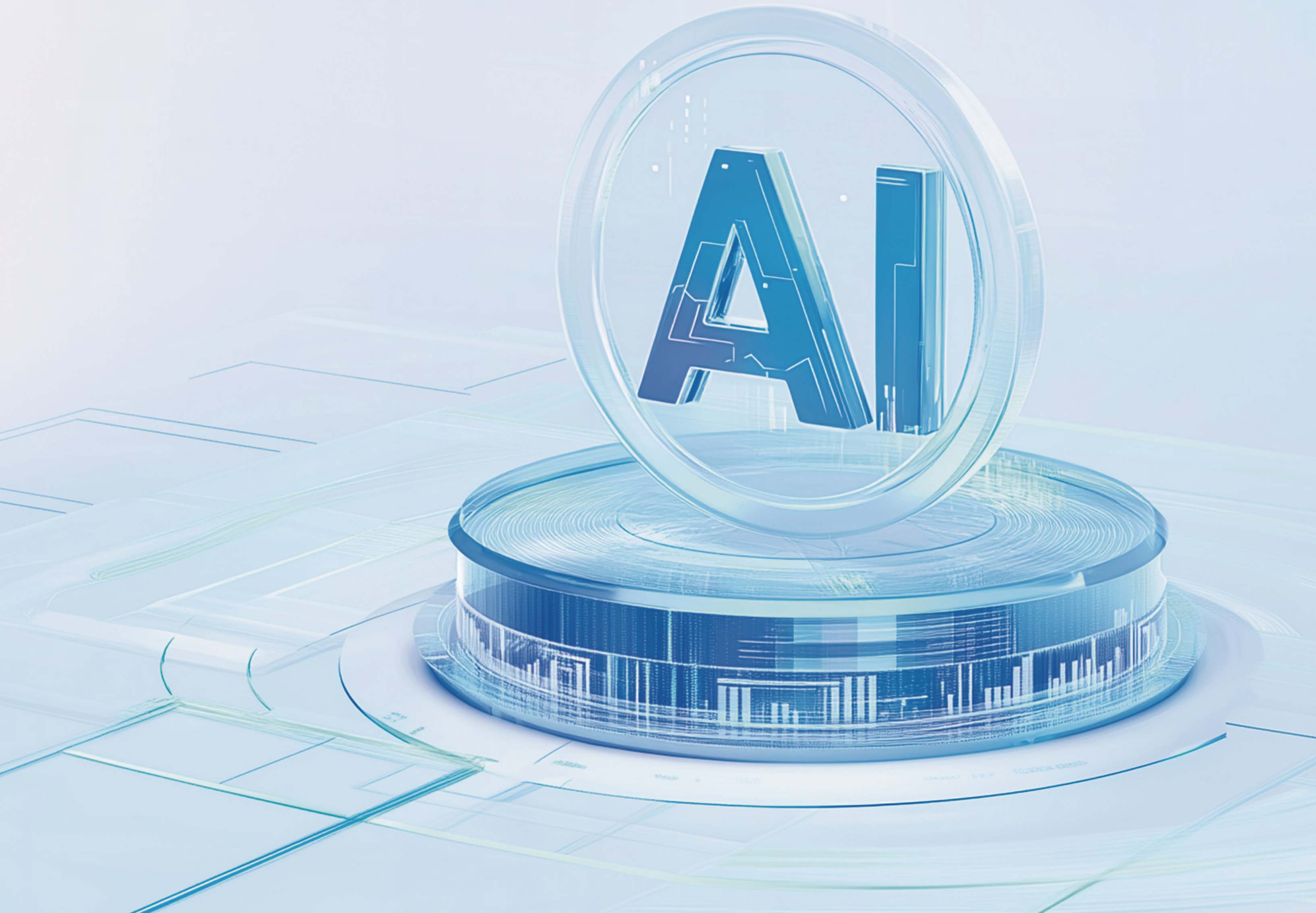
ZTE

AIR RAN White Paper

AI RAN Pioneer Empowering All-Scene Business Expansion and Innovation



Contents

Foreword

In an era where information technology is advancing by leaps and bounds, artificial intelligence (AI), as the core driving force of the Fourth Industrial Revolution, is reshaping the global industrial landscape with unprecedented depth and breadth. From intelligent diagnosis and precision treatment in the medical field to smart risk control and high-efficiency trading in the financial sector, the innovation and application of AI have become the key forces driving the transformation and upgrading of various industries, profoundly altering the operational dynamics of socio-economic systems.

Amid this wave of transformation, the pervasive development of AI is continuously propelling the deep integration of wireless access networks (RAN) and AI. On one hand, AI serves as the core engine for elevating mobile networks to a higher level of intelligence (AI for RAN); on the other hand, significant enhancements in network capacity, such as bandwidth, data rates, and connection density, coupled with the convergence of communication, perception, and computing, provide a broader stage for ubiquitous AI applications (RAN for AI).

ZTE Corporation has released the AIR RAN (AI Reshaped RAN) solution, building on its deep expertise in both AI for RAN and RAN for AI, ZTE is exploring the infinite possibilities of intelligent network innovation. Specifically, in terms of experience enhancement, differentiated service assurance lays the foundation for transforming traffic management into experience management; regarding operational efficiency, the use of intelligent algorithms enables precise fault prediction and rapid repair, significantly improving network reliability and stability; in energy efficiency, real-time analysis of network traffic and user behavior allows for dynamic adjustment of base station power, effectively reducing energy consumption and promoting the green and sustainable development of the communications industry; and in expanding new AI applications, emerging areas such as intelligent agents, embodied intelligence, and the Internet of Vehicles are being fostered, meeting the increasingly diverse and personalized needs of users and transforming mobile networks from mere data transmission channels into enablers of intelligent services.

This white paper closely follows the developmental trajectory of the integration of AI and wireless access networks (RAN), offering an in-depth analysis of the technical challenges and opportunities involved. It elaborates on integration trends, key technologies, and commercial explorations with the aim of providing industry professionals with insightful perspectives, helping them seize emerging opportunities, and collectively advancing the industry towards a new, intelligent future.

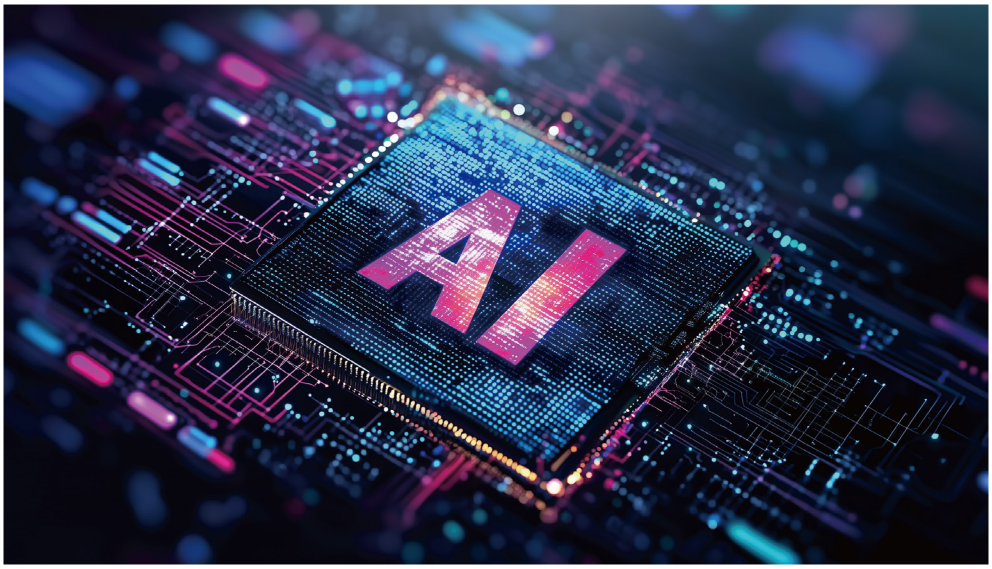
01 AI Innovations: Advancing Technology and Commercial Scale

In recent years, artificial intelligence has made tremendous progress, particularly in key technological areas such as machine learning, natural language processing, and computer vision. These innovations are rapidly driving AI out of research laboratories and into practical applications, accelerating its commercialization across various industries. Notably, large language models like GPT-o1 are at the forefront of this transformation, fundamentally changing the interaction between AI and language, knowledge, and decision-making. These large models not only lay the foundation for the widespread adoption of AI technologies but also unleash unprecedented value in multiple sectors including healthcare, finance, manufacturing, and entertainment.

The development of large language model technology has provided the essential groundwork for this revolution. The rapid advancements in these models have greatly enhanced machines' abilities to understand and generate human-like language, revolutionizing fields from customer service to the creative industries. The latest large models boast hundreds of billions or even trillions of parameters, demonstrate remarkable capabilities in contextual understanding, tone perception, and intent recognition, thereby making AI-driven applications more efficient, intuitive, and user-friendly. AI is increasingly recognized as a key driver of value creation across various domains. In healthcare, significant progress has been achieved through AI, aiding in more accurate diagnoses, predictive analytics, and personalized treatments. According to McKinsey, AI is projected to generate over USD 150 billion in annual value in the healthcare sector by 2026 [1]. In finance, AI-powered tools are streamlining risk management, fraud detection, and algorithmic trading, thereby enhancing efficiency and profitability. The impact of AI in manufacturing, retail, and logistics is also substantial, where AI-driven automation helps reduce operational costs, improve customer experiences, and boost productivity.

The communications network sector is similarly benefiting from artificial intelligence, with system vendors and operators jointly pushing for revenue growth and cost reductions through AI applications. By leveraging AI in areas such as predictive maintenance, network optimization, and customer service, telecom companies are able to significantly enhance efficiency. AI enables mobile operators to predict network congestion, optimize bandwidth allocation, and even reduce downtime by anticipating and addressing technical issues. Data shows that in 2021, the global AI market in the telecom industry was valued at USD 1.2 billion, and it is expected to grow to over USD 6.3 billion by 2026, representing a compound annual growth rate of 38% [2]. AI-driven chatbots and virtual assistants have also improved the quality of customer service, boosting user satisfaction while reducing labor costs. Moreover, AI-based data analytics allow mobile operators to offer personalized services, tap into new revenue streams, and enhance customer loyalty.

