JUN 2026 | VOL. 28 • NO. 3 • ISSUE 224

ZTE TECHNOLOGIES

Expert Views

Current Status and Future Evolution of L4 Autonomous Network in the AI-Native Era

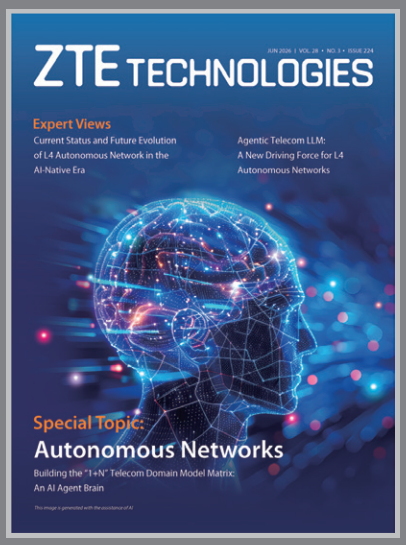
Agentic Telecom LLM: A New Driving Force for L4 Autonomous Networks

Special Topic:

Autonomous Networks

Building the "1+N" Telecom Domain Model Matrix: An AI Agent Brain

This image is generated with the assistance of AI



ZTE TECHNOLOGIES

JUN 2026 | VOL. 28 • NO. 3 • ISSUE 224

Sponsor: ZTE Corporation

Published by ZTE Technologies Editorial Office

General Editor: Lin Xiaodong

Deputy General Editor: Lu Dan

Editor-in-Chief: Liu Yang

Circulation Manager: Wang Pingping

Editorial Office Address: No. 55, Hi-tech Road South, Shenzhen, P.R. China

Circulation Office Address: 12F Kaixuan Building, 329 Jinzhai Road, Heifei, P.R. China

Website: www.zte.com.cn/global/about/publications

Email: magazine@zte.com.cn

Building a New Autonomous Network Ecosystem with Full-Stack AI Innovation

Zhang Wanchun

SVP of ZTE

In 2026, driven by the large-scale commercial deployment of 5G-Advanced, deeper integration of large AI models, and maturing digital twin technologies, autonomous networks are entering a critical stage of value realization. They are evolving toward a full-link intelligent closed-loop, becoming a key engine for operators' revenue growth, business innovation, and low-carbon development. The industry has also entered a critical stage of advancing toward L4 autonomous networks (AN).

The traditional "partial AI deployment" model has long constrained industry development. With AI concentrated at the O&M layer, network elements lack native intelligence, causing delayed data sensing, weak cross-domain collaboration, and inaccurate effectiveness measurement. Meanwhile, single-domain autonomous use cases have yet to scale. Intelligent network element upgrades have become key to breaking through industry bottlenecks.

ZTE's AIR Net high-level autonomous network solution establishes a full-stack AI paradigm spanning network elements, platforms, and applications. Built on the data engine, large AI model engine, and digital twin engine, it integrates the Nebula Telecom Large Model "1+N" matrix and key capabilities including knowledge graph, Co-Sight Agent Factory, and Co-Claw. Supported by network element-native intelligent hardware, the solution creates an intelligent digital employee that enables full-process intelligence from real-time perception and proactive analysis to autonomous decision-making and automatic closed-loop. It also promotes the standardization of the A2A-T protocol and upgrades traditional KPIs to a value-oriented KBI–KEI–KCI framework.

This issue presents ZTE's in-depth practices and forward-looking insights in autonomous networks. From innovative technical solutions to benchmark commercial cases, it showcases ZTE's AN L4 capabilities and large-scale deployment strengths, providing replicable paths for industry evolution.

Looking ahead, ZTE will continue to deepen full-stack AI innovation and collaborate with global partners to build an open and win-win autonomous network ecosystem. Leveraging high-level autonomous intelligence capabilities, ZTE will support operators' digital transformation and contribute to the high-quality development of the digital economy.

CONTENTS

Expert Views

Success Stories

Special Topic: Autonomous Networks

Current Status and Future Evolution of L4 Autonomous Network in the AI-Native Era

Guan Kai

Director of Data Intelligence & Services Product Planning, ZTE

Jiang Xianzhong

Chief Engineer of Data Intelligence & Services Product Planning, ZTE

Autonomous networks (AN) are entering a critical stage toward L4 high-level autonomy, representing a key inflection point for the communications industry to shift from the "connectivity era" to the “AI-native era.” With the large-scale commercialization of 5G-Advanced and accelerated R&D of 6G, network nodes, service scenarios, and data traffic are growing exponentially. The traditional "manual + script" O&M model can no longer meet the operational demands of complex networks, becoming a major bottleneck for industrial upgrading. Meanwhile, the rapid iteration of new technologies such as large AI models, digital twins, and agentic AI is expanding the capability boundaries of traditional networks, driving their evolution from passive connection carriers to active intelligent service entities.

According to TM Forum, over 60% of the world’s leading operators have deployed L3 autonomous capabilities. Pioneers including China Mobile, Deutsche Telekom, Vodafone, and China Telecom have achieved L4 closed-loop operation in multiple high-value scenarios, signifying that autonomous networks have moved from concepts to large-scale deployment. Faced with the declining traffic dividends and the dilemma of "traffic growth without revenue increase", traditional models built on scale and cost efficiency are reaching their limits. Improvements in O&M efficiency alone are insufficient to offset stagnant revenue growth, and the core demand for autonomous networks has shifted from cost reduction to revenue generation.

Current Status of L4 Autonomous Networks

In 2025, industrial practices of L4 autonomous networks have broken through the limitations of traditional "automated scripts + rule engines", enabling a leap from "isolated intelligence" to "systematic intelligence". Based on TM Forum standards and practices of global top operators and vendors, the L4 phase now shows three core features.

First, AI deployment is evolving from scattered applications to systematic collaboration, with a full-lifecyle closed-loop as a key feature. L4 autonomous networks do not rely on the stacking of isolated AI modules; instead, they establish an autonomous, end-to-end closed-loop of perception, analysis, decision, and execution. This has been verified in the practices of leading global operators. China Mobile has achieved automated scheduling of 300 Tbps of daily IP network traffic and built cross-domain autonomous closed-loops based on its self-developed "Jiutian" network large model, reducing MTTR by 25%. Deutsche Telekom, in collaboration with Google Cloud, launched MINDR, a multi-agent system. Its early-deployed RAN Guardian Agent reduced the time needed to manage major events from hours to around a minute, an improvement of more than 95%. During the February Carnival season, it pre-checked 611 mobile sites and enabled automatic peak-load adjustments. Vodafone partnered with Cyient to build the VISMON™platform, which integrates multi-market and multi-vendor data to realize multi-agent collaborative optimization for radio access networks, improving spectrum efficiency by 19% and solving the pain point of traditional fragmented network management.

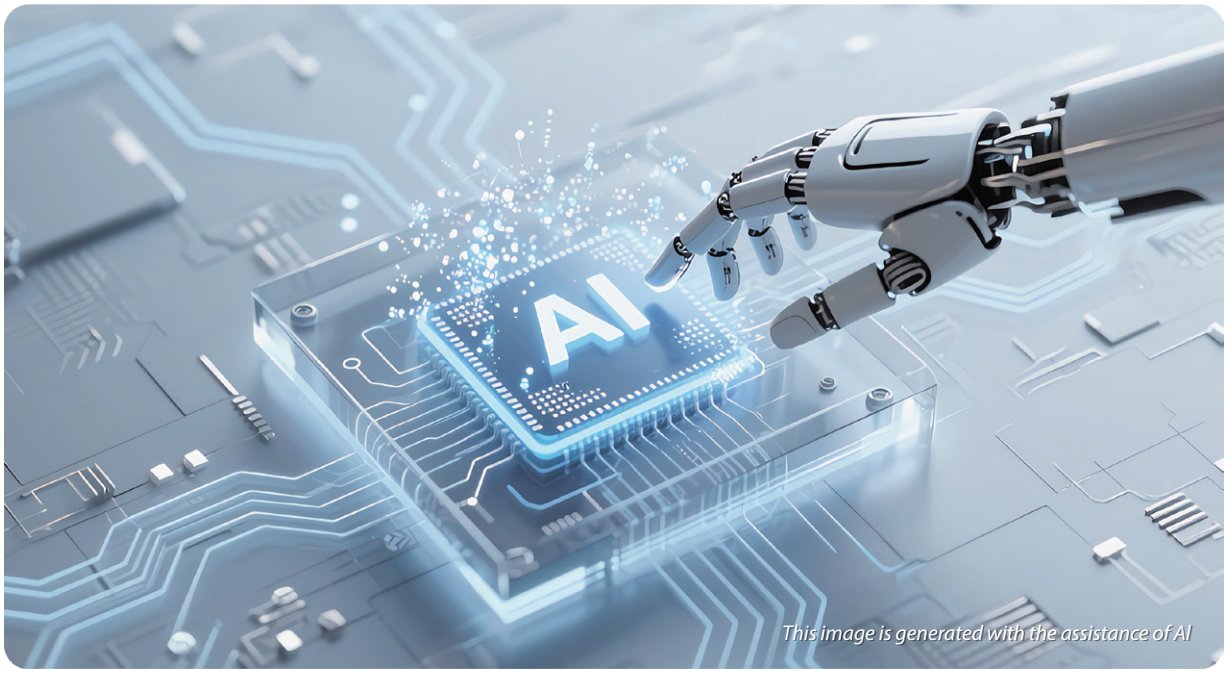
Second, agents are becoming a key enabler of AN L4 implementation, helping establish an "AI-led, human-supervised" paradigm. Unlike traditional AI that relies on manually defined rules, autonomous network L4 agents possess closed-loop capabilities, including goal orientation, autonomous planning, environmental adaptation, and continuous learning. Especially in multi-agent collaboration scenarios, they can efficiently complete complex tasks such as cross-domain fault self-healing and dynamic resource scheduling. ZTE and China Mobile jointly established the "Co-Innovation+" lab, building a collaborative system of expert-type and copilot-type intelligent digital employees, achieving over 90% accuracy in cross-domain fault root cause localization and a 21.34% reduction in MTTR. TM Forum's Agent-to-Agent for Telecom (A2A-T) protocol provides a common language for agents across vendors and domains, reducing cross-domain integration cycles from months to days, driving multi-agent interaction from "manual integration" to "adaptive integration".

Third, the value measurement system is evolving from O&M KPIs to an operations-oriented KBI-KEI-KCI system, shifting focus from cost to profit. Operators' goals are also expanding from basic O&M indicators such as shorter fault handling time and lower labor costs to the monetization of network capabilities. China Mobile Guangdong used NWDAF-based hierarchical assurance to provide low-latency cloud gaming packages for VIP users and improve Douyin data rates by 35.6%, directly driving ARPU growth. China Mobile Fujian's home broadband complaint agent achieved over 90% demarcation accuracy and improved customer satisfaction by 22%. China Telecom empowered vertical industries through network large models, transforming autonomous capabilities into differentiated service advantages and winning multiple TM Forum innovation awards. These clearly show that AN L4's commercial value lies not only in improving O&M efficiency, but also in enabling breakthroughs in operational revenue growth.

Future Trends of L4 Autonomous Networks

The real leap toward L4 AN lies in the integration of three paradigms: large models as the "cognitive brain," agents as the "execution hub," and digital twins as the "decision sandbox." Large models overcome the limitations of traditional rule engines by enabling natural language understanding, expert experience accumulation, and long-horizon reasoning, driving the shift from "executing instructions" to "understanding intent." Agents are goal-oriented with closed-loop perception, decision, execution and learning. Multi-agent collaboration supports cross-domain fault self-healing and dynamic resource scheduling. Digital twins provide a safe environment for validating AI decisions, ensuring reliability and reversibility through simulation.

Based on TM Forum standards, the practices of global leading operators, and the natural trajectory of technological evolution, autonomous networks are expected to evolve along three core trends over the next 3-5 years. Built on the three technical paradigms



This image is generated with the assistance of AI

of large models, agents, and digital twins, these trends will reshape the underlying logic of network intelligence, advance the large-scale deployment of autonomous networks, and drive the industry toward high-quality development. The three core trends are as follows:

Trend 1: Architecture Evolves from Partial Intelligence to Full-Stack AI, with NE-Native Intelligence as the Strategic Core.

Today, most vendors' autonomous solutions focus on AI at the NMS layer, resulting in delayed perception, slow response and cross-domain fragmentation. The essence of full-network L4 lies in NE-native intelligence: Every NE becomes an intelligent agent node with local decision-making, forming a distributed intelligent collaboration system and breaking away from centralized management.

ZTE proposes a four-layer agent architecture—NE, Network, Service, and Business—representing full-stack AI. In the future, the network will no longer be a "managed object", but a "living entity" with self-regulating capabilities. For example, wireless base stations equipped with built-in intelligent hardware and lightweight structured large models can perceive user mobility trajectories in real time and dynamically adjust beams. Integrating OTDR functions into optical modules enables second-level fault localization of fronthaul links, while embedded gSDU units support load-aware automatic on/off, achieving "zero load, zero power consumption".

Going forward, those that first achieve NE-level AI-native capability and embed intelligence into every network node will gain a competitive edge in L4 high-level autonomous networks. This is also the inevitable direction for the architectural evolution of autonomous networks.

Trend 2: Operations Leap from O&M Efficiency to Revenue Growth, with Network-as-a-Service (NaaS) Unlocking a Second Growth Curve.

Against the backdrop of diminishing traffic dividends and intensifying homogeneous competition, operators are shifting from scale-driven to value-driven growth, with L4 autonomous network capabilities serving as a key enabler. The operational focus of autonomous networks is also shifting toward experience monetization. Networks will evolve from a bandwidth pipeline to a tradable, customizable intelligent service offering. Three major monetization paths are gradually taking shape.

- First, monetization through differentiated experience packages. For example, customized

services, such as low-latency guarantees for cloud gaming, uplink acceleration for live streaming, and SLA visualization for industrial leased lines are offered to different user groups, with charging based on experience quality.

- Second, monetization via vertical industry empowerment. In sectors such as the low-altitude economy and the industrial Internet, 5G-A integrated communication and sensing networks provide drones with integrated “communication + sensing + navigation” services, with charging based on route and duration.
- Third, monetization through data elements. After desensitization, network data can empower scenarios such as automaker site selection, retail heat analysis and financial risk control, building a “data bank” that unlocks data value under compliance.

The NWDAF-driven tiered and graded experience assurance solution jointly launched by ZTE and China Mobile Guangdong delivers an average of 3.65 assurance actions per user per day, providing personalized protection for tens of millions of users. This demonstrates the feasibility of turning network capabilities into commercial product and provides a replicable path for operators’ business transformation.

Trend 3: The Ecosystem Is Evolving from Single-Vendor Closed-Loop to Open Collaboration, with Standards Shaping Industry Competitiveness.

L4 autonomous networks span multiple domains including radio access, core network, transport, IP, services, and business. No single vendor can provide full-stack capabilities, so open collaboration has become essential. Over the next three to five years, competition in autonomous networks is expected to shift from products to ecosystems, with standards playing the key enabling role.

Global standardization is accelerating. TM Forum is promoting multi-agent collaboration and the A2A-T protocol, addressing coordination challenges through structured templates, event subscription, and fine-grained authorization. Leading operators, including China Mobile, Vodafone, and Orange, are actively involved. CCSA is advancing standards for large models and agents, building a "general + specialized" framework. 3GPP is enhancing AI security specifications, providing a foundation for AI-native security and standardized safeguards for trusted agent interactions and global collaboration.

ZTE, together with China Mobile, has established the "Co-Innovation+ Autonomous Network Lab" to explore cross-vendor technical cooperation. It promotes the compatibility of the A2A-T protocol with mainstream frameworks such as MCP and ACP and carries out practical verification to enable "plug-and-play" for cross-vendor agents. Looking ahead, those who build an open agent factory, shared test suites, and a unified governance framework will lead the ecosystem of autonomous networks, shaping the future of large-scale, trusted, plug-and-play cross-vendor and cross-domain agents.

Conclusion

The large-scale deployment of L4 autonomous networks marks the industry’s transition from an era dominated by manual operations to a new AI-native stage. We are at a historic inflection point: AI has evolved from a network add-on tool into the neural center spanning NEs, architecture and operations. The network is no longer a passive pipeline but a digital living entity powered by agent collaboration, capable of perception, reasoning, decision-making, and evolution.

Autonomous networks will follow three pillars—full-stack AI, open collaboration, and value-driven development—enabling real-time perception, autonomous decision-making, and continuous evolution. L4 is not the destination, but the starting point for autonomous networks to evolve toward higher-level intelligence. We call on global operators, vendors, standards organizations and vertical industry partners to collaborate in deepening full-stack AI innovation, building an open and win-win ecosystem for autonomous networks, and advancing autonomous networks toward higher intelligence, empowering the high-quality development of the digital economy and industries. ZTE TECHNOLOGIES

Agentic Telecom LLM: A New Driving Force for L4 Autonomous Networks

Gao Yanqin

Chief R&D Planning Engineer for AIM Telecom Large Model and Agents, ZTE

Du Yongsheng

Chief Technical Engineer for AIM Telecom Large Model and Agents, ZTE

AI is rapidly evolving from general-purpose generative LLMs toward agentic AI systems equipped with autonomous decision-making, environmental interaction, and multi-agent collaboration capabilities. In professional scenarios such as telecom networks, which feature strong engineering attributes, strict constraints, and closed-loop operations, agentic capabilities have become key to technology deployment and value creation.

Meanwhile, autonomous networks are transitioning from L3 to L4 autonomy. A dual-model collaborative architecture, combining agentic telecom LLMs and domain-specific structured data models, is becoming a key technical path toward L4 autonomy and self-governance.

Paradigm Shift of Autonomous Networks: Toward Agentic L4 Closed-Loop

AI systems are undergoing a profound paradigm shift from passive response to proactive execution, and from single-model inference toward multi-agent collaborative evolution (Fig. 1). Autonomous agentic AI frameworks such as Manus, Co-Sight, and OpenClaw are emerging, enabling capabilities such as goal decomposition, persistent memory, tool invocation, and environmental adaptation, marking the advent of the agentic AI era.

Agentic AI systems centered on LLMs and agents are also shifting from traditional object- and service-oriented design toward capability- and agent-oriented design. Under this paradigm, reusable reasoning becomes a new factor of production; iterative evolution drives intelligent growth; and collaborative interaction enables deep integration between models and real-world environments. The key challenge is to enable continuous intelligence evolution while ensuring rigorous and reliable agentic execution.

