ZTE Advocates a Pragmatic and Cost-Effective AI-RAN Approach

Author: Rémy Pascal

October 2025

In partnership with:

ZTE

Contents

AI is only starting to transform the RAN

Why is AI-RAN getting so much attention?

Driven by remarkable AI advancements and the increasing complexity of cellular networks —more technologies, more spectrum, more cell sites—interest in AI-augmented radio access networks (AI-RAN) has surged over the past two years.

AI is expected to soon penetrate all parts of the RAN, from air interface and radio algorithms to baseband processing and management systems. It will play a role at all stages of the network lifecycle, from initial design and planning to deployment and operations.

Numerous AI for RAN applications have emerged, particularly around spectral efficiency, performance optimization, energy savings, and operations and maintenance (O&M).

The anticipated benefits for communication service providers (CSPs) include reduced opex, enhanced network performance, and significant improvements in sustainability, reliability, and security. For users, AI promises better and more personalized experiences.

Early activities show that different approaches are possible

Operators around the world have started exploring the potential of AI for networks.

In Europe, for example, early activities by Deutsche Telekom, Orange, and Vodafone focus on improving operational efficiency.

In Japan and South Korea, operators such as SoftBank and SK Telecom are evaluating graphics processing unit (GPU)-based RAN architectures and the possibility of leveraging the same compute platform for both RAN workloads and edge AI.

Meanwhile, in China, China Mobile is deploying plug-in intelligent computing boards on existing baseband units (BBUs) at thousands of cell sites to optimize performance and efficiency, while also monetizing premium experiences.

ZTE's vision for AI-RAN is layer- and application-specific

While the use cases and applications for AI are virtually limitless, implementation should only occur where and when clear improvements are achievable. AI is not effective at all tasks, and it does not always bring clear benefits in domains governed by explicit rules, deterministic logic, and high interpretability. On the contrary, AI tends to excel in non-linear fields.

The RAN consists of multiple layers, and the potential gains from AI vary across them. In general, the higher the layer, the greater the expected impact. However, there are exceptions, and as mentioned previously, the impact also depends on the nature of the task.

When defining an AI-RAN strategy, the starting point should be identifying specific areas and use cases where AI delivers clear improvements over existing solutions and traditional methods.

Recognizing the specificities and uniqueness of each use case, it is advisable to equip each layer with tailored AI implementations and capabilities. In practice, this means using different AI techniques, tools, and model sizes for each use case.

ZTE's AIR RAN portfolio prioritizes network efficiency and enhanced user experience while minimizing disruption

ZTE envisions a future where AI and connectivity are deeply integrated, transforming the traditional communication platform into a multi-purpose communication-computing-intelligence platform.

Like other vendors, ZTE aims to leverage AI to deliver superior experiences and enhance efficiency. What sets ZTE apart, however, is that it aims to achieve these goals while avoiding a complete overhaul of the existing infrastructure, minimizing the disruption to existing RAN systems.

This vision has been reflected in ZTE's BBU product roadmap over the years. In 2020, the vendor introduced NodeEngine, its first BBU with an embedded general processing board for computing. In 2023, ZTE introduced UniEngine, an all-in-one solution that integrates RAN, core, and computing, designed for lightweight private 5G deployments. Most recently, in 2024, ZTE unveiled AIREngine, enabling the transformation from traditional BBU to native-AI BBU.

ZTE's AIR RAN aims to deliver higher network efficiency and enable monetizable enhanced experiences simply by adding a plug-in AI card into the existing BBU, rather than deploying entirely new hardware.

This approach is not only simpler and more cost-effective; ZTE also argues that there are advantages in keeping computing resources separated (rather than shared), owing to the differing characteristics and requirements of RAN and AI workloads.

Efficiency-focused RAN

ZTE highlighted three main areas of efficiency improvement, as outlined in **Table 1**.

Table 1: Efficiency-focused RAN—areas of application and expected benefits

Area	Description	Benefit
Spectrum efficiency	Frequency division duplex (FDD) massive multiple input, multiple output (mMIMO) beam adaptation with dynamic signal transmission adjustments to optimize performance	15–20% cell capacity gain
Energy efficiency	Optimization of energy savings by service type, time, and target. Less latency-sensitive applications, such as messaging, can, for example, be moved and processed into certain time slots to free up more time slots, enabling greater savings.	Additional 5–10% energy efficiency on top of traditional energy efficiency features, such as symbol/channel/carrier shutdown, deep sleep, and hibernation
O&M efficiency	• A vast amount of useful data from previous use cases is collected offline and desensitized before use in fine-tuning of the existing LLM. • This data is also applied for training use-case-specific small AI models (e.g., alarms model, network monitoring model, troubleshooting model, and scheduling model). • AI agents are built on ZTE's telecom large model, Nebula, to handle specific tasks, such as network assurance agent, fault handling agent, and so on. These agents collaborate to improve and maximize O&M efficiency.	Augment O&M efficiency, thanks to AI agents