On the other hand, as AI applications continue to evolve, create greater value, and become more inclusive, they will undoubtedly impose higher and more diverse demands on communication networks, particularly mobile networks. This will enable AI to be accessed and utilized across various scenarios, reshaping user habits and transforming lifestyles and production methods. The further proliferation of AI and the development of diverse AI applications will, in turn, drive increased requirements for network bandwidth, latency, and reliability, fostering diversified collaboration across cloud, edge, and terminal infrastructures, and ultimately promoting innovation and advancement in network technologies.



02 AI and RAN: Convergence at Full Speed

2.1 AI Adoption in Wireless Networks

The GSMA's “Mobile Economy 2024” report [3] indicates that by 2030, the proportion of 5G connections is expected to reach 56%, with mobile data traffic growing at a compound annual growth rate (CAGR) of 23%. This suggests that mobile operators will need to maintain continuous investment in 5G deployment to accommodate evolving network traffic. The large-scale deployment of 5G MIMO multi-antenna technology will incur enormous investment costs, with total capital expenditures expected to reach USD 1.5 trillion by 2030. However, operators' return on investment is forecast to decline from 19% in 2022 to 14% in 2030. To reduce costs and enhance efficiency, operators urgently need to accelerate the automation of networks and services. On one hand, this involves pushing AI technologies down into network operations for optimization—using artificial intelligence to enhance network maintenance, improve network quality, provide precise optimization insights, and intelligently identify and prevent potential faults. On the other hand, by combining intelligent solutions with business recognition techniques, operators can achieve refined energy savings, stimulate key business traffic, and offer differentiated experience guarantees through initiatives such as new 5G-A packages.

From a technological development perspective, the convergence of communications, cloud computing, artificial intelligence, and digital twin technologies has become a key trend in the evolution of next-generation networks. As 5G commercial deployment deepens, expanding integrated computing-network services based on base station capabilities has become a focal scenario for operators. For consumer (B2C) applications, pooling and sharing base station computing resources to form a network can meet the ultra-low latency and high-bandwidth demands of local computing applications, such as local XR content rendering, embodied robotics, and immersive interactive applications. For enterprise (B2B) applications, localized integrated communication and computing services enable flexible and agile deployment of services while effectively reducing the overall cost of industrial applications. Typical scenarios include industrial visual AI inspection and intelligent collision prediction in the Internet of Vehicles.

2.2 AI-Driven Wireless Network Transformation

The rapid development of new AI services is driving a comprehensive upgrade of mobile networks. In terms of interaction, the evolution is shifting from solely person-to-person interactions to interactions between humans and AI agents, among AI agents themselves, and between AI agents and other devices. In terms of content, the fusion of multiple modalities and the deep integration with the real world are spawning richer and higher-quality generative AI (GenAI) content formats. These changes impose entirely new requirements on network performance and architecture, pushing mobile networks to transform from traditional connectivity services into intelligent services that combine high-performance networking and computing.

Transformation 1: GenAI Sparks Explosive Growth in Content Quantity, Video Quality, and User Engagement, Fueling Network Traffic Surge

Applications of AI in content generation are expanding rapidly. The market for AI-generated video is expected to grow at a CAGR of 19.9% from 2024 to 2030 [4]. With advances in AI algorithms, the computational efficiency required to generate high-resolution video has significantly improved. For instance, technologies such as GANs (Generative Adversarial Networks) can now efficiently generate high-resolution video content. Currently, most AI-generated short videos achieve 2K resolution, while advanced models like StyleGAN3 have already surpassed 8K resolution for static images, with video generation resolution steadily advancing toward similar levels.

In addition, GenAI is reshaping immersive extended reality (XR) and 3D content generation by breaking through the traditional bottlenecks of low efficiency and high cost in content creation, paving the way for substantial growth. Immersive content demands ultra-high-speed data transmission; for example, an XR application for a single user at 8K resolution requires at least 200 Mbps of downlink bandwidth, and in the future 6G era, the per-user bandwidth demand could exceed 20 Gbps.

AI-driven hyper-personalization and enhanced experiences are greatly increasing user engagement. For example, Spotify uses AI to provide personalized recommendations, resulting in a 20% increase in premium subscription registrations; Snapchat has launched a new AR experience integrated with GenAI that generates real-time AR content, leading to a 40% increase in daily active users utilizing AR filters.

According to GSMA's predictions, mobile network traffic will grow more than fourfold by 2030. The disruptive transformation brought about by AI could very well break the existing traffic growth patterns, ushering in even more changes.

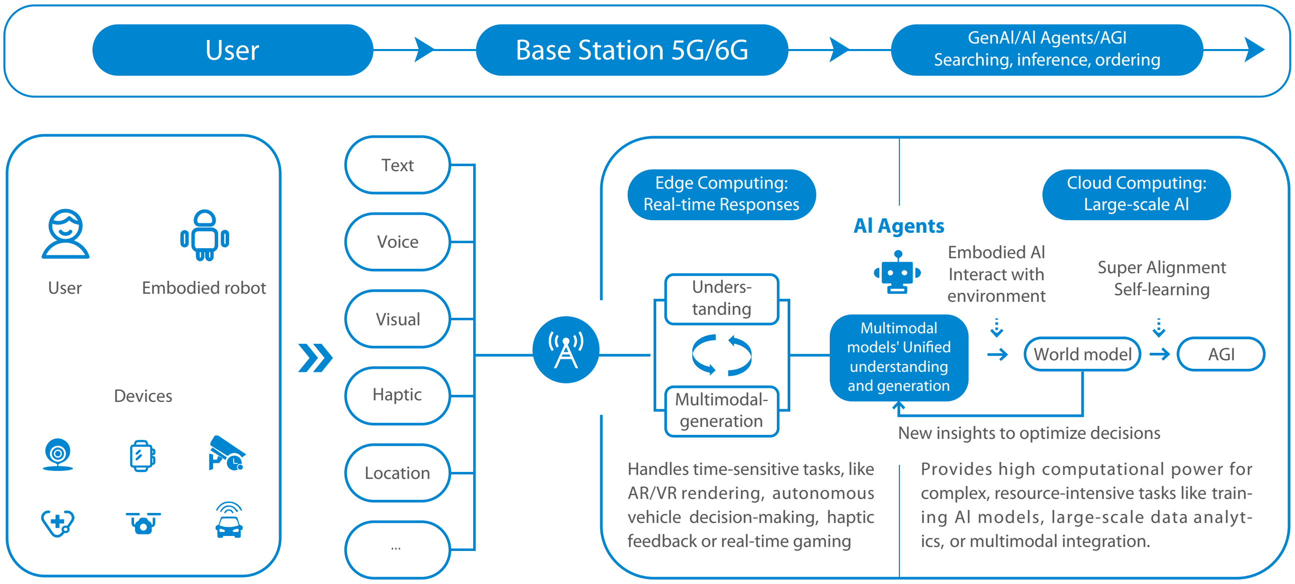
Transformation 2: AI Agents Will Become Ubiquitous Applications, Necessitating Simultaneous High Bandwidth and Low Latency, and Triggering Multi-Directional Data Flow Pattern Changes

From personal assistants to enterprise smart customer service, and even decision engines in industrial automation, the real-time interactions in these scenarios impose stringent requirements on network performance. The network data flow is evolving from the traditional model of undifferentiated user data to a model where business data is transmitted along a single path from the user to the server, upgrading to a multi-path transmission model that involves multiple interactions between users, edge servers, AI agents, cloud servers, and various nodes. Typical real-time interactions involving AI agents require simultaneous support for high-speed uplink and downlink data transmission, as illustrated in the table below:

AI Agent Types	Throughput	Delay	Use cases
Personal Assistant	10-20Mbps（UL）	<50ms	Smart speakers, mobile assistants
Enterprise Customer Service	20-50Mbps（UL）	<20ms	Customer service, business support
Industrial Automation AI	<100Mbps（DL&UL）	<10ms	Unmanned workshops, machine collaboration, remote maintenance
Complicated multi-modalityAI	20Mbps-100Mbps (UL) 100Mbps-500Mbps (DL)	<10ms for multi-modalitysensing <200ms for audio/video streaming	Holographic communication, virtual humans, embodied robot

Transformation 3: The Collaborative Model of Cloud-Edge-Terminal Integrated Computing is an Exploratory Direction for AI Service Development

The diversity and complexity of AI services make traditional, single-tier computing architectures inadequate. From real-time processing to computational scale, as well as data privacy and collaborative efficiency, the demands for computing resources in AI services are growing multidimensionally.


User
Base Station 5G/6G
GenAI/AI Agents/AGI
Searching, inference, ordering
User
Embodied robot
Devices
Text
Voice
Visual
Haptic
Location
...
Edge Computing:
Real-time Responses
Cloud Computing:
Large-scale AI
AI Agents
Embodied AI
Interact with
environment
Super Alignment
Self-learning
Unders-
tanding
Multimodal-
generation
Multimodal
models' Unified
understanding
and generation
World model
AGI
New insights to optimize decisions
Handles time-sensitive tasks, like AR/VR rendering, autonomous vehicle decision-making, haptic feedback or real-time gaming
Provides high computational power for complex, resource-intensive tasks like training AI models, large-scale data analytics, or multimodal integration.

Cloud-Dominated Collaboration: Suitable for scenarios with high computational demands but relatively low latency requirements. In this model, the cloud handles the bulk of the computation, edge devices manage task distribution and data caching, and terminals perform lightweight computing. For example, in complex AI model training, the cloud provides massive computational power, while the edge handles localized responses to user requests.

Edge-Dominated Collaboration: Ideal for applications with extremely high latency sensitivity, also saving significant wired transmission resources. Examples include autonomous driving and embodied robotics. In this model, real-time task processing is primarily carried out by edge devices, while the cloud is responsible for long-term data storage and model optimization. For instance, through deep real-time collaboration in data, computing power, and models between edge devices and smart vehicles, tasks such as road condition recognition, joint monitoring of hazardous events among multiple vehicles, and edge-based decision-making can be achieved, while the cloud optimizes global routing.

Terminal Intelligent Collaboration: Terminal devices handle simple computing tasks, while the edge and cloud work together to complete more complex computations. For example, in AR/VR scenarios, AR glasses can perform basic local rendering, with complex environmental modeling delegated to edge computing, and the cloud generating a global environmental model.