For autonomous networks to advance to L4, they must achieve unattended operation and predictive automatic closed-loop control in target task

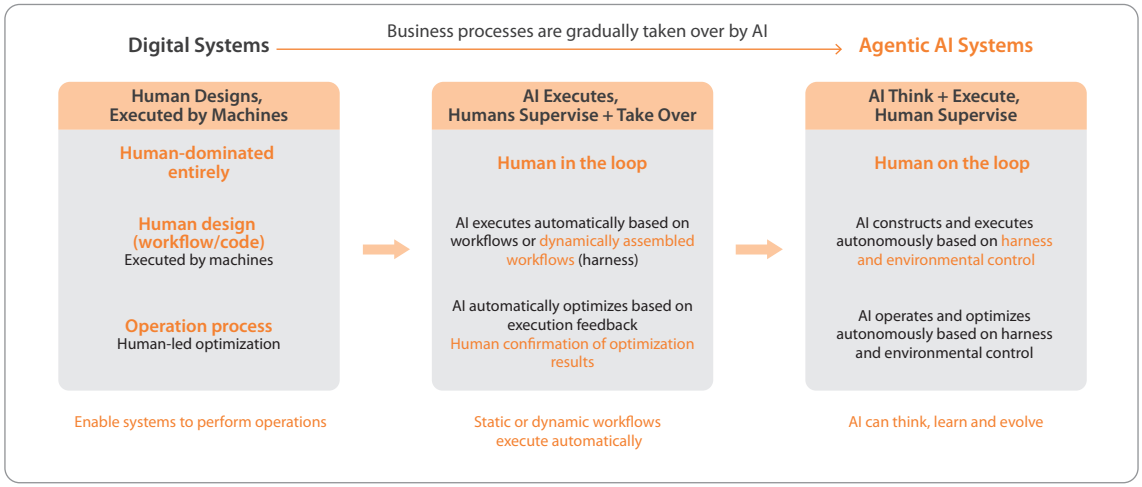


Fig. 1 The evolution path of agentic AI systems.

scenarios. This requires six key capabilities: autonomous perception, intelligent analysis, cross-domain decision-making, automatic execution, effect verification, and continuous evolution. Traditional LLMs can hardly satisfy both unstructured intent understanding and highly reliable structured data processing in the communications field, nor meet the requirements of low latency, high security, and lightweight deployment in production scenarios.

Therefore, agentic LLMs for communication production environments should adopt a dual-model collaborative architecture. In this framework, the agentic LLM is responsible for intent understanding, knowledge reasoning, natural interaction, and task planning, while the structured data model handles highly structured information such as network topology, KPI indicators, signaling flows, and configuration rules. Together, they form a reliable autonomous network core and provide a new paradigm for achieving the L4 generational leap.

Core Technologies of Agentic Telecom LLMs

Autonomous networks must support intent-driven policy coordination across RAN, FN/BN, Core, and cloud-edge domains, while ensuring stable convergence within minutes to hours. This requires unified cross-domain intent semantics, constrained policy solving, replayable execution traces, and KQI/KPI-based closed-loop verification. Therefore, agentic AI for autonomous network scenarios must meet strict requirements such as low cost, low latency, high reliability, easy deployment, and iterability. Under these constraints, an integrated solution is formed through lightweight design, domain-specific optimization, dual-model collaboration, and multi-agent autonomous cooperation (perception–analysis–decision–execution–verification). Combined with controllable tool-based scheduling and a continuous evolution mechanism, it supports the practical engineering deployment of L4 autonomous network capabilities.

Key technologies of agentic telecom LLMs include:

- **Autonomous agent planning for communication scenarios**

Decompose tasks based on communication protocols, signaling, and O&M processes, supporting a closed loop of perception–decision–execution–feedback and the autonomous generation of execution logic for NE interaction, scheduling, and fault handling.

- **Multi-agent communication and collaboration**

Define standardized communication interfaces and negotiation mechanisms for cross-domain and cross-NE agents, supporting task division, state synchronization, and resource scheduling.

- **Agent-based communication context awareness and memory**

Perceive network status, session links, and topology in real time, support both short-term

session memory and long-term experience accumulation, and adapt to continuous communication sessions and dynamic scheduling requirements.

- **In-depth integration of telecommunication domain knowledge**

  Incorporate communication protocols, 5G/6G architecture, O&M specifications, and signaling flows to construct high-quality annotated data and domain knowledge.

- **Trustworthiness, controllability, security, and compliance**

  Establish agent communication authentication, data encryption, and behavior auditing mechanisms to meet telecom security, privacy, and regulatory requirements.

- **End-edge-cloud collaborative deployment architecture**

  Adapt to the telecommunication network architecture, with cloud-based training and iteration, as well as lightweight inference locally or at the edge, balancing model capability with deployment resource constraints.

- **Scenario-based closed-loop iteration**

  Continuously improve task execution in real-world scenarios such as intelligent O&M, agent collaborative communication, fault self-healing, and network optimization.

A dual-model collaboration mechanism is established between the agentic LLM and the structured data model. The LLM handles natural language intent parsing, execution flow planning, and decision semantic output, while the structured data model performs predictive analysis, decision evaluation, and rule matching based on structured network information. A unified scheduling engine enables low-latency interaction between the two models, ensuring that both decision accuracy and response efficiency meet requirements in production scenarios.

A closed-loop mechanism supports real-time network data backflow, incremental training, performance evaluation, and automatic iteration. Simulation verification using digital twins effectively mitigates model degradation and hallucination issues. Long-term memory modules accumulate fault cases, optimization solutions, and collaboration experience, enabling continuous model evolution toward L4 network self-optimization, self-healing, and self-evolution.

Next-Generation Autonomous Network Operation Mode Based on Agentic Models and Its Application Scenarios

Centered on LLMs and agents, an agent architecture with autonomous closed-loop control, exploration, and continuous evolution can be established to support multi-agent distributed collaboration and full-process autonomous closed-loop operations in next-generation autonomous networks (Fig. 2). Perception agents collect network status in real time; analysis agents perform root cause localization and risk prediction; decision agents generate cross-domain optimization schemes; execution agents invoke automation tools to complete configuration delivery and resource scheduling; verification agents conduct closed-loop evaluation on execution effects; and learning agents realize knowledge accumulation and continuous model iteration.

Typical application scenarios include:

- **Autonomous network fault healing:** Perception agents capture real-time alarms such as base station outages, link congestion, and poor-quality cells. The LLM interprets fault types and impact scope, while the structured data model quickly correlates topology, metrics, and historical cases to locate root causes. Execution agents trigger cell handovers, parameter adjustments, link switching, and other operations, reducing fault recovery time by more than 80%.
- **Cross-domain intelligent resource scheduling:** For high-concurrency scenarios such as major sports events and concerts, agents across radio, bearer and core networks collaborate. The LLM parses service assurance intents, while the structured data model performs traffic prediction, timeslot allocation, and routing calculation to optimize global resource configurations, ensuring peak-period user experience.
- **Intent-driven automatic service provisioning:**

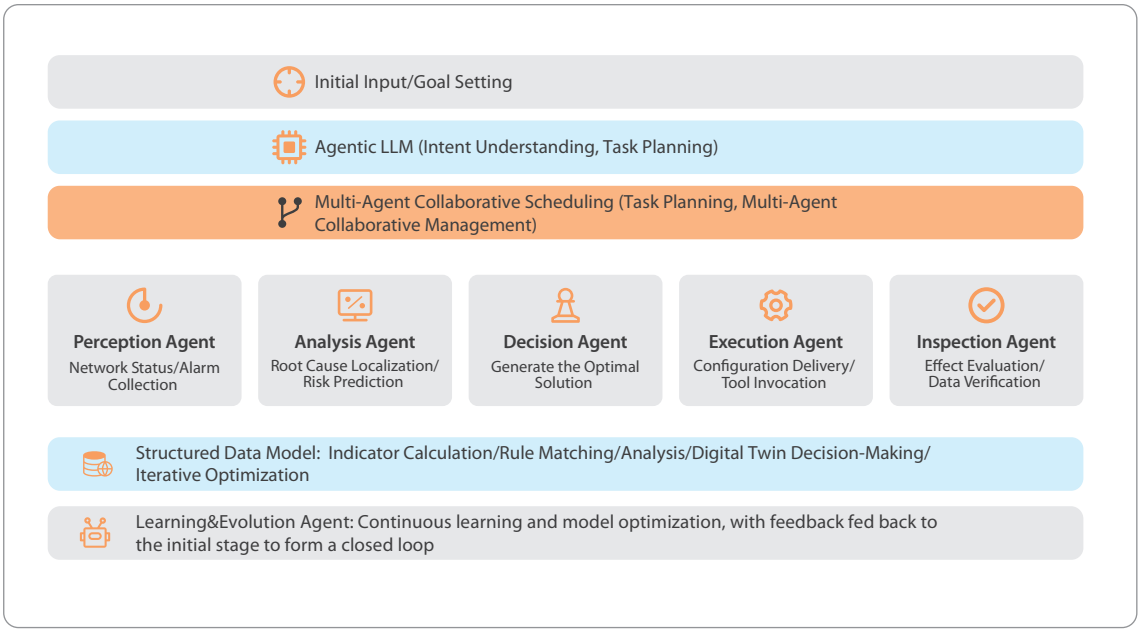


Fig. 2 Agentic dual-model and multi-agent auto-collaborative closed-loop.

Users specify bandwidth, latency and reliability requirements in natural language. The agentic LLM converts them into network-executable instructions, while the structured data model completes resource verification, network slicing orchestration, and configuration generation, enabling one-click activation and reducing provisioning time from days to minutes.

- **Autonomous energy consumption optimization:** Based on traffic patterns and equipment loads, agents dynamically adjust base station power, board status, antenna tilt, and other parameters to reduce energy consumption while ensuring user experience, achieving a balance between energy savings and network performance.

This operational mode eliminates process breakpoints and data silos in traditional O&M, shifting the network from reactive response to proactive prediction and autonomous governance, providing a feasible engineering path for L4 autonomous networks.

Vision for L4 Autonomous Networks Based on Agentic AI

Agentic telecom LLMs will reshape the capability boundaries of autonomous networks, driving L4 autonomy from pilot verification to large-scale commercial deployment. Future autonomous networks will achieve full-scenario closed-loop operations, global collaborative intelligence, and full life-cycle self-evolution, significantly reducing operating costs while improving network reliability and service provisioning efficiency.

From an industrial perspective, agentic telecom LLMs will help establish standardized agent protocols, an open tool ecosystem, and domain-specific model foundations. They will accelerate collaboration among operators, equipment vendors, and technology providers, and promote the maturity of L4 autonomous network standards and their global implementation. Toward 6G, they will serve as core intelligent units for space-air-ground-sea integrated networks, supporting the vision of future networks with ubiquitous connectivity, extreme experience, and native intelligence.

Agentic AI is far more than a simple technological upgrade; it represents a paradigm shift from digitalized and intelligent networks to autonomous networks. With agentic telecom LLMs as the core engine, autonomous networks will advance to L4 autonomy, opening a new chapter in the intelligent transformation of the communications industry. ZTE TECHNOLOGIES

Building the "1+N" Telecom Domain Model Matrix: An AI Agent Brain

Zheng Peng

Vice General Manager of RAN Products, ZTE

Zhang Wenshuan

Chief Engineer of Agentic AI Architecture for Autonomous Networks, ZTE

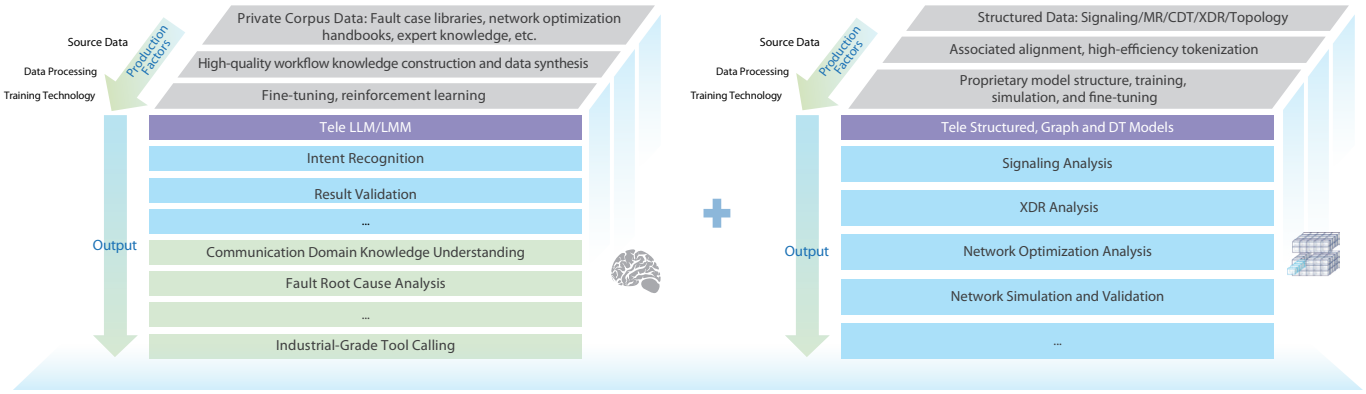
Jin Ningdi

Chief Engineer of Digital Twin Planning for Autonomous Networks, ZTE

The core feature of L4 autonomous networks lies in end-to-end, closed-loop collaboration enabled by intelligent hardware that supports real-time perception, precise analysis, intelligent decision-making, and automated execution. As a key enabler of autonomous networks, AI agents are transitioning from theoretical exploration to large-scale deployment, driving the evolution of network operation and maintenance (O&M) models from passive response to proactive prediction, adaptive optimization, and autonomous closed-loop operation.

ZTE's autonomous network AI agent, with the self-developed Nebula Telecom Large Model as its core engine, deeply integrates network domain knowledge. It is capable of autonomous planning and intelligent decision-making in complex scenarios, and can

This image is generated with the assistance of AI



▲ *Fig. 1 ZTE Nebula Telecom Model "1+ N" matrix.*

complete end-to-end closed-loop execution in high-value scenarios. The agent aims to achieve "Three-Zero, Three-Self" autonomous network goals: (zero waiting, zero failure, zero contact; self-configuration, self-repair, and self-optimization) .

Large models serve as the "super brain" of AI agents. They must be able to understand ambiguous natural-language intent, accumulate expert knowledge for root cause analysis and optimization of complex networks, and drive a paradigm shift in O&M—from "executing commands" to "understanding intent", "autonomous reasoning", and "long-term planning".

To address this, ZTE has launched the Nebula Telecom Model "1 + N" matrix (Fig. 1), where "1" represents a general-purpose large model for the communications domain, and "N" represents multiple specialized models, such as structured large models for communications, graph models, and digital twin models. The collaboration within the "1 + N" model matrix enhances the identification and handling of long-tail problems and edge cases, builds an intelligent cognitive foundation tailored to the operational mechanisms of communication networks, and supports the sustainable evolution of L4 autonomous networks.

General-Purpose Large Model for Telecom

This model adopts an innovative hybrid reasoning architecture designed for high-order reasoning and autonomous decision-making under complex business logic. By leveraging massive high-value telecom corpora—including typical fault cases, standardized O&M procedures, and technical specifications—while internalizing expert knowledge and integrating synthetic data generated from digital twin simulations, the model enables a deep integration of general generative "fast thinking" (System 1) and rule-based, symbolic logic-driven "slow thinking" (System 2).

In high-value scenarios of autonomous networks, such as the Radio Access Network (RAN), Core Network, and Transport Network, this hybrid reasoning capability demonstrates strong business intent parsing ability, complex toolchain orchestration ability, and cross-domain analytical ability. It effectively overcomes the dual bottlenecks of "insufficient domain-specific depth" and "limited logical reasoning breadth" that general-purpose large models face in the communications domain. Practical tests show that, in key tasks such as O&M metric trend analysis, user O&M intent classification, fault diagnosis tool invocation, alarm root-cause correlation, and 4G/5G network parameter optimization, accuracy improves by approximately 30% on average compared to open-source baseline models of the same parameter scale.

Structured Large Model for Telecom

Structured data in communication networks, such as signaling messages, measurement reports (MRs), XDR detailed records, and KPI metric sequences, carries core information about network operational status and user behavior, and is characterized by strong temporal properties, protocol specificity, and high field density. Although general-purpose large models can process structured data through textualization, such conversions often lead to information distortion, semantic ambiguity, and increased token consumption and reasoning latency due to sequence expansion, making it difficult to meet the stringent real-time requirements of production-grade deployments.

In response to this challenge, ZTE has developed a native structured large model for communication scenarios, using a specialized Tokenizer to achieve efficient semantic compression. It directly maps raw structured data into compact, low-redundancy token sequences, improving reasoning efficiency severalfold while ensuring data integrity. Furthermore, this model can collaborate with a general-purpose LLM through multimodal alignment mechanisms, bringing native structured data understanding capabilities into LLMs and supporting downstream generative applications such as KPI anomaly detection and signaling flow reconstruction.

Given that structured models rely on high-quality labeled data, the manual acquisition of which is costly and time-consuming, we propose a self-evolving training paradigm called "dual-brain collaborative learning" :

- **The Cognitive Brain** is based on the general-purpose large model for the communications domain, and is responsible for global cognitive coordination and deep logical reasoning. It focuses on offline processing of complex samples, knowledge distillation, counterfactual reasoning, and synthetic corpus generation, continuously providing the Execution Brain with high-quality training data and strategic guidance.
- **The Execution Brain** employs a lightweight structured large model, focusing on real-time anomaly detection, root cause classification, and optimization plan generation in specific scenarios. It supports strategy simulation and validation within digital twin environments, ensuring that decisions are traceable and evaluable.

This paradigm establishes a "cognition-execution" closed-loop evolution, significantly improving model iteration efficiency and generalization capability. In real-world root cause identification tasks for wireless network quality degradation, the identification accuracy improved by 30%, model update cycles were shortened by 50%, and data labeling efficiency increased by 50%, providing comprehensive support for network self-optimization and self-healing.

Graph Model

Graph models serve as a trustworthy external memory for large models, enabling structured, traceable, and verifiable knowledge accumulation and dynamic updates. They provide interpretable reasoning evidence chains, supporting compliance verification and audit traceability for high-risk operations. Their core technical capabilities cover two major modules: Knowledge Graph Construction and Evolution, and Graph-Augmented Hybrid Retrieval and Reasoning. Together, they build a full-chain knowledge hub that spans from enterprise multi-source heterogeneous data to intelligent applications.

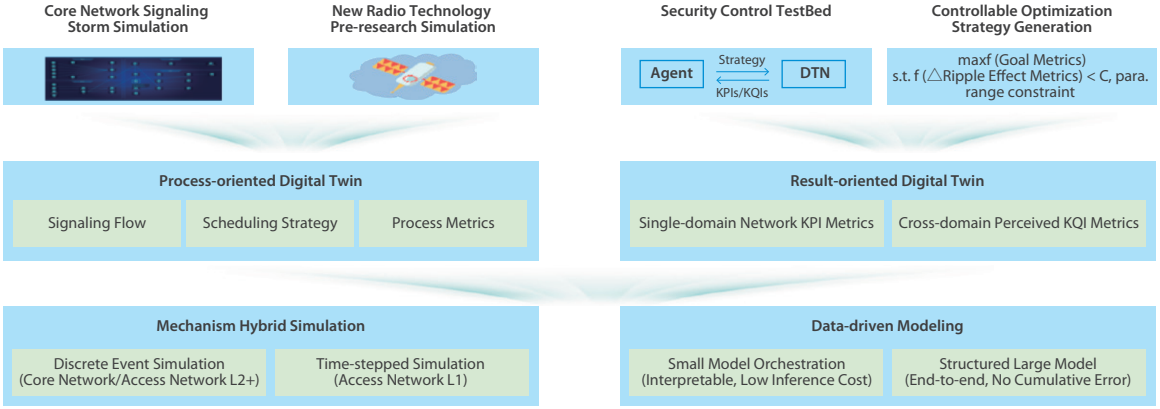
- **Knowledge Graph Construction and Evolution** breaks through traditional static knowledge extraction modes by establishing a dual-channel knowledge construction system of "static extraction + dynamic update." It introduces a self-verification and reflective evolution framework, enabling continuous self-optimization and closed-loop evolution of the knowledge graph.
- **Graph-Augmented Hybrid Retrieval and Reasoning** leverages the semantic connectivity and topological relationships of graph structures to achieve logical chaining of discrete knowledge points, eliminating reasoning gaps caused by information fragmentation.

