumption of H800 (700W), data migration consumes 420W. To solve this problem, the memory and computing integration technology fully integrates memory and computing, avoiding data migration and greatly improving energy efficiency.

Regarding interconnection buses, after 3D parallel splitting of AI models, data transmission between chips becomes essential. Tensor parallel (TP) transmission dominates transmission time, exceeding 90%. Test data shows that using the same number of servers to train GPT-3, compared to PCIe, reduces the transmission time of one micro-batch between adjacent GPUs from 246.1 ms to 78.7 ms, and overall training time from 40.6 days to 22.8 days. Therefore, the bandwidth of the interconnected bus becomes crucial.

Storage for AI Computing

At various stages of end-to-end development for AI model, innovation requirements have been proposed for storage, including:

- **Multi-data storage:** Multimodal datasets such as video, image, and audio bring requirements for diverse storage formats such as blocks, files, objects, and big data, as well as for protocol interworking.
- **Massive storage:** To ensure the precision of AI model training, the dataset is usually 2–3 times the size of the parameter values. In the current era of rapid development for AI models, expanding from 100 billion to one trillion, storage capacity serves as a crucial indicator.
- **High concurrent performance:** In the AI parallel training scenario, multiple training nodes need to read datasets simultaneously. During the training process, each training node needs to periodically save checkpoint to ensure system resilience for breakpoint training. The high performance of these read/write operations can greatly improve the efficiency of AI model training.

Therefore, for AI computing storage, it is necessary to provide multi-data storage capability and multi-protocol interworking capability for block (iSCSI), file (NAS), object (S3), and big data (HDFS). Performance can be improved through comprehensive software and hardware optimization. Hardware acceleration methods involve offloading storage interface protocols via DPUs, and performing deduplication, compression, and security operations, as well as automatic data tiering and partitioning based on popularity. Software optimization methods include distributed caching, parallel file access systems, and

private clients. In addition, NFS over RDMA and GPUs can greatly reduce data access latency.

Lossless Network

The parallel computing nature of AI model training brings a large amount of communication overhead, making the network a key factor that restricts training efficiency. Lossless network is thus essential, requiring zero packet loss, high throughput, large bandwidth, stable low latency, and ultra-large-scale networking.

Currently, lossless network protocols are divided into IB and RoCE. The IB network, originally designed for high-performance computing (HPC), boasts low latency, high bandwidth, SDN topology management, rich networking topologies, and high forwarding efficiency. However, its industrial chain remains closed. RoCE, designed for a unified transport network, offers high bandwidth and network flexibility. It provides strong support for cloud-based services and promotes ecological openness. Therefore, RoCE is essential for localization. Nevertheless, its network performance and technology maturity lag behind IB, and latency requires further optimization based on chips.

Traditional network congestion and traffic control algorithms operate independently on both the client side and the network side. The network provides only coarse-grained congestion marking information, making it challenging to prevent congestion, packet loss, and queuing delays in high-throughput, full-load scenarios. Therefore, it is necessary to implement accurate and fast congestion control and traffic scheduling algorithms through client-network coordination to further enhance network performance.

In network topology, Fat-Tree CLOS and Torus rail multi-plane topologies are two mainstream solutions for addressing network congestion issues in network design. The Fat-Tree CLOS network enhances traditional tree network by maintaining a low convergence ratio of 1:1 for uplink bandwidth to downlink bandwidth, ensuring no blocking path between any two nodes. The Torus rail multi-plane network connects GPUs at the same location on different servers to the same group of switches, forming a rail plane. GPUs at different server locations are connected to different switches, creating multiple rail planes.

Resource Task Scheduling Platform

Unlike general computing resource management platforms that distribute resources to multiple tenants using virtual cloud technology, intelligent computing scenarios concentrate on aggregating computing power. In AI task training, hundreds of

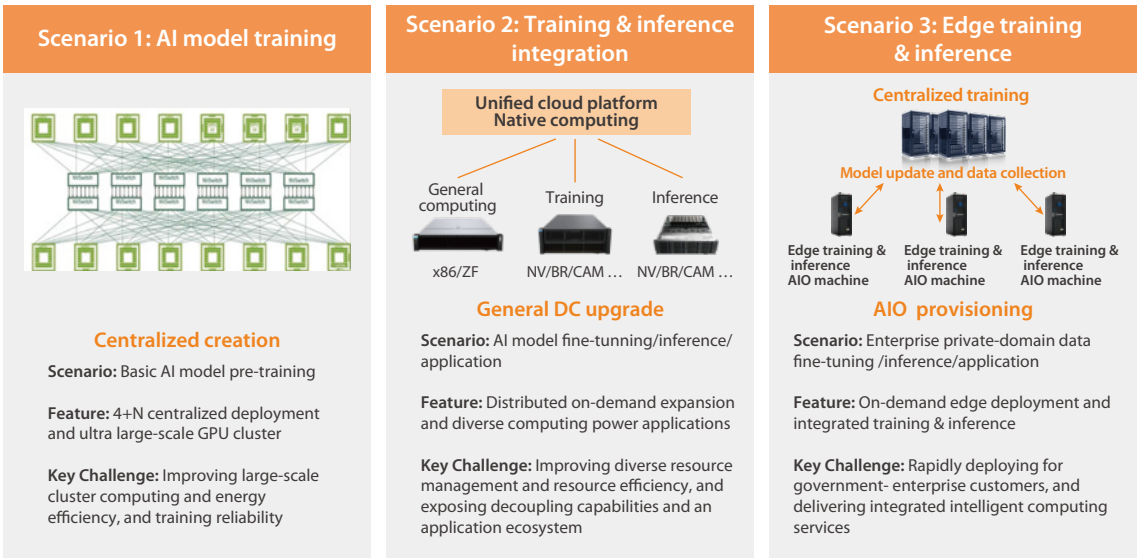


Fig. 2. Hierarchical deployment of the intelligent computing center.

tasks and thousands of nodes may run simultaneously. Utilizing the task scheduling platform optimally matches tasks with available resources, minimizing queue wait time, maximizing parallel task operations, and achieving optimal resource utilization. Currently, there are two mainstream scheduling systems: Slurm and Kubernetes.

Slurm, primarily for task scheduling in HPC scenarios, is widely used by supercomputers (including Tianhe Computer) and computer clusters around the world. Kubernetes, a container orchestration platform, is used to schedule, automatically deploy, manage, and extend containerized applications. At present, Kubernetes and wider container ecosystems are increasingly mature, shaping a general computing platform and ecosystem.

In AI task scheduling, Slurm and Kubernetes face different challenges. The deep learning workload shares similarities with HPC, making Slurm suitable for managing machine learning clusters. However, Slurm is not part of the machine learning ecosystem developed around containers, so it is difficult to integrate AI platforms like Kubeflow into such environments. In addition, Slurm is complex to use and maintain. Conversely, Kubernetes is easier to use, integrates well with common machine learning frameworks, and sees increasing adoption for big model training. Yet, scheduling GPUs with Kubernetes may lead to prolonged resource idle time, resulting in low average cluster usage (about 20%). Resources can only be scheduled by card, lacking the ability to split, schedule by card type, or queue them.

Deployment Scenario

Due to the varying requirements of computing characteristics and deployment locations for pre-training basic AI models, fine-tuning industry AI models, and adapting AI models to customer scenarios, the intelligent computing center adopts a three-level deployment model (Fig. 2). This includes the Hub AI model training center, provincial training and inference integration resource pool, and edge training and inference AIO, aligned with the hierarchical architecture of operators' computing data centers.

Operators bear the responsibility of improving innovation in key software and hardware technologies and building intelligent computing infrastructure. ZTE offers a full range of products spanning from IDCs, chips, servers, storage, data communications, to resource management platforms. Leveraging its extensive experience in the telecom and government-enterprise sectors, ZTE is poised to assist operators in realizing their ambitions in intelligent computing technology innovation and development. ZTE TECHNOLOGIES

Insights into AI Agent Technology in 2024

Du Yongsheng

Chief Architect of ZTE AIM/Wireless UME AI Model Products

Gao Yanqin

General Planning Engineer of ZTE AIM/Wireless UME AI Model Products

Less than three months after the release of ChatGPT 3.5 in early 2023, the paper on large language model (LLM)-powered autonomous agents was published, igniting immediate interest in AI agent technology. GPTs or customized versions of ChatGPT were launched at the OpenAI Developer Conference in November 2023. At present, agents have become a mainstream product of AI models.

Compared to general-purpose AI model products, agents, whether role-playing or task-specific, are technically easier to control. They offer more accurate outputs and are easier for users to understand and adjust. We will conduct an agent review to guide our follow-up research and work direction.

Current Technologies and Principles of Agents

From its initial structure-based definition to its current multi-modal model, the concept of an agent has gone through the phases of tools, social agents, workflows, multi-agent cooperation, and multiple modalities. Currently, an agent is defined as a virtual role based on the AI model with the ability to learn, remember, perceive the environment, recall past experiences, plan target tasks and execute them to influence the environment.

Each part of the agent, as shown in Fig. 1, is described as follows:

- **Thinking and skills:** The agent receives task objectives from users, plans tasks using the LLM, and maps sub-tasks to the corresponding skills. Skills include indirect instruction communication for other agents' production

tasks and instructions directly executed in the production environment. Task decomposition can occur after overall planning or through iterative decomposition.

- **Environment and perception:** The analysis and execution of a task depend on the context of its environment. To understand the environment, a kind of modeling mode is required first to convert real-world information into a machine-readable language. For example, Metaverse is an environment modeling mode for the physical world, while the digital twin solution in the communication industry is an environment modeling mode for the communication network.
- **Memory and learning:** The agent learns from other agents and environment feedback through imitation or reinforcement learning, and stores these learning outcomes into memory, enabling them to resolve similar problems in the future. The ability to learn and adjust to changing environments plays a vital role in the self-evolution of agents.

Agent Value Analysis

First, the current agent's value mainly relies on LLMs, essentially a conditional probability generation model. Utilizing different prompts, such as text generation, task disassembly, logical inference, and scenario understanding, LLMs generate different types of outputs. Based on the output capability of LLMs, agents build humanized roles to serve within the production field.

Second, from a mainstream industry perspective, the value of agents is embodied by a composite expert team composed of both human and multiple virtual members. This setup massively improves the scope of work, enabling one person to do the work traditionally requiring multiple persons. The paradigm has shifted from using tools to orchestrating multiple agents, which then use tools to complete tasks. Compared with conventional tools, LLM-based agents can provide greater generality and flexibility in judgment and decision-making.

Experimental Results of ZTE

At present, ZTE has developed four types of agents based on its understanding of of LLMs and the communications industry. These include assurance assistant, intelligent Q&A, fault assistant,

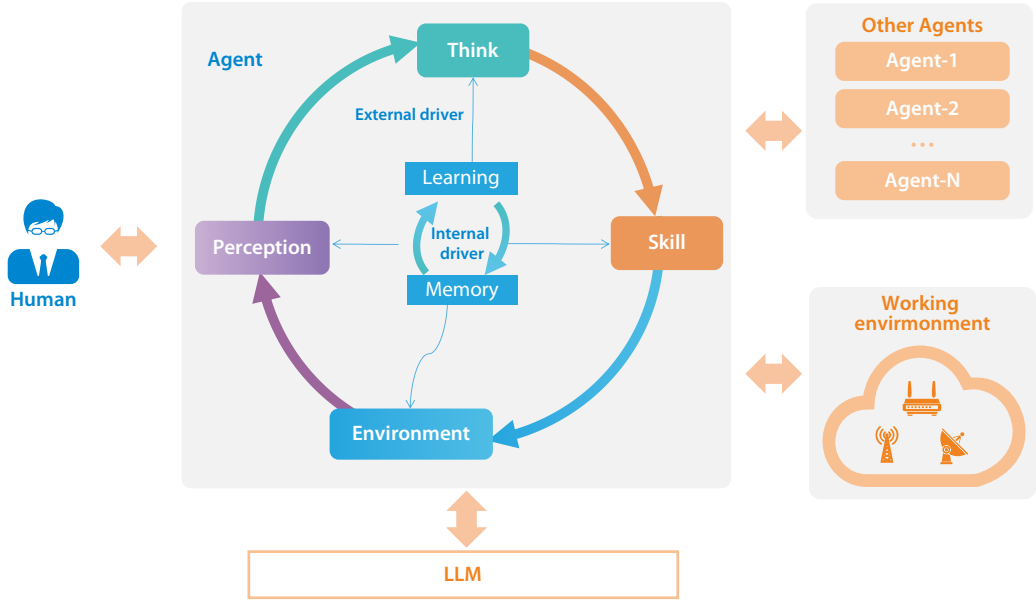


Fig. 1. Agent deployment and definition.

and network observation assistant.