2.3 Telecom Operators' Edge AI Advantages

Telecom operators' mobile networks are inherently wide-area distributed networks, which naturally align with the distribution of mobile and AI services. Leveraging the advantages of the network edge, operators have long been exploring edge computing; Mobile Edge Computing (MEC) is a prime example. By pushing MEC deeper and more peripherally into the network, operators can achieve localized deployment of services, thereby effectively reducing latency, bandwidth costs, and terminal expenses, while enhancing service quality and data security. ZTE, for instance, has been at the forefront of exploring edge computing. In 2020, it introduced the NodeEngine computing engine, which provides edge computing capabilities through a board at the base station, enabling local traffic offload and flexible deployment of edge applications. In 2023, ZTE launched the industry's simplest UniEngine all-in-one device, which, in addition to delivering computing power, achieved full integration of 5G private network elements. In 2024, ZTE unveiled the AIREngine intelligent computing engine, which integrates AI directly into the base station through an intelligent computing board.

Compared with remote cloud computing, the edge computing capabilities of telecom operators offer distinct advantages in several scenarios:

Localized Scenarios: Typical examples include private networks in various industries. Enterprises often have high requirements for data sensitivity and security, necessitating that data remains on-premises. As operators build private networks, providing integrated computing-network solutions facilitates on-site deployment of applications—including AI applications—within enterprise campuses. In industrial control, for instance, traditional ISA-95 is evolving toward cloud-edge-terminal architectures. With the introduction of scenarios such as machine vision for quality inspection and predictive maintenance, industrial sites generate a significant amount of both real-time control commands and non-real-time data. This gives rise to the concept of edge control—a new form of control that integrates industrial device edges, telecom network edges, and computing edges. Operators can leverage integrated communication and computing infrastructures to provide IT/CT/OT converged equipment for industrial sites. For example, ZTE's NodeEngine base station/UniEngine integrated device, applied in automated logistics sorting, significantly reduced the number of on-site industrial PCs by incorporating built-in computing power, thus centralizing and cloudifying PLC control.

Wide-Area Scenarios: These involve applications with high real-time requirements. Services aimed at human users are generally less sensitive to latency and benefit more from cloud-based processing, whereas control-oriented IoT services are extremely latency-sensitive. A typical example is the integration of vehicle-road-cloud scenarios, where roadside sensing (including cameras, radars, and traffic light data) is analyzed and then transmitted via V2X cloud control to vehicles for driving assistance. This scenario is latency-sensitive, and by offloading computing to the base station to perform roadside sensing analysis and decision-making, driving safety can be better ensured. Similarly, as large models become more prevalent in immersive devices and embodied intelligence applications, offloading rendering and control functions to the edge to reduce latency will provide a significantly improved real-time experience.

03 AI and RAN Fusion: Unlocking Potential in a Dynamic Landscape

3.1 Challenges in AI RAN

While the integration of AI and RAN brings unprecedented opportunities to the communications industry, it also introduces corresponding challenges in data collection, algorithm design, and the utilization of computational resources within RAN.

In terms of data granularity, AI RAN must handle a diverse array of data types—such as those from the Internet of Vehicles and XR applications—where the data collection cycles must be real-time to satisfy the stringent latency requirements of AI services. This sharply contrasts with traditional networks, which typically feature fixed data types and longer collection cycles. For instance, in vehicular networks, evasive maneuvers (like "ghost head" avoidance decisions) require real-time judgments in environments with high-speed movement and uneven latency distribution, thereby imposing even higher demands on the timeliness of data collection.

Regarding algorithm matching, traditional model designs emphasize interpretability and ease of inference implementation; however, they often entail complex parameter configurations and high manual overhead. In contrast, AI model designs, though less interpretable, feature simpler parameter settings that can achieve a statistically optimal global solution. Their drawback, however, is that AI models may require retraining when application scenarios change—leading to potential performance degradation and additional overhead. For example, in autonomous driving, AI models must be dynamically adjusted according to varying road conditions and traffic situations to ensure safety and reliability.

The efficiency of computing resource usage is another significant challenge for AI RAN. AI terminal forms are diverse—from XR glasses to automobiles and embodied intelligent devices—with computational requirements ranging from roughly N×10 TOPS to N×100 TOPS. Additionally, the tidal distribution characteristics of traffic impose higher demands on dynamic resource allocation within the network. For instance, the differing traffic volumes between residential areas and central business districts during daytime versus nighttime require dynamic resource allocation to maximize returns. These changes necessitate new approaches to the evolution of base station computing power and resource allocation. Base station capabilities are expected to evolve from dedicated hardware toward heterogeneous computing solutions to meet differentiated computational needs and diverse application scenarios. Examples include expanding general-purpose computing boards on dedicated hardware and employing intelligent scheduling algorithms to optimize resource distribution, ensuring efficient utilization during both peak and off-peak periods.

The deep integration of AI and RAN is propelling the communications industry toward greater intelligence and efficiency. By continuously optimizing data processing, algorithm design, and computing resource utilization, telecom operators are evolving from merely empowering wireless networks with AI to achieving a dual collaboration between AI and RAN. In this transformation,AI RAN becomes a critical pillar for enhancing user experiences, delivering more convenient and efficient services. In the future, as technology further matures and application scenarios continue to expand,AI RAN is poised to become the core driving force behind industry transformation and will likely spawn entirely new commercial opportunities.

3.2 Accelerated Capitalization and Expansion AI Benefits in RAN

As AI technology continues to evolve, mobile operators have already leveraged AI to empower wireless network operations, yielding significant improvements in user experience, energy efficiency, and operational maintenance efficiency, while also opening up new application scenarios.

Emergence of New Services: With the rise of live streaming, gaming, and other new services, users now demand a more diversified network experience. Although traffic volumes are increasing, the growth in ARPU (average revenue per user) remains slow—posing challenges in customer acquisition and revenue growth. AI-driven, end-to-end service packages can precisely sense business needs and allocate resources, thereby enhancing user experience and network resource utilization, which in turn helps operators boost revenue and drive industry-wide intelligent upgrades.

Energy Consumption Challenges: The communications industry is highly energy intensive, and the widespread deployment of 5G has exacerbated the issue of wireless network energy consumption. By incorporating AI into energy management for fine-grained control, operators can improve energy utilization while maintaining service quality—thereby promoting greener development. For example, intelligent energy-saving controls can be applied to base stations in commercial areas.

Revolutionizing Operational Maintenance: Intelligent agents are transforming the field of network operations by streamlining maintenance processes, enhancing system capabilities, and enabling multi-agent collaboration with precise fault prediction. This results in significantly improved operational efficiency and drives the evolution toward higher-level intelligent maintenance. The industry is also exploring new opportunities that arise from the integration of RAN and AI.

Under current technological trends, the industry is actively exploring two main scenario lines: one focused on RAN connectivity and the other on AI applications. In future scenarios of deep AI RAN integration, these can be primarily divided into:

Scenarios Dominated by RAN Connectivity: In these scenarios, research on how AI can enhance RAN performance is gradually becoming more focused and shifting in direction—from the wireless application layer (L3) and link layer (L2) toward the physical layer (L1). The IMT-2030 (6G) Promotion Group [5] has initiated analyses and research on applying AI technologies across the 6G wireless physical, link, and application layers, especially exploring the optimization potential and value extraction of deep learning at the physical layer. Simultaneously, the decision between adopting multiple modular small AI models or a single end-to-end model at L1 will have a trend-setting impact on RAN product architectures.

Scenarios Centered on AI Applications: In these scenarios, cloud collaboration has become an inevitable trend, and the industry is actively examining the value of edge-side AI and the necessity of cloud-edge-terminal collaboration. Leveraging their extensively distributed wireless networks, mobile operators can efficiently deploy computing resources at the edge to achieve deep collaboration in data, models, and computing with terminal AI services. This approach meets requirements for data privacy, low latency, efficient real-time responses, and rapid, cost-effective deployment. However, for AI applications that demand mandatory cloud-edge-terminal collaboration, such as embodied robotics, immersive multi-modality interaction, and autonomous driving; the industry is still exploring deterministic value in niche scenarios. Moreover, in scenarios that focus primarily on AI applications while also considering RAN connectivity, questions remain, such as whether to adopt the SoftBank led universal architecture [6] and whether transforming widely deployed wireless network infrastructures will yield a deterministic return on investment. These issues warrant further in-depth study.

3.3 Commercialization Trends in AI RAN

Artificial intelligence is spearheading a profound technological transformation, in which mobile networks are not only providers of connectivity but are also evolving into active enablers of innovation by opening up their computing power and AI capabilities. With the continued development of 5G and 6G technologies, mobile networks are set to expose their robust computational resources and AI processing capabilities to a wide range of applications, thereby creating entirely new business models and service scenarios and driving the rapid growth of the digital economy.