Currently, ZTE's graph model technology has achieved end-to-end application closure in scenarios such as network intelligent optimization, automated fault diagnosis, and fault localization at China Mobile's Co-Innovation+ Lab. In wireless network fault root cause localization tasks, this technology achieved over 90% localization accuracy and reduced average fault handling time by 10%, validating its advancement and commercial feasibility in complex O&M scenarios.

Digital Twin Model

The "zero-touch" requirement of autonomous networks demands high system reliability, while the "black-box risk" of AI decisions and their potential impact on live network operations are currently the biggest obstacles to closed-loop implementation. Digital twins provide a security foundation for autonomous decision-making and serve as a critical basis for validating network decisions.

ZTE has constructed a complete panoramic



◀ *Fig. 2 ZTE's panoramic digital twin architecture.*

architecture for digital twins in the telecom domain (Fig. 2). From a technology stack perspective, a dual-driven approach combining mechanism-based and data-driven methods is adopted. On one hand, high-performance mechanism-driven simulation engines, such as discrete-event simulation and time-stepped simulation, are developed. On the other hand, innovative data-driven modeling methods are introduced, combining lightweight small-model orchestration with end-to-end structured large models. This hybrid architecture allows each technology to maximize its effectiveness in the most suitable value scenarios.

From the perspective of the objects represented by digital twins, ZTE has achieved full-chain coverage from "process" to "results." For process-oriented twins, we focus on fine-grained simulation of signaling flows and scheduling strategies, successfully applying them to complex scenarios such as signaling storm simulation and NTN air-interface simulation pre-research. For result-oriented twins, we are committed to quantitative pre-assessment of single-domain network-level KPI metrics and cross-domain perceived KQI metrics. Building on this foundation, we explore systems capable of autonomous iterative closed-loop operation, achieving flexible configuration of optimization objectives, quantifiable control of ripple effects, and precise constraints on parameter adjustments, ultimately generating truly manageable and controllable intelligent optimization strategies. Through digital twin models, we promote the transformation from "high-risk repeated manual operations" to "low-risk one-time self-optimization," achieving advanced autonomous intelligence.

Looking ahead, as generative AI matures and computing power becomes increasingly edge-deployed, we will gradually integrate digital twin and generative simulation technologies to build a world model for the telecom domain with counterfactual reasoning and implicit dynamics understanding capabilities. This model not only fully internalizes structured knowledge, such as communication protocols, channel characteristics, and network topologies, but also moves beyond a purely data-driven paradigm, enabling core capabilities such as millisecond-level forward-looking spatial reasoning and direct output of optimal control strategies.

ZTE's "1+N" telecom model matrix constructs the AI agent brain for L4 autonomous networks through the deep integration of general-purpose large models with structured, graph, and digital twin models. This multi-model fusion architecture enables a closed-loop capability spanning intent understanding, multimodal analysis, trustworthy decision-making, and secure simulation, while advancing network operations from reactive response to proactive prediction and autonomous optimization. Currently, this solution has delivered tangible results in scenarios such as end-to-end fault handling across the entire domain. As world models and counterfactual reasoning technologies continue to evolve, multi-model collaboration will further support the realization of the "Three Zero and Three Self" goals for autonomous networks. ZTE will continue to deepen AI-network integration and innovation, driving telecom networks toward higher autonomy. ZTE TECHNOLOGIES

Toward AN L4: Creating Digital Employees via LLM–Knowledge Graph Synergy

Gu Xiaotao

Autonomous Network Product Planning Manager, ZTE

Han Song

Autonomous Network Product Planning Manager, ZTE

The ultimate goal of autonomous networks (AN) is to transform network service delivery from human-led operations to system-led autonomy, achieving level 4 (L4) high autonomy and progressively advancing toward L5 full autonomy. AN L4 requires systems to autonomously close the loop of intent/experience, awareness, analysis, decision-making, and execution in the vast majority of scenarios without human intervention—a capability that relies fundamentally on AI empowerment. However, current AI applications in telecommunications operations still face three major bottlenecks:

- Alarm data is overwhelming and lacks structured correlation, making it extremely difficult to identify root-cause alarms.
- Fault propagation paths are highly complex, and traditional rule-based engines have limited coverage.
- AI decision-making lacks interpretability, hindering trust from operations personnel.

In response to the above bottlenecks, the industry has begun exploring AI agent systems endowed with human-like cognition and collaboration capabilities. China Mobile, in collaboration with ZTE, has pioneered the concept of "Digital Employee" and applied it in real-world deployments. Centered on the deep synergy between the large language model (LLM) and knowledge graphs, this architecture enables a transition from "statistical prediction" to "causal reasoning."

Solution: Collaboration Between LLM and Knowledge Graphs

Based on two static knowledge graphs—resource topology and alarm propagation (Fig. 1)—real-time alarms are integrated to form a dynamic root-cause reasoning graph, which is pruned via graph search algorithms to generate the minimal root-cause reasoning subgraph (Fig. 2). Finally, the minimal root-cause reasoning subgraph is converted into a prompt and fed into the LLM to enable precise root cause localization, delivering three major benefits: accuracy, stability, and timeliness.

- Accuracy is reflected in the graph's ability to accurately represent the existing network failure conditions and to use the pruned subgraph with higher relevance as the LLM's input.
- Stability refers to unified data representation, enhanced by prompt optimization, which provides consistent inputs to the LLM and suppresses hallucinations.
- Timeliness is demonstrated through real-time alarm updates, dynamically delivering the latest network fault situations to the LLM.

The key components of this solution are described as follows:

- **Resource topology knowledge graph:** Fine-grained modeling of cross-domain objects (such as cells, RRUs, boards, optical ports, equipment, and rooms) and their physical/logical connections to create a digital map of the network.
- **Alarm propagation knowledge graph:** Leverages the LLM to automatically extract knowledge and

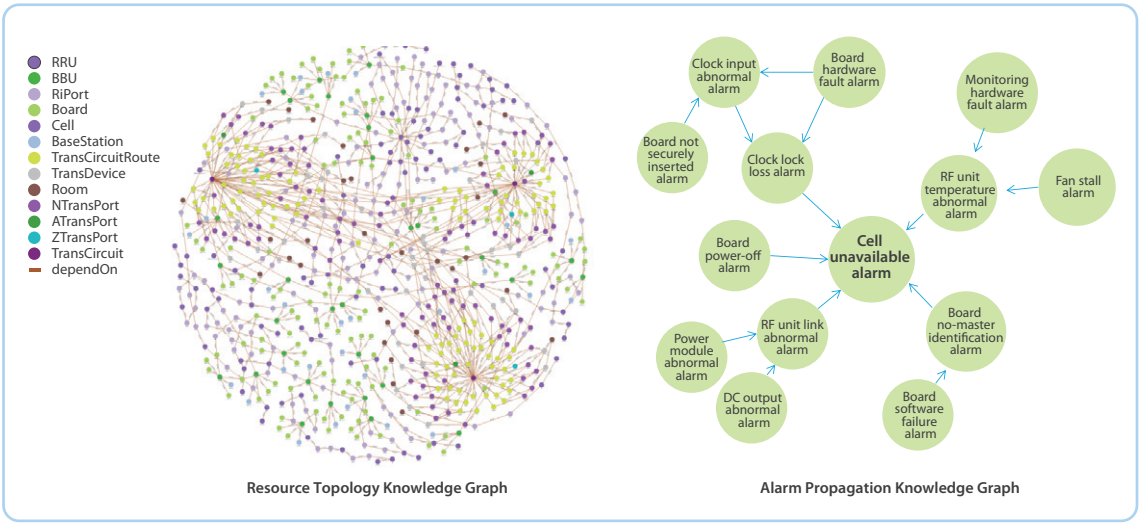


Fig. 1 Resource topology knowledge graph and alarm propagation knowledge graph.

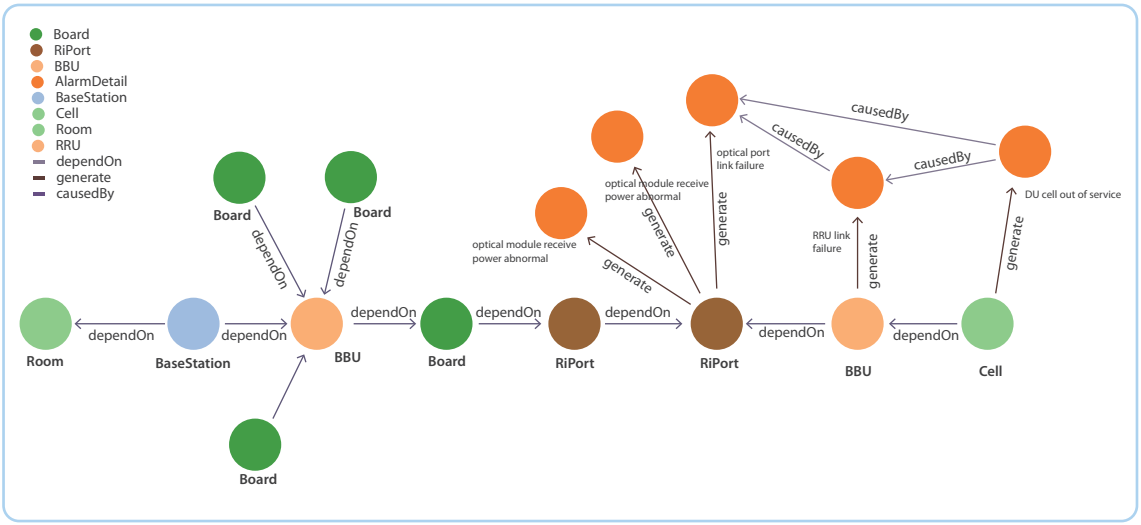


Fig. 2 Minimal root-cause reasoning subgraph.

all possible propagation relationships from vast alarm manuals and historical case documents.

- **Dynamic root-cause reasoning graph:** Integrates static topology with real-time alarms to form a dynamic root-cause reasoning graph, serving both as a real-time dashboard for network fault paths and a fact database for LLM reasoning.
- **Minimal root-cause reasoning subgraph:** Uses graph search algorithms to prune and compress the dynamic root-cause reasoning graph, generating a minimal root-cause reasoning subgraph with higher fault relevance.
- **LLM reasoning:** Uses the LLM as the decision-making center to perform deep reasoning based on minimal root-cause reasoning subgraph, precisely outputting fault root causes, fault locations, affected network elements, and solutions. Meanwhile, engineering optimizations—such as prompt refinement, input token compression, and concurrent queue management—enhance model performance.

Practice: Embedded into Operators' End-to-End O&M Processes to Improve AI Decision-Making Accuracy

In the "Co-Innovation+" Autonomous Network Lab jointly built by China Mobile and ZTE, the intelligent decision-making "Digital Employee" focuses on the high-value autonomous network scenario of fault

handling. Grounded in knowledge graphs and powered by an LLM, it builds an AI-driven decision-making system. This successfully transforms the traditional, inefficient processes that rely on human expertise and rule-based matching into an automated, high-precision, and interpretable AI decision-making workflow. The key breakthrough lies in using knowledge graphs to address inaccurate and unstable inputs, while leveraging the LLM to tackle complex reasoning challenges. The synergy between these technologies enables root-cause identification accuracy to exceed 90%. Below, we dissect this process layer by layer across three critical stages.

- **Fault Knowledge Graph Construction**

In traditional processes, fault alarms originate from multiple vendors and different domains (radio, transmission, power), resulting in inconsistent formats and ambiguous semantics. Manual analysis requires extensive review of equipment manuals and historical work orders, heavily depending on expert experience.

Introducing a knowledge graph solution can significantly address the above historical issues. A resource topology graph extracts 13 types of network entities (e.g., cell, RRU, optical port, board, machine room, transmission route) to build cross-vendor, cross-domain topological relationships. The alarm propagation graph leverages an LLM to automatically parse over 200 equipment alarm manuals, extracting over 12,000 alarm propagation relationships—such as "RRU link failure → cell outage." Graph neural networks (GNNs) are used to infer implicit causal relationships like "optical port link failure → RRU link failure." Further, a dynamic root-cause reasoning graph precisely maps real-time alarms to corresponding nodes in the resource topology, seamlessly integrating static topology with dynamic alarms to form a cross-vendor, cross-domain fault propagation knowledge base.

Standardizing inputs by unifying heterogeneous alarms and resource data into structured graph nodes and edges makes the LLM's input interpretable and traceable.

- **Minimal Root-Cause Reasoning Subgraph Generation**

In traditional processes, a single cell outage alarm may be associated with hundreds of irrelevant alarms. Manual filtering is time-consuming and error-prone. Feeding raw alarm streams directly into an LLM causes token explosion, reasoning confusion, and frequent hallucinations.

Introducing a graph search solution enables identification of more relevant fault paths. Based on a dynamic root-cause reasoning graph, it identifies all possible propagation paths and pinpoints the network elements affected by the fault. Graph search algorithms perform depth-first traversal combined with dynamic pruning strategies to eliminate redundant nodes and irrelevant paths, generating a minimal root-cause reasoning subgraph.

The reasoning scope is significantly narrowed—


This image is generated with the assistance of AI

> The "Co-Innovation+" Autonomous Network Lab embedded the "Digital Employee" into fault-handling workflows by leveraging the synergy between the LLM and knowledge graphs, achieving a breakthrough in AI decision-making accuracy

the LLM focuses on the most likely fault paths, filtering out 90% of irrelevant information in practice. By integrating alarms, topology, and propagation relationships into a structured graph, token consumption is reduced by 20% compared to previous approaches.

- **LLM Root-Cause Reasoning**

In traditional processes, fixed rule libraries cannot cover "long-tail faults" (e.g., multi-point concurrency, cross-domain coupling). Even with LLM technologies, high input noise and missing context result in an overall accuracy rate below 60%.

Introducing an LLM + Graphs collaborative solution can significantly improve reasoning accuracy. For input optimization, only the pruned root-cause reasoning subgraph—not raw alarm logs—is fed into the LLM. For prompt engineering, a prompt scoring system is established, with prompt templates customized by vendor and domain. A real-world network case dataset of over 1,400 entries from the Changping and Chaoyang districts of Beijing is built and continuously used for training and iteration. A "fast-slow thinking" mechanism is implemented: simple scenarios (e.g., single-point power outage) use a structured, rapid-response mode, while complex scenarios employ deep reasoning.

Root-cause localization accuracy exceeds 90%, with 91.7% achieved in Changping and 90.4% in Chaoyang District. Explainability is enhanced: outputs include "root-cause network element," "fault location," and "resolution recommendations," each mapped precisely to graph paths for easy human verification.

Summary and Outlook

In 2025, the "Co-Innovation+" Autonomous Network Lab embedded the "Digital Employee" into fault-handling workflows by leveraging the synergy between the LLM and knowledge graphs, achieving a breakthrough in AI decision-making accuracy. Rather than relying on a single new technology, this success was driven by an operator-centric business process, transforming the paradigm from "people searching for information" to "intelligent systems identifying root causes," setting a verifiable and replicable benchmark for achieving AN L4.

In 2026, development will continue along both technical and business fronts. On the technical side, the network graph model will serve as the core, with cross-vendor compatibility and synergy among knowledge graphs, graph search, and LLM further enhanced. On the business side, building upon the existing macro station connectivity scenario, the system will expand into indoor distribution connectivity scenarios to improve analytical coverage, while extending its scope to transmission networks, with a focus on pinpointing root causes for high-level transmission fault tickets. Concurrently, based on improved root-cause identification, the system will enhance fault resolution recommendations and further integrate into the fault handling workflow, accelerating the evolution of fault handling across all scenarios toward L4 autonomous networks. ZTE TECHNOLOGIES

1 Expert + 2 Copilots: AIMind Redefines Fault Management Experience

Zhao Song
Service Tools Product Planning Manager, ZTE

Li Daoru
Service Tools Product Planning Engineer, ZTE

Amid accelerating digital transformation, increasingly complex network architectures, and rising demands for business continuity, traditional operations and maintenance (O&M) models face unprecedented challenges: alert fatigue, difficulty identifying root causes, delayed responses, heavy reliance on manual intervention, and inefficient cross-domain collaboration—resulting in persistently high mean time to repair (MTTR) and escalating O&M costs.

To address these industry pain points, ZTE launched the AIMind cross-domain fault agent, introducing an innovative "1 Expert + 2 Copilots" intelligent O&M paradigm centered on a Cross-Domain Fault Expert agent, working synergistically with two supporting agents: the NOC Copilot and the field maintenance engineer (FME) Copilot (Fig. 1). Together, they create a closed-loop intelligent O&M system spanning the entire lifecycle—from perception, analysis, and decision-making to execution, evaluation, and feedback—enabling a fundamental leap from reactive response to proactive prediction and from isolated operations to collaborative intelligence, thereby redefining the standard for network operations.

Cross-Domain Fault Expert: The Intelligent Core of Fault Closure

The Cross-Domain Fault Expert is guided by the network event management process and leverages large language model (LLM) capabilities to integrate key functions—perception, analysis, decision-making, execution, and evaluation—within the event closure lifecycle. Using a multi-agent collaboration mechanism, it enables end-to-end closed-loop management of network faults. The system comprises four core agents: identification, analysis, scheduling, and evaluation, covering the complete workflow from event detection to closure validation.

By integrating AI models and knowledge graphs, the system intelligently detects network anomalies, enhancing fault discovery efficiency. Employing chain-of-thought reasoning, it accurately pinpoints root causes and enables precise fault isolation. Powered by an intelligent decision engine, it autonomously generates resolution strategies and dispatches work orders, significantly reducing response times. Through multi-dimensional validation mechanisms, it ensures effective resolution and confirms true fault remediation. Meanwhile, it automatically extracts and codifies fault insights, forming a knowledge feedback loop that continuously enhances intelligence.

The Cross-Domain Fault Expert transforms traditional fault handling through intelligent technologies, delivering comprehensive O&M value.