The assurance assistant, utilized in major activity support scenarios, has a high degree of automation. It replicates real workflows to a virtual space, where key support experts, assistants, and troubleshooting experts work together to automatically complete workflows. They communicate with people through summarization, reporting and risk assessment. This is a complex type of job agent, developed with the aim of achieving L5 fully autonomous network.

The other three agents are technically task agents:

- **Intelligent Q&A:** This agent is to build a ToB-oriented knowledge base application based on the RAG+agent technology.
- **Fault assistant:** Facilitated by the fault knowledge bases and APIs, this agent assists O&M personnel in quickly troubleshooting faults.
- **Network observation assistant:** Utilizing both large and small models, multiple agents perform network analysis across various dimensions. They then send their findings to the general network insight agent, which summarizes them and outputs network observation results.

Agent Development Trends and Technology Breakdown

At present, the mainstream types of agents in the academic community are consistent with the experimental results of ZTE. They are as follows:

- **Logic agent:** This kind of agent generates languages and multi-modal outputs based on its understanding of input languages and multi-modal data.
- **Task agent:** Designed for specific tasks, this agent breaks down plans and performs corresponding operations. It lacks long-term memory during the process.
- **Job agent:** Oriented to abstract work responsibilities and overarching objectives, this agent perceives the environment, remembers process status, and generates sub-objectives to advance the work.

From the perspective of development trends, self-evolving agents are also important, as they can self-learn.

Mainstream agent products are categorized according to their technical level, as shown in Table 1.

We conduct further paper scanning and research on the technologies mentioned in Table 1, and find the following:

- **Technology maturity analysis:** Despite many papers on the technologies marked with underlines in the table, there is a lack of mature solution in industrial environments.
- **Technical problems analysis:** Environment modeling and self-learning technologies pose the most significant challenges to solve. Despite being put forward early on, there are no good real-world solutions in physical production. In addition, their association with AI models is weak, and advancements in AI models have little impact on these technologies.
- **Technical potential analysis:** Self-adaptive organization, exploration, intelligent prompt words, memory, and dialogue have potential for further development, which may be crucial for creating the gap between agent levels in the short term.

Table 1. Mainstream agent technology breakdown. ▶

	Logic Agent	Task Agent	Job Agent	Self-Evolving Agent
Environmental Perception	1. Text 2. Role profile	1. Text 2. Role profile	1. Text 2. Mature role profile 3. Multi-mode	1. Text 2. Mature role profile 3. Multi-mode 4. Self-execution process perception
Environmental Understanding	Context	Long short-term memory	1. Long short-term memory 2. Environment modeling	1. Long short-term memory 2. Environment modeling 3. Valid abstract memory
Problems Thinking	Thinking chain technology	Thinking chain technology	Intelligent prompt words	Intelligent prompt words
Solution Model	None	1. Service 2. Small model 3. SQL	1. Service 2. Small model 3. SQL 4. Prompt driving other expert agents	1. Service 2. Small model 3. SQL 4. Prompt driving other experts 5. Code generation and automatic exploration
Feedback Learning	Search + RAG	Search + RAG	1. Search + RAG 2. Imitation learning	1. Search + RAG 2. Imitation learning 3. Reinforcement learning
Multi-Agent Coordination	None	Add	1. Add 2. Negotiation 3. Inspiration	1. Add 2. Negotiation 3. Inspiration 4. Inspiration on valid exploration
Multi-Agent Coordination Model	None	Human controlling agent	Agent controlling agent	Adaptive organization
Product Classification at Home and Abroad	RAG repository, voice	1. Automation 2. General assistant 3. Development 4. Software and hardware combination	1. Digit staff 2. Workflow	Priority given to experimentation and study

> In the upcoming year, we expect a rapid increase in the number of ordinary agents, and the phenomenon of group intelligence may emerge before that of powerful agents.

- **Development trend analysis:** Based on the above analysis, task agents involve only one less mature technology. Job agents involve five immature technologies, including one difficult technology, environment modeling. Key technologies involved in self-evolving agents are all difficult. Therefore, task agents may experience the fastest development speed and offer the highest value.
- **Analysis of current products:** Major products both domestically and internationally focus on the task agent type.

Insights About Agent Trends

Based on the above analysis, we can draw the following conclusions:

- At the current stage, the product focuses on simple task agents, which utilizes mature technologies and can be easily replicated and promoted. This aligns with our product experiments, and we anticipate a rapid increase in the number of such agents.
- Under the conditions described above, the technologies that widen the gap between agents are memory and dialog.
- Developing a powerful individual agent is difficult because it involves enhanced learning and environment modeling technologies. This is consistent with the input cost and outcomes observed in our experiments with environment modeling.

Leveraging AI models, a simple task agent can provide inspiring information for other agents. If a certain number of task agents can be reached, one of the two necessary conditions for the emergence of group intelligence can be met. Second, with the abstract summarization capability of AI models, an agent within a team can combine multiple highly correlated information fragments from different agents, fulfilling another necessary condition of group intelligence. Once these two necessary conditions are met, the phenomenon of group intelligence may start to emerge.

To sum up, through academic tracing, product experimentation, and technical decomposition of different types of agents, we have derived an insight: in the upcoming year, we expect a rapid increase in the number of ordinary agents, and the phenomenon of group intelligence may emerge before that of powerful agents.

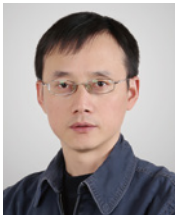
Building on this insight, we need to further consider the following aspects:

- Develop framework technologies with low learning costs and low technological thresholds to quickly generate agents; establish multi-agent collaboration and intelligent control centers to manage the emergence of group intelligence.
- Track the key capabilities driving agent evolution, including environment modeling, memory, learning and design, and thoroughly explores the potential of memory.
- Develop agent products related to enterprise data analysis and SOP workflows to bring benefits to enterprises. ZTE TECHNOLOGIES

ZTE AI Full-Stack Intelligent Computing Solution: Empowering Various Industries

Wang Weibin

Chief Scientist of ZTE Product Planning



Lu Guanghui

Chief Architect of ZTE CCN Product

To date, AI has experienced three major shifts and two downturns. In November 2022, OpenAI released ChatGPT and its generative AI technology, employing transformer algorithms and pre-trained AI models. This propelled the third wave of AI development to unprecedented heights, marking a turning point and peak of excitement in AI models.

The generative AI technology can create new content, imitate human creativity and innovation, and play an important role in numerous fields, driving the prosperity and advancement of AI. Large-scale miracles are being achieved as bigger models bring greater intelligence. As AI technologies evolve, various industries will use AI to enhance operational efficiency, create business value, and move from a digital realm to a digital and intelligent world.

Faced with opportunities and challenges in generative AI technology, ZTE steadfastly maintains its role as a driver of the digital and intelligent economy, striving to be the ultimate AI company and serve as a leading AI enterprise model. Additionally, ZTE is committed to helping industries build end-to-end intelligent computing infrastructure and digital transformation solutions. Leveraging its universal computing solution, ZTE has launched the Nebula intelligent computing solution, guided by openness, efficiency, intelligence, and security concepts, and oriented towards training and inference scenarios. The solution spans intelligent infrastructure, AI platforms, AI models, and applications, empowering operators to build intelligent computing centers and drive digital intelligent transformation across industries (Fig. 1).

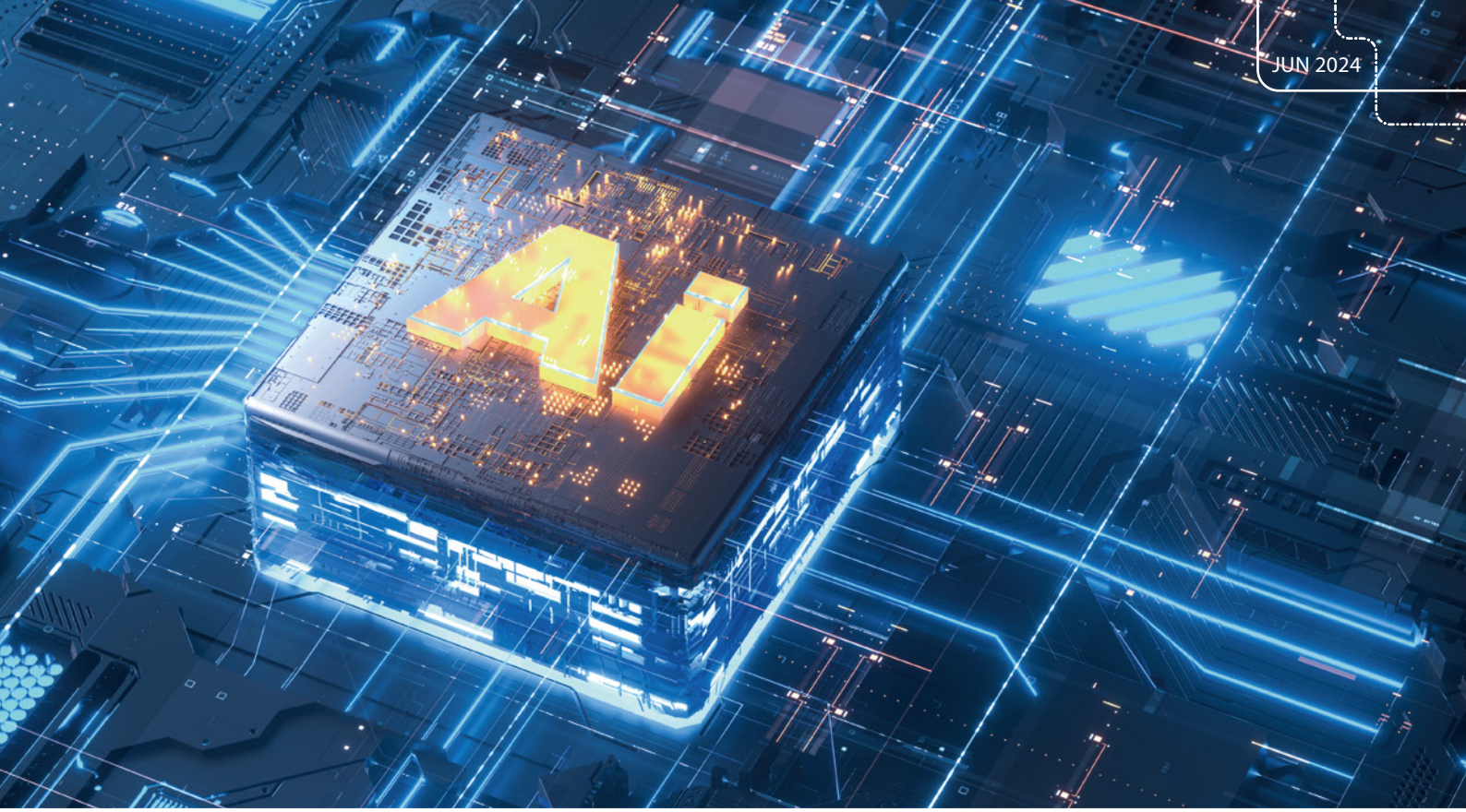
Intelligent Computing Infrastructure: Efficient and Secure

ZTE intelligent computing infrastructure layer, including IDCs, AI computing, integrated storage, lossless network, and resource management platform, supports the construction of diverse and multi-layered intelligent computing infrastructure. It ranges from AI model training & intelligent computing centers, hybrid training & intelligent computing centers, to edge training integrated computers. Each layer caters to specific performance, cost, and service needs in different scenarios. This multi-layer design enhances adaptability, flexibility, and user choice.

Efficiency Is the Priority

The cost of single AI model training is high, so an efficient intelligent computing infrastructure is required. ZTE builds such infrastructure through hardware, resource management, and product solutions.

In hardware selection, processors with high



computing power, large video memory, high-speed interconnection and high-performance concurrent multiple storage are chosen to improve system parallel operation rates, thereby boosting cluster computing power. Additionally, the independently developed DPU smart NIC provides a lossless network with ultra-large bandwidth and ultra-low latency, enhancing overall reliability and efficiency.

Regarding the resource management platform, multiple heterogeneous hardware devices are connected to meet efficient resource management needs for AI model training and inference. ZTE's AI resource management platform, TECS, provides job scheduling and intelligent computing cluster management, including computing enhancement (such as vGPU technology), storage enhancement (such as supporting high-performance file storage), network enhancement (such as supporting integrated communication technology), and cluster management scheduling.