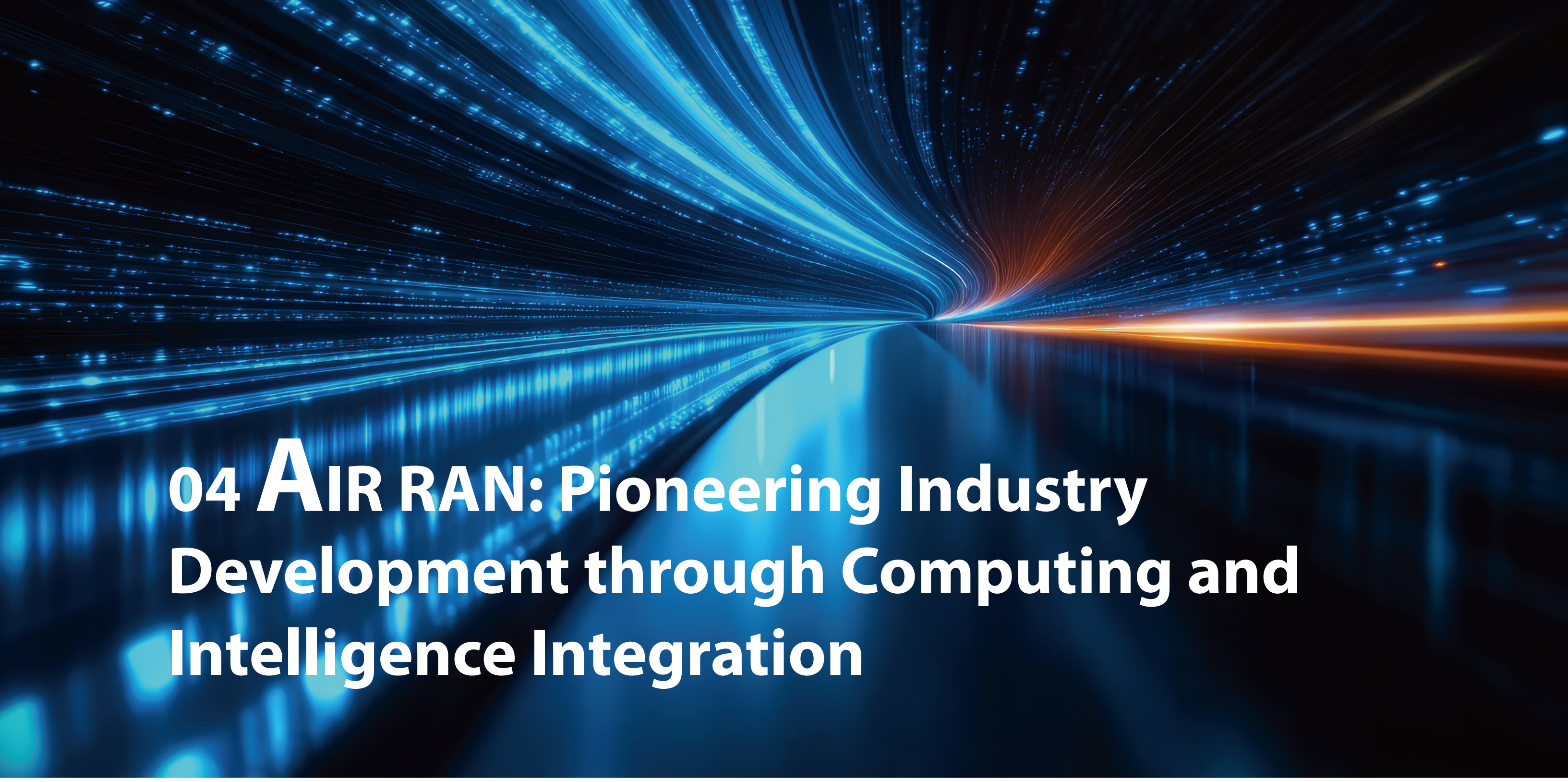
In numerous fields, mobile networks are moving beyond merely offering high-speed connectivity. They are beginning to open up their edge computing and AI inference capabilities to support industry applications, helping enterprises lower deployment barriers and accelerate digital transformation. With the support of network-side computing power, these industries can achieve intelligent upgrades more economically and efficiently.

Robotics: For robotic applications, mobile networks can expose capabilities such as AI-based visual analysis and real-time decision-making to assist robots in performing complex recognition and planning tasks. Edge nodes in the network can provide nearby computing support to ensure ultra-low latency and reduce terminal costs. Ubiquitous network coverage further enables robots to access these capabilities at any time, facilitating intelligent operation in any scenario.

Extended Reality (XR): In the XR domain, mobile networks not only provide high-bandwidth transmission but can also offer cloud rendering, spatial computing, and other advanced capabilities to help XR devices deliver more realistic virtual experiences. Network-side computing support can significantly reduce the hardware requirements for XR terminals, enabling even lighter XR devices to present high-quality content.

Autonomous Driving: For autonomous driving applications, mobile networks can expose their AI sensing and decision-making capabilities to vehicles and traffic management authorities to support real-time analysis in complex environments. Edge nodes can offer supplementary computing power to vehicles, thereby enhancing the safety and reliability of autonomous driving, while ubiquitous network connectivity ensures that vehicles can access these capabilities whenever needed.

By opening up their computing power and AI capabilities, mobile networks are evolving from mere connectivity providers into true enablers of innovation. This transformation not only meets the growing demands for computing power, low latency, and ubiquitous connectivity posed by emerging applications, but also stimulates further innovation by making these capabilities widely accessible—thereby creating entirely new commercial value. As technology continues to evolve, the deep integration of mobile networks with AI will jointly drive industries toward a smarter, more efficient future.



04 AIR RAN: Pioneering Industry Development through Computing and Intelligence Integration

The deep integration of artificial intelligence (AI) and wireless access networks (RAN) not only drives technological innovation and transformation but also spawns entirely new business models and market opportunities. However, this process inevitably faces numerous challenges. In response, ZTE has taken the lead in technological exploration within the two key areas of AI for RAN and RAN for AI, launching the AIR RAN (AI Reshaped RAN) solution. This solution builds a multi-scenario intelligent computing and communication convergence product system and establishes a comprehensive digital intelligence foundation. By building an all-encompassing digital-intelligent foundational platform, and by continuously advancing both the intelligent value of RAN and the integrated digital-intelligent upgrade for ToB industries, ZTE is driving technological innovation and practice to lead the industry's development.

4.1 ToC Network Applications

Against the backdrop of slowing mobile internet traffic growth, declining ARPU, and increasingly homogeneous service packages, operators face the challenge of enhancing their competitiveness. ZTE's AIREngine—by introducing AI into 5G-A BBU—offers a breakthrough approach that not only optimizes the user experience but also promotes green, low-carbon development and improved operational efficiency. Through intelligent scheduling and energy consumption optimization, AI effectively reduces the energy use of wireless access network equipment, helping to achieve carbon neutrality. In complex operational environments, AI enhances field testing efficiency in scenarios such as high-speed rail and maritime areas, thereby reducing time and labor costs. At the same time, AI empowers network fault tracing and security analysis, making data processing more efficient and response times faster, which in turn ensures network stability and reliability. The deep application of AI technology injects innovative momentum into the communications industry, opening up new pathways for green, efficient, and differentiated development.

4.1.1 Superior User Experience - Enhancing Connection Value

Delivering an exceptional user experience is key to boosting customer stickiness and market competitiveness. In today's fiercely competitive communications market, users are no longer satisfied with merely basic services; they increasingly demand additional value and high-quality experiences. By leveraging AI technology, operators can intelligently identify user scenarios and dynamically optimize strategies to ensure that popular entitlements (such as video or gaming platform memberships) are delivered smoothly, thereby enhancing user satisfaction and loyalty. Simultaneously, operators must transition from competing solely on low prices to adopting value-based pricing. By empowering exclusive benefits and intelligent services through AI, they can clearly align costs with value—stabilizing the customer base and increasing market share, which is set to become the core direction of future competition in communication services.

Business Differentiation: The network can provide differentiated guarantees based on the unique characteristics and requirements of different services to achieve a service quality that matches each business. For video services, for example, bandwidth can be prioritized while protocols and scheduling strategies are optimized to reduce buffering and enhance the viewing experience; for gaming—which is sensitive to latency—the focus is on lowering delay and strengthening connection stability; for voice services, dynamic adjustment of transmission paths is necessary to ensure smooth interaction; and for live streaming, which demands high uplink rates, substantial uplink bandwidth must be ensured. By intelligently identifying typical services and accurately evaluating key quality indicators (KQIs), the network can balance multiple service experiences and significantly enhance overall service quality. In parallel, the network can evaluate service experience from multiple dimensions, thereby helping to transition network metrics from traditional key performance indicators (KPIs) to business key quality indicators (KQIs).

User Personalization: Operators have defined various user categories based on different user demands for network differentiation. This requires that the network be capable of providing customized services according to each specific user classification. Moreover, given the random distribution of various user types within the network, when resources are limited, intelligent allocation of network resources is needed to maximize overall assurance effectiveness.

Experience Continuity: By embracing a "user-centric" philosophy, deep, intelligent collaboration among cells centered on the user can be achieved to deliver a consistent experience. Intelligent collaborative solutions dynamically cluster users based on their mobility trajectories, thereby gaining insights into user patterns, service coverage, and cell performance—and even predicting user movement. Based on these mobility trajectories and service demands, dynamic clustering is performed, with coordinated uplink/downlink scheduling within clusters and multi-point collaboration among clusters to reduce or avoid interference.

Service Certainty: For specific government/enterprise users, emerging services, or particular industry application scenarios, the network's connection quality (QoS) demands are both higher and more precise. Clear assurance standards are set in terms of data rate, latency, and reliability—standards that are particularly critical in certain industry scenarios. The network can learn service characteristics and activate assurance strategies that match these features. By enabling intelligent agents to analyze the real-time transmission properties, the network can formulate and optimize decision-control strategies. Through forecasting service volume changes, optimizing resource scheduling, and balancing loads, the network coordinates to meet the needs of priority users, ensuring the delivery of high-quality services.

4.1.2 Superior Energy Efficiency - Reducing Carbon Emissions

One of the main challenges in reducing energy consumption at base stations is how to minimize power use while maintaining network performance. Traditional methods often rely on fixed energy-saving strategies, which are ill-suited to adapt to complex and dynamic network environments. AI technology, through deep learning and big data analytics, enables scene recognition and the dynamic adjustment of energy-saving strategies to effectively tackle these challenges. Specifically, AI technology can achieve the following:

Scene Recognition and Load Prediction: Different application scenarios exhibit distinct traffic distribution patterns in cells. For example, traffic trends in universities and office buildings often show higher loads during weekdays and lower loads on weekends, whereas malls and parks experience peaks during weekends and evenings. By applying the K-Means algorithm to cluster time-series waveforms, different load models can be accurately identified. This not only aids in predicting future load changes but also provides a basis for developing personalized energy-saving strategies.

Dynamic Energy-Saving Strategy Generation: Based on the predicted trends of cell traffic, AI technology can generate optimal energy-saving strategies. For instance, during low-load periods, certain base stations can be automatically shut down or have their transmission power reduced, while during high-load periods, normal operation is maintained. Moreover, AI can differentiate between capacity and coverage layers, selecting co-covered cells as compensatory cells to share the load. This dynamic adjustment not only reduces unnecessary energy consumption but also ensures that network performance remains uncompromised.

Real-Time Evaluation and Self-Optimization: By monitoring network KPIs in real time, AI technology can promptly evaluate the effectiveness of energy-saving measures and make second-level (sub-second) adjustments based on actual feedback. For example, if a particular energy-saving strategy degrades network performance, the AI system will automatically adjust the strategy to achieve the optimal balance between energy consumption and performance. This closed-loop optimization mechanism makes energy-saving strategies both flexible and efficient.