- **Fault detection phase:** The system combines AI models with rule-based recognition, drastically reducing fault identification time. For complex scenarios spanning multiple disciplines and network elements, AI recognition significantly outperforms traditional rule-based methods.
- **Fault analysis phase:** The system leverages chain-of-thought reasoning and expert troubleshooting experience, together with a knowledge graph, to analyze fault propagation paths. This improves root cause localization accuracy, reducing manual analysis time from 30–60 minutes to just 3–5 minutes.
- **Dispatch and handling phase:** The system makes intelligent scheduling decisions based on the

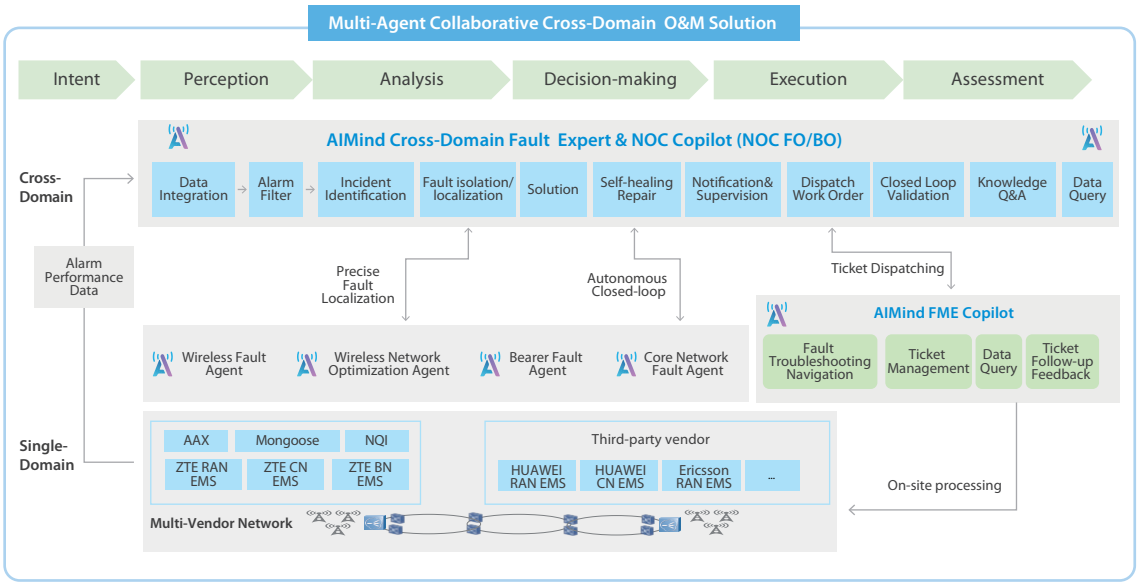


Fig. 1 AIMind "1 Expert + 2 Copilots" reshapes the experience of fault management.

Co-Sight framework, automatically generating handling strategies, dispatching work orders, and issuing commands. This cuts the average dispatch response time from 10 minutes to under 3 minutes.

- **Effectiveness evaluation phase:** The system ensures true fault closure through multi-dimensional automated verification and automatically generates fault reports, reducing manual data entry workload by over 80%. More importantly, with continuous learning via knowledge recycling and model fine-tuning, the system improves performance over time, forming a positive feedback loop that makes it smarter the more it is used.

NOC Copilot: A Unified Entry Point Empowering the Operations Team with an "Intelligent Secretary"

As the core human-machine interaction mechanism for operations personnel, the NOC Copilot deeply integrates backend multi-agent capabilities to deliver a unified interactive experience for data queries, knowledge inquiries, and event handling. Accessible via PC and mobile app, it adapts to the usage scenarios for different roles, including monitoring supervisors, monitoring staff, and expert support, enabling operations personnel to stay informed about the status of incident handling anytime, anywhere.

The NOC Copilot provides three core functions: event monitoring, knowledge-based Q&A, and data querying, achieving intelligent operations and maintenance through natural language interaction.

- **Event monitoring and collaborative response:** The system delivers end-to-end event monitoring and collaborative response capabilities. Events are categorized into four lists based on response stages: active, pending scheduling, scheduled, and archived, enabling operations staff to quickly locate events of interest via PC or mobile devices. During the response process, operators can issue critical scheduling commands directly through natural language dialogue. The information dissemination feature pushes event notifications, response progress, or risk alerts to designated roles or groups, while the task follow-up initiates manual reminders for delayed or overdue tickets, automatically linking them to responsible parties and SLAs. Additionally, the ticket assignment feature intelligently recommends or automatically creates and assigns response tickets to the appropriate teams based on event type and impact scope.
- **Knowledge Q&A capability:** Leveraging large model-based knowledge Q&A, this feature enables operations staff to instantly access professional knowledge on-site. It integrates telecommunications fundamentals, including general technologies, equipment and network

knowledge; fault knowledge, covering symptoms, causes, inspection methods, and resolution techniques; and emergency response plans addressing service quality issues. The system utilizes retrieval-augmented generation (RAG) to retrieve relevant knowledge from the knowledge base and generate responses using large models. It supports multi-turn dialogues, follow-up questions, and cross-document correlation analysis.

- **Data query capability:** This feature supports natural language queries for event-related data and presents results in visual formats. The query scope covers multiple dimensions, including event data, alert data, performance data, topology data, and log data. Operators can describe their query needs in natural language, and the system automatically generates query parameters, invokes appropriate tools to execute the query, processes the returned data, and presents it in easily understandable, visual formats, such as charts.

The NOC Copilot provides a unified interaction portal that integrates event monitoring, knowledge Q&A, and data query capabilities, lowering the barrier to operation. It replaces traditional UI interactions with natural language dialogue, while enabling efficient human-AI agent collaboration. The NOC Copilot also adjusts presented content based on role permissions and preferences.

FME Copilot: The Intelligent Partner for On-Site Operations

The FME Copilot is an intelligent troubleshooting support system designed for FMEs. Leveraging large AI models, it provides intelligent support for every stage of on-site operations, effectively reducing technician workload and improving efficiency. The system covers five key phases: work order acceptance, site preparation, en-route, on-site troubleshooting, and work order closure and return. It simplifies operations through natural language interaction, lowers skill barriers via knowledge Q&A and interactive troubleshooting guidance, optimizes site planning with intelligent scheduling, enhances complex problem-solving through human-AI collaboration, and reduces manual effort with automated form-filling.

- **Work order acceptance phase:** The system employs intelligent recommendation algorithms to automatically filter and recommend the most suitable work orders, considering multiple factors such as technician skills, location, and work order urgency. This improves matching efficiency and eliminates the inefficiencies associated with manual queries and experiential judgment.
- **Site preparation phase:** Intelligent planning algorithms provide optimal site sequences, route planning, and material lists, saving average preparation time. Automated material requisition further reduce the time cost of cross-departmental coordination.
- **On-site troubleshooting phase:** The work order briefing function allows technicians to quickly understand the status of a work order. Interactive troubleshooting guidance provides step-by-step instructions, enabling novices to achieve the efficiency of experienced engineers, effectively addressing skill gaps. The real-time collaboration feature allows technicians to quickly connect with back-end experts, accelerating the resolution of complex issues.
- **Work order closure and return phase:** Automated form-filling reduces manual data entry time.

The FME Copilot significantly enhances on-site troubleshooting efficiency and quality through end-to-end intelligent assistance.

From Tool to Partner: Ushering in a New Era of Intelligent Operations

AIMind "1 Expert + 2 Copilots" is an intelligent O&M system characterized by seamless collaboration, complementary capabilities, and a closed-loop data flow. The Cross-Domain Fault Expert "thinks deeply," the NOC Copilot "sees comprehensively," and the FME Copilot "acts accurately." Together, they achieve data integration, intent coordination, and experience feedback through a unified agent framework and knowledge platform.

Beyond improving fault resolution efficiency, reducing manual costs, and enhancing system stability, AIMind transforms the work experience of O&M personnel—freeing them from repetitive tasks to focus on high-value decision-making and innovation, shifting them from passive fire-fighters to commanders and collaborators of intelligent systems. ZTE TECHNOLOGIES

An End-to-End Intelligent Solution for Mobile Service Complaint Handling in Autonomous Networks

Xia Lin

Chief Engineer for Autonomous Network Solutions, ZTE

As the communications industry transitions toward digitalization and intelligence, autonomous networks (AN) have become a common goal for global operators and vendors. Mobile service complaint handling, as a key indicator of user experience, impacts operators' competitiveness and brand image.

However, current mobile service complaint handling faces three major challenges: delayed fault detection, difficult cross-domain fault localization across RAN, core network, transmission, and service platform, and a lack of quantitative evaluation for closed-loop governance. To address these challenges, ZTE has developed a comprehensive solution for mobile service complaint handling in line with the evolution path of autonomous networks. Centered on digital and intelligent empowerment, the solution establishes a full lifecycle closed-loop system covering perception, decision-making, execution, and optimization through in-depth collaboration between the VMAX cross-domain platform and single-domain agents.

Solution Overview

The essence of ZTE's end-to-end mobile service complaint handling solution (Fig. 1) is to build a multi-dimensional collaborative system that covers operators' complaint production processes, cross-domain expert analysis systems, and single-domain expert systems. It reshapes the complaint handling workflow through full-scale data aggregation and multi-dimensional AI algorithms.

Core Architecture: Three-Tier Collaborative Loop

The solution consists of three key tiers:

- **Business process tier:** Manages the full complaint process, including complaint acceptance, preprocessing, solution formulation, centralized optimization, and archiving.
- **Cross-domain VMAX capability tier:** Enables automatic fault boundary determination through a cross-domain big data platform, breaking down data silos and performing correlation analysis across radio access, core, and bearer network data.
- **Single-domain tool/agent tier:** Includes RAN agents (AAX/NGI/NQI), core network agents (CNIA), and bearer network agents (BigDNA), which perform domain-specific root cause analysis using expert systems and AI models.

Key Process

The solution transforms the traditional process into four key stages through intelligent technologies.

- **Complaint prediction and interception:** AI algorithms detect network faults in advance and perform correlation analysis using user perception indicators (KQI). Before users report faults, the customer service system generates prediction results to reduce potential complaints at the source.
- **Intelligent automatic boundary determination:** Leveraging the cross-domain VMAX platform, the system automatically triggers boundary determination after receiving work orders and attributes issues to specific network layers or service dimensions.
- **Root cause location and collaborative optimization:** Through agent collaboration, RAN, core, and bearer network agents perform parallel

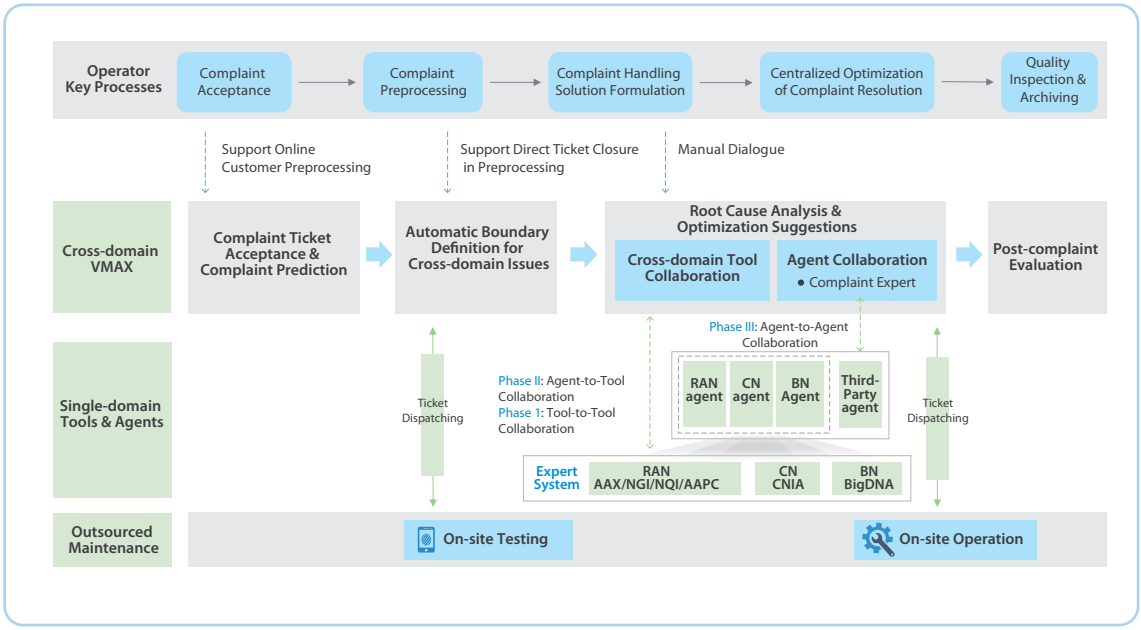


Fig. 1 ZTE's end-to-end mobile service complaint handling solution.

analysis, provide root cause location suggestions, and generate optimization schemes or remote operation instructions.

- **Mobile app-assisted O&M:** A mobile support system helps front-line O&M staff quickly retrieve information and submit test records with one click, improving on-site efficiency.

Core Features

The solution is characterized by the following three key features:

Shifting from Passive Response to Proactive Interception

Traditional complaint handling follows a post-incident resolution approach, whereas this solution emphasizes pre-incident prevention. By building a perception and prediction model, the system monitors the correlation between network KPI fluctuations and user experience degradation in real time, reducing filed work orders through customer-service preprocessing.

Hierarchical Decoupling and Intent-Driven

The solution evolves from automation to autonomy across three levels, with current work at Level II:

- **Level I (tool collaboration):** Automation of standalone tools.
- **Level II (agent-assisted tools):** AI-assisted expert decision-making.
- **Level III (full agent collaboration):** Establishment of an intent-driven self-healing network, where the system automatically adjusts network resources to meet targets such as a 20% reduction in user complaint rates.

Closed-Loop Evaluation System (KBI-KEI-KCI)

Building on autonomous network evaluation systems from international standard organizations, telecom operators, and vendors, this solution constructs a multi-dimensional indicator system to quantitatively assess complaint-handling effectiveness.

- **Key business indicators (KBI):** Overall effectiveness, including reduced personnel requirements, lower labor costs, and shorter processing latency.
- **Key efficiency indicators (KEI):** Scenario-level effectiveness, including complaint interception rate, average boundary determination duration, and on-site fault handling efficiency.
- **Key capability indicators (KCI):** Scenario-level capabilities, including boundary determination

coverage, secondary complaint ratio, and automation effectiveness.

User Value and Commercial Effectiveness

This solution has been implemented on a large scale by multiple operators in China, forming replicable complaint governance practices. Taking a typical deployment of an operator as an example, an end-to-end autonomous network capability system is constructed around the whole complaint process, with the following implementation scenarios and outcomes:

- **T0 online customer service:** Builds intelligent interception capabilities for three types of complaints—fault-related, planning-related, and common issues—achieving accurate reduction of complaints at the source. The system is invoked over 3,000 requests per day, with an average of 240 valid complaint interceptions per month, greatly reducing low-value work orders.
- **T1 complaint pre-analysis:** Establishes capabilities for intelligent boundary determination, classification, and automated order dispatching for 16 scenarios, including both network and non-network causes. The system handles over 600 requests per day, achieving an analysis accuracy of ≥86% while reducing complaint analysis time by 60%.
- **T2 front-line on-site handling:** Develops three types of mobile app tools for complaint analysis, dial testing, and network element queries, realizing efficient closed-loop on-site handling through mobile terminals. Mobile apps enable front-line staff to reduce the processing time of a single on-site work order by 0.5 hours, significantly improving O&M efficiency.

By optimizing the full process—from complaint interception at the source, through intelligent boundary determination, to frontline handling—the solution delivers three core values:

- **Ultimate operational efficiency improvement:** With an automated effectiveness analysis module, the solution realizes automatic comparison of complaint governance performance across time and organizational dimensions. It shortens the cross-domain complaint handling cycle from days to hours, reduces reliance on expert intervention, reduces the need for both in-house and outsourced maintenance personnel, cuts labor costs by over 30% , and frees up human resources to focus on high-value network planning.
- **Precise user perception management:** KCI indicators monitor the pass rate of work order verification to ensure that complaints are truly resolved. A post-complaint evaluation mechanism further verifies user perception after repair, eliminating secondary complaints.
- **Data-driven refined governance:** End-to-end work order statistics and an automated effectiveness analysis dashboard provide insights into governance performance across different branches and scenarios, supporting decision-making for resource allocation and improving the targeting of network construction and expansion.

Conclusion

ZTE's mobile user complaint handling solution for autonomous networks effectively improves the 5G network Q&M efficiency by reconstructing the whole process across four dimensions: prediction and interception, intelligent boundary determination, collaborative optimization, and effectiveness evaluation.

Beyond providing a set of technical tools, the solution establishes a scientific operational methodology. Its three-tier indicator system (KBI/KEI/KCI) provides an objective framework for operators to evaluate the maturity of autonomous networks.

With the maturity of generative AI (AIGC) and large language model (LLM) technologies, complaint handling is expected to evolve toward "Conversation as a Service." ZTE will continue to drive advancements in autonomous networks and explore the deeper application of large models in intelligent complaint assessment, root cause location, and closed-loop resolution, helping operators build more resilient, intelligent, and user-friendly mobile communication networks. ZTE TECHNOLOGIES

Co-Sight Pro: From Hard-Coded Experience to Knowledge Evolution

Liao Kaimeng

Wireless AI Agent Product System Architect, ZTE

Ni Hua

Wireless AI Agent Product System Architect, ZTE

As 5G network complexity surges, traditional network optimization modes relying on "expert experience + hard-coding" can no longer address core challenges such as experience silos, data black boxes, and evolutionary gaps. This paper proposes a knowledge lifecycle-driven autonomous network optimization agent and focuses on how the Co-Sight Pro dynamic planning framework leverages knowledge engineering methods, such as panoramic graphs, test-driven intelligent refinement (TIR) evaluation sets, and dynamic network memory, to enable a shift from single-point tool invocation to multi-step autonomous decision-making.

Pain Points of Network Optimization Decision-Making

In current network optimization workflows, although automation tools have partially intervened, three critical gaps remain when handling complex work orders:

- **Experience silos and rigid decision-making:** Expert experience is difficult to replicate, while traditional hard-coded tools lack flexibility and adaptability in dynamic network environments, and rely on manual knowledge transfer.
- **Data black boxes and unexplainability:** Large-model hallucinations lead to unreliable decisions. A lack of transparency in decision-making undermines expert trust, and model outputs do not fully align with network optimization goals, requiring manual intervention.
- **Evolutionary gaps and missing feedback loops:** Bad cases in the runtime environment cannot be automatically fed back into the production system, causing long iteration cycles, recurring issues, and limited improvement over time.

To solve these pain points, it is necessary to shift from a traditional "rules + tools" paradigm to a knowledge-driven paradigm.

Knowledge-Driven Autonomous Decision-Making Architecture

ZTE's autonomous decision-making network optimization agent uses Co-Sight Pro as its dynamic planning engine and knowledge engineering as the foundation to build a closed loop of perception-analysis-decision-execution-evolution. Its logic follows a knowledge lifecycle:

- **Knowledge production:** Raw data and expert experience are transformed into structured knowledge through the automated construction of a panoramic knowledge graph, TIR evaluation sets, and dynamic network memory.
- **Knowledge-driven:** Models are fine-tuned through domain knowledge injection. The knowledge graph drives Co-Sight planning, while chain of thought (CoT) reasoning enables explainable intelligent decisions.
- **Knowledge feedback:** Runtime bad cases are automatically mined to drive TIR code iterations and incremental model training, forming a knowledge flywheel.