TECS is an enhanced product based on the original self-developed general computing resource management product, tailored for AI model training and inference requirements. It can function separately from the original product and integrate seamlessly to achieve unified management and orchestration of general and intelligent computing, or can be independently deployed to manage and orchestrate intelligent computing resources.

In product solution, ZTE has launched an all-in-one (AIO) out-of-the-box training machine to accurately meet the requirements of industry secondary training and real-time inference service scenarios, as shown in Fig. 2. This all-in-one machine integrates computing, storage, network devices, and AI platform software, supporting mainstream AI frameworks. It minimizes costs for training and inference of private domain models while lowering technical thresholds. This means that users do not need complex deployment and configuration procedures, and can be put into operation quickly, achieving flexible allocation of training and inference resources.

Security Is the Basis

Among the three basic elements of AI—computing power, algorithms, and data, computing power is the core element and primary driver for the comprehensive advancement and rapid application of AI systems. Therefore, it is critical to provide secure and reliable computing power. ZTE focuses on developing intelligent computing power, striving to establish multi-channel supply chains both domestically and internationally, specifically tailored to AI model training and inference scenarios. ZTE provides a complete set of mature solutions based on high-performance AI servers and IB switches sourced

from leading GPU manufacturers worldwide. Additionally, through extensive collaboration with domestic head-end GPU manufacturers, ZTE has undertaken significant self-development efforts, providing diverse end-to-end intelligent computing solutions. These include high-performance AI servers utilizing chips from the leading GPU manufacturers, box-type and frame-type RoCE switches, and distributed storage servers supporting high-performance, multi-dimensional storage (such as files, objects, blocks, and big data).

In addition, AI model training with tens of billions of parameters is time-consuming due to large training data. To ensure stability and reliability, and avoid interruptions caused by hardware faults, ZTE's TECS resource management platform offers a secure visual management tool for automatic monitoring. It also provides breakpoint training renewal service, minimizing interruption time and greatly reducing losses.

AI Platform: Openness and Decoupling

ZTE AI platform, centered on openness and decoupling, offers complete AI products. It provides a unified programming environment and tool chain, minimizing model development and migration costs while facilitating ecosystem construction.

To help developers and users better develop, train, evaluate, implement, and update AI models, ZTE provides a component-based AI platform—AI studio (AIS). Integrated into the intelligent AIO computing system or AI application, the AIS covers workflows such as data collection, data annotation, model training, model fine-tuning, knowledge base management, compilation optimization, and inference deployment. Supporting PyTorch and other mainstream AI frameworks, it delivers end-to-end intelligent computing center solutions, model capabilities and operational engines for AI applications.

AI Models and Applications: From Universal to Dedicated

ZTE has articulated its strategy for empowering enterprises' digital transformation with large-scale models as "1+N+X", transitioning from "universal" to "dedicated".

"1" Base AI Model Series

Using its engineering capabilities, ZTE independently develops the Nebula base AI models series, including NLP and multi-mode models. Leveraging vast training data and using unsupervised or self-supervised learning methods, ZTE demonstrates exceptional understanding and expression capabilities in different tasks and domains.

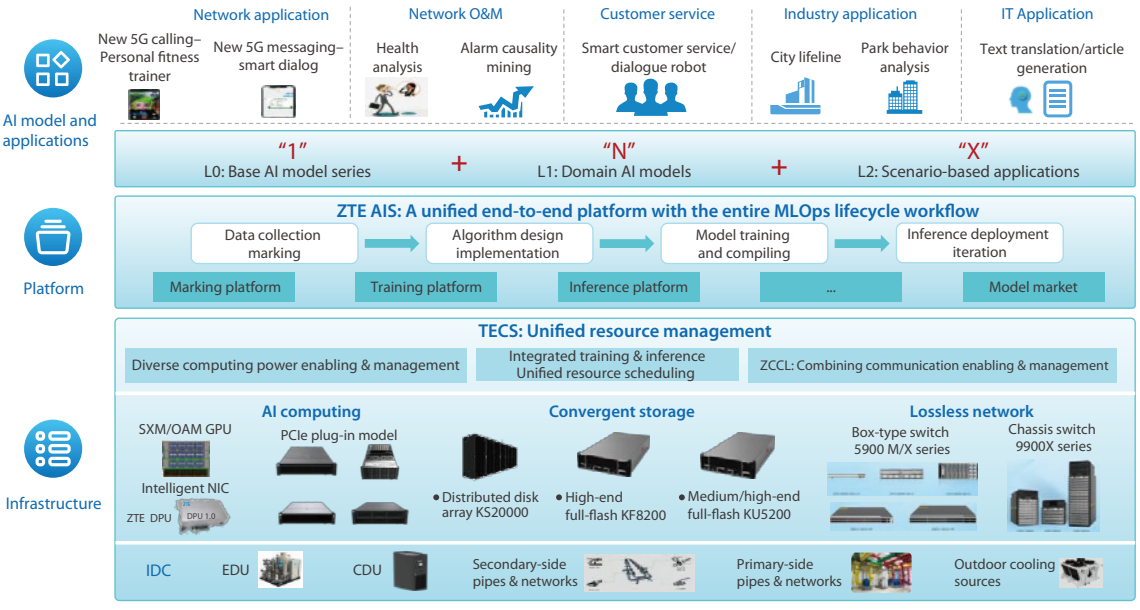


Fig. 1. ZTE Nebula intelligent computing solution.

Fig. 2. ZTE AiCube X5000 solution.

"N" Domain Models

Based on the base AI model, the domain AI model improves professional capabilities through incremental pre-training on the domain Know-How. In the R&D field, since 2022, ZTE has utilized AI model technologies to enhance R&D efficiency, assisting R&D personnel in requirement analysis, product design, programming, testing, version release, and documentation. At present, ZTE's coding AI models rank among the top assessed by HumanEval, boasting industry-leading capabilities in diverse coding languages and Chinese proficiency. In the telecom realm, ZTE leverages vast amounts of high-quality network O&M and service operation data to enrich AI telecom models, surpassing others in telecom knowledge. These AI models support multi-mode data in the communication field, addressing complex issues such as coverage, capacity, performance reports, and network presentation. Moreover, ZTE's telecom AI models incorporate a robust intent engine integrated with autonomous networks, enhancing operators' network operation efficiency through efficient workflow cascades.

"X" AI Scenario-Based Applications

ZTE has developed various subdivision applications based on domain AI models such as computer vision-based AI models. These include the urban lifeline solution for comprehensive urban security risk detection and early warning across critical infrastructure like water, power, gas, heat, and transportation. Additionally, ZTE offers a one-stop AI development assistant covering the entire R&D process, built upon the AI coding model. Utilizing the AI network model, ZTE has developed a range of O&M tools, such as fault O&M robots, tailored to support different scenarios. Moreover, ZTE has created the SMS anti-fraud service application based on the AI language model.

Rich Applications: Empowering Customers' Digital Intelligent Transformation

To help operators and partners build end-to-end intelligent computing infrastructure and digital transformation solutions for enterprises, ZTE launches the open and decoupled Nebula intelligent computing solution. This solution provides AI full-stack products and is deployed across various sectors including intelligent computing centers, R&D efficiency improvement, communications, anti-fraud governance, and urban governance. In the communications sector, ZTE released the industry's first AI-model-based SMS anti-fraud governance system in 2023. In industry domains, ZTE cooperates with numerous partners, signing strategic agreements and implementing multiple projects spanning machine vision, industrial production, and more. As a cornerstone technology of digital transformation, AI models play a pivotal role in the evolution and commercial prosperity of numerous industries in the new era. ZTE is fully poised to seize this significant opportunity with its partners, ensuring that AI can catalyze benefits across thousands of industries. ZTE TECHNOLOGIES

ZTE's GPU Server Solution: Driving Digital Economy

Zhou Zanxin

Chief Engineer of ZTE Server and Storage Products

The AI field is undergoing a new round of rapid development, and the demand for generative AI computing power has skyrocketed. This trend is poised to become a new growth point and accelerator in the AI computing market.

In 2023, China's GPU server market continued its rapid growth. According to IDC, the accelerated server market in China reached US$9.4 billion in 2023, an increase of 104% over the previous year, with shipments totaling 326k units. GPU-accelerated servers accounted for 92% of this market, reaching US$8.7 billion. IDC forecasts that by 2028, China's accelerated server market will reach US$12.4 billion.

Requirements of AI Applications for GPU Servers

Compared with general servers, GPU servers offer several distinctive features:

- **High-performance CPUs:** A large number of computing resources are required for AI training and inference, necessitating high-performance CPUs to meet the processing requirements of large datasets.
- **GPU accelerator cards:** Compared with CPUs, GPUs excel in parallel computing , enabling them to accelerate the training and inference for deep learning models. A PCIe GPU can meet the requirements of most small and medium model training and inference applications. A single server usually supports four to eight GPU cards for parallel processing, enhancing computational performance and efficiency.
- **Large-capacity memory:** Sufficient capacity memory accelerates data flow and algorithm processing speed.
- **High-bandwidth network interface:** A high-speed network bandwidth (100GE or above) is required to transmit a large amount of data during the training process.

The rise of AI models brings force higher requirements for GPU servers. In particular, a large-scale model requires a huge amount of computing power to train, exceeding the capabilities of a single GPU. In this case, a single-server multi-card setup or multi-server clusters are needed to implement parallel training techniques, including tensor parallelism (TP), data parallelism (DP) and pipeline parallelism (PP). The specialized requirements of large models for GPU servers include:

- **High-performance GPUs with large memory:** A large model requires massive parallel computing capability, and a large number of parameters and gradient information need to be stored. Therefore, high-performance GPUs with large memory are required for training and inference.
- **High-speed interconnection of GPUs within the server:** The single-server multi-card setup utilizing the TP technique has exceptionally high requirements for the communication bandwidth between multiple GPUs within the server. An SXM/OAM GPU accelerator card supporting high-speed interconnection channels is required to facilitate high-speed interconnection among eight GPUs within a server, accelerating data transmission and model synchronization.
- **High-performance interconnection network between servers:** In multi-server clusters, the

inter-machine parameter plane interconnection network needs to utilize a high-speed multi-track traffic aggregation architecture to give full play to the computing resources of GPU clusters. On the one hand, PCIe 5.0 slots are required to support 200/400G high-performance and low-latency IB/RoCE NICs. On the other hand, at least 10 NIC slots are required, with at least two NICs dedicated to the management and storage plane. GPUs and parameter plane NICs are configured in a 1:1 ratio to ensure that the parameter plane NICs connected to GPU cards in the same position across multiple GPU servers belong to the same switch, optimizing communication efficiency and accelerating parallel transmission.

- **High-speed memory and storage:** During the training of large AI models, rapid data read and write operations are crucial. It is necessary to support high-speed components such as DDR5 memory and NVMe SSD to enhance data transmission speed and reduce latency, thus improving the training efficiency.
- **Liquid cooling:** The ultra-high computing power density of SXM/OAM GPUs causes the power consumption of GPU servers to increase dramatically. Air cooling solutions restrict the computing power density of intelligent computing data centers, and fail to meet energy-saving and consumption reduction requirements. Liquid cooling becomes necessary.

Given the special requirements of AI model training and inference on GPU servers, dedicated GPU servers needs to be designed to support high-speed intra-sever and inter-sever networking. These servers should be appropriately configured and optimized to continuously adapt to new challenges and requirements.

ZTE GPU Server "3+2+3" Solution

To cope with the rapid development of AI, ZTE has launched the "3+2+3" GPU server solution, meeting the full-scenario AI application requirements of various customers (see Fig. 1).