Multi-Dimensional Energy Efficiency Evaluation: Traditional energy efficiency evaluations typically focus on a single traffic metric; however, with the continual emergence of new services and scenarios, a single metric can no longer comprehensively reflect network energy consumption and efficiency. AI technology can construct multi-dimensional energy efficiency models that continuously expand the evaluation system from scenario-based, normalized, and multi-dimensional perspectives. This not only helps in obtaining a more comprehensive picture of network energy efficiency but also guides the selection of subsequent energy-saving strategies.

By introducing AI technology into 5G-A, operators can achieve extreme energy-saving goals while maintaining network performance—with overall network energy consumption expected to drop by 15% to 25% throughout the day. This not only helps reduce operating costs but also contributes significantly toward achieving green and low-carbon objectives.

4.1.3 Superior O&M Efficiency - Optimizing Network Quality and Costs

During network maintenance and optimization, challenges such as insufficient fault analysis data, imprecise fault localization, and high technical demands for on-site maintenance often result in low accuracy and efficiency with traditional methods. By leveraging deep learning and big data analytics, AI technology can perform scene recognition and dynamically adjust strategies to effectively address these challenges. Key solutions include:

Fault Assistant: To minimize the learning curve and technical requirements for infield maintenance personnel while boosting fault troubleshooting efficiency, the Fault Assistant uses natural language as its interaction method and combines both large and small models to perform fault-handling tasks. It offers features such as knowledge Q&A, dynamic data querying, guided troubleshooting, and a fault dashboard.

- It supports the identification of fault phenomena and provides handling recommendations—from standard protocols, operator specifications, to equipment manufacturer product models and features—as well as answers based on relevant case knowledge.

- It allows users to diagnose online equipment modules (hardware, link, clock, etc.) via natural language queries, thus retrieving the latest device status and operational data to aid maintenance personnel in comprehensive judgment and fault localization.

- It provides system-level guidance for troubleshooting processes; maintenance personnel can engage in multi-turn natural language dialogues with the Fault Assistant to gradually uncover the root cause of faults.

Network Optimization Expert: This large-scale model leverages multi-objective joint optimization and digital twin technology to automatically correlate alarm risks, operation logs, and other information. It performs comprehensive analyses on cells with poor performance at both the cell and grid levels, intelligently outputting precise root causes and parameter self-optimization recommendations for issues related to coverage, interference, and capacity. Furthermore, through twin simulations, it predicts future network demands and provides accurate proposals for network expansion and enhancement.

Network Monitoring Expert: Offering interactive, multi-dimensional network monitoring and analysis capabilities, the Network Monitoring Expert converts natural language queries into search statements and, based on service types, gathers insights from multiple intelligent agent experts. It then summarizes network observations and provides professional recommendations for planning, optimization, and maintenance. Relying on the outcomes of multi-dimensional network insight tasks—and by combining observed phenomena with a knowledge base—it can suggest solutions. For example, by merging business perception with radio access quality data in a single domain, it can accurately identify indicators such as video latency and, using AI perception models, assess service quality; it also supports playback of user trajectories over time, visualized with GIS grid mapping, correlating with primary serving cells and distinguishing between indoor and outdoor environments to provide visual support for maintenance.

Knowledge Assistant: By integrating a large model with an external knowledge base (including maintenance help documents and network element service documentation), the Knowledge Assistant provides users with a rich, up-to-date, and reliable Q&A service. Users can query alarms, performance data, and network element information from the maintenance system using natural language, and the system summarizes the retrieved data to present the results in a more user-friendly format.

Assurance Expert: Tailored for wireless network assurance scenarios, this large-model operation and maintenance solution is applied in major events (e.g., concerts), emergency responses, and service traffic surges. Utilizing business recognition technology, it can deeply analyze network traffic to accurately identify service types—such as short videos, cloud gaming, long videos, web browsing, QR code scanning, instant messaging, and mobile gaming—and trigger the large model to initiate user experience assurance. In conjunction with smaller models, it intelligently generates assurance plans, reducing manpower inputs by 30%.

Through the integration of 5G and AI, practical tests have shown that the efficiency of analyzing poor-quality scenarios can be improved up to nine times. This solution also supports virtual drive tests and a new model for opening up data and capabilities.

4.2 ToB Industry Applications

As industrial digital transformation deepens, smart factories face multiple challenges including large-scale data processing, real-time monitoring and control, AI algorithm support, and edge as well as secure computing. ZTE's NodeEngine computing base station and UniEngine integrated computing-network device, through the deep collaboration of communications, computing power, and AI, bring intelligent upgrades to the ToB industry—comprehensively enhancing production efficiency and competitiveness.

4.2.1 Intelligent Factory Upgrades

In the steel industry, unmanned overhead crane systems achieve automated material handling through wireless communications and edge computing, addressing issues such as aging equipment, labor shortages, and safety risks. Relying on real-time data collection, machine vision technology, and AI algorithms, the system performs task scheduling, path optimization, and fault diagnosis. The unmanned overhead crane system leverages AI and machine learning to integrate data from various process stages and devices—along with parameters such as material weight, shape, and location—to calculate the appropriate hook height, speed, and angle, thereby preventing material sway or collisions. Based on production plans, material inventory, and equipment status, it dynamically adjusts the priority and sequence of transport tasks to avoid both material backlog and shortages.

In the 3C manufacturing sector, which heavily relies on automation and precise quality management, machine vision inspection systems employ image processing and AI algorithms to efficiently detect issues such as surface defects and assembly errors. These systems must process large volumes of image data with high efficiency and low latency, meeting real-time requirements through parallel computing and optimization algorithms. The inspection data is used not only for immediate feedback but also as a basis for production optimization and quality traceability—helping enterprises achieve continuous, data-driven improvement.

5G communications provide smart factories with high bandwidth, low latency, and highly reliable data transmission, while edge computing reduces latency by processing tasks locally to ensure real-time responsiveness. AI technology further enhances system intelligence—from data insights to equipment control—injecting innovative momentum into industrial scenarios, whether it is unmanned overhead cranes in the steel industry or visual inspection systems in 3C manufacturing.

4.2.2 Optimized Multi-Stream Video Transmission

In scenarios featuring densely deployed IP cameras, concurrent multi-stream video transmission may lead to the simultaneous transmission of multiple I-frames (referred to as "I-frame collisions"), which causes a sudden surge in bandwidth demand far exceeding the typical bit rate of the video streams. This phenomenon can trigger congestion in the transmission links, resulting in dramatic increases in latency and video stuttering that severely impact transmission quality and service stability.

To ensure stable and smooth video transmission, an I-frame detection AI algorithm service can be deployed on the integrated computing-network device. By employing a pre-trained model to identify I-frame characteristics in the video streams, the system can detect in real time the probability of I-frame collisions for each camera and adjust those with high collision probabilities accordingly. Adjustment commands are then sent to the cameras via the camera management protocol, thereby effectively avoiding I-frame collisions, reducing network bandwidth demands, and ensuring video transmission quality and link stability.

4.3 Comprehensive Intelligent Digitalized Site

With the continuous deep integration of AI and RAN, sites, being the critical hardware foundation of communication networks, are also undergoing significant transformation toward digital and intelligent evolution, driving the network toward a higher level of self-intelligence.

A high-level self-intelligent network needs to adjust site resources in real time according to user and service demands in order to enhance service perception and operational efficiency while reducing energy consumption. The hardware foundation for this is that sites must be capable of real-time sensing and multi-dimensional adjustment. Looking at the entire site, the core hardware comprises the BBU, RRU, and antenna, along with supporting components such as power supplies, transmission systems, and cabinets. While the BBU and RRU have already achieved a digital-intelligent foundation, components such as antennas, power supplies, transmission systems, and cabinets still need to evolve toward digital and intelligent integration to establish a comprehensive intelligent foundation for the network.

4.3.1 Antenna Digitalization

Traditional antennas, which are widely deployed as passive devices, are "dumb" devices with very limited sensing and adjustment capabilities—typically only allowing for vertical tilt adjustment. Parameter acquisition and antenna adjustments are mainly performed manually, which is inefficient and unable to adapt to changing service environments. As the sole link between base stations and end users, the digital evolution of antennas is an inevitable trend. Key features include:

Real-Time Operational Parameter Acquisition: Accurately and continuously obtaining site parameters such as azimuth, mechanical tilt, and geographic coordinates.

Multi-Dimensional Beam Adjustment: Enabling adjustments in horizontal angle, vertical angle, horizontal beamwidth, and vertical beamwidth. The greater the adjustable range and the more frequency bands that can be independently adjusted, the more effective the network optimization.

Intelligent Network Perception: Automatically acquiring the mapping relationships among the antenna array, RRU channels, and cells.

4.3.2 Transmission Digitalization

With the increasing deployment of CRAN (Centralized RAN), the distance between the BBU and RRU has grown, rendering the fronthaul optical paths more complex. The fronthaul, being a passive link, is monitored by traditional optical modules that can only detect changes in optical power—not the overall health of the link. When a fronthaul fault occurs, it has traditionally required manual, point-by-point troubleshooting on-site, which is both time-consuming and labor-intensive. Digitalizing the fronthaul link is an inevitable path to resolving these issues. By deploying intelligent optical modules in conjunction with smart software algorithms, automatic fault diagnosis and intelligent root-cause analysis for the fronthaul can be achieved, allowing issues to be precisely resolved in a single site visit—greatly improving maintenance efficiency and enhancing network availability.

4.3.3 Power Supply Digitalization

By adopting gSDU, power supplies and main equipment are digitally and intelligently integrated to achieve green, low-carbon site operations.

Power distribution port names can be defined by software, automatically recognizing the relationships between load devices and cells. Based on the specifications and models of load devices, circuit breaker overcurrent protection thresholds are automatically adjusted so that one circuit breaker can adapt to various types of base station equipment—ensuring uniform power distribution. Each port can be remotely switched on or off.

Intelligent Business Sensing: Achieving intelligent start/stop functions for AAU/RRU, thereby enabling dual benefits in energy saving and system sensing; simultaneously, the system can intelligently predict site equipment power consumption and dynamically adjust the number of power rectifiers in sleep mode to ensure maximum conversion efficiency.

Hierarchical Backup Power: The smallest backup unit is at the equipment level. As more equipment is added, battery capacity does not necessarily need to increase. The system manages communications with third-party power supplies, obtaining information on voltage, current, battery capacity, and remaining battery life, intelligently predicting backup duration, and dynamically adjusting cutoff voltage for smart backup power.