Co-Sight Pro: The "Brain" of Autonomous Decision-Making

Co-Sight Pro is a core component of agentic AI and differs from traditional fixed workflows. Unlike rigid

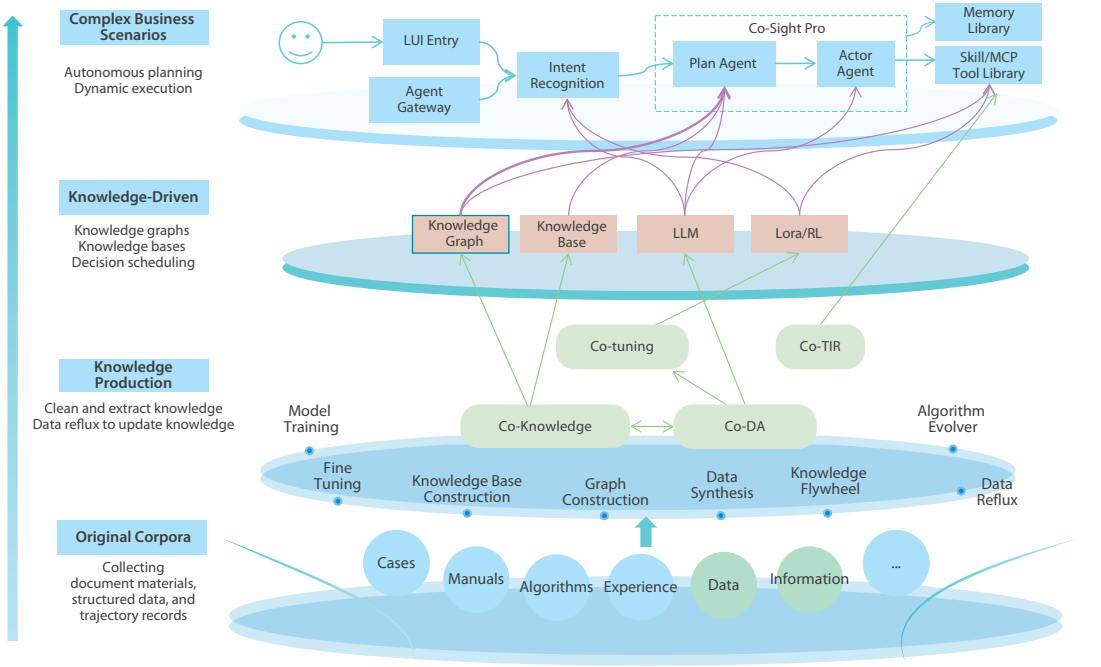
workflows that cannot adapt to sudden network changes, Co-Sight Pro enables dynamic planning and execution, while leveraging TIR technology for precise tool use (Fig. 1).

- **Graph-driven dynamic planning:** Upon receiving a task (e.g., handling 4/5G load imbalance), the Planner Agent first mounts the knowledge graph. It queries the knowledge graph in natural language to retrieve event and root cause subgraphs, providing a holistic view of the correlation among phenomena, events, and root causes. Based on these results, it autonomously orchestrates a directed acyclic graph (DAG) for concurrent execution.
- **Reflection and replanning:** The Actor Agent executes the planned diagnostic workflow while monitoring results and network status. When unexpected changes occur (e.g., a tool returns null or a new alarm appears), it reflects, replans, adjusts next steps, or terminates the task when needed.
- **Synergy with TIR Tools:** Within the diagnostic chain, Co-Sight invokes root cause diagnostic tools trained on TIR technology. These tools codify expert experience, are GPU-independent, and support high concurrency. If a TIR tool cannot reach a conclusion, a fine-tuned model serves as a backup to ensure diagnostic coverage.

Knowledge Engineering: The Foundation for Autonomous Decision-Making

The autonomous decision-making capability of Co-Sight Pro relies on a robust knowledge engineering system.

- **Panoramic knowledge graph—"prior knowledge" for decisions:** The graph addresses the massive volume of parameters, lengthy algorithm documents, and complex cross-references in network optimization. It automatically builds associations among phenomena, root causes, and events. By querying the graph, Co-Sight Pro quickly identifies the necessary checks, providing structured input for planning.
- **TIR evaluation set—"accuracy guarantee" for**



◀ *Fig. 1 Knowledge-driven Co-Sight Pro workflow.*

Table 1. Co-Sight Pro marks a significant advancement over traditional optimization workflows.

Process	Traditional Pain Points	New Mode Driven by Co-Sight Pro	Technical Support
Work Order Recognition	Diverse semantics, difficult object extraction	Interactive intent guidance, multi-turn clarification	Model fine-tuning via domain knowledge injection
Problem Review	Labor wasted on false anomalies/ self-healed issues	Automatic TIR tool invocation to scan real-time status and filter	TIR algorithm verification + Model backup
Root Cause Diagnosis	Black-box reasoning, untrustworthy	Agentic Reasoning + TIR tools for explainable diagnostic chain	Graph-driven Co-Sight planning + CoT reasoning
Solution Generation	Conflicting root causes, hard to merge	Multi-objective optimization, auto-merging of conflicting solutions	Large/small model fusion for multi-objective optimization + Digital twin verification
Execution Loop	Rigid execution, risk of negative optimization	Real-time KPI monitoring, auto-rollback if KPIs degrade	One-click parameter application + Intelligent rollback
Experience Accumulation	Recurring issues	Bad case reflux, driving online evolution of TIR code/models	Dynamic network memory + Closed-loop feedback

diagnosis: The TIR evaluation set addresses issues such as insufficient algorithm accuracy, high false alarm rates, and high GPU inference costs. TIR technology converts expert experience into high-precision code through intelligent data generation and code self-evolution. In 4/5G load imbalance scenarios, an F1 score of 100% can be achieved, enabling Co-Sight Pro to obtain reliable, efficient, and fully explainable conclusions when utilizing diagnostic tools.

- **Dynamic network memory—"experience nutrients" for evolution:** Dynamic network memory prevents the loss of runtime experience and shortens long iteration cycles. It records the context of agent operations, including tool invocation history, strategy success or failure, and status changes. These unstructured trajectories are transformed into structured CoT corpora for model fine-tuning, allowing Co-Sight Pro to improve path planning based on historical experience.

Business Value and Outlook

Co-Sight Pro's autonomous decision-making capabilities have significantly improved the entire network optimization workflow, including work order recognition, problem review, root cause diagnosis, solution generation, execution loop, and knowledge accumulation. (Table 1).

Field data show that the work order optimization rate exceeds 70%, with a target of over 90% by 2026. The work order self-closure rate currently stands at 20% and is expected to to surpass 30% by 2026. Work order processing latency is under four hours, with a goal to reduce it to below two hours. Additionally, the root cause diagnostic accuracy via TIR has achieved over 99% in target scenarios.

The autonomous decision-making network optimization expert based on Co-Sight Pro marks a transition from using knowledge to learning from and evolving knowledge. Currently, Co-Sight Pro has enabled dynamic planning and root cause diagnosis in capacity and quality scenarios, resolved the conflict between algorithm accuracy and resource consumption via TIR, and established the foundation for system evolution with dynamic network memory.

Looking ahead, we will continue advancing toward L4 autonomous networks featuring high agent autonomy. Technical challenges such as long-chain CoT reasoning, online TIR code updates, and multi-agent collaboration will be addressed. In terms of scenarios, we will cover coverage, quality, and interference scenarios, with cross-UME and multi-vendor adaptation. Ultimately, we aim to build a self-sensing, self-diagnosing, self-optimizing, self-evolving, and self-healing network, enabling a new "human-on-the-loop, process-supervised" Q&M paradigm. ZTE TECHNOLOGIES

Co-TAP: Three-Layer Agent Interaction Protocol

An Shunyu

User Experience Designer, ZTE

Mao Zhiyong

Wireless AI Agent Product System Architect, ZTE

Zhou Guiyue

AI Algorithm Engineer, ZTE

With the rapid advancement of large language models (LLMs), LLM-based agents have emerged as key drivers toward artificial general intelligence (AGI). Individual agents have limited capabilities, but multi-agent systems (MAS) leverage task decomposition and collaboration to solve complex problems. MAS face challenges such as poor interoperability, inefficient collaboration, and weak knowledge sharing. Co-TAP (T: Triple, A: Agent, P: Protocol) addresses these issues via three layers: Human-Agent Interaction, Unified Inter-Agent Communication, and Memory-Knowledge Management, providing a standardized framework for next-generation distributed AI systems.

The Demand for Multi-Agent Collaboration and Existing Barriers

LLMs have greatly enhanced agents' capabilities in text generation, coding, and basic reasoning. However, in complex real-world scenarios such as large-scale software development, cross-disciplinary research, and urban traffic scheduling, no single agent can cover all required knowledge and capabilities. MAS, through division of labor and complementary strengths, has therefore become a natural solution, aiming to achieve collective intelligence beyond individual capabilities.

Despite their potential, current MAS still face key challenges:

- **Interaction coordination:** The lack of unified interaction standards in user–agent collaboration makes instruction delivery, state synchronization, and exception handling uncertain, causing misunderstandings, unclear progress tracking, and resource conflicts, hindering seamless and reliable collaboration.
- **System interoperability and integration:** Agents often use different frameworks, platforms, and communication mechanisms, making cross-system collaboration difficult. Developers often need costly point-to-point integrations for cross-platform interactions, which are hard to maintain and vulnerable to protocol changes, limiting scalability and stability.
- **Knowledge accumulation and transfer:** Knowledge acquired by agents is often limited to a single task or session. Differences in data schemas and system-coupled storage hinder knowledge sharing, causing redundant relearning and limiting continuous improvement.

Co-TAP addresses these challenges by providing efficient, reliable, and scalable "universal languages" and behavioral guidelines for loosely coupled agents.

Co-TAP Core Architecture: Decoupling and Synergy Across Three Layers

Co-TAP organizes its core functions into a three-layer protocol stack that is clearly decoupled yet closely coordinated (Fig. 1). Each layer addresses a specific class of problems and collaborates through well-defined interfaces.

Human-Agent Interaction Protocol (HAIP): Enabling Seamless Human-Agent Collaboration

The HAIP provides a real-time, synchronous, structured, and semantically rich framework for user-agent interactions.

- **Bridge between backend agents and frontend interfaces:** HAIP imposes no constraints on the technology stack of agents. It converts the internal states, actions, and outputs of backend agents into a structured stream of events understandable by the frontend, allowing

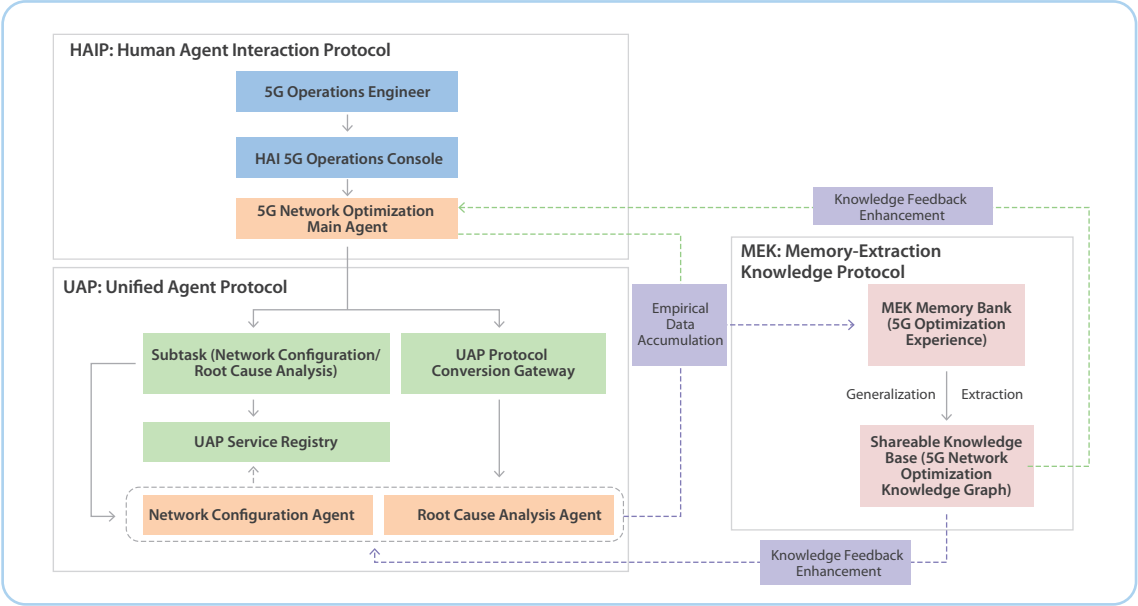


Fig. 1 Example of Co-TAP three-layer protocol collaboration.

developers to focus on agent capabilities rather than communication details.

- **Event-driven and real-time capabilities:** HAIP replaces traditional request-response communication with an event-stream architecture based on technologies such as server-sent events (SSE) This aligns perfectly with the inherent streaming nature of AI applications. Continuous events can be pushed to the frontend for token-by-token text generation or step-by-step task progress updates, enabling real-time interaction.
- **State sharing and collaborative controllability:** HAIP provides state synchronization mechanisms to quickly restore consistent context after network interruptions. The principle of collaborative controllability is embedded throughout. Through standard events, users can control the full lifecycle of agent tasks, including start, pause, resume, and terminate, ensuring human control in the human-in-the-loop (HITL) framework.

Unified Agent Protocol (UAP): A Standardized "Lingua Franca" for Agent Ecosystems

UAP serves as the hub of the Co-TAP ecosystem. Rather than a simple aggregation of communication protocols, it is an infrastructure framework centered on service governance and semantic interoperability, aligned with agents' cognitive patterns. Following a "modular decomposition + ecosystem development" strategy, UAP decouples multi-agent collaboration capabilities into independent modules such as AI gateways and registration centers, significantly enhancing system flexibility and maintainability. Built on Unified Registration & Discovery, Protocol Translation & Bridging, and Advanced Collaboration Primitives, UAP enables agents to introduce themselves, understand others, negotiate cooperation, and reach consensus, laying the foundation for an open and evolvable "Internet of Agents."

- **Capability description and registration:** Establishes a unified service capability description model covering basic service information, capability descriptions, and protocol-specific extensions. These "digital business cards" are registered in a centralized service registry, forming a globally queryable "agent social directory."
- **Agent and tool discovery:** Agent discovery is managed through the registry, allowing clients to query target service addresses and capabilities by service name, protocol type, tags, or service descriptions. Agents can identify potential partners through semantic queries or tag matching, enabling autonomous service discovery and dynamic composition, thereby marking the end of the era of hard-coded calls.
- **Protocol translation and adaptation bridging:** The UAP gateway supports protocol adaptation to standardized interfaces such as HTTP/REST

and gRPC, allowing legacy systems to integrate into the Internet of Agents via adapters. It also handles heterogeneous protocol conversion and centrally manages protocol translation plugins through a hierarchical registration and centralized management mechanism, significantly reducing adaptation overhead for individual agents. During protocol negotiation, the gateway determines whether to establish direct communication (when protocols are compatible) or route traffic via the gateway (when protocols are mismatched), ensuring interoperability across heterogeneous agent services.

Memory-Extraction-Knowledge (MEK) Protocol: Enabling Continuous Evolution of Agent Collectives

The MEK protocol enhances agents' intrinsic learning capabilities, enabling them to continuously learn from individual experience and improve through collective interaction.

- **Memory:** Converts structured experience from the perception-understanding-storage pipeline into a scalable long-term knowledge repository. By enforcing a unified memory-unit schema, the protocol ensures that each memory item is traceable and relationally linked.
- **Extraction:** Derives generally applicable knowledge from raw memory through filtering, de-identification, generalization, and standardization, converting personalized experience into standardized knowledge units that are reusable across platforms.
- **Knowledge:** Defines standardized patterns for knowledge sharing and assimilation across heterogeneous agents. A unified knowledge-unit schema ensures consistency throughout storage and transmission, enabling secure cross-agent knowledge transfer. This reduces redundant exploration and improves system-wide problem-solving efficiency, and helps MEK build a "supra-individual" agent system with continuous learning and collective intelligence.

Application Value and Future Outlook

The value of the Co-TAP three-layer protocol lies not only in its technical sophistication but also in its profound impact on industrial practice.

- **Enhancing robustness and maintainability of system engineering:** Through its layered design and decoupling, Co-TAP enables independent upgrades and replacements of system components, significantly reducing the development and management complexity of complex MAS.
- **Fostering a prosperous agent ecosystem:** Unified interaction standards, akin to the TCP/IP protocol of the internet, enable "plug-and-play" capabilities for agents from different vendors and with diverse functionalities. This holds the promise of catalyzing an open and vibrant market for agent-based applications.
- **Accelerating deep integration of AI in critical domains:** In sectors demanding ultra-high reliability—such as intelligent manufacturing, smart cities, and financial risk control—the standardized and controllable collaboration framework provided by Co-TAP serves as an essential infrastructure for the large-scale industrial deployment of AI.

Looking ahead, as agent capabilities continue to strengthen and application scenarios expand, foundational protocols like Co-TAP will become critical technical pillars for building a truly scalable and autonomous AI society. Future work may focus on optimizing protocol performance, ensuring security and privacy, and enhancing adaptive capabilities in more dynamic and open environments.

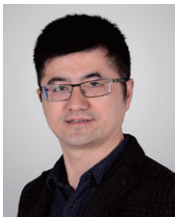
Conclusion

Co-TAP provides a forward-looking framework for addressing the challenges in the evolution of multi-agent systems, including interoperability, collaborative control, and knowledge evolution. Through the intricate design and synergistic operation of its three layers—HAI, UAP, and MEK—Co-TAP provides a powerful and flexible blueprint for achieving efficient, reliable, and evolvable collective intelligence. More than merely filling existing technological gaps, Co-TAP contributes to next-generation distributed AI ecosystems and points toward a new era of human-machine symbiosis and swarm intelligence. ZTE TECHNOLOGIES

Core Network Complaint Agent: An Efficient Approach for Complex Complaint Handling

He Wei

Chief Engineer of Computing & Core Network Intelligent O&M Product Planning, ZTE



Chen Chun

Intelligent O&M Product Planning Manager of Computing & Core Network, ZTE

With the continuous evolution of 5G-A networks and the acceleration of digital transformation across industries, users' requirements for network quality have shifted from "being able to use" to "being reliable and intelligent". This shift drives continuous optimization of user experience, which is at the core of communication services. In this context, the network complaint handling system is becoming increasingly valuable as a key tool for operators to perceive user experience and achieve closed-loop service management. It helps operators build more intelligent and efficient user experience assurance systems.

Pain Points in Traditional Complaint Mode

Traditional complaint handling faces several limitations that reduce efficiency. Complaint location efficiency is low. The network complaint system needs to integrate multi-dimensional data, such as network status, configuration, and signaling for comprehensive analysis, while root cause inference is complicated.

Moreover, high skill requirements further limit efficiency. Manual complaint signaling analysis requires familiarity with extensive service procedures and signaling protocols, and complex scenarios—such as signaling sequence conflict or protocol compatibility issues—demand advanced skills.

Limited scenario coverage is another key issue. The existing system mainly relies on rule libraries and fixed workflows, making it unsuitable for new service scenarios and difficult to support the evolving requirements of complaint analysis.

Technology Drives a New Paradigm for Complaint Analysis

The evolution of AI technologies is transforming network complaint analysis. Large language model (LLMs) provide powerful capabilities in semantic understanding and knowledge integration. By combining pre-trained foundational models with chain-of-thought reasoning, large models offer a new approach for network complaint signaling analysis.

Agent architecture serves as the "decision-making center" of AI. Leveraging large models, agents can perform dynamic inference and task planning, autonomously break down complaint handling processes, and efficiently complete complex, closed-loop tasks.