Based on Three Major CPU Platforms

Tailored to different customer needs, ZTE has

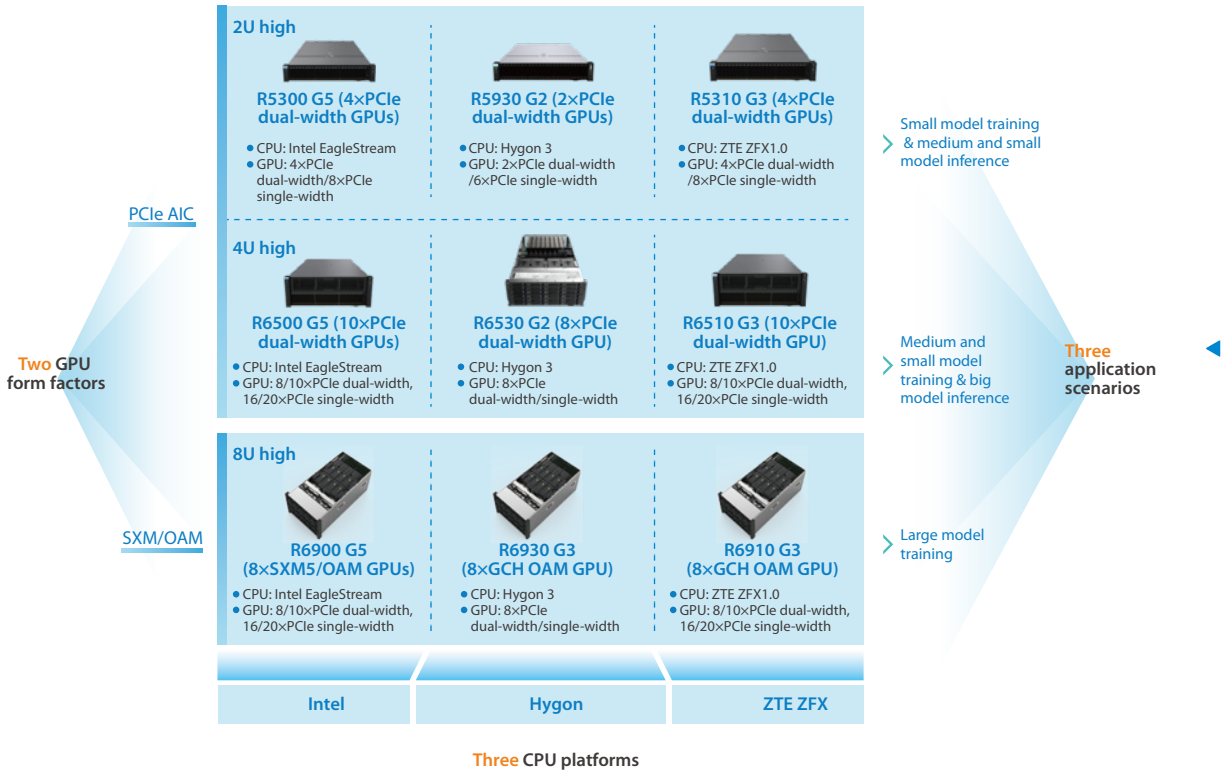


Fig. 1. ZTE GPU server "3+2+3" solution.

launched different types of GPU servers built on three major CPUs including mainstream X86 architecture CPUs, domestically-produced X86 architecture CPUs, and ZTE-developed ZFX CPU platforms.

Supporting Two GPU Form Factors

The ZTE GPU server supports PCIe AIC GPUs as well as SXM/OAM GPUs designed for high-speed interconnection between cards, such as Nvidia SXM GPU accelerator cards or OAM GPU accelerator cards (Biren and Cambrian).

Oriented to Three Application Scenarios

ZTE series GPU servers offer multiple configurations to meet the requirements of large-scale, medium-scale, and small-scale AI model training and inference scenarios.

For small model training and medium/small model inference scenarios, a general rack server is used. A single server can be configured with four dual/single-width full-height GPUs or six/eight single-width half-height GPUs, corresponding to the ZTE R53xx/59xx series servers.

In medium/small model training and large model inference scenarios, a dedicated PCIe AIC GPU server is employed. A single server can be configured with eight or 10 double-width, full-height and full-length GPUs or 16 or 20 single-width, full-height and full-length GPUs, corresponding to the ZTE R65xx series GPU servers.

For large model training scenarios, a dedicated SXM/OAM GPU server is used. A single server can be configured with eight SXM/OAM GPUs. To meet the multi-node cluster computing requirements, the GPU, parameter plane interconnection NIC and NVMe SSD are configured in a 1:1:1 setup, corresponding to the ZTE R69xx series GPU servers.

Conclusion

The GPU server market has become a high-growth segment within the server market, with its compound growth rate expected to remain high in the next few years. ZTE series GPU servers offer users high-quality and efficient computing power solutions, contributing to the establishment of a solid intelligent computing infrastructure that could further drive the growth of the digital economy. ZTE TECHNOLOGIES

High-Performance Network Designed for AI Model Training

Yang Maobin

Chief Engineer of ZTE Cloud & AI Network Planning

The popularity of ChatGPT has accelerated AI development from decision-making to generation, driving the need for high-performance networks for training AI models with billions of parameters. AI model training relies on distributed parallel computing, including data, pipeline, and tensor parallelism. To fully leverage GPU computing power, communication time overhead must be limited to within 5%. This necessitates a high-performance network for AI model training, characterized by zero packet loss, low latency, high throughput, large bandwidth, and large-scale networking.

Mainstream Solutions for High-Performance Networks

The two main high-performance network technologies used in AI model training scenarios are InfiniBand (IB) networks and RDMA over converged Ethernet version 2 (RoCEv2) networks.

The IB network, originating in the 1990s to replace the PCI bus technology, has become unexpectedly popular and widely used in high-performance computing and AI data centers. It implements packet lossless transmission through the credit flow control mechanism, and provides QoS for specific traffic optimization. Despite its advantages, its complex configuration, maintenance, and expansion, along with the need for special hardware and subnet managers, incur high costs. Therefore, the IB network is not as popular as Ethernet.

The RoCEv2 network, built upon Ethernet, allows remote direct memory access via encapsulated RDMA frames in IP/UDP packets. Data packets arriving at the RDMA NIC of the GPU server can be directly transmitted to the GPU memory, bypassing the CPU to reduce the delay. In addition, congestion control solutions like DCQCN are deployed to reduce RoCEv2 congestion and packet loss. Designed as a unified transport network, the RoCEv2 network caters to high bandwidth and elasticity needs, offering better support for cloud services and scalability, which is crucial for domestic high-performance networks.

RoCEv2 Network Congestion and Flow Control Analysis

In the RoCEv2 network, DCQCN is the most commonly used congestion control algorithm. It detects and indicates network congestion through the ECN flag on the switch. When congestion is detected, the switch adds ECN flags to data packets. RDMA NIC adjusts data transmission rates based on these flags via CNPs. The DCQCN algorithm is fair and efficient, ideal for high throughput, low-latency scenarios in high-performance computing and AI learning.

However, DCQCN also has the following disadvantages that cause network throughput to fluctuate between 50% and 60%:

- **Inaccurate congestion indication:** The 1-bit ECN flag lacks precision in distinguishing different levels of congestion.
- **Slow and inaccurate rate adjustment:** Only

CNPs are used to adjust the rate and there is no feedback from other networks.

- **No optimization based on traffic characteristics:** The diverse characteristics of long and short flows as well as scheduling interval cycles are not considered.
- **No consideration for multi-path balanced scheduling:** More traffic is unevenly distributed, and multi-path bandwidth resources of the AI network are not fully utilized.

ZTE's Innovative Solution for RoCEv2 End-Network Collaboration

Traditional DCQCN networks are difficult to avoid congestion, packet loss, and delay issues in high-throughput, fully-loaded networks due to imprecise congestion flag data and separate end-side and network-side flow control mechanisms. To improve transmission performance in high-performance networks, ZTE proposes an innovative solution for RoCEv2 end-network collaboration. This solution implements accurate and fast congestion control and traffic scheduling algorithms through end-network collaboration, boosting RoCE network throughput to more than 90% (Fig. 1). It pioneers end-network collaboration and innovation in congestion control and accurate flow control.

New Congestion Control for End-Network Collaboration

The network devices promptly and accurately deliver link congestion information to the end side through fast CNP and in-band telemetry technologies to implement new congestion control.

- **Fast CNP:** In a traditional DCQCN network, when a network device is congested, related link data packets are marked with an ECN flag. Upon receiving the ECN flag, the destination NIC sends a CNP packet to the source NIC, which then adjusts the rate. This process takes a long time, leading to delayed rate adjustments. Therefore, the fast CNP solution is introduced. When detecting congestion, the intermediate switch immediately sends CNP packets containing detailed congestion information to the source NIC. The source NIC can use this information to adjust traffic accurately and quickly, thus rapidly alleviating network

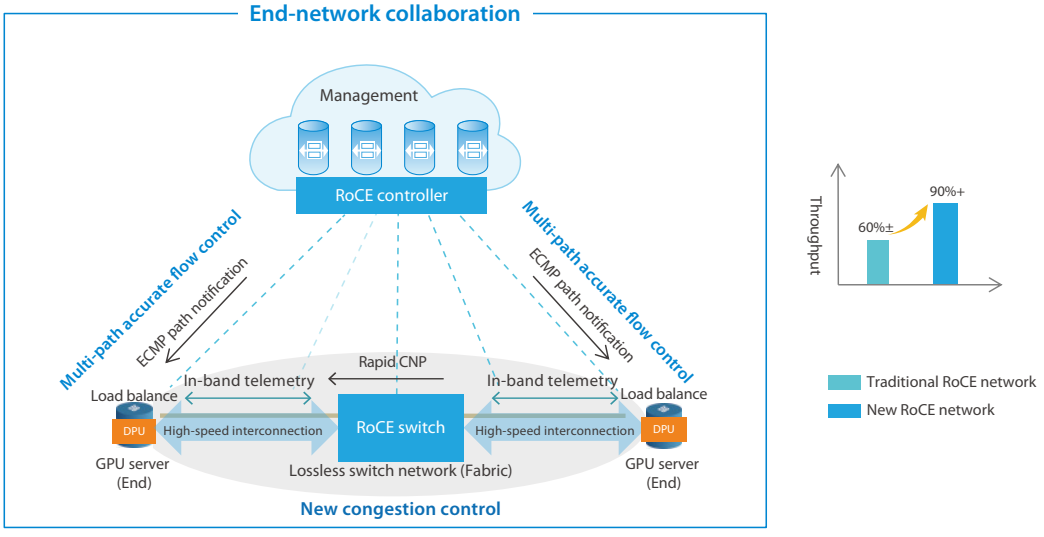


Fig. 1. ZTE's innovative solution for RoCEv2 end-network collaboration.

congestion.

- **Accurate congestion control based on in-band telemetry:** In the traditional DCQCN, the 1-bit ECN congestion indication fails to accurately convey link congestion levels, hindering accurate traffic control at the source. Therefore, an in-band telemetry-based solution is proposed to carry more path load information. The intermediate device fills available bandwidth, queue depth, timestamps, and sent byte counts in the telemetry packet. After collecting telemetry date from all network devices on the path, the end adjusts traffic accurately in real time based on a trained and optimized traffic scheduling algorithm. This optimization aims to achieve high throughput, low latency, and congestion-free end-to-end path traffic.

Multi-Path Accurate Flow Control for End-Network Collaboration

The network side collaborates with the end side, leveraging the RoCE network's equal-cost multi-path routing (ECMP) and multiple load balancing technologies to improve data transmission efficiency.

- **ECMP end-network collaboration notification:** The RoCE network in the AI model training data center uses the fat-tree CLOS architecture and has abundant ECMP paths. The RoCE controller comprehends the network topology, and synchronizes ECMP data to the end side to optimize data transmission and improve network efficiency.
- **Load balancing tailored to traffic characteristics:** The end selects load balancing technologies based on traffic characteristics (e.g., mouse flow, and elephant flow), routing packets through packet or source port hashing, and adjusts policies in real time to enhance data transmission efficiency.

As AI model parameters increase from 100 billion to one trillion and AI chip computing power remains limited, a 10,000-card intelligent computing cluster network becomes inevitable. Therefore, accurate end-network congestion control in large-scale networking scenarios poses a pressing industry challenge. ZTE's innovative solution for RoCEv2 end-network collaboration aims to improve RoCE network throughout, enhance AI model training network performance, unlock additional AI computing power, and boost model training efficiency. ZTE TECHNOLOGIES

Opening Doors to Diversity in AI Chips

Gao Zhenzhong

Chief Engineer of ZTE Computing and Core Network Hardware

In 1956, at the Summer Seminar of Dartmouth College in the United States, scientists like McCarty and Minsky first proposed the concept of AI. Over the past 60 years, AI has undergone extensive development and exploration. By 2015, AI surpassed human visual recognition precision and entered large-scale commercial use in video applications. In 2022, ChatGPT emerged as a groundbreaking product, propelling the adoption of AI models in industrial applications.

As a crucial cornerstone of AI development, AI chips have undergone two major phases. Before 2012, AI research and applications primarily relied on CPUs. In 2012, Alex Krizhevsky from the University of Toronto pioneered the use of GPUs in AI, achieving a groundbreaking victory in the ImageNet competition using only four NVIDIA Geforce GTX 580s. This event astonished academia and opened the door to diversity in AI chips.