Intelligent Green Power: Based on weather forecasts to predict solar generation, the system can intelligently and dynamically adjust the output power of base station equipment and the charging current for lithium batteries, thereby extending site service time.

Intelligent Peak Shaving: This feature supports charging lithium batteries during periods of low electricity prices and using them during peak price periods, thereby reducing overall electricity costs. By forecasting site power consumption from historical data, the system calculates the battery capacity available for discharge during peak pricing periods and intelligently adjusts the lithium battery discharge duration to ensure the safe operation of the base station.

By integrating wind-liquid hybrid near-end cooling technology, the BBU and cabinet temperature control systems work in tandem. They dynamically track and analyze changes in cabinet energy consumption in response to climate variations and perform intelligent energy-saving optimizations. Additionally, the cabinet becomes visible, perceptible, and controllable via wireless network management.

05 AIR RAN: Growing a Robust Ecosystem through Collaboration

As analyzed earlier, the core objective of the AIR RAN architectural evolution is to support ubiquitous intelligence in the network, promote intelligent interconnection among network agents, and ultimately transform the network into an intelligent hub that connects the physical and digital worlds.

At the product architecture level, the industry currently primarily adopts two forms. One is the heterogeneous XPU architecture—evolved from RAN-dedicated accelerators—that integrates general-purpose and specialized computing, while the other is based on general-purpose CPU+GPU architectures. Each form has its own characteristics. The heterogeneous architecture employs customized designs that, depending on the specific RAN module, flexibly selects the most suitable computing architecture—especially in modules that are less amenable to AI while accelerating those that are. This approach can deliver higher computational efficiency in resource scheduling and performance optimization and offers clear advantages in cost and power consumption. In contrast, the advantage of the general-purpose architecture is its ability to support rapid deployment of third-party AI applications. However, when implementing an intelligent RAN, operators must switch between different hardware units to coordinate classic RAN modules with certain AI modules, which may result in significant resource waste. Overall, although the general-purpose architecture has some exploratory value at this stage, the heterogeneous architecture—with its advantages in overall solution maturity, compatibility, customization, and power efficiency—appears to be the more promising long-term direction for AIR RAN evolution.

At the technical architecture level, regardless of the hardware product form chosen, the following dimensions should be considered:

AI for RAN: Integrate AI into the RAN protocol stack to inject precise intelligence into layers L1, L2, and L3. AI applications to enhance wireless network performance encompass optimization strategies at the lower, middle, and upper layers. At the lower layer, high performance is driven by combining the strengths of traditional algorithms with AI algorithms; at the middle layer, multi-task driving enables multi-task collaboration through localized AI task optimization; and at the upper layer, strong intelligence is driven by large models and natural language understanding to improve network control intelligence.

RAN for AI: Through deep fusion of communication and AI technologies, the RAN not only fulfills traditional connectivity functions but also extends its capabilities via perception and computation. This empowers applications such as embodied intelligence, vehicle-infrastructure collaboration, and low-altitude unmanned aerial vehicles. Base stations, by leveraging integrated communication and sensing technologies, can perceive their environment and dynamic targets. Combined with AI, they enable real-time optimization, enhance resource utilization and service quality, and support the intelligent upgrade of scenarios across the board.

Collaborative Intelligence: The evolution of network intelligence is moving from isolated agents to collaborative intelligence, where cooperation among agents decomposes complex problems and optimizes global decisions. This collaborative approach significantly enhances system adaptability, robustness, and innovative capability. AI for RAN provides performance support, while RAN for AI contributes data and interaction. The two mutually reinforce each other, driving the overall intelligence upgrade.

5.1 AI for RAN: Hierarchy Intelligence for Network Performance

Overall, the designs of the various RAN layers differ in their theoretical foundations, optimization objectives, and computation methods. Therefore, when injecting intelligence via AI, it is necessary to tailor optimization to the distinct characteristics of the lower, middle, and upper layers.

5.1.1 High-Performance Lower Layer

For local modules that cannot be accurately expressed with a unified mathematical model or do not conform to a typical mathematical distribution—for example, time-frequency domain estimation of non-pilot channel segments in multiple scenarios, channel compression, and fitting for non-idealities of different hardware—the traditional algorithms struggle to obtain a theoretical optimum or incur excessively high complexity in doing so. Based on the characteristics exhibited in the time and frequency domains, AI's "broad experience" can be leveraged to improve accuracy. However, given the computational complexity of AI and the real-time requirements—as well as constraints on model scale imposed by hardware and power consumption—the question remains whether AI's generalization can meet performance demands, which currently presents a significant challenge.

The following table summarizes the two generalization approaches:

	Traditional Algorithm	Algorithm
Generalization Approach	Based on classic mathematical principles such as digital signal processing, matrix theory, and probability theory, and designed with assumptions of typical mathematical distributions (i.e., performing a priori generalization deductions based on mathematical principles and distribution characteristics). In different times, locations, and site types, as long as the mathematical laws or distribution characteristics remain unchanged, predictions in new scenarios remain valid.	Uses data correlations as its premise, extracting and storing the correlation patterns as parameters from various types of data. In new scenarios, AI computes the correlations from new data and matches them to similar existing data to quantify a priori prediction probabilities.
Advantages	Mathematical models and distribution assumptions ensure consistency, with good interpretability and stability. Direct one-step mapping to the target space with little or no data dependency, low cost, and low inference complexity.	Learns the correlation relationships across various scenarios from training data, allowing for a broader and more fine-grained match to scenario-specific distributions. Provides a more comprehensive summary of existing scenario patterns.
Disadvantages	For scenarios where it is difficult to unify mathematical expressions or precisely assume uniform mathematical distributions (e.g., variations in the time-frequency domain of channels under different environments and interference, or non-idealities in different hardware that can only be modeled in a coarse manner), the granularity is low and accuracy suffers.	Lacks an invariant assumption, making it difficult to quantitatively confirm whether the various data combinations in a new scenario are already represented in the existing data—resulting in lower certainty of generalization. Requires small step sizes and long-term iterative search with large amounts of data and high-dimensional computations, leading to high costs and inference complexity.

For example, wireless signals exhibit different characteristics in the time and frequency domains (e.g., multipath effects in the time domain, and oxygen decay phenomena in the frequency domain). Classic lower-layer frameworks and algorithm designs are built on these observations, using Fourier transforms to flexibly switch between time and frequency domains and locate the optimal domain for signal processing. In contrast, if one were to directly perform end-to-end AI inference by feeding raw air interface time-domain data as input, the clearly visible frequency domain features would be diffused and hidden within the time domain, and the cost of feature extraction would be significant. Moreover, while traditional algorithms can directly map to the target space through formula derivation, AI algorithms must extensively accumulate rules from data. This results in AI models having an inference complexity that is several times, or even 2-3 orders of magnitude, higher than traditional algorithms—leading to a dramatic increase in hardware scale and power consumption under time constraints.

Thus, for the lower layer, the more promising strategy is to fuse the strengths of traditional algorithms with those of AI algorithms. By leveraging the theoretical generalization advantages of traditional methods alongside AI's scene-learning capabilities—and balancing high performance with complexity constraints—a hybrid algorithm can be constructed to achieve optimal performance across different scenarios.

5.1.2 Multi-Task Coordinated Upper Layer

Classic middle-layer algorithms primarily manage system control based on different channel, service, and scenario characteristics, resulting in control decisions and actions. If AI is to be employed for this purpose, two major aspects must be considered:

Learning Space: The learning space for middle-layer control is enormous, involving a combination of numerous input features and output action dimensions that are practically impossible to exhaustively traverse in engineering practice.

Input Feature Space: Even when raw data streams are compressed into relevant features based on empirical knowledge, the input feature space remains vast. For example, for channel estimation, each terminal may involve factors such as received signal strength, interference, temporal variations, multipath effects, and terminal-specific implementations. The combinations across all terminals grow geometrically.

Decision/Action Space: Even when the action space is narrowed by defining frame structures and standard processes, each minimal unit of time, frequency, or spatial resource can be allocated to different terminals for various modulation schemes or channel measurements. This can result in an action space numbering in the millions.

Evaluation System: On one hand, due to the high sensitivity of electromagnetic wave phases, there is currently no twin environment that aligns input feature dimensions with output action dimensions to real-world conditions. On the other hand, in an actual transmission environment, feedback from exploring unknown action spaces may introduce issues into the live network.

In summary, replacing the entire middle layer or large modules with AI faces the challenges of an excessively large learning space and the difficulty of establishing a reliable evaluation system. A more viable approach is to reduce the task scope of individual modules—thereby shrinking the learning space—by splitting the overall middle-layer system into a multi-task collaborative architecture based on expert experience. This allows for localized AI task optimization while enabling traditional and AI-driven tasks to collaborate within the overall middle-layer framework.

5.1.3 Strongly Intelligent Top Layer

Based on statistical prediction and signaling language understanding—and combined with human intent and accumulated operational experience—the upper layer can achieve more intelligent network control. On one hand, the data at the upper layer, built on statistical significance, often exhibits strong regularity; on the other hand, since upper-layer deterministic processes involve more human interaction, much information has already been compressed into signaling or bitstream spaces. This results in a smaller learning space, and mature expert knowledge is often already documented in text form. These factors make the upper layer particularly suitable for AI integration. It is worth noting that the RAN upper-layer applications, which rely primarily on model inference, have additional room for quantization and model compression. Therefore, based on application performance, more cost-effective computing methods should be chosen to support strong intelligence.

5.2 RAN for AI: Connectivity, Perception, and Computing

With the deep integration of AI and communication technologies, networks are advancing toward full-scale intelligence and collaboration. Base stations are no longer merely hubs connecting terminals to the network; they also play a central role in perception and computation, providing efficient support for AI-driven services. With the introduction of wireless edge computing and multi-modal perception technologies, base stations have transitioned from simple data transmission nodes to platforms capable of intelligent decision-making. This significantly enhances network performance, privacy protection, and resource utilization efficiency. On this basis, base stations are empowering frontier scenarios such as humanoid robotics, vehicle-infrastructure-cloud integration, non-terrestrial networks (NTN), the low-altitude economy, and maritime communications—accelerating the innovative applications of wireless networks.