Finally, the RAG knowledge base decouples knowledge storage from models, allowing users to upload new content independently. Uploaded content can be retrieved and invoked immediately without retraining the models, greatly improving the agility of knowledge delivery.

Innovative Complaint Agent Solution

ZTE's core network complaint agent solution builds an end-to-end intelligent complaint analysis system based on the application layer, orchestration layer, and model layer, achieving technological breakthroughs from raw signaling parsing to intelligent reasoning (Fig. 1).

- **Application layer:** Provides visual signaling sequence diagrams, complaint analysis conclusions, and typical cases to support fast,

closed-loop complaint resolution.

- **Orchestration layer:** Integrates planning, execution, and memory functions to automatically understand complaint intent, plan task objectives, and dynamically orchestrate tool chains, automating end-to-end complaint analysis.
- **Model layer:** Integrates a large signaling model and a large language model to accurately understand the logic of professional protocols such as 5GC and IMS, enhancing signaling analysis capabilities in complex scenarios.

Intelligent Orchestration for Automatic Complaint Handling

The orchestration layer of the complaint agent acts as the "intelligent center" of the system. Based on a large multimodal foundation model, the agent automatically retrieves historical experience after classifying complaint intent, and performs in-depth inference according to O&M manuals and tool descriptions in the knowledge base. A targeted complaint handling solution is then generated and automatically executed.

For example, when an international roaming user cannot access the Internet, the orchestration layer analyzes the original signaling inference results and dynamically performs key diagnostic actions, such as international roaming configuration checks and user status detection, to quickly and accurately locate the root cause of the complaint.

Large Multimodal Models for End-to-End Intelligent Signaling Analysis

Signaling data in network complaint scenarios has dual features: it contains highly structured protocol fields, while also involving service information that requires in-depth semantic understanding. To address this, the solution integrates large signaling model and LLM technologies to build an analysis system with end-to-end automated signaling analysis and intelligent inference capabilities.

By introducing an attention mechanism, ZTE Nebula Telecom Large Model can effectively capture service logic in signaling data and extract key information such as sequence characteristics and abnormal parameter patterns in 5GC signaling. To address the differences between structured signaling and natural language semantics, the solution innovatively uses vector alignment technology to construct mapping relationships between signaling semantics and service rules, achieving unified representation of data in different protocol formats and significantly reducing the model's adaptation complexity for heterogeneous data.

To cope with analysis challenges in complex signaling scenarios, the large model integrates native structured thought chains, enhancing inference capability for hidden faults. It can intelligently identify signaling sequence conflicts and cell logic exceptions, significantly improving the capability to locate complex issues.

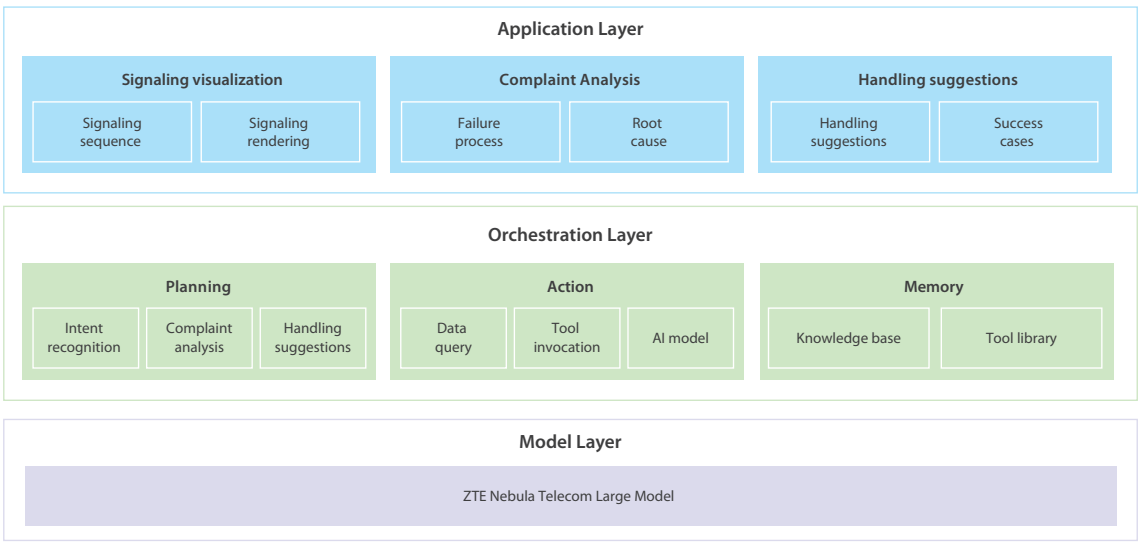


Fig. 1 Core network complaint agent architecture.

> In the future, complaint agents will evolve into core engines for intelligent O&M with self-learning and self-optimization capabilities.

Plug-and-Play Tools to Drive On-Demand Expansion of Agent Capabilities

The model context protocol (MCP) standardizes how an agent understands, stores, and uses contextual information, aiming to establish a unified communication interface between the agent and external data sources or tools. Through MCP and various tool interfaces, the complaint agent can flexibly invoke capabilities covering NE configuration, user status, and service quality, enabling more effective interaction with the network environment and efficient task completion.

For example, in the case of a "slow internet speed" complaint, the system automatically invokes a service quality assessment algorithm model to process key performance indicators of user packets in real time, including latency and packet loss. In this way, the system analyzes trends within the data, intelligently selects the optimal algorithm to improve analysis accuracy, and assists the large model in determining whether the issue is on the wireless side or the network side.

Complaint Agent: A Case Study

In 2025, ZTE and Jiangsu Mobile successfully completed the verification of key technologies for complaint agents based on large multimodal models. The complaint agent has significantly improved international roaming user experience, reduced O&M costs, achieved up to 90% accuracy in analyzing signaling plane issues for international roaming, and shortened the duration of 5GC signaling analysis and root cause location. This deployment sets a benchmark for the implementation of complaint agents.

- **Efficient and accurate signaling analysis:** The large multimodal model can perform complaint signaling analysis within minutes, accurately locating root cause failure points and key fields based on the end-to-end inference capability. This greatly improves automated closed-loop complaint handling.
- **In-depth coordination of O&M processes:** An innovative Tier 1 (first-line) + T2 (second-line) dual-agent coordination mechanism is used to handle complaints by stages. In the T1 phase, the network complaint platform generates a preliminary solution based on XDR. Complicated issues are automatically transferred to the T2 phase, where ZTE complaint agents complete in-depth analysis and handling of complaint work orders.
- **Agile knowledge base empowerment:** Leveraging the external knowledge base of the RAG architecture, international roaming complaint O&M cases can be uploaded in real time, allowing rapid identification of scenario-specific issues without retraining the model. This greatly improves the flexibility of knowledge updates.

In the future, complaint agents will evolve into core engines for intelligent O&M with self-learning and self-optimization capabilities. By deeply integrating large-model cognition and inference with specialized communications network knowledge, an end-to-end intelligent diagnosis system can be constructed to continuously improve root cause localization accuracy in complicated complaint scenarios, enabling automated and unmanned closed-loop complaint handling. ZTE TECHNOLOGIES

AI Plug-in Solution for Legacy Systems: Enabling Full AI Evolution of RAN Product O&M

Liu Yang

Wireless AI Agent Product System Architect, ZTE

Yue Shubin

Wireless AI Agent Product System Architect, ZTE

As industrial standards for autonomous networks evolve and large language model (LLM) technologies develop rapidly, network O&M is shifting from the traditional "users looking for functions" mode to an AI-driven intelligent agent mode. As a core scenario of network O&M, the radio access network (RAN) faces growing complexity, including inefficient and fragmented operations, high maintenance expertise requirements, non-reproducible experience, limited reasoning for long-tail scenarios, and difficulty in system-wide reconstruction. ZTE has launched an AI plug-in solution for traditional network management, enabling non-intrusive AI enhancement and building a knowledge-driven, end-to-end automated intelligent O&M paradigm for RAN.

Core Challenges of RAN O&M Systems During Autonomous Network Transformation

At present, operators' RAN O&M unified management expert (UME) systems face bottlenecks in evolving toward advanced L3/L4 autonomous network capabilities, mainly in four areas:

First, capabilities are fragmented and operations are inefficient. Atomic business capabilities are scattered across apps and interfaces without intelligent connection, requiring users to switch modules and repeatedly input parameters, resulting in long operation paths.

Second, expert experience is difficult to retain and accumulate as reusable knowledge assets. O&M decisions rely heavily on individual experts, while valuable knowledge is often shared through manuals or oral instructions rather than being digitized as reusable assets.

Third, system transformation costs are high. Fully reconstructing traditional network management systems to integrate AI capabilities requires large investment, long cycles, and may risk business interruption, making non-intrusive AI enhancement more urgent.

Fourth, scenario generalization and reasoning remain limited. When faced with complex and changing O&M demands, existing automation depends mainly on hard-coded rules and fixed scenarios, lacking the reasoning and generalization capabilities needed for scenario-based closed loops in autonomous networks.

Non-Intrusive AI Enhancement for Intelligent O&M Upgrades

To address these pain points, ZTE's AI plug-in solution adopts a "non-intrusive AI enhancement" design philosophy. It introduces the UME Intelligent Partner as an AI enhancement layer, overlaying full-process AI capabilities on the legacy system without modifying the core architecture of the existing UME system or intruding into business code. This enables intelligent assembly of business capabilities and builds an end-to-end automated O&M closed loop covering perception, analysis, decision-making, execution, and verification. The overall architecture is shown in Fig. 1.

The solution adopts a decoupled design for flexible expansion. Front-end decoupling is achieved by using the O&M Expert front-end Playwright plug-in as a stateless execution engine, separating business logic from the execution environment through standardized JSON instructions. At the business process level, the

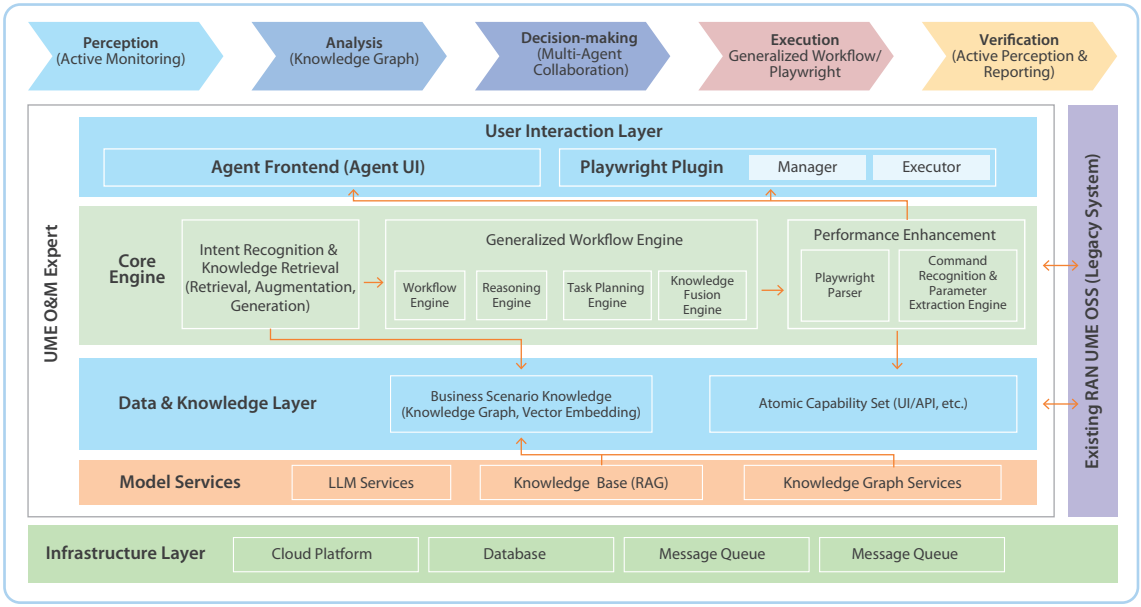


Fig. 1 Overall architecture of the UME O&M expert system.

system shifts from rule-based, hard-coded processes to knowledge-driven dynamic workflows, where atomic capabilities are service-oriented, while the workflow engine handles only scheduling rather than specific business logic. At the knowledge level, business rules, expert experience, and operation guidelines are separated from code logic and transformed into independently manageable and evolvable knowledge assets.

The core design philosophy is "enhancement rather than replacement". Through plug-and-play deployment, the solution achieves two main capabilities:

- **Scenario-level end-to-end assistance:** Automatically decomposes tasks, generates generalized workflows, and completes automated execution based on business objectives input by users in natural language.
- **Intelligent assistance for discrete operations:** Provides one-click access to functions, data, and knowledge through aggregated search, offering users guided operations.

Core Technological Innovations: Resolving Key Challenges in AI Implementation

Focusing on core bottlenecks in the intelligent upgrade of existing systems, the solution realizes low-cost, highly adaptable, and generalizable implementation of AI capabilities through four core innovations:

Knowledge-Driven Generalization of Dynamic Workflows

The solution breaks away from the fixed-process model of traditional hard-coded systems. Through AI search and knowledge graph technology, it extracts business rules, expert experience, and operational guidelines from code logic and transforms them into independently iterable knowledge assets. Based on users' vague business objectives, the system retrieves relevant knowledge in real time and dynamically generates executable workflows. This allows flexible orchestration of atomic capabilities and generalized adaptation to business scenarios.

Zero-Adaptation Front-End Process Automation

The solution introduces a stateless front-end execution plug-in based on the Playwright automation framework. Through standardized JSON instructions, it separates business logic from the execution environment. With "iframe container isolation + dual-channel reverse proxy," it follows the three-zero principles: zero code intrusion, zero cross-domain configuration, and zero framework dependency,

enabling seamless integration across IP domains and technology stacks, as well as adaptation-free automated execution of legacy functions.

Three-Tier Gradient Domain Knowledge System

To address the lack of vertical domain knowledge in general large models, the solution constructs a three-level knowledge system: L1 basic general knowledge, L2 scenario application knowledge, and L3 advanced reasoning knowledge. It preserves expert experience digitally and supports accurate reasoning, decision-making, and workflow generation by large models.

Full-Process Controllable Multi-Agent Collaboration Technology

The solution builds a multi-agent architecture with three core modules: intent classification, task planning, and execution engine. Through a transparent process covering intent understanding, plan generation, step-by-step execution, and feedback optimization, it enables end-to-end controllability of AI-assisted processes. Users can monitor workflow generation and execution, and confirm, edit, or interrupt operations at any time, ensuring operational security and improving trust in AI capabilities.

Commercial Value and Practice

To date, ZTE's AI intelligent plug-in solution for O&M systems has completed large-scale implementation verification. Based on knowledge graph and AI search capabilities, it supports generalized end-to-end workflow generation and execution across 400 functions, 23 major scenarios, and 125 sub-scenarios in RAN network management, including high-frequency O&M scenarios such as base station commissioning and cell capacity expansion or reduction.

In practice, O&M personnel only need to describe requirements in natural language, and the system can automatically complete intent recognition, task planning, step-by-step execution, and result verification. For example, in cell capacity expansion, the traditional process requires switching across multiple modules, querying parameters, and manually executing commands, which is time-consuming and error-prone. With the AI plug-in, the process can be completed automatically within minutes, improving efficiency while reducing human errors.

For operators, the solution brings three core values:

- **Significant cost reduction and efficiency improvement:** Through a non-intrusive deployment mode, it greatly reduces the cost and cycle time of intelligent upgrades for live network O&M systems. Meanwhile, the end-to-end automated closed loop improves O&M efficiency by more than 80%.
- **Long-term accumulation of knowledge assets:** The three-level knowledge system enables digital, asset-based management of expert O&M experience, preventing knowledge loss and allowing large-scale replication of high-quality O&M capabilities.
- **Smooth evolution of autonomous network capabilities:** By providing a progressive upgrade path, the solution helps operators advance the O&M system from L2 passive automation to L3 active closed loop and L4 advanced autonomy, laying a solid foundation for fully intelligent network O&M in the 6G era.

As 6G native intelligence research advances and networks move toward L4/L5 autonomy, fully AI-driven, end-to-end closed-loop network O&M is set to become an industry trend. With AIOS at the core, ZTE will continue to optimize its AI plug-in solutions for legacy systems and further integrate large models and multi-agent technologies into telecom O&M, consolidating its three-in-one core capabilities: dual-domain global self-perception, intent-driven self-O&M, and secure, controllable self-evolution.

ZTE will extend perception from network-native states to cross-domain business ecosystems, shift O&M from post-event handling to proactive prevention, and enhance closed-loop execution for more efficient, reliable, and secure automation. Through continued collaboration with global operators, ZTE will drive the AI-native evolution of RAN O&M, laying a solid foundation for autonomous networks and opening a new chapter of native intelligent O&M for wireless networks in the 6G era. ZTE TECHNOLOGIES

OTN Holographic Technology: Digital Twin Practices for L4 Autonomous Optical Networks

Ming Zhengqin

OTN Product Planning Manager, ZTE

OTNs are rapidly evolving toward 400G/800G and beyond, as well as full-mesh topologies, leading to exponential growth in network scale and complexity. However, optical network O&M still faces several key challenges, including difficult fault localization due to unquantifiable optical signals, unpredictable performance caused by optical power, dispersion, and nonlinearity, and hidden degradation risks such as fiber aging and loose connectors. To address these challenges, ZTE has introduced OTN holographic optical technology and autonomous optical network solutions.

OTN Holographic Optical Technology

By combining OTN optical-layer digital twins with performance evaluation algorithms, holographic optical technology extracts and analyzes network-wide optical parameters to enable rapid fault localization, reliable performance assurance, and early warning of hidden risks.

High-Precision Digital Twin Modeling

Physical components such as optical modules, filters, amplifiers, and fibers are modeled with high precision to build an end-to-end digital twin of the optical system (Fig. 1). By dynamically collecting data on optical power, loss, and bit error rates, the twin remains synchronized with the physical network in real time, providing a basis for performance evaluation, risk prediction, and solution simulation.

Optical Performance Evaluation Algorithms

- **Holographic optical channel performance evaluation algorithm:** Based on the physical fiber-link model, calibration parameters, and measured data, this algorithm performs high-precision OSNR calculations using more than 30 collected optical parameters. It requires no instruments or additional hardware, supports online measurement, avoids service interruption, and features low deployment costs.
- **Holographic fiber performance evaluation algorithm:** Based on the optical time domain reflectometer (OTDR), this algorithm captures characteristic waveforms of Rayleigh scattering and Fresnel reflection, incorporates state of polarization (SOP) monitoring data, and employs machine learning classification algorithms to identify fiber breaks, degradation, co-cables, and external interference. It enables early warning and root cause determination.

Typical Scenarios for Holographic Optical Technology

Holographic optical technology plays an important role in optical network O&M.

Reliable Provisioning of Optical Channels

Traditional optical path provisioning methods lack precise transmission quality assessment, which can lead to excessive margins, increased costs, or repeated debugging. By using the holographic optical channel performance evaluation algorithm and introducing OSNR constraints, the selected path can be verified to meet OSNR requirements, thereby improving the provisioning success rate.