Key Requirements

AI chips can be divided into training and inference chips. Training involves providing a large amount of labeled or unlabeled data to adjust mode parameters through optimization algorithms, so that the model can learn related modes and laws from the data. Inference refers to applying a trained model to real-world scenarios for prediction, classification, or decision-making.

The key requirement of training chips is to enhance AI computing power and reduce model training duration. With the rise of AI models, their numbers and scales have exponentially increased within a few months. The emergence of 10-billion-level AI models and the birth of trillion-parameter AI models have intensified the demand for computing power during training, surpassing the pace of Moore's Law for chip processing. As a result, the training time of AI models have consistently extended. For instance, OpenAI's GPT-3 model with 175 billion parameters required approximately 1,024 A100 GPUs for a single run in 2022, while the GPT-4 model with 1.8 trillion parameters demanded about 25,000 A100 GPUs for a single run in 2023. Comparatively, the training time for GPT-4 nearly doubled that of GPT-3.

The requirements of inference chips vary based on service scenarios. For instance, in online Q&A scenarios, the computing power of AI chips needs to match reading speeds (250 words per minute, maximum 1,000 words). In 5GC scenarios, voice codec and image processing capabilities are necessary additions to AI computing power.

Deployment Location

AI chips are deployed primarily on the cloud and terminal sides. The cloud side refers to a cloud data center, while the terminal side refers to devices like mobile phones or PCs that can be accessed or used locally by individuals.

Taking the operator's intelligent computing center as example, cloud data centers consist of group nodes, central nodes, and edge nodes (Fig. 1). Group nodes manage intelligent operations, central nodes handle training and non-real-time inference, while edge nodes and edge clouds are combined for real-time inference.

The terminal-side AI, with security, independence, low latency, and high reliability, efficiently handles diverse AI inference tasks. At present, multiple AI models have launched "miniaturization-based" and

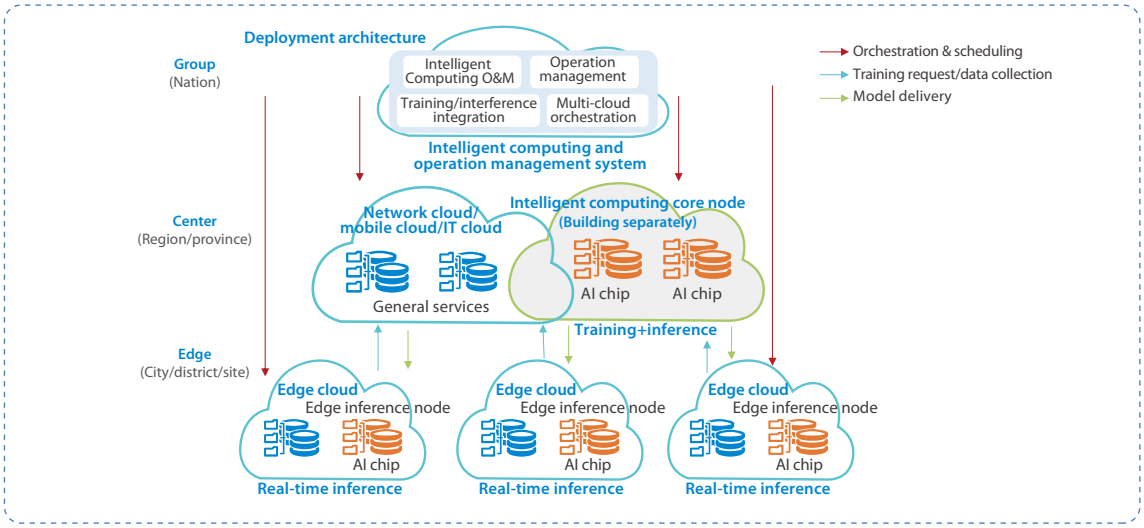


Fig. 1. Deployment architecture of the operator's intelligent computing center.

"scenario-based" versions, providing a basis for operation at the terminal side.

Technology Paths

The two leading AI chip technology paths are represented by the general parallel-computing architecture, exemplified by GPU, and the dedicated custom architecture tailored for accelerating AI tasks.

GPU is originally designed for graphic rendering, which involves extensive repetitive computing tasks. To meet the large-scale and parallel feature of AI operations, GPU chips deploy thousands of image computing cores capable of processing multiple tasks simultaneously.

Unlike GPUs, AI-dedicated chips are processors designed specifically for AI operations, utilizing AI-dedicated cores inside. They sacrifice video rendering and high-performance computing capabilities for advantages in power consumption and size. However, due to their dedicated and customized design, they have longer development periods and lack the universality and programmability of GPUs, resulting in many AI-dedicated chips being less powerful.

Looking into Future

The reality of trillion-parameter models is here, and even larger models may emerge soon. As the scale of models grows, both GPU and AI-dedicated chips face bottlenecks in performance and power consumption, resulting in the expansion of cloud data centers and the need for liquid cooling to manage heat dissipation. Additionally, devices at the terminal side like AI mobile phones and AI PCs are impacted by power consumption issues, affecting user experience. To solve these problems, the next-generation AI chip design focuses on the following directions:

- In computing architecture, the introduction of in-memory computing aims to reduce power consumption. Currently, mainstream GPUs and dedicated AI chips use the Von Neumann architecture, which separates computing from storage. However, 60% to 90% of chip energy is consumed during data migration. The in-memory computing architecture fully integrates memory and computing, avoiding data migration and greatly reducing power consumption.
- On the chip implementation layer, Chiplet and 3D stacking technologies are used to improve chip yield and performance. Chiplet divides a chip into multiple dies with specific functions (such as computing and storage), selecting the most suitable semiconductor process for each die to optimize yield. The dies are interconnected through a high-speed bus, and finally integrated and encapsulated into a single chip. 3D stacking expands chips from two-dimensional to three-dimensional, increasing the number of dies vertically, and improving chip performance while maintaining the original encapsulation size. ZTE TECHNOLOGIES

ZTE Intelligent Computing AI Platform: Facilitating AI Model Training and Inference

Zhou Xiangsheng
AI Platform R&D Manager, ZTE

Sun Wenqing
AI Algorithm Engineer, ZTE

In the context of explosive data growth, continuous improvement in algorithm performance, and the ongoing iteration of computing products, we are in a phase where AI is leading all-round industrial transformation. In this process, the AI platform plays a critical role. Through intensive management of data, computing, algorithms, and services, the AI platform converts workshop-style, discrete algorithm research into standardized, automated production processes, avoiding redundant efforts and allowing users to focus on high-value challenges in intelligent services.

AI Platform: Key Infrastructure for Enterprise Intelligent Transformation

As a key bridge connecting computing and algorithms, the AI platform not only systematizes and formalizes the common requirements in the algorithm development process, but also offers users customized capabilities and services. In addition, the platform should have features such as sharing and multiplexing, efficient training and inference, fast delivery, and continuous iteration. To address these needs, ZTE has developed a platform for heterogeneous computing management and AI model training and inference—the intelligent computing AI platform. This platform consists of the infrastructure layer, engine layer, service layer, and capability layer.

- **Infrastructure layer:** Thousands of GPUs and CPUs provide computing power, supporting both international mainstream graphics cards and domestic graphics cards.
- **Engine layer:** The engine layer includes machine learning (ML) engine, hyperparameter tuning engine, training engine, compilation engine, and inference engine. This layer integrates multiple high-performance training and inference engine frameworks, such as Tensorflow, Pytorch, Oneflow, and Deepspeed.
- **Service layer:** The service layer consists of dataset management, data labeling, model training, hyperparameter tuning, as well as model evaluation, compiling, and inference, covering end-to-end services of the AI model.
- **Capability layer:** The capability layer provides various built-in algorithm and inference packages to solve practical problems, available for direct deployment and calling.

From basic computing and scheduling technologies to deep learning frameworks and engines, as well as perception and cognition capabilities such as NLP, CV, audio processing, and AI model, the AI platform serves as a key infrastructure for enterprise intelligent transformation. It not only integrates computing hardware and software tools, but also provides R&D interfaces for AI algorithms. Through this comprehensive integration, the AI platform greatly improves resource utilization efficiency and accelerates AI implementation.

Entering the Era of AI Models

Currently, in the AI implementation scenario, many small models that solve intermediate tasks or specific field tasks are being replaced by more universal AI

models, leading to the full transformation of artificial intelligence into artificial general intelligence (AGI). Aadditionally, there is a growing demand for AI models with comprehensive, stable, and efficient data storage and cleaning, as well as training and inference skills, along with cluster resources. This poses new challenges to the construction of AI platforms.

The emergence of AI modules brings about a unified model structure and a training-inference paradigm. First, the transformer structure remains the preferred choice for the basic components of the backbone model. Second, concerning training and inference methods, using the AI module as an example, the training methods (including pre-training, instruction fine-tuning, and reinforcement learning fine-tuning) and inference methods (such as random sampling decoding) initially proposed by OpenAI continue to be mainstream solutions for AI model training and inference.

However, the unification of this structure and application paradigm does not close the gap between the industry's average level and the leading AI companies. Instead, it shifts the focus of AI competition from algorithm R&D innovation to the competition in scale and efficiency of engineering AI model training and inference. This makes it a primary requirement for AI platform construction to integrate key technologies for training and inferring AI models.

Key Engineering Technologies for AI Model Training and Inference

The key technologies in the AI model training and inference process include distributed training, AI model inference acceleration, AI model evaluation, and AI model data engineering.

- **Distributed training:** The distributed training technology can extend training to multiple AI hardware products, breaking the limits of single hardware memory and computing power. The intelligent computing AI platform integrates 3D hybrid parallel technology and has independently developed automatic parallel tools. These tools support AI model training technologies such as data parallelism (DP), tensor parallelism (TP), pipeline parallelism (PP), and activation re-computation, automatically adjusting parallel hyperparameters based on clusters and model characteristics.
- **AI model inference and acceleration:** The AI model inference acceleration technology is a comprehensive technique for reducing memory consumption and computational delay during the inference process. The intelligent computing AI platform improves inference efficiency through various means, such as service scheduling, memory optimization, and quantization compression. In ZTE's industry-leading "Zhiyu" SMS anti-fraud governance system based on AI models, the inference solution provided by the intelligent computing AI platform reduces inference delay by 30% compared to the industry's general solution.
- **AI model evaluation:** The AI model evaluation method differs greatly from traditional approaches. Therefore, the AI platform provides a comprehensive objective evaluation dataset to evaluate the performance of AI models from multiple dimensions. Additionally, the platform integrates a model-based evaluation mechanism to evaluate the semantic accuracy and logical consistency of the generated contents.
- **AI model data engineering:** High-quality training data can mitigate AI model hallucination and shorten the training cycle. The intelligent computing AI platform provides intelligent data engineering pipelines such as model-in-the-loop data marking, SFT data generation and expansion, data cleaning and deduplication, quality evaluation, and privacy protection.

With the support of key engineering technologies for AI models, ZTE's intelligent computing AI platform has achieved preliminary success in collaboration with ZTE and Chinese telecom operators. At the company level, the AI platform supports the training of AI models across multiple domains including telecommunications, coding, computer vision (CV), and multi-modal areas. For telecom operators, the AI platform has established training and inference clusters in 31 provinces of an operator group, offering nine core functions such as model training, management, and inference services. It has become an important tool cloud for operators' AI development. ZTE TECHNOLOGIES

AI Model Empowers Intelligent Operation and Maintenance for Efficiency Enhancement

He Wei

Chief Engineer of ZTE MANO Product Planning

With the acceleration of digital transformation in the industry, intelligent operation and maintenance (O&M) requirements within the telecom field are becoming increasingly complex. Consequently, intelligent O&M has emerged as one of the key factors for maintaining competitiveness in the digital era. However, due to the rapid development of services and continuous technological updates, traditional O&M methods are inadequate to meet the evolving O&M requirements of communication devices. The advent of AI model has brought breakthroughs in the field of intelligent O&M. It offers a more user-friendly man-machine interaction mode, processes massive structured data, delivers high-precision analysis and prediction, and empowers advanced O&M capabilities.

Applications of AI Model in Intelligent O&M

The AI model technology has been widely used in intelligent O&M in the telecom field, which includes O&M knowledge Q&A, fault and exception detection, root cause location, and fault prediction and prevention.