Enhanced Connectivity Services: In scenarios such as embodied intelligence and interactive AI, AI computing capabilities are further pushed toward the terminal. The collaborative nature of distributed AI computation drives increased connectivity transmission bandwidth. Moreover, embodied intelligence scenarios require frequent interactions, posing new requirements on network latency and reliability.

All-Scenario Perception Empowerment: By integrating base station sensing (through joint communication and sensing techniques) with AI, precise environmental and target perception and dynamic optimization can be achieved in NTN, maritime, and low-altitude scenarios. This significantly enhances communication coverage, resource utilization, and user experience, thereby promoting a comprehensive upgrade of intelligent network services.

Expanded Computing Services: Building upon connectivity and perception, edge computing can empower industrial robots by acting as their "brain." In autonomous driving scenarios, edge-deployed applications can centrally manage data from multiple vehicles and roadside sensors, avoiding redundant AI inference at each endpoint, and protecting individual driving privacy. In this way, robust computing and reliable connectivity services are provided to the endpoints.

5.2.1 Connectivity for Embodied Intelligence and Cloud-Edge Collaboration

To promote the development of robotics, part of the computational load must be offloaded to the edge or cloud to extend operational endurance, reduce computing costs, and enhance multi-robot collaboration. By deploying multi-modal large models and intelligent agent services on theAIR RAN, robots not only benefit from wireless connectivity but are also endowed with superior natural language understanding and communication, integrated perception, decision-making analysis, and task planning capabilities. This supports their transition from localized intelligence to embodied and collective intelligence. Embodied intelligence, which combines a physical body with intelligent algorithms, emphasizes enhancing perception and decision-making through interaction with the environment. Deploying the "brain" for embodied intelligence at the RAN edge, in collaboration with a "little brain" on the robot's end, can achieve low-latency decision-making while saving power and protecting data privacy.

For example, consider a humanoid robot with 28 degrees of freedom—14 rotational joints, 8 linear joints, and 6 in the head. Its dexterous hands possess 22 degrees of freedom, and the entire robot is equipped with 28 actuators (including high-torque motors and brushless gear motors), along with a depth vision camera and LiDAR. The total data volume transmitted to the edge computing platform is roughly >30Mbps upstream, with mainly control commands and voice data downstream (resulting in relatively low traffic). This data is transmitted over the RAN, which must guarantee low delay and minimal jitter. For instance, to meet the end-to-end voice interaction requirement of under 2 seconds, the network delay should be kept below 20ms.

Purpose	Content/Command	Uplink Speed (Mbps)	Downlink Speed (Mbps)
Control	Body Motor	1M(State)	0.1M(Cmd)
	Hand	1M(State)	0.05M(Cmd)
	Odometer Information	0.5M	--
Voice	Voice Report	1M	1M
Video	Video Report	10M	--
LiDAR	LiDAR Report	20M	--
Total		33.5M	1.15M

5.2.2 All-Scenario Perception in Low-Altitude, Maritime, and NTN

With rapid advances in communication technology, base station functionality has expanded from traditional communication services to the new domain of integrated communication and sensing. Base station antennas can sense the external environment by analyzing reflected signals—thereby monitoring the positions, trajectories, and environmental characteristics of objects in real time. When combined with AI, these sensing capabilities demonstrate tremendous potential in NTN, maritime communications, and low-altitude UAV scenarios.

In non-terrestrial networks (NTN), base stations can detect the movement and environmental characteristics of ground user devices through signal reflections, using AI to dynamically optimize beam directions and spectrum allocation. This approach not only enhances network coverage but also supports dynamic navigation and emergency communications. In maritime environments, base stations monitor ocean conditions, ship positions, and weather patterns. AI predicts the stability of communication links and load demands, allowing for dynamic resource allocation that effectively supports long-distance maritime communications, route planning, and collision avoidance. In low-altitude UAV operations, drones require highly accurate, real-time updates on geographical information and obstacles—needs that static maps and onboard radar or cameras often cannot fully meet. Base station antennas can sense a UAV's altitude, speed, and trajectory, while AI processes this data in real time to optimize spectrum usage and flight path planning. This ensures efficient coordination and reliable communication among drone fleets.

For instance, in a low-altitude scenario, integrated communication and sensing technologies combined with AI play a crucial role across multiple domains. In logistics, they help drones achieve precise positioning, autonomous obstacle avoidance, and safe, rapid delivery—thus reducing costs and increasing efficiency. In smart transportation, real-time monitoring of aerial traffic helps avoid congestion and accidents, providing precise navigation for aerial vehicles and optimizing flight paths for improved convenience and safety. In low-altitude security, the system can identify illegal drones, monitor and counteract them in real time, and also provide predetermined route monitoring and protection for legitimate networked drones.

Through AI-driven real-time decision-making, the base station's integrated sensing capabilities transform the RAN from a traditional communication platform into an intelligent sensing platform. This enables dynamic optimization and efficient resource utilization in complex environments, comprehensively empowering the intelligent upgrade of all scenarios.

5.2.3 Integrated Connectivity, Perception, and Computing for Autonomous Driving

In a 5G-based vehicle-infrastructure-cloud integrated solution, roadside sensing is a key component. To fully leverage the advantages of 5G, it is essential to deeply integrate 5G network performance with roadside sensing AI analysis and edge computing technologies.

Multi-Modal Sensing Fusion: Real-time responsiveness is critical for roadside intelligent systems. Roadside units must rapidly integrate data from multiple sensors (such as millimeter-wave radars and cameras). With edge-deployed real-time AI performing data fusion, precise information—such as a vehicle's 3D position, speed, and heading—can be provided to support quick decision-making in complex intersections. Compared with traditional cloud-based processing, which may suffer from network congestion and increased delay due to long transmission distances, deploying processing at the 5G base station can drastically shorten the transmission path, significantly reduce latency, and enable faster utilization of fused data for vehicle control. For example, while a vehicle is traveling at high speed, it can adjust its state in a timely manner based on roadside information to ensure safe and efficient driving.

Dynamic Target Tracking: The roadside sensing system must offer real-time, low-latency responses to rapidly changing conditions among vehicles and pedestrians. Edge real-time AI, using techniques such as RNNs and LSTMs, can quickly detect and identify targets, predict their trajectories, and update their statuses—helping vehicles to plan paths in advance (for instance, braking or evading when a pedestrian crosses the road). Unlike cloud-based processing, which may experience delays due to transmission and processing, edge deployment allows local data processing with extremely low latency, ensuring that vehicles can monitor target dynamics in complex traffic scenarios in real time and effectively safeguard driving safety.

Currently, a joint pilot project between ZTE and Jinan Mobile has been deployed for a vehicle-infrastructure-cloud integrated scenario covering a 5-km route and 14 intersections (8 key intersections). The "5G + Internet of Vehicles System," based on 5G, roadside sensing AI analysis, and edge computing technologies, is fully deployed. Vehicles can receive real-time information about traffic lights, beyond-visual-range accident warnings, and pedestrian intrusion alerts. Field tests have shown that the system reduces road accident rates by approximately 37%, improves traffic situational awareness by about 20%, and achieves beyond-visual-range accident warning detection distances of over 3km. The cloud platform also enables real-time traffic supervision, providing dynamic foundational data on vehicle operations, infrastructure, and traffic management for smart vehicles, users, and regulatory bodies.

5.3 Collaborative and Ubiquitous Intelligent Agent Services

Traditional communication AI small models primarily address individual communication tasks. Their construction and training rely on human experts, rendering their behavior relatively passive, static, and localized. In contrast, future wireless systems will face multi-target, complex tasks spanning communication, sensing, computing, intelligence, and information.

An intelligent agent is a high-level intelligent entity that can actively perceive, understand, and plan autonomously. Based on a specific large model, it completes tasks and self-improves. The "Network Intelligent Agent" (NW Embedded AI Agent, NEA) is an intelligent agent deployed on the network side—either integrated within RAN nodes or as an independent node. It not only serves the node in which it resides but also collaborates with other NEAs to accomplish complex tasks.

NEAs address broadly generalized, multi-domain problems by proactively collecting data, deeply perceiving the environment, and making decisions without relying on external expert guidance. Their behavior is proactive, dynamic, and globally oriented. By combining stored knowledge, experience, and tools, NEAs offer more explainable, robust, and trustworthy decision-making compared to traditional communication AI small models.

In the pursuit of higher-level autonomous network operation and maintenance in 6G wireless systems, the evolution is shifting from "add-on autonomy" to "inherent intelligence." With the assistance of NEAs, future network operation and maintenance will merge. Wireless systems will be better able to adapt precisely to their environments and requirements, evolving toward optimal performance. In network planning, NEAs are dedicated to environmental situation self-discovery, flexible on-demand networking, and trusted ubiquitous access; in network maintenance, they focus on intelligent strategy generation and validation, task orchestration and collaboration, session management, and autonomous target optimization. With NEA's assistance, the future integration of network operation and maintenance will enable wireless systems to precisely adapt to environmental demands and evolve toward optimization.

Looking forward, NEAs will evolve from single-domain, individual intelligent agents to cross-domain collaborative intelligence. Through collaborative decomposition of complex problems and optimized decision-making, the limitations of individual agents can be overcome. Different agents will complement each other, enhancing the system's multi-objective decision-making capability and robustness, and potentially giving rise to collective intelligence that generates innovative solutions for complex network environments.

06 Future Horizons: Shaping the Long-Term Value of Wireless Network

In the coming 6G era, mobile networks will not only serve as the carrier of ubiquitous connectivity but will also evolve into an intelligent platform that enables seamless interactions among people, machines, objects, and intelligence. The core technological vision for 6G is to achieve "integrated communication, sensing, computing, and intelligence," that is, to deeply integrate and apply communication, perception, computing, and AI intelligence at the system architecture level, thereby driving the digital and sustainable development of society as a whole. As the deep evolution of wireless access networks into intelligent networks, AIR RAN is a crucial component in realizing this vision.

6.1 AIR RAN Vision

Aiming to realize the 6G aspirations of "ubiquitous intelligent connectivity, digital twin, and green efficiency," 6G RAN must support ultra-high bandwidth, ultra-low latency, massive connectivity, and complex dynamic environments, making RAN intelligence an indispensable technological cornerstone. AIR RAN endows wireless access networks with multidimensional capabilities such as deep intelligent perception, autonomous decision-making, and dynamic optimization, including:

Intelligent Communications: AIR RAN leverages AI-driven RAN optimization and wireless resource management to achieve more efficient communication capabilities.