Ensuring Quality of Backup Channels

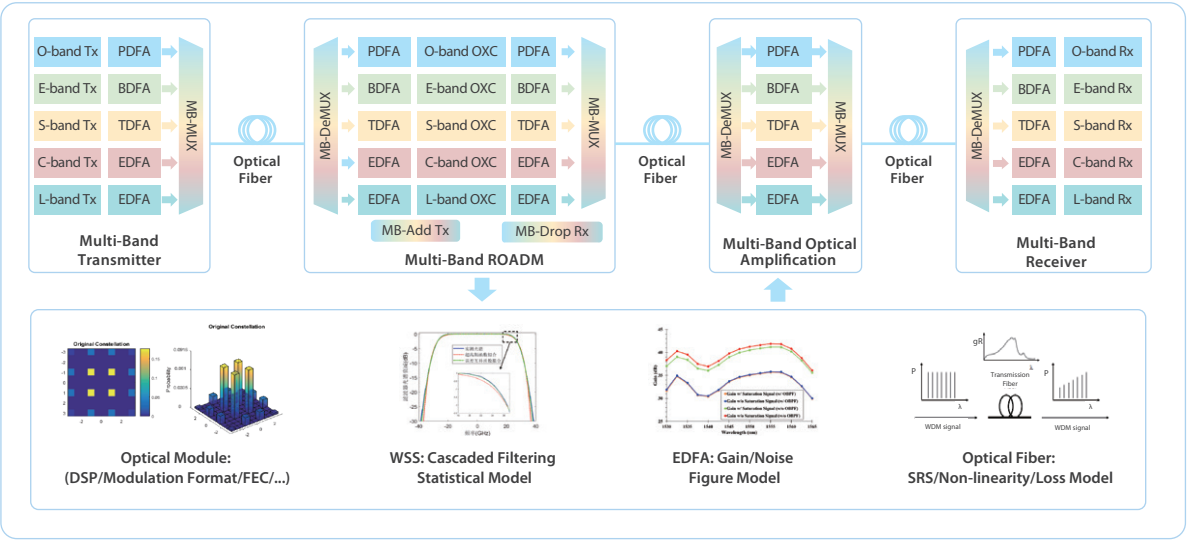


Fig. 1 Holographic optical digital twin modeling.

Optical backup channels often remain idle for long periods and lack effective monitoring, which may result in switching failures when the primary channel fails. Periodic holographic optical channel performance evaluations can be performed on backup channels to detect hidden issues in advance and enable timely repairs.

Precise Localization of Fiber Breaks

Manual segment-by-segment troubleshooting after fiber interruptions is time-consuming and labor-intensive. With the holographic optical channel performance evaluation algorithm, equipment alarms can automatically trigger OTDR measurements upon a fiber interruption. The algorithm analyzes abrupt attenuation changes to accurately locate the fault point, thereby reducing the mean time to repair (MTTR).

Co-Cabling and Shared Routing Risk Analysis

In existing networks, working and protection fibers are often laid in the same cable, so a fiber cut may cause service interruptions. The holographic fiber performance evaluation algorithm can automatically detect physical co-cabling scenarios by analyzing overlaps in OTDR measurement traces, eliminating safety hazards associated with shared routes.

Early Perception of Fiber Degradation

Hidden issues such as fiber bending, compression, connector contamination, splicing anomalies, and construction damage continuously erode system margins and may eventually impact services. The holographic fiber performance evaluation algorithm performs in-depth analysis of splicing points, bending points, and connection points in OTDR traces. By comparing them against historical baselines, it provides early degradation warnings to prevent failures before they occur.

ZTE's Holographic Autonomous Optical Network Solution and Use Cases

ZTE's holographic autonomous optical network solution leverages holographic optical technology to build an intelligent evaluation and optimization system based on a "Digital Twin + AI" framework. At its core is a high-fidelity digital twin of the physical optical network that is synchronized with the live network in real time, enabling intelligent evaluation, diagnostic analysis, and automated closed-loop management throughout the entire network life cycle.

Solution Architecture

The overall solution is structured into two primary layers (Fig. 2):

- **Intelligent network element layer:** This layer consists of OTN equipment capable of high-precision optical performance data collection, monitoring, and pre-processing.
- **Intelligent control & management layer:** Built upon digital twins and large models, the holographic

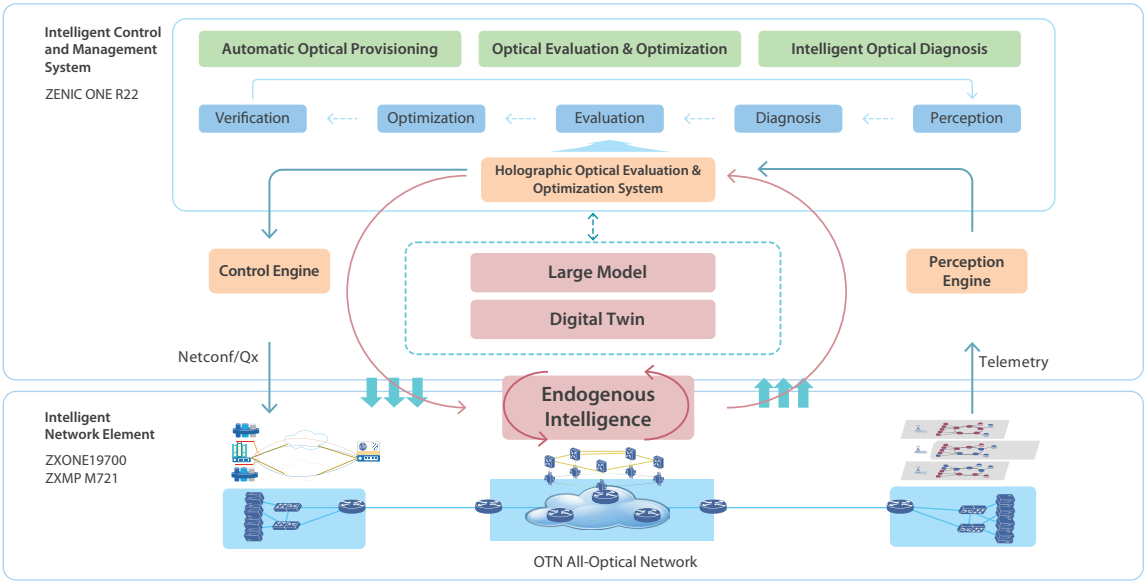


Fig. 1 Architecture of ZTE's holographic optical autonomous network solution.

optical evaluation and optimization system includes three core modules: automatic optical provisioning, optical evaluation & optimization, and intelligent optical diagnosis. Together, these modules form a complete closed loop of "perception, diagnosis, evaluation, optimization, and verification."

Application Cases

The holographic autonomous optical network solution has been successfully deployed by multiple telecom operators.

- **OLP backup channel evaluation at Hubei Mobile**

To ensure that services in the existing network are not affected during optical line protection (OLP) switchover, manual switchover tests had to be performed at night. With the holographic optical assessment and optimization subsystem, OLP backup-channel performance can now be evaluated online without switching equipment. Verification time per NE was cut from over 30 minutes to under 10 minutes, improving O&M efficiency by more than 200% and eliminating service interruptions.

- **Intelligent Co-Cabling Analysis at Sichuan Mobile**

In the live network, many working and protection fibers are co-located in the same cable, significantly reducing service redundancy. After the holographic optical intelligent diagnosis subsystem was enabled, co-cabled fibers can be identified among outgoing fibers from the same network element, ring fibers, optical layer primary/backup protection fibers, and SNCP transport-layer fibers. In actual engineering environments carrying live traffic, co-cable identification reached 90%. The system recommended routes based on shared risk link group (SRLG) policies, effectively avoiding co-cabling and shared-routing scenarios.

Outlook

ZTE's holographic optical technology and autonomous network solutions effectively address optical-layer O&M pain points by enabling accurate evaluation, intelligent diagnosis, and preventive maintenance, greatly improving O&M efficiency and service assurance.

Future enhancements will focus on three areas: improving digital twin accuracy by integrating non-linear simulation and machine learning; expanding application scenarios to include optical module degradation warnings, fiber degradation analysis, construction warnings, and geological monitoring; and evolving toward L3 Conditional Autonomy and L4 High Autonomy.

Through continuous innovation and scenario-driven development, ZTE's OTN holographic autonomous optical network solution will help global operators build intelligent optical networks with self-perception, self-diagnosis, self-prediction, and self-optimization capabilities. ZTE TECHNOLOGIES

Home Broadband Complaint Agent Solution

Lu Yun

R&D Planning Expert (Wireline), ZTE

Chen Aiming

Chief OLT Planning Engineer (Wireline), ZTE

Zhang Rong

Chief Wireline System Architect, ZTE

With the continuous development of the home broadband business (HBB), users have increasingly higher requirements for network quality and service experience. At present, home broadband complaint handling faces pain points such as slow response, difficult fault location, and long repair cycles, resulting in a persistently high number of complaint work orders. According to Q4 2025 data, home broadband complaints accounted for 41.7% of all network fault tickets, of which about 68% were soft faults caused by improper configuration or operation of user-side equipment. The average handling time reached 3.2 hours, straining O&M resources.

To overcome these bottlenecks, ZTE has developed a home broadband complaint handling intelligent agent. It leverages large AI models, end-to-end quality analysis engine, and automatic control capabilities to transform the service mode from passive response to active perception, intelligent diagnosis, and automated repair. Its core objectives are to achieve a 20% year-on-year reduction in home broadband complaints, shorten the average complaint handling time to less than 120 minutes, and increase the self-service resolution rate to over 40%, supporting simplified O&M and customer experience upgrade.

Disadvantages of Traditional Complaint Handling Methods

Currently, the complaint handling process has four major structural defects that seriously restrict service efficiency and user experience:

Fault Location Relies on Manual Experience and Lacks End-to-End Analysis

In the traditional mode, fault delimitation depends on segment by segment troubleshooting by O&M personnel, covering the user terminal, home gateway, indoor cabling, splitter, OLT, and MAN. Due to the lack of a unified QoS view, it is difficult to accurately map "user-perceived lag" to abnormal network-layer indicators. For example, "video stuttering" may be caused by Wi-Fi channel congestion, ONU optical power thresholds, or mismatched MAN QoS policy. However, the current system only provides device-level alarms, such as "ONU offline", and cannot identify soft faults, resulting in low fault delimitation efficiency and more work orders.

Soft Faults Require On-Site Support

About 68% of home broadband complaints are related to soft faults, such as Wi-Fi channel interference, DNS configuration errors, user PC faults, and router faults. These problems usually do not require hardware replacement. However, the existing system lacks remote intervention capabilities, leading to unnecessary on-site visits. For example, slow network speeds due to Wi-Fi 2.4 GHz channel congestion can be solved remotely by adjusting the channel or guiding users to switch to the 5 GHz frequency band, but in practice, it still follows the "on-site repair" procedure.

High Skill Requirements

Troubleshooting highly relies on the experience and knowledge of O&M personnel, resulting in long training periods for new employees, and high cross-domain coordination costs. In scenarios integrating multiple technologies such as PON, IP, Wi-Fi, and IPTV, it is difficult for a person to master all issues, resulting in repeated diagnosis and a high misjudgment rate. In addition, the lack of standard diagnosis processes may cause different personnel to reach different conclusions when handling the same

problem, affecting service consistency.

Low User Participation and Lack of Self-Service

After a fault occurs, users can only submit complaints through the customer service hotline or app, and cannot obtain network status information or perform a preliminary self-check. Even simple user-side issues, such as router restart, ONT restart, or loose network cables still require manual intervention. This generates a large number of low-value work orders. Users also lack transparency regarding handling progress, increasing dissatisfaction.

Key Advantages of Home Broadband Complaint Agent

By integrating large AI models, multi-source data, automated control, and user interaction capabilities, the home broadband complaint agent builds a closed loop of "perception, analysis, decision-making, execution, and feedback", comprehensively solving the pain points of traditional complaint handling and enabling service model innovation.

End-to-End Service Quality Analysis and Precise Fault Delimitation

The home broadband complaint agent connects full-link data from the FTTH access network (FAN domain), core network, home terminals, and application layer to implement end-to-end service quality monitoring and analysis. Based on large AI models, service indicators such as HTTP response delay, video stuttering rate, and game packet loss rate are modeled and mapped back to network layer KQIs/KPIs, enabling accurate fault delimitation from user-perceived deterioration to root cause. Algorithm models are used for automatic association analysis to quickly locate the faulty link and determine whether the issue lies in the operator's PON network, the home side, the transmission network or the application side. Fault delimitation accuracy can reach more than 92%, greatly shortening the troubleshooting time.

Remote Handling and Automatic Optimization of Soft Faults

The agent integrates remote control capabilities and supports automatic operations on devices such as the home gateway and ONT. For identifiable soft faults, the system can automatically trigger repair policies, such as remotely switching to the optimal channel or instructing users to switch to the 5 GHz band for Wi-Fi issues, automatically adjusting DBA parameters or issuing optical module gain optimization commands for critical optical power issues, remotely resetting DNS configurations or switching to the standby server for DNS abnormalities, and adjusting service QoS configurations to guarantee differentiated services for video and gaming stuttering. Pilot data shows that 62% of soft faults can be remotely fixed without dispatching installation and maintenance (I&M) work orders.

Reducing Skill Requirements and Simplifying O&M

The agent uses an expert knowledge base and diagnosis decision tree to standardize and automate complex troubleshooting processes. O&M personnel only need to follow the root cause analysis and troubleshooting suggestions provided by the agent without fully understanding the underlying protocols or topologies. For example, the system can output: "Fault cause: Wi-Fi channel 3 seriously overlaps with the adjacent AP. Recommended operation: Switch to channel 9 remotely or instruct users to switch to the 5 GHz band." This approach reduces the skill threshold for front-line maintenance personnel by 40%, shortens the onboarding period of new employees from three months to two weeks, and supports the transformation to minimalist O&M.

App-Based Self-Diagnosis and Self-Repair

For users, the agent provides capability invocation interfaces integrated into the HBB mobile app, offering network health detection, one-click speed tests, Wi-Fi optimization, and fault self-healing functions. Users can view the home network quality score, Wi-Fi coverage heat map, and optical power status at any time. When a network fault is detected, the app pushes a diagnosis report and provides a one-click repair button, enabling operations such as restarting the ONT, switching the

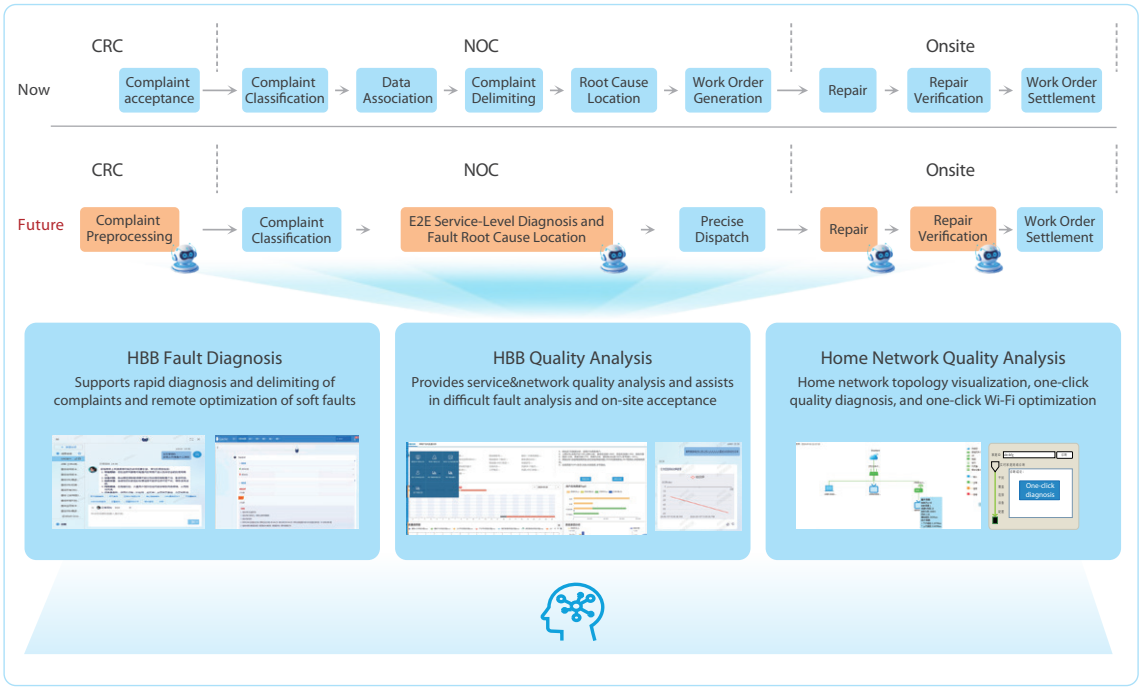


Fig. 1 Home broadband complaint agent solution.

Wi-Fi channel, clearing the DNS cache, and enabling prioritized assurance for video and gaming services. These self-operations allow users to solve problems independently, significantly reducing low-value complaint tickets. In addition, the app displays processing progress and repair results in real time, improving service transparency and user satisfaction.

Home Broadband Complaint Agent Solution

ZTE's home broadband complaint agent is currently deployed on the UME system. It provides northbound interfaces and interconnects with the telecom operator's customer response center (CRC), network operations center (NOC), and I&M assistant (Fig. 1). It supports complaint preprocessing, fault diagnosis, and service quality analysis, enables remote complaint processing and precise fault ticket dispatching, and helps front-line I&M personnel improve on-site efficiency.

The home broadband complaint agent can assist in three phases of complaint handling.

- **Acceptance phase:** The agent provides end-to-end complaint pre-diagnosis capabilities, identifies faults such as poor quality or fiber disconnection, accurately delimits the faults, and supports precise complaint ticket assignment. In addition, it can remotely repair soft faults with one click, reducing fault tickets.
- **Analysis phase:** The agent accurately locates faults in the PON network segment, assisting with fault dispatching and rapid repair. For complex faults, it assists NOC personnel in performing quality traceback at both service and network levels, enabling efficient analysis.
- **Maintenance phase:** The agent supports the I&M assistant in home network fault analysis, as well as Wi-Fi analysis and self-optimization. It helps I&M personnel perform fast repair and verification.

By deploying the home broadband complaint agent, operators can enable end-to-end service quality analysis, automatic soft fault repair, simplified O&M, and user self-service, establishing a new-generation closed-loop system for home broadband fault handling. This solution significantly improves O&M efficiency and customer experience, supports operators' evolution towards L4 autonomous networks, and serves as the core engine for HBB digital transformation. ZTE TECHNOLOGIES

Cross-Domain Synergy: ZTE and Guangdong Mobile Build an Autonomous Network Demonstration Zone

Guo Ruicheng

Project Director of Computing & Core Network, ZTE

Shao Mengfei

Wireless R&D Planning Expert Engineer, ZTE

With the rapid development of the digital economy, 5G, gigabit optical networks, cloud computing, and industry-integrated applications continue to evolve, while network scale expands and service models become increasingly complex. As a leading Chinese communications operator, China Mobile Guangdong (Guangdong Mobile) is facing unprecedented challenges.

In 2025, Guangdong Mobile and ZTE jointly launched the "Autonomous Network Demonstration Zone" project, promoting the transition of network O&M from "passive response" to "active prevention" and accelerating the evolution from L3 to L4 autonomous networks.