- **O&M knowledge Q&A:** When the AI model is utilized in the telecom field for knowledge questions and answers, its capabilities in storage, memory, understanding, and application become especially crucial. By analyzing a large amount of communication data and technical documents, the AI model can deeply understand various communication devices, protocols, and network topologies. This comprehensive understanding allows the AI model to efficiently integrate contextual information when addressing complex communication issues, and quickly and accurately generate answers to questions. Additionally, the AI model also has the capability to continuously update and refine its knowledge base based on real-time communication data and the latest industry trends to keep pace with the latest knowledge and provide timely and reliable support and guidance for operation and maintenance personnel.
- **Fault and exception detection:** Utilizing AI-model intelligent algorithms and models, the system processes and analyzes collected data to identify abnormal data or behaviors that are inconsistent with the normal status. This typically involves feature extraction, data modeling and classification, and formulation of abnormal judgment standards. In practice, to enhance detection accuracy and robustness, considerations for data temporal characteristics, spatial correlations, and trends over time are often necessary. Moreover, tailored algorithms and models may be required to accommodate diverse data characteristics and business needs in various domains and application scenarios. Furthermore, continuous updates and optimizations of algorithms and models are essential to ensure system reliability and stability, enabling adaptation to new fault types and changes in scenarios.

- **Root cause location:** Based on fault detection, ZTE conducts further analysis of abnormal data to infer the cause and location of the fault, thereby identifying the specific type and location. This requires various diagnosis technologies and methods, such as fault tree analysis and expert systems. Through root cause analysis, the system can pinpoint the origins of issues more accurately and take effective measures to resolve faults, thereby enhancing system reliability and stability.
- **Fault prediction and prevention:** The AI model can learn from vast amounts of historical O&M data to identify patterns and trends in fault occurrence, establishing a fault occurrence model. Utilizing this model, the AI model monitors and analyzes real-time data to predict potential fault risks and send early warnings, providing O&M personnel with sufficient time to take preventive measures and reduce the fault rate. This predictive maintenance approach not only reduces the impact of sudden failures but also maximizes system stability and availability, enhancing operational efficiency and resource utilization.

Compared with traditional AI for IT operations (AIOps), the AI model provides enhanced intelligent O&M capabilities, such as simpler interaction, broader knowledge coverage, fault self-learning, and a more flexible model architecture. It provides O&M personnel with lower maintenance threshold and continuously generalizes O&M capabilities.

Architecture and Key Technologies of ZTE CCN AI O&M Model

ZTE CCN AI O&M model is based on the Nebula model, developed by ZTE, in the telecom field. It uses high-quality corpus for fine-tuning the base models and generating an AI O&M model oriented to the core network and network cloud (Fig. 1). There are three types of AI O&M model applications: intelligent interaction (CoPilot-I), intelligent analysis (CoPilot-A), and intelligent generation (CoPilot-G).

- **CoPilot-I:** It provides the functions such as professional knowledge Q&A, network health querying, and key indicator information querying.
- **CoPilot-A:** It includes fault analysis assistance, network optimization assistance, and inspection report.
- **CoPilot-G:** It generates inspection reports, operation solutions, and network reports.

To meet the above-mentioned AI O&M model capabilities, all AI O&M model products targeted for ZTE core network and network cloud incorporate the currently popular key technologies such as retrieval-augmented generation (RAG) and

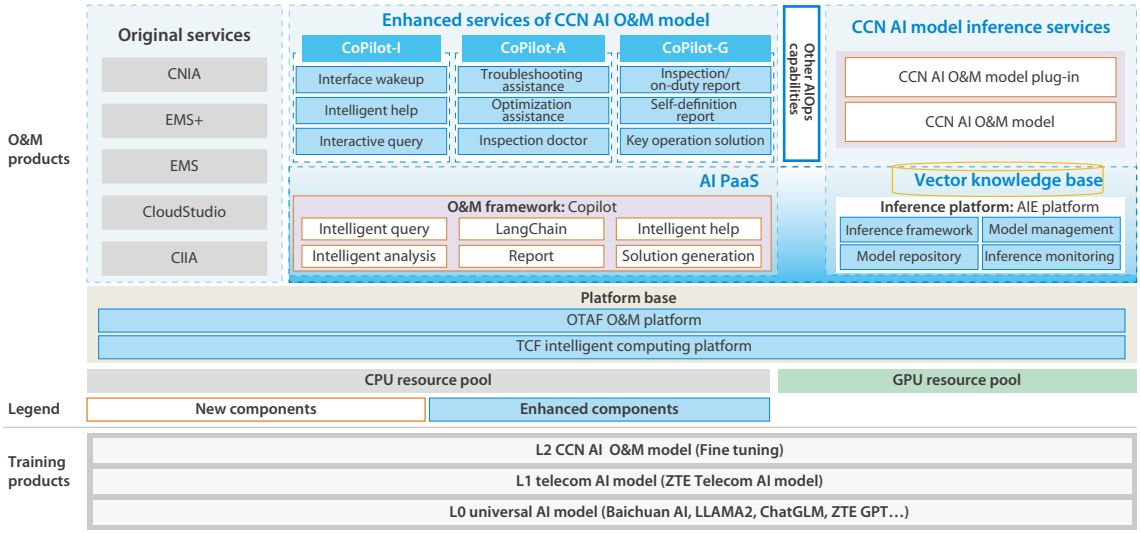


Fig. 1. ZTE CCN AI O&M model architecture.

multi-agent collaboration.

Retrieval-Augmented Generation

To accomplish more complex and knowledge-intensive tasks, it is necessary to build a more accurate and reliable system and alleviate the hallucination issue of AI models. RAG stands out as a key technology for AI models. It helps large language models generate answers by retrieving information from data sources. The RAG technology can greatly enhance the accuracy and relevance of content, effectively addressing hallucination issues, accelerating knowledge updates, and improving the traceability of content generation. RAG has become the most popular system architecture for AI models to obtain new external knowledge.

Multi-Agent Collaboration Architecture

Multi-agent collaboration refers to the process in which multiple agents communicate and collaborate with each other in a shared environment to achieve common goals. Each agent possesses a level of autonomy and intelligence, enabling perception, decision-making, and execution based on environment information. Through mutual interaction and cooperation, the system can benefit from the advantages and strengths of each agent to achieve more efficient and intelligent decision-making and action. Leveraging the multi-agent collaboration architecture, independent agents, such as knowledge experts, fault experts, on-duty experts, and solution experts, can be collaboratively created to build an intelligent network O&M system architecture.

Challenges and Future Development of AI Models in Intelligent O&M

Although AI models have promising application prospects and advantages in the intelligent O&M field, challenges still exist, including improving the model's adaptation capability, reducing complexity, and addressing data privacy and security issues.

In the future, with the continuous development of technologies and application scenarios, AI models will be widely and deeply applied in the intelligent O&M field. For instance, with the proliferation and advancement of edge computing, AI models will gradually migrate to the edge to achieve more efficient and real-time intelligent O&M. At the same time, AI models will be closely integrated with machine learning, deep learning, and other technologies to further enhance the efficiency and precision of intelligent operation and maintenance. AI models will encounter both challenges and opportunities in handling more extensive data. Therefore, we need to continually explore, innovate, and apply practices tailored to specific scenarios and requirements. Additionally, further strengthening research and development in related technologies is necessary to promote the advancement and progress of intelligent O&M technologies. ZTE TECHNOLOGIES

AI Model + 5G: Empowering Industry Intelligence

Wang Chaoying

Architect of ZTE CCN Product Planning

Liu Xiliang

Chief Engineer of ZTE CCN Product Planning

Since the second half of 2022, the popularity of AI models, represented by ChatGPT, has swept the world, signaling the official entrance of AI technology into the era of AI models. However, in 2023, the model development and applications have followed different trajectories. While the models are rapidly evolving, their applications have not kept pace, lagging behind the advancement of AI models. As a new technology, AI models strongly promote the development of basic intelligent science. However, their wide applications need to be integrated with the industry and other technologies. This paper discusses the new paradigm of applying AI models combined with 5G in various industries and explores the prospects and evolution of the "AI model + 5G" to promote the development of industry intelligence.

In China, the government held six consecutive sessions of the "Blooming Cup" 5G application competitions to promote in-depth integration of the digital economy and the real economy. Throughout this process, various issues have been exposed, such as the high threshold of 5G network technologies, maintenance and operation difficulties for industry users, and the failure to realize the value of network data. Additionally, model applications in the industry encounter numerous challenges, such as acquiring high-quality datasets and achieving rapid industrial deployment, which remain significant concerns.

The development of industry intelligence requires that AI model and 5G complement each other, with both being jointly promoted: AI model enables 5G and drives the digital and intelligent transformation of the industry, while 5G empowers AI model and accelerates its applications, as illustrated in Fig. 1.

AI Model Enables 5G

The challenges encountered in the development of 5G within the industry include complicated network O&M, unguaranteed service level agreements (SLAs), and insufficiently differentiated processing methods for specific services. These issues can be addressed by introducing AI models into the 5G network.

AI Model Facilitates Intelligent 5G Network Operation

After industry customers deploy 5G networks, professional knowledge and personnel are required for service provisioning and routine O&M, thereby increasing network costs and investment. To mitigate this, AI models are applied in 5G O&M to achieve intelligent network operation, significantly reducing industry investment. These AI models aid industry customers in converting service requirements into network planning and configuration, enabling intent-driven services and improving service provisioning efficiency. By analyzing a vast amount of network logs and alarm data, AI models can identify or predict faults and provide corresponding solutions. Through in-depth analysis of customers' behavior data, AI models can predict network requirements and service preferences, generate related service bundles, and facilitate better network operation.

AI Model Guarantees Network SLA

Applications in various industries have unique demands concerning network bandwidth and latency. For example, video surveillance services in the intelligent manufacturing field necessitate

substantial bandwidth, while production control services prioritize low latency. Therefore, 5G networks need to provide SLA guarantees tailored to different users or services. On the core network side, AI models can be used to evaluate and predict user service quality and generate service guarantee policies in time. On the wireless side, these AI models analyze wireless signals and sensor data to implement more accurate resource allocation and scheduling, thereby enhancing network efficiency and quality.

AI Model Enables Intelligent Service Analysis

In industrial application scenarios, stringent security protocols often dictate that data must remain confined within specific areas to mitigate potential risks. However, the emergence of advanced AI models and algorithms has heralded a transformative solution to this challenge. With the introduction of AI models and algorithms, the 5G network can analyze and intelligently classify data traffic in real time without compromising security protocols. This innovative approach enables industries not only to uphold confidentiality protocols but also to extract valuable insights from the data generated within their confines. By deploying AI models at the edge of the network, where data is generated and consumed, the system can harness the potential of intelligent service analysis. The integration of AI models into the 5G network empowers industries to tailor their services according to specific requirements, ensuring optimal performance and responsiveness. Through continuous monitoring and analysis, the system can dynamically adapt to changing conditions, providing differentiated service guarantees tailored to the unique needs of each industrial sector.

5G Empowers AI Model

As an information infrastructure in the digital era, the 5G network boasts high bandwidth, low latency, and massive access, greatly enhancing the industry's data collection capabilities and enriching AI model datasets. At the same time, as 5G networks are widely used in industries, AI models will play a pivotal role in the digital and intelligent transformation of these industries.