Real-Time Sensing: AIR RAN will possess environmental intelligent sensing capabilities, providing the data foundation required for the realization of digital twins.

Intrinsic Computing: In AIR RAN, computing is an inherent function. By leveraging edge computing and distributed AI, computing delays are reduced and efficiency is enhanced.

Intelligent Evolution: A core new feature of AIR RAN is "self-learning, self-optimization, and self-management," which will transform RAN from a passive responder into one that proactively perceives, predicts, and evolves.

Extended Computing Capability: Through deep edge computing and distributed computing architectures, AIR RAN provides computational support for AI inference and training at the access network level.

Data Processing Hub: AIR RANwill serve as a data processing node for AI applications, offering low-latency data processing services for applications with extremely high real-time requirements (e.g., autonomous driving, industrial control, remote healthcare).

Intelligent Coordination Center: By means of multidimensional resource scheduling, task distribution, and model collaboration, AIR RAN will provide an efficient operating environment for distributed/edge AI applications.

Integrated Intelligent Ecosystem: AIR RAN will evolve into an intelligent hub that spans multiple industries and scenarios, empowering digital transformation across vertical sectors.

6.2 AIR RAN Roadmap

The development and large-scale application of AIR RAN will not occur overnight. It must progress through a "short, mid, and long term" development journey, specifically:

Short-Term (2025-2027): Broadly introduce AI to achieve deep optimization of RAN, addressing traditional network pain points such as low spectrum utilization efficiency, high energy consumption, and underutilized computing resources. This period will focus on building a foundational RAN AI algorithm framework that imparts preliminary self-learning and optimization capabilities to RAN. AIR RAN will not only satisfy users' demand for high-quality connectivity but also enhance the flexibility and adaptability of RAN. Meanwhile, validation tests for applications such as vehicle-infrastructure-cloud integration, embodied intelligence, and edge computing will also be completed during this period.

Mid-Term (2028-2030): Achieve intelligent coordination and dynamic management within RAN. At this stage, AIR RAN will become the core hub for communication, computing, model, and data interactions, supporting multi-element network collaboration and serving as an intelligent platform with deep sensing and real-time optimization capabilities. It will also function as an AI edge server, assisting both the endpoint and cloud sides in jointly constructing an AI ecosystem.

Long-Term (Post-2030): Fully transition toward a network ecosystem characterized by "integrated communication, sensing, computing, and intelligence," realizing intelligent connectivity across all scenarios and regions. RAN will be capable of autonomous sensing and real-time decision-making to support complex multidimensional interactions. AIR RAN will truly become the core infrastructure for building a digital twin world, driving society toward an intelligent future of full connectivity.

6.3 AIR RAN Ecosystem

The development of AIR RAN relies not only on technological breakthroughs but also on close collaboration across the entire industry. The evolution of the AIR RAN ecosystem is a comprehensive process that spans technology, experimentation, industry application, vision, practice, and returns.

The vision for AIR RAN is highly aligned with the 6G system's goal of "integrated communication, sensing, computing, and intelligence," and is key to the 6G network and ecosystem. Through AIR RAN, future 6G networks will transition from being "connection-driven" to "intelligence-driven." In light of the opportunities and challenges that lie ahead for AIR RAN, all industry stakeholders are called upon to actively participate and jointly accelerate the industrialization process by:

Strengthening R&D: Enhance innovation in RAN-side AI algorithms, edge computing, and network architectures to drive AIR RAN technologies from the laboratory to commercial deployment, supporting intelligent base stations capable of both AI training and inference.

Enhancing Cross-Industry Collaboration: Communication equipment vendors, operators, cloud service providers, and AI companies need to form deep collaborations to jointly drive the development of the AIR RAN ecosystem and create edge intelligent computing services based on AIR RAN.

Fostering Open Joint Innovation: Establish open AIR RAN testbeds and experimental platforms to encourage participation from more small and medium-sized enterprises and research institutions in the technological development and practical implementation of AIR RAN.

Promoting Standardization Cooperation: Advance the formulation of standards related to AIR RAN, establish unified technical frameworks and interface specifications, and promote the standardization of AI model deployment in RAN to ensure interoperability among vendors.

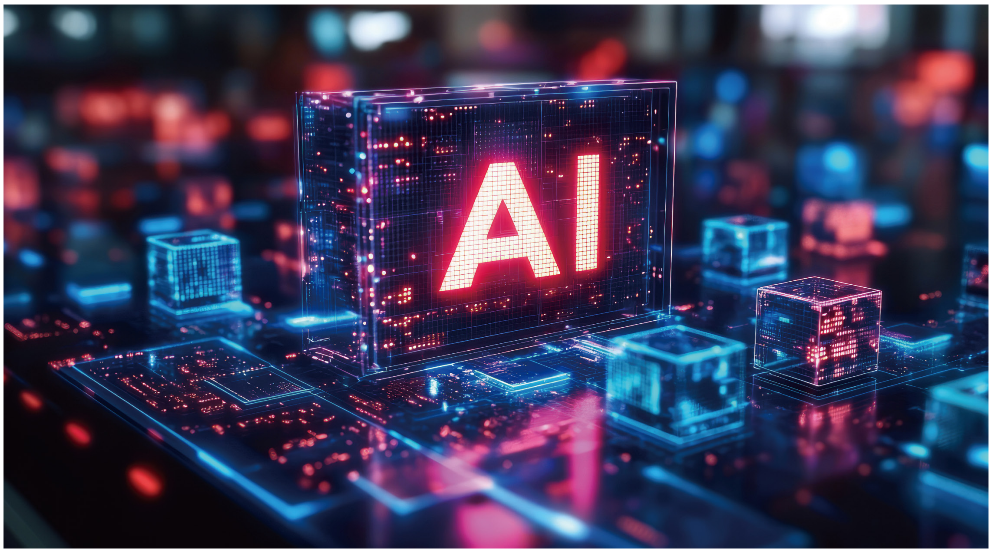
Exploring Business Models: Investigate commercialization paths for AIR RAN, such as realizing value through "Network-as-a-Service" (NaaS) models, and promote deep collaboration with vertical industries to jointly build a new AI commercial ecosystem.

Building a Green Ecosystem: Work together to construct an open, win-win, and environmentally friendly industrial ecosystem, thereby driving the broader application of AIR RAN.

As AIR RAN evolves, it will gradually unlock customer benefits and industry value, driving the transformation of networks from "bringing intelligence with the network" to "empowering the network with intelligence," and laying a solid foundation for a fully intelligent society.

Looking ahead, with continuous technological innovation and breakthroughs, data processing will achieve deep fusion and real-time analysis of cross-domain, multi-source heterogeneous data, ensuring that AI applications can respond precisely and operate efficiently in diverse scenarios. In algorithm design, new-generation AI models that combine efficiency, interpretability, and strong adaptability are expected to be developed, greatly enhancing the intelligence of network resource management and service scheduling. Meanwhile, computing architectures will further evolve toward heterogeneous, distributed, and collaborative paradigms, fully leveraging the complementary advantages of edge and cloud computing to form an omnipresent and elastically scalable computing network that meets the explosive growth in AI business computational demands.

Globally, ZTE will work hand in hand with industry partners to form a close-knit community of technological innovation and value creation, jointly establishing unified industry standards and norms, and promoting cross-regional and cross-platform sharing and collaboration. This will accelerate the global proliferation and application of AIR RAN technology, narrow the digital divide, and ensure that intelligent network services benefit every corner of the world—ushering in a future where everything is intelligently connected, efficiently coordinated, and infinitely innovative.



07 Glossary

ABBREVIATIONS	FULL NAME
AI	Artificial Intelligence
RAN	Radio Access Network
5G	5th Generation Mobile Communications Technology
6G	6th Generation Mobile Communications Technology
MIMO	Multiple Input Multiple Output
XR	Extended Reality
AR	Augmented Reality
VR	Virtual Reality
MEC	Multi-access Edge Computing
KPI	Key Performance Indicator
KQI	Key Quality Indicator
QoS	Quality of Service
L1	Layer 1, Physical Layer
L2	Layer 2, Data Link Layer
L3	Layer 3, Network Layer
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
5G-A	5G-Advanced
NTN	Non-Terrestrial Network
NEA	Network Embedded AI Agent
NaaS	Network as a Service
OPEX	Operating Expense
CAPEX	Capital Expense
GSMA	GSM Association
IMT-2030	International Mobile Telecommunications-2030
GPT	Generative Pre-trained Transformer
V2X	Vehicle-to-Everything
BBU	Baseband Unit
RRU	Remote Radio Unit
K-Means	K-Means Clustering Algorithm
GAN	Generative Adversarial Network
3C	Computer, Communication, and Consumer Electronics

References

[1] "AI: Healthcare's new nervous system", Accenture, 2020
https://www.accenture.com/au-en/insights/health/artificial-intelligence-healthcare

[2] "The Worldwide AI in Telecommunication Industry is Expected to Reach $6.3 Billion by 2026", Research and Markets, 2021
https://www.prnewswire.com/news-releases/the-worldwide-ai-in-telecommunication-industry-is-expected-to-reach-6-3-billion-by-2026--301401190.html

[3] "The Mobile Economy", GSMA, 2024
https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/wp-content/uploads/2024/02/260224-The-Mobile-Economy-2024.pdf

[4] "AI Video Generator Market Size, Share & Trends Analysis Report By Component (Solution, Services), By Application (Marketing, Education), By Organization Size, By Source, By Region, And Segment Forecasts, 2024 - 2030", Grand View Research, 2024
https://www.grandviewresearch.com/industry-analysis/ai-video-generator-market-report

[5] "无线人工智能（AI）技术研究报告", IMT-2030（6G）Promotion Group, 2022
https://www.imt2030.org.cn/

[6] "AI RAN：Telecom Infrastructure for the Age of AI", SoftBank, 2024
https://www.softbank.jp/corp/set/data/technology/research/story-event/Whitepaper_Download_Location/pdf/SoftBank_AI_RAN_Whitepaper_December2024.pdf

ZTE CORPORATION

NO.55, Hi-tech Road South, ShenZhen, P.R. China

Tel: +86-755-26770000 Postcode: 518057

Fax: +86-755-26771999 Website: www.zte.com.cn