A Full-Stack Intelligent Evolution Path

To address common pain points in network intelligence transformation, including data silos, fragmented intelligence, and broken loops, ZTE and Guangdong Mobile have adopted a cooperation philosophy of joint research, scenario co-building, and ecosystem collaboration to develop systematic solutions.

Leveraging its self-developed Nebula Telecom Large Model and full-stack intelligent solutions, ZTE effectively breaks down technical barriers and promotes the evolution of network O&M toward high-level autonomy. Anchored by a "1+N" strategy—where "1" represents the unified telecom large language model, and "N" refers to diverse auxiliary models, including structured models, graph models, digital twin models, and other industry-specific models—the system employs a three-tier architecture comprising the network element, single-domain, and cross-domain layers. This framework integrates Copilot-assisted O&M with agent-based automated handling to achieve an end-to-end intelligent closed loop. Building on this foundation, an intelligent decision engine is introduced, featuring foundation-model flexibility, task autonomy, dynamic knowledge, and rigorous reasoning, driving the network's transition from pre-defined rules to autonomous cognition.

The two parties focus on four core domains—

wireless network, core network, transmission network, and network cloud—and integrate AI agent technologies into network O&M scenarios, significantly improving O&M quality and efficiency.

Wireless Network: Building the Industry's First General Fault Expert Agent

To solve the issues of traditional wireless troubleshooting, such as heavy reliance on manual experience, time-consuming fault delimitation, and low coordination efficiency, ZTE and Guangdong Mobile jointly launched the industry's first general fault expert agent. Deeply integrated with Guangdong Mobile's "micro-grid" production management system, the agent reconstructs the wireless maintenance process.

The project builds a unified perception layer covering multiple sources of data such as alarms and KPIs to achieve panoramic fault visualization. It simulates expert diagnosis logic to automate the full process of "abnormality identification–root cause location–handling recommendation." By deploying intelligent agents, the system can automatically generate work orders, trigger parameter adjustment, and perform remote recovery operations, forming a closed-loop process.

As a result, the accuracy of wireless fault root cause diagnosis increased to over 90%, the accuracy of knowledge Q&A increased to over 90%, and MTTR was reduced by 21.34%. The project received recognition as a TM Forum Innovation Hub Pioneer Project 2025 for "Multi-Agent Collaboration: Driving E2E Closed-Loop RAN Faults Management."

Core Network: Building an AI-Agent-Based Automated Fault Handling System

As the service center, the core network has extremely high requirements for stability. However, fault causes are complicated, and traditional monitoring methods cannot respond in a timely manner.

The project builds a core network fault agent, establishes a cross-NE correlation analysis model, and integrates fault knowledge graphs with deep learning algorithms. This improves the timeliness of fault risk identification, provides fault delimiting conclusions and recommended actions, and enhances efficiency of closed-loop work order processing. For the first time in Guangdong, an innovative one-click handling capability has been introduced and workbench integration has been completed, enabling one-click operations such as isolation, switchover, and restart.

In this project, both the coverage rate and accuracy rate of intelligent fault diagnosis in the core network exceeded 90%, realizing intelligent and efficient fault handling.

Transmission Network: Driving Energy-Efficient and Low-Carbon SPN Operation

As a key bearer platform for government-enterprise private lines and cloud-network integration, the SPN network features high device density and high power consumption, making energy efficiency optimization an urgent task.

By combining AI capabilities deployed at the network element level with centralized AI capabilities at the network level, dual-layer intelligent management is enabled to automatically formulate optimal energy-saving strategies, achieve

endogenous intelligence, and support capability exposure.

Based on the core architecture of "dual-layer intelligence and end-to-end energy saving", the project adopts a hierarchical, collaborative design to build a full-link energy-saving ecosystem and a service-lossless policy engine. Using long- and short-term prediction algorithms, the solution analyzes network traffic in real time and dynamically adjusts the energy-saving status of devices, which enables full-process automated energy saving while supporting service assurance, intelligent energy-saving prediction, and energy consumption visibility and controllability. Power savings for a single device exceeded 15%, and annual average electricity costs can be reduced by more than RMB 1 million. A replicable green network operation template has been established.

Network Cloud: CUCD Enables Efficient and End-to-End Change Management

Due to the high frequency and high risk of network cloud changes, as well as the need for manual coordination, hidden faults caused by improper human operations must be strictly prevented.

By opening up OMC real-time data reporting capabilities and automatic network operation interfaces, the system is fully interconnected with the network cloud workbench. Centered on the three stages of "pre-event prevention, in-event control, and post-event verification", the system strengthens self-warning and self-isolation capabilities for change risks at key nodes, including joint review of change solutions, approval of work orders, task preparation, result testing and verification, and on-duty observation during cutover. This establishes an intelligent change monitoring system, enabling end-to-end change automation and ensuring that risks remain manageable and controllable.

In this project, the automation rate of the entire change process exceeded 90%, significantly reducing the risk of human errors, comprehensively improving the standardization, security, and efficiency of change operations, and providing powerful support for building a highly reliable and autonomous network O&M system.

In addition, the two parties completed the deployment of pilot sites for 12 high-value scenarios, including network cloud fault monitoring, transmission network fault monitoring, wireless performance optimization, wireless energy efficiency optimization, and transmission change monitoring, initially building a capability foundation for L4 autonomous networks.

From Technology Breakthroughs to Ecosystem Leadership

The Autonomous Network Demonstration Zone was built over one year and has achieved remarkable results. It has not only significantly improved key O&M indicators, but also established a replicable model.

ZTE and Guangdong Mobile jointly participated in the technical achievement verification of the "Co-Innovation+" Autonomous Network Lab and applied for the Science and Technology Award of the Guangdong Communications Association. Multiple achievements were showcased at the Digital Transformation World (DTW), PT Expo China, the China Mobile Partner Conference, and TM Forum Innovate Asia 2025 in Bangkok.

Based on the Autonomous Network Demonstration Hall, the two parties have built an open platform for achievement demonstration, technical exchange, and talent training, continuously empowering the industrial ecosystem.

This successful practice is a milestone in the in-depth collaboration and joint innovation between ZTE and Guangdong Mobile, and an important step toward higher-level autonomy in the communications industry. From pilot exploration to benchmark demonstration, and from partial optimization to global intelligence, ZTE will continue to work with Guangdong Mobile to develop the demonstration platform into a hub for technological innovation, business value creation, and talent development, contributing to the high-quality development of the digital economy. ZTE TECHNOLOGIES

L4 Wireless Autonomous Network Exploration: Joint Deployment of a Demonstration Zone by ZTE and Fujian Mobile

Liu Yang

Expert Engineer of Wireless Product R&D Planning, ZTE

Song Ziyi

Project Delivery Director, Fujian Branch of ZTE

With the continuous expansion of 5G and the increasing diversity of service scenarios, network O&M is evolving from traditional equipment-, network element-, and KPI-oriented management to a broader emphasis on service quality, user experience, and operational value. Meanwhile, network issues are becoming increasingly cross-domain, dynamic, and complex in their impact propagation, making manual, single-domain O&M approaches insufficient for fault localization, accurate handling, and end-to-end closed-loop management.

Against this background, autonomous networks (AN) have emerged as a key direction for operators. In line with group-level requirements for AN Level 4 (L4) development, China Mobile Fujian (Fujian Mobile), in collaboration with ZTE, has established a wireless autonomous network demonstration zone. By piloting AI agents and an integrated perception platform and gradually incorporating these capabilities into live networks, the project has explored a practical path from technical verification to scaled deployment.

Overall Construction Strategy

The construction of Fujian's wireless autonomous network has focused on live-network production scenarios, particularly high-frequency use cases such as network optimization, energy saving, fault management, and user experience assurance. The work follows the principle of demand-driven deployment, process integration, data enablement, and coordinated evolution.

Capability deployment is driven by practical production needs, aiming to improve O&M efficiency, reduce operational cost, and enhance user experience. Agent capabilities are embedded into key operational stages, including problem identification, analysis and handling, execution, and feedback, progressively establishing closed-loop workflows. In parallel, the integrated perception platform is leveraged to aggregate and correlate

multi-source data from network elements, terminals, signaling, services, and customer complaints, thereby providing unified support for network optimization and user experience assurance.

On this basis, ZTE and Fujian Mobile have established a dual-track development framework of "AI agents + integrated perception platform." The AI agents mainly target O&M efficiency improvement and fault handling, while the integrated perception platform focuses on end-to-end service experience evaluation, user-perception assurance, and operational support. Together, they form the core of the wireless autonomous network demonstration zone in Fujian.

Deployment and Performance of Wireless Autonomous Network Agents

Fujian Mobile and ZTE have established wireless autonomous network demonstration zones in Fuzhou and Quanzhou. Based on intelligent boards, gSDU, beam-tracking antennas, OTDR, and related facilities, scenario verification has been completed in representative areas such as Quanzhou West Street and Fuzhou University Town, forming an application model integrating AI agents, intelligent hardware, and production workflows.

Performance Optimization Agent

The performance optimization agent has been deployed in Quanzhou West Street and Fuzhou University Town and interfaced with the customer operations workbench. It can automatically identify issues based on multidimensional data, including coverage, traffic load, and service quality degradation, generate optimization recommendations, and track execution, enabling closed-loop work-order management. By the end of 2025, the automated closed-loop rate of performance optimization work orders reached 34.6%, up 14.8 percentage points year on year. In some scenarios, optimization schemes could be generated within minutes, demonstrating strong adaptability to live networks.

Energy Efficiency Optimization Agent

The energy efficiency optimization agent has been piloted across approximately 100 BBUs in conjunction with intelligent power-saving units. The agent dynamically adjusts energy-saving strategies based on service load and site operating conditions. Pilot results indicate an overall energy reduction of 10.71%, which validates the effectiveness of fine-grained, load-aware energy-saving policies.

Experience Assurance Agent

The experience assurance agent has mainly been applied to scenarios such as traffic stimulation and business-network coordinated subscriber-activation forecasting. Subscriber-activation forecasting has been verified in scenarios such as Fuzhou Metro and Quanzhou West Street. The overall provisioning estimation accuracy reached 98.12%, and the user experience fulfillment rate reached 95.17%. These results show that the agent is transitioning from conventional passive assurance toward proactive demand prediction and operational support.

Fault Monitoring Agent

The fault monitoring agent, integrated with OTDR capabilities, has been piloted in Fuzhou and Quanzhou, achieving an 18% reduction in mean time to repair (MTTR). In addition, maintenance personnel can use a mobile application to perform one-click inquiries, intelligent diagnosis, and spare-part pre-configuration, thereby establishing an operation model characterized by pre-arrival fault localization and immediate handling upon dispatch.

By advancing fault knowledge, localization paths, and spare-part recommendations forward to the dispatch stage, this model reduces uncertainties in on-site troubleshooting and improves both frontline maintenance efficiency and operational standardization.

Integrated Perception Platform: From Network Metrics Analysis to End-to-End Experience Perception Enablement

Traditional network optimization has mainly focused on network-element indicators and cell-level

> The practice in Fujian indicates that the key to autonomous network development lies in establishing an end-to-end closed loop encompassing perception, analysis, decision-making, execution, and evaluation

performance, with limited visibility into actual user experience. In particular, challenges have persisted in complaint analysis, cross-domain demarcation, and root-cause localization. To address these issues, Fujian Mobile has also piloted an integrated user-experience perception platform that supports perception discrepancy identification, cross-domain demarcation, root-cause analysis, and complaint handling across data services, voice services, and customer complaints.

For problem demarcation, the platform correlates multi-source data from network elements, terminals, signaling, and services to construct an end-to-end view from user terminals to service servers, addressing issues of unclear fault boundaries and inaccurate responsibility attribution in traditional single-domain analysis.

For root-cause analysis, the platform combines signaling traceback with high-precision grid-based positioning to reconstruct complaint events in both temporal and spatial dimensions, improving root-cause localization accuracy from the cell level to the 100-meter level.

For closed-loop processing, the platform has been integrated with a mobile application, enabling frontline personnel to perform online data queries, submit feedback, track processes, and report results, facilitating standardized, closed-loop complaint handling.

The platform transforms user experience into a analyzable, measurable, and traceable object. It provides a unified data foundation for both the experience assurance agent and the network optimization agent, promoting the evolution of network optimization from a sole focus on network performance toward a balanced emphasis on both network performance and user experience.

Conclusion

The wireless autonomous network and integrated perception platform jointly developed by ZTE and Fujian Mobile have achieved phased results in network optimization, energy saving, fault handling, user experience assurance, and operational support. Specifically, the deployment has realized a 34% self-closed-loop rate for performance optimization work orders, a 10.71% reduction in 5G site energy consumption, and an 18% reduction in MTTR. In addition, promising potential has been demonstrated in scenarios such as traffic stimulation, subscriber-activation forecasting, and complaint handling support.

The practice in Fujian indicates that the key to autonomous network development lies in establishing an end-to-end closed loop encompassing perception, analysis, decision-making, execution, and evaluation, enabling the effective integration of AI agents, data platforms, and production systems. As coordination mechanisms, perception capabilities, and standardized interfaces continue to mature, this practice is expected to evolve into a replicable and scalable paradigm for autonomous network development, providing a useful reference for operators pursuing digital and intelligent network transformation. ZTE TECHNOLOGIES

Grameenphone Partners with ZTE to Launch an Agentic AI Pilot for Network Fault Management

Ru Le

Wireless R&D Planning Expert Engineer, ZTE

Grameenphone, a subsidiary of the multinational telecom group Telenor, is Bangladesh's largest telecom operator, with approximately 85 million subscribers and about 45 percent market share. Since March 4, 2025, ZTE has assumed full responsibility for Grameenphone's end-to-end network operations and maintenance (O&M).

Following the takeover, ZTE analyzed live network data and found that wireless network faults accounted for over 70% of all network incidents, making them the primary focus of the NOC. Consequently, both parties decided to prioritize the deployment of an autonomous network pilot targeting cross-domain and wireless fault management, with a focus on wireless cell outages and base station outages. The initiative covers more than 4,000 ZTE-managed 4G sites in the region and introduces two AI agents—the cross-domain fault expert and the wireless fault Copilot—to empower key fault management processes (Fig. 1).

- In the fault identification phase, the cross-domain fault expert employs a hybrid large-small model architecture, where the small model handles dynamic data aggregation and the large model generates concise event summaries, enabling minute-level fault identification.
- In the root cause localization phase, the cross-domain fault expert leverages chain-of-thought (CoT) reasoning over multi-dimensional data from engineering, environment, equipment, and network management systems, constructing a knowledge graph to support root cause identification, generate fault conclusions, and autonomously infer fault propagation paths.
- In the handling recommendation phase, based on CoT reasoning, optimal fault handling recommendations are provided. For faults caused by wireless equipment, the wireless fault agent can be invoked to trigger automatic restart and recovery to achieve self-healing. For faults that cannot be self-healed, the cross-domain fault expert automatically analyzes the faults and dispatches work orders to the first-line maintenance (FLM) personnel for on-site handling.
- In the on-site operation phase, the wireless fault

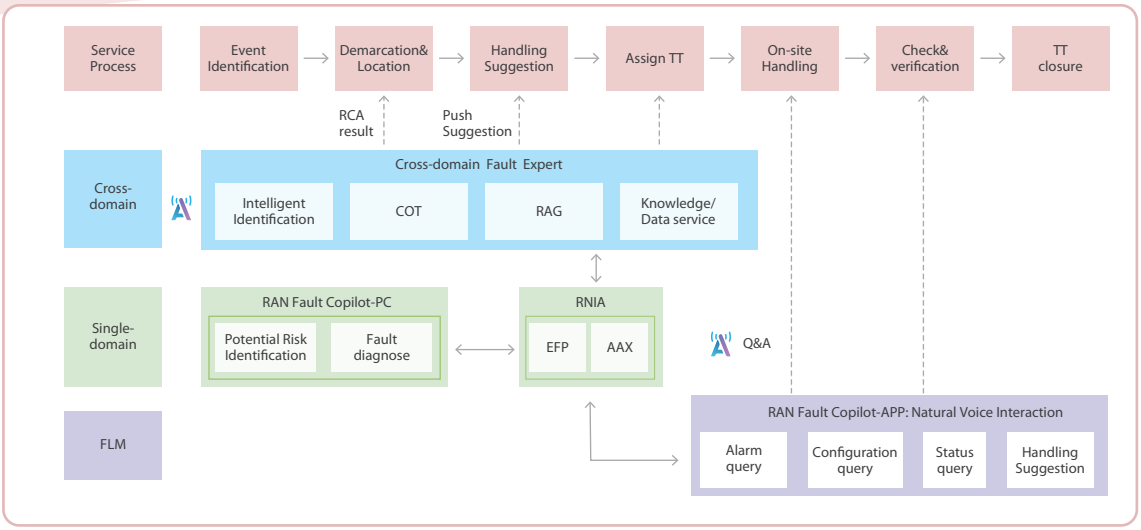


Fig. 1 LLM-based fault management solution.

Copilot brings fault knowledge, backend data, and atomic capabilities to mobile devices. Upon arrival at the site, FLM personnel can use the Copilot app for real-time voice interaction and fault information queries. The Copilot presents site topology, the root cause, and corresponding solutions through multimodal visualization. After handling the fault, FLM personnel can query and verify the repair status on site, significantly reducing troubleshooting complexity, enabling one-time task completion, and improving FLM fault handling efficiency.

By embedding large model and agent capabilities into the existing network O&M system, minute-level intelligent identification of wireless cell and site service outage faults can be achieved, with fault localization and root cause analysis accuracy exceeding 90%. Through agent collaboration, manual workload is reduced, driving more efficient fault closure. The solution is expected to achieve a 70% increase in Q&A efficiency, a 3% reduction in wireless fault tickets, and a 20-minute reduction in wireless fault MTTR. Additionally, with the introduction of these two agent applications, the autonomous level of Grameenphone's wireless fault management scenario will be elevated to L3+.

At MWC 2026, Grameenphone and ZTE signed a Memorandum of Understanding (MoU) to strengthen collaboration on large-model and agent technologies aligned with TM Forum's latest autonomous network specifications and Grameenphone's three-year network development goals.

Looking ahead, ZTE and Grameenphone will further expand their autonomous network collaboration by leveraging the AIR Net advanced autonomous network solution to comprehensively scale up high-value scenarios in wireless network fault management. They will accelerate the deep integration of the cross-domain fault expert and wireless fault Copilot into Grameenphone's live network O&M operations, maximizing business value for Grameenphone.

Additionally, the two parties will collaborate on other high-value autonomous network use cases, including wireless network optimization, customer complaint handling, and core network change management, deploying a series of AI agents—including network optimization experts, customer complaint experts, and core network fault experts. Leveraging the agentic AI architecture and A2A-T protocol, the collaboration will strengthen single-agent closed-loop and multi-agent collaboration capabilities to enable cross-domain and cross-layer intelligent orchestration, significantly improving network operational efficiency and ensuring optimal user experience. Meanwhile, they will jointly incubate TM Forum-related Catalyst projects to validate and promote key autonomous network technologies, providing a replicable AI-enabled blueprint for the global telecommunications industry. ZTE TECHNOLOGIES

ZTE

To lead in connectivity and intelligent computing, enabling communication and trust everywhere