5G Network Enhances AI Data Sets

In the industry intelligence process, data stands as the core element shaping the competitiveness of AI models. The 5G network supports multiple access modes like IP/LAN, overcoming mobility restrictions

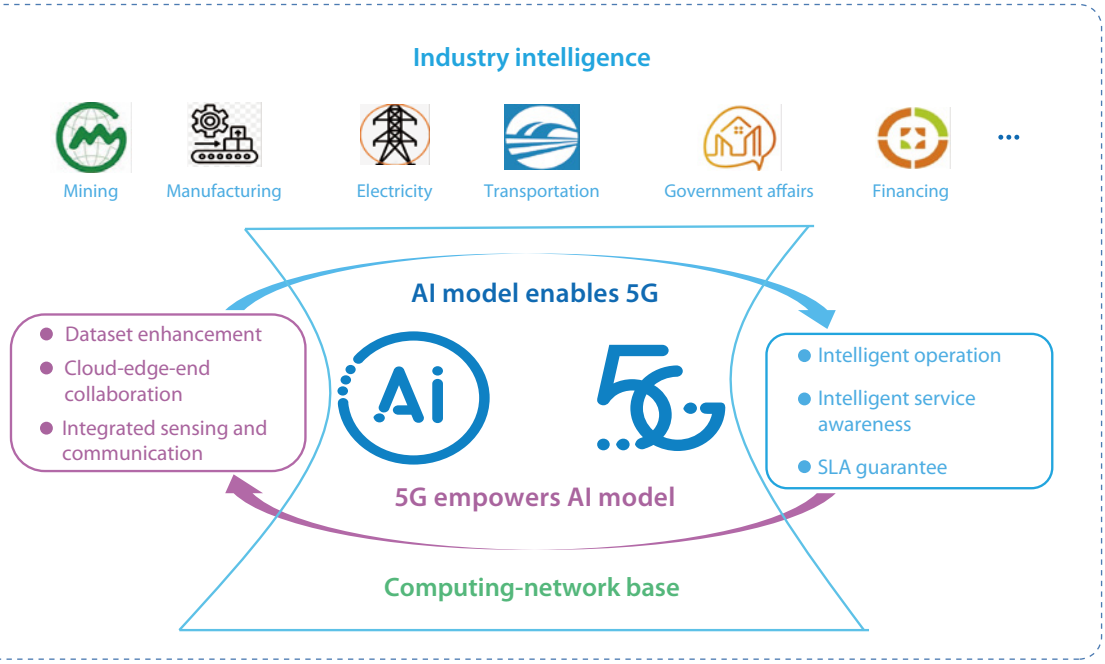
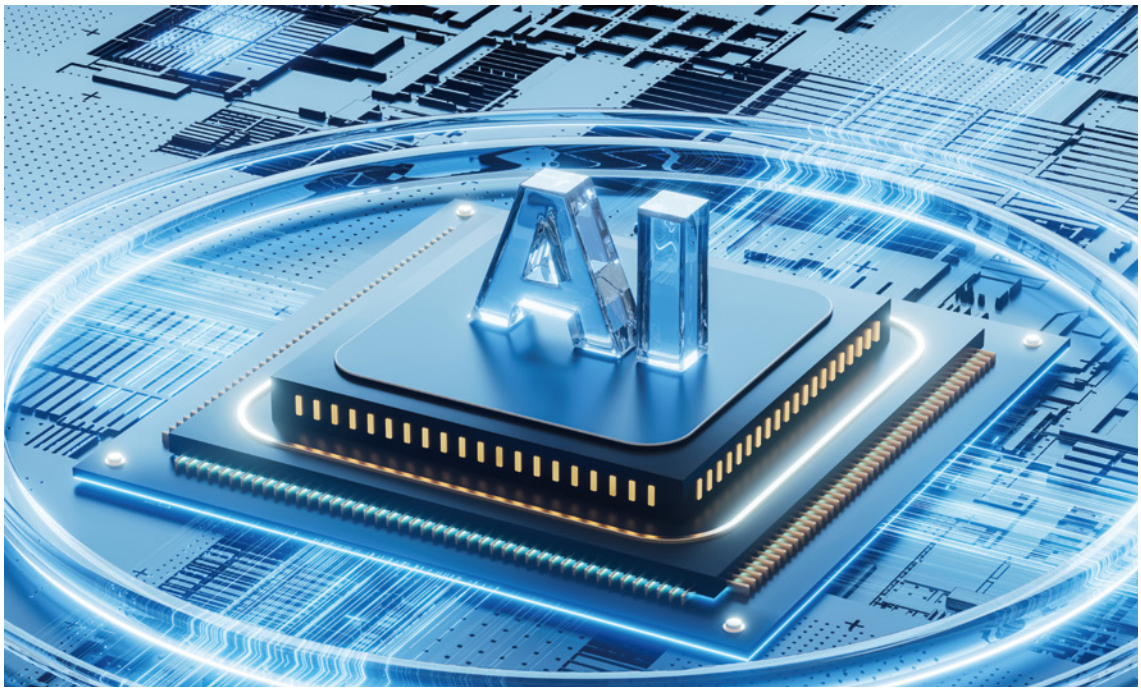


Fig. 1. "AI model + 5G" collaboration framework in the industry.



and facilitating access to voice, image, and video data from anywhere in the industry, thereby greatly enriching the datasets of AI models. In addition, the 5G network provides user authentication and transmission encryption technologies to ensure the security and effectiveness of data access, thus greatly enhancing the quality of AI model datasets.

5G Network Expands AI Model Application Scenarios

5G, as a dynamic innovation catalyst, is rapidly gaining ground across industries, expanding the application scenarios for AI models. On the one hand, through the deployment of 5G private networks, heterogeneous and massive connections of monitoring devices, such as sensors and cameras at front-line production sites, can be supported. These new scenarios also pose additional requirements for AI models in the industry. Leveraging the multi-mode analysis capabilities of AI models, intelligent and accurate fault warning and risk management can be achieved, significantly improving production efficiency. On the other hand, 5G networks also promote the collaboration among AI models. Currently, AI models primarily collaborate between the cloud and edge. With the applications of 5G networks, AI agents can be extended to 5G intelligent terminals to establish a comprehensive cloud-edge-end collaboration system.

Conclusion

To foster the industrial applications of "AI model + 5G", ZTE has launched the AiCube all-in-one machine. The AiCube consists of computing hardware, a cloud platform, and an AI platform. The computing hardware is compatible with CPUs/GPUs from various manufacturers and models. The cloud platform implements unified resource management, allocates computing resources on demand, and can deploy both 5G networks and AI models. The AI platform provides operators and enterprise customers with tools such as data management, model development, model training, model inference, and application statistics to enhance usability and efficiency.

As a full-stack intelligent computing solution provider, ZTE will work closely with operators and industry customers to continually advance "AI model + 5G" applications, co-create a new intelligent computing ecosystem, embrace a future of intelligent computing, and inject new impetus into the development of the digital economy. ZTE TECHNOLOGIES

ZTE's Tangxue: 5G+AI for Integrated Communication and Computing

Source: RCRWireless.com

Tang Xue, Vice President of ZTE

"The integration of 5G and AI offers both opportunities and challenges for 5G network development and monetization. It unlocks innovative applications like deterministic connections, edge rendering, and V2X applications."

The rapid advancement of artificial intelligence(AI) is revolutionizing various industries, ushering in a new era of innovation. Pioneering AI models such as ChatGPT and Sora by OpenAI, and Gemini by Google are at the forefront of this transformation, providing users with an expanded realm for creativity and imagination and opening up unprecedented possibilities. AI models are fundamentally reshaping social structures and daily routines by improving the ability to generate new ideas, solve complex problems, and automate routine tasks. Tang Xue, Vice President of ZTE, recently shared valuable insights and experiences with experts from diverse fields at the GSMA panel on Convergence of 5G & AI at the Edge during MWC2024 in Barcelona, Spain.

Tang Xue emphasized that the integration of 5G and AI is inevitable, ushering more possibilities for new business models and applications. As the edge node closest to users, performing AI processing at 5G network can effectively reduce data transmission delay, improve data processing efficiency, and reduce the cost of AI applications. While 5G networks provide large bandwidth, high speed, and low latency, making it ideal for supporting the connection and computing requirements for distributed AI computing. The integration of 5G and AI not only can achieve flexible collaboration of cloud and network, but also provide stronger intelligent capabilities, allowing 5G to adaptively optimize and adjust according to data characteristics and scenario requirements, improving network performance and efficiency.

"5G and AI integration, 5G-A new capability growth and new scenario exploration need a lot of computing power. Therefore, an integration of communication and computing infrastructure is required", Tang said. ZTE 5G-A BBU is just this novel infrastructure, which is hardware ready for communication and computing integration, further explores computing power of RAN system, implements computing power orchestration adapting to full scenario requirements, including site computing power enhancement, inter-site computing power sharing and computing power integrated in computing force network, flexibly empowers B2C new services, B2B application in-depth development and low-altitude new economy. For V2X case, a built-in intelligent computing board can be added in 5G-A BBU for a heterogeneous computing power support and enabling more differentiated intelligent applications, such as 3rd party platform deployment.

In addition, ZTE has been engaged in AI application innovation for a long time. RAN Composer is the industry's first native-AI based RAN intelligence solution. On top of 5G-A BBU, the integration of 5G and AI at edge nodes brings traditional network resource management revolution, achieves user-centered experience with precise resources allocation based on network serving capability, service requirements and UE capability, enabling supreme user experience, higher energy efficiency and O&M efficiency.

With 5G-A BBU, RAN Composer helps vertical industries achieve deterministic experience guarantees, adjusting resource allocation and scheduling strategies in real time through precise analysis of flows and packets. In WISCO (Wuhan Iron and Steel Corporation) case, remote control of crane is a typical application with large-bandwidth video to be transmitted in the uplink and highly reliable control instructions need to be transmitted in the downlink. A differentiated scheduling strategy will be performed based on AI learning and recognition of video data and control instructions. For example, video data identifies I frames (key frames of the picture) and uses smooth scheduling to avoid "I frame" collisions; control instructions use ML to predict the data packet sending for an accurate scheduling, which not only ensure the reliability of the control instructions, but also reduce the occupation of system resources.

5G-A BBU also supports the deployment of 3rd party applications, such as 5G-based V2X application. ZTE exhibited a 5G-based V2X case in cooperation with Tianyi Transportation this year. On an open road of 162 square kilometers, 5G connection and AI-based edge computing are used to achieve precise collaboration of roads, vehicles and clouds. The autonomous driving network reduces congestion rate by 20%, decreases carbon emissions by 12%, and cuts down accident rate by 8%.

In summary, the integration of 5G and AI presents significant opportunities and challenges for the development and monetization of 5G networks. It enables innovative applications and business scenarios, fully unleashing the value of 5G networks, such as deterministic connection guarantees, edge rendering, and V2X applications. These new applications and services can significantly increase the revenue of operators and service providers, serving B2C, B2B, and new economy development.

2024 is the first year of 5G-Advanced (5G-A) commercialization. Driven by new technologies, new services and new scenarios, the boundaries of communication networks continue to expand. 5G-A is not only the key of 5G development in the next 10 years, but also the key to shape future digital society. "Facing 5G-A, ZTE 5G-A BBU and RAN Composer are ready to unlock 5G value with high efficiency", Tang added. For B2C scenario, connecting the virtual and real world for a digital life; For B2B scenario, facilitating 5G in core production and low-altitude economy for a digital society. ZTE TECHNOLOGIES

Protecting People's Lives Through "Smart Safeguard" AI Anti-Fraud System

Huang Xiaobing

Chief Planning Engineer of ZTE Messaging Products

Wang Wei

ZTE Product Planning Expert Engineer

According to public information released by China's National Anti-Fraud Center, telecom fraud has become the crime with the largest number of cases, the fastest growth rate, and the widest coverage in recent years. By the end of 2022, China's public security departments had cracked 1.156 million telecom fraud cases, arrested 1.553 million suspects, and intercepted over 916.5 billion yuan of funds related to fraud cases. The increasing prevalence of telecom fraud pose a significant threat to personal and property safety.

Difficulties and Challenges in SMS Fraud Monitoring

SMS fraud stands out as one of the most common types of telecom fraud. Fraudsters constantly alter and adapt SMS contents to bypass the SMS monitoring system of telecom operators. Some common tactics employed by fraudsters include:

- Circumventing keyword rules by utilizing combining mutations, escape characters, homophones, and similar shapes.
- Using a combination of Chinese characters, symbols, and digits to express standard URLs and numbers, evading regular expression monitoring policies implemented in the existing network.
- Evading traffic and keyword thresholds through utilizing a vast pool of numbers.
- Making breakthroughs through methods such as dialing tests to enable the sending of massive messages.

The traditional fraud management solution with a long upgrade period faces huge challenges. Overly lax policies may lead to low interception efficiency, while overly strict policies affect normal communication.

AI Model Enables Technological Revolution

On November 30, 2022, OpenAI launched ChatGPT, which obtained 100 million users within two months after its launch. Built on the

transformer neural network architecture, ChatGPT, a large language model (LLM), has made major breakthroughs across multiple deep learning fields, including large-scale natural language processing, sequence data analysis, and target detection. Trained on extensive corpora, LLMs can acquire generalized knowledge and a deep understanding of languages and dialogues. Moreover, targeted training allows LLMs to solve problems in specific fields and rapidly adapt to new tasks and scenarios.

Accurately identifying fraudulent SMS messages requires a deep understanding of natural languages. Furthermore, it is necessary to classify sensitive information and identify the real intentions conveyed in the content. Lastly, given the evolving nature of fraudulent SMS messages, it is necessary to learn from samples and dynamically upgrade knowledge and models. These are the technologies where transformer-based LLMs excel. It is worthwhile to develop new SMS anti-fraud technologies and products utilizing AI models through prototype testing and exploration.

Rapid Technical Breakthrough Helps Tackle Difficulties

In the early stage of the project, we faced several challenges in selecting LLMs:

- **Uncertainty in model selection:** It is challenging to determine the most appropriate model while ensuring legal compliance.
- **Uncertainty in corpus and training solutions:** The quality, quantity, format, and prompts of the corpus were unknown. We started from scratch with training and inference solutions.
- **High GPU and server costs:** Inference performance was low in the medium term, and the number and costs of GPUs required for handling large service traffic were too high.

To achieve rapid technical breakthrough, we dared to try different approaches, make mistakes and adjust solutions promptly.

In terms of model selection, during the initial exploration phase, we tried models with less than 100 million parameters to 340 million, 7 billion and 13 billion parameters. This process encompassed four parameter scales and included six different models from both domestic and international sources, including self-developed ones. We evaluated a total of more than 20 combinations.

In terms of corpus and fine-tuning, we obtained a first-hand, high-quality corpus compliant with regulations, tried various fine-tuning solutions, and finally devised the most effective approach: "special prompt words+sample fine-tuning", greatly improving recognition accuracy and recall rate.

To address the challenges of high GPU quantity and high costs, we designed a multi-layer architecture with cache acceleration at the front and utilized a combination of small models and large models. Additionally, we implemented inference acceleration to achieve optimal performance.

After evaluating the effects and cost indicators of the models, we selected the most optimal solution and passed legal compliance review.

Perfect Combination of Communications and AI

Through continuous innovation, ZTE has successfully released the industry's first anti-fraud big model system called “Smart Safeguard” (Fig. 1). With its out-of-the-box functionality, the system automatically identifies illegal SMS messages without policy configuration. This greatly reduces the complexity and workload of on-site policy O&M, while enhancing the accuracy of illegal SMS message identification and recall rate. It enables integrated management of identifying, preventing, and controlling junk and fraudulent SMS messages.

Currently, the system has implemented the industry’s first LLM-based SMS anti-fraud management pilot in pilot offices of operators A and B in China and quickly transitioned into commercial use.

- **Operator A’s achievements:** Since the system was deployed in the provincial company, there has been a significant increase in the interception rate of fraudulent messages. The

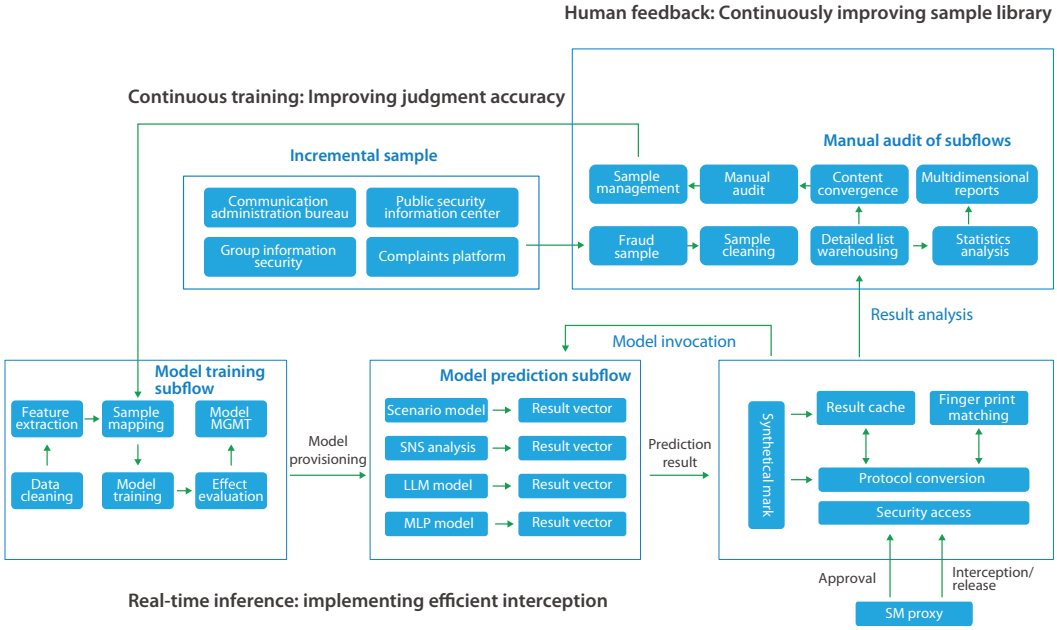


Fig. 1. ZTE "Smart Safeguard" anti-fraud big model system.

daily average number of overseas junk SMS messages dropped from 500,000 to 600,000 to 20,000 to 30,000. The forecast success rate and interception accuracy rate can reach up to 99%. In addition, the number of fraud-related cases has significantly decreased. In August 2023, the month-on-month ratio of overseas fraud-related cases decreased by 64%. Following the office's provisioning, the system received high acclaim from the operator and the anti-fraud center of the province.

- **Operator B's achievements:** The total number of mobile-originated (MO) messages initiated by domestic terminals is 4 million per day, all of which are monitored by the Smart Safeguard system. The average daily interception success rate for junk and fraudulent messages has increased from 57.25% to 93.60%. The false interception ratio has reduced from 42.75% to 6.4%.

In addition, ZTE's AI anti-fraud technologies have been chosen by China's Ministry of Industry and Information Technology as an innovative technology application for preventing and controlling telecom fraud, and they have been promoted nationwide.

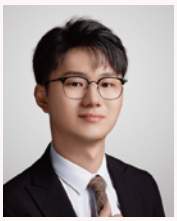
Future Evolution and Prospect

The introduction of anti-fraud model marks the beginning of AI model application in the communication sector. The Smart Safeguard series models will be developed, evolved, and applied across multiple domains, including service scope, media capabilities, and industrial applications.

- **Field expansion and capability openness:** ZTE will achieve capability replication and openness, further exploring the application of anti-fraud governance in 5G communication, IT, and content release.
- **Computer vision model:** Besides SMS text content anti-fraud, multimedia content is a fast-growing form of telecom fraud. To ensure that media content is trustworthy, secure, and reliable in the new 5G communications era, the Smart Safeguard model must efficiently identify and combat fraudulent multimedia content in the future.
- **Industrial model:** 5G industrial customers have diverse requirements including intelligent dialogues, industrial knowledge services, and enterprise applications. By supporting L0/L1/L2 AI models and integrating them into platforms on the new communications network side, such as the 5G messaging platform, ZTE "Smart Safeguard" model can rapidly meet the AI capability requirements of 5G industrial communication, effectively serving government and enterprise customers. ZTE TECHNOLOGIES

True and ZTE Build Thailand's First FTTR Community

- True and ZTE have collaborated to create an industry-leading all-optical Wi-Fi solution for the Wyndham Royal Lee Phuket, Thailand
- ZTE's innovative FTTR solution has a wide range of applications in traditional home scenarios and enterprise settings such as hotels and retail establishments
- True Online's 2 Gbps bandwidth packages, coupled with ZTE's FTTR solution, have transformed the Internet experience offered to guests at Wyndham Royal Lee Phuket



Xia Dezhi
CPE Product Planning Manager, ZTE

Thailand, a tropical country in Southeast Asia, hosts approximately 14.27 million fixed broadband (FBB) users, boasting a household penetration of around 58.96%. Notably, 95% of these users rely on fiber to the home (FTTH) technology. True Corporation (True for short), as Thailand's largest fixed-network and video operator, is dedicated to leveraging cutting-edge technologies and providing appealing service packages. This dedication is geared towards enhancing market competitiveness and elevating user satisfaction.

Wyndham Royal Lee Phuket stands as a high-end serviced-apartment-style community meticulously crafted by Wyndham Hotels and Resorts in Thailand. After years of operation, the network infrastructure of accommodation had become outdated, failing to cater to the evolving Wi-Fi usage needs of its guests. Recognizing the imperative for an upgraded solution, True and ZTE joined forces during the refurbishment of the Wyndham Royal Lee Phuket. This collaboration resulted in the development of an industry-leading, all-optical Wi-Fi solution tailored to meet the specific needs of Wyndham Royal Lee Phuket.

Wi-Fi: A Critical Pain Point Wyndham Royal Lee Phuket Urgently Needs to Address in Operations

With the widespread adoption of smart home devices and the rapid development of services such as high-definition video and gaming, the very high-speed digital subscriber line (VDSL) 100 Mbps bandwidth access solution used by Wyndham Royal Lee Phuket was no longer able to meet the high-speed Internet needs of its guests. This growing demand for network services poses greater challenges for hotel network infrastructure, necessitating urgent upgrades to provide faster and more stable Internet connections.

The VDSL solution involved sharing one Wi-Fi gateway for every 2–3 rooms, resulting in poor Wi-Fi coverage and low bandwidth—the actual bandwidth accessed by guests was less than 10 Mbps. Guests frequently reported Wi-Fi connectivity failures and unstable Wi-Fi connections, which degraded the user experience and harmed the premium image of the hotel.

After astutely identifying Wi-Fi as a pain point of Wyndham Royal Lee Phuket, True worked with ZTE to hold discussions with the customer about the latest fiber to the room (FTTR) all-optical networking solution in the industry, which would deliver gigabit bandwidth to every room. The solution gained high praise from Wyndham Royal Lee Phuket, and the three parties agreed to replace its existing VDSL solution with True Online's broadband access service and ZTE's FTTR solution when Wyndham Royal Lee Phuket officially reopens after the refurbishment. This proactive approach would address the low bandwidth and poor Wi-Fi coverage in the guests rooms, providing guests with fast and stable Internet service. Consequently, it enhances the guests' stay experience and elevates Wyndham Royal Lee Phuket's high-end image.

FTTR All-Optical Networking Solution

Before the Wyndham Royal Lee Phuket project, True Online had already started large-scale deployment of dual-band Wi-Fi 6 optical network terminals (ONTs) and mesh access points (APs) to ensure its competitiveness in bandwidth and coverage. To maintain its leading position in the Thai operator market, True Online actively conducted FTTR network trials with ZTE and was very satisfied with the trial results. ZTE's innovative FTTR solution has a wide range of applications in traditional home

True teams up with ZTE to build Thailand's first FTTR community

scenarios and enterprise settings such as hotels and retail establishments. In light of the building structure, the three parties decided to deploy one main ONT and 26 room ONTs on each floor, ensuring that every room had an independent room ONT. The deployment locations of the FTTR devices were meticulously selected by ZTE engineers to guarantee full Wi-Fi signal strength in every corner of the rooms. Leveraging the high-bandwidth capability of the FTTR devices, True Online provided 2 Gbps high-speed packages to boost the Internet experience and service satisfaction of guests significantly.

With the ZTE FTTR solution, Wyndham Royal Lee Phuket achieved its goal of offering high bandwidth and premium Wi-Fi. The deployed solution has four highlight features: gigabit-plus bandwidth and superfast services, full coverage and no dead zones, smart roaming and seamless handover, as well as large space and multiple connections.

- **Gigabit-plus bandwidth and superfast services:** The solution employs a combination of optical fiber and Wi-Fi 6 technology to remove the performance constraints of the traditional Ethernet cable and prevent speed degradation, ensuring smooth gaming, video streaming, and conferencing experiences.
- **Full coverage and no dead zones:** Fiber extends from the information box to every room to implement whole-home Wi-Fi coverage. The high data rates and strong anti-interference capabilities of the solution ensure that every corner of the home enjoys full signal strength.
- **Smart roaming and seamless handover:** An in-house smart roaming algorithm enables millisecond-level imperceptible roaming handover without disruptions as guests move about in the hotel building.
- **Large space and multiple connections:** Each floor has a full-coverage network comprising up to 26 Wi-Fi hotspots and connecting a maximum of 256 smart devices.

Creating a New All-Optical Hotel Experience

True Online's 2 Gbps bandwidth packages, coupled with ZTE's FTTR solution, have transformed the Internet experience offered to guests at Wyndham Royal Lee Phuket. Wyndham Royal Lee Phuket has expressed great satisfaction with the deployment and plans to deepen collaboration with True and ZTE to further enhance hotel network technology and provide its guests with a better service experience.

True Online's FTTR service launched in Thailand has garnered significant industry attention due to its versatile applications, marking it as a new growth area for the operator. As the leader in Thailand's FBB market, True is committed to strengthening its collaboration with ZTE in exploring new technologies and solutions, expecting to bring the Thai people an even more dynamic and enriching digital life experience. This innovative service has sparked excitement among consumers and industry experts alike, positioning True as a pioneer in delivering cutting-edge telecom solutions tailored to meet the evolving needs of the Thai market. ZTE TECHNOLOGIES

To enable connectivity and trust everywhere