Source: ZTE

Experience-centric RAN

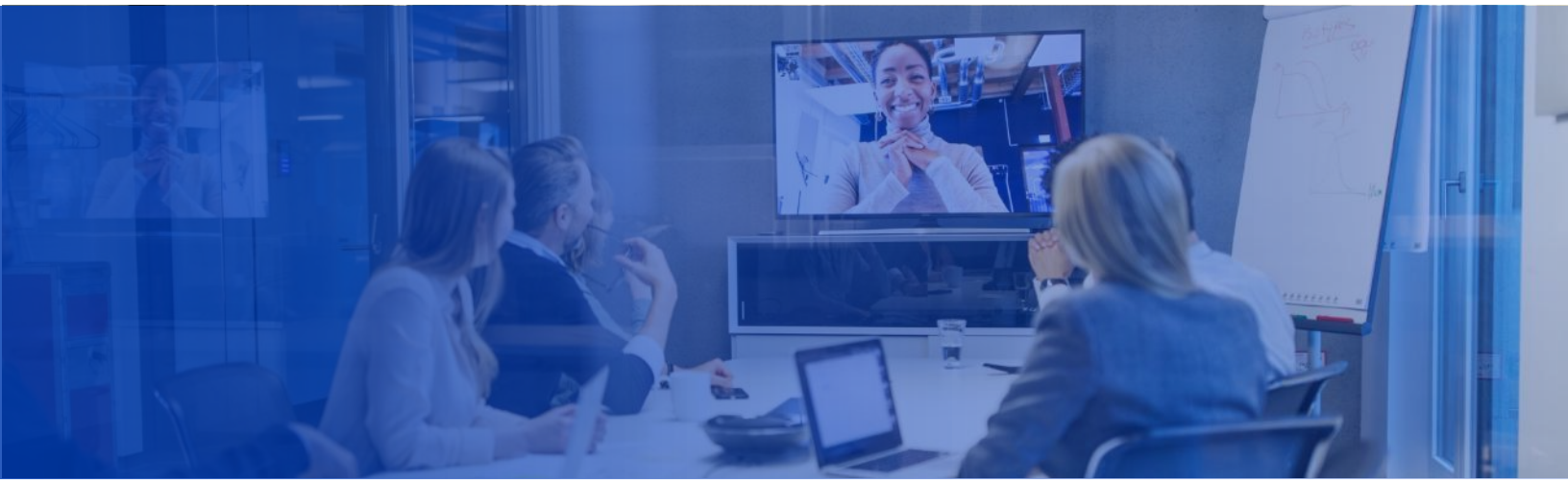
AI-RAN also enables CSPs to bring differentiated and premium customer experiences to market and to monetize these experiences through new innovative 5G-Advanced (5G-A) plans and optional add-ons.

New practices and initiatives for enhanced experience monetization include

- **New 5G plans** with multi-dimension privileges, such as speed (guaranteed uplink and downlink), benefits (video or game memberships), and scenarios (e.g., airports and stadiums)
- **Close-loop experience guarantees** with continuous KQI evaluation against KQI targets

- **User experience and status visualization** including customized logo displayed on end-user device screens (e.g., 5G-A VIP logo)
- **Multiple marketing policies** (e.g., call promotions, SMS promotions, and QR codes)

According to ZTE, AI-based experience monetization could enable operators to upsell customers and generate higher ARPU while simultaneously improving retention rates.



Conclusions

There are various strategies and approaches available to CSPs when considering the adoption of AI in mobile networks. Some options require a fundamental network transformation and significant investment in new RAN compute resources and architectures.

What makes ZTE's AIR RAN approach attractive is that it builds on existing RAN infrastructure and proposes a pragmatic and cost-effective integration of AI in the RAN. AIR RAN is primarily software-focused and requires limited new hardware investment. Moreover, it does not limit itself to the use of large language models (LLMs); instead, it builds on the deep integration of software algorithms with existing hardware.

ZTE is confident in the advantages of its AI-RAN approach, especially in terms of delivering higher network efficiency and enabling enhanced and monetizable user experiences.

Omdia commissioned research, sponsored by ZTE

Appendix

Rémy Pascal, Practice Leader, Mobile Infrastructure
askananalyst@omdia.com

Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa TechTarget, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

Get in touch

www.omdia.com
askananalyst@omdia.com

Copyright notice and disclaimer

The Omdia research, data, and information referenced herein (the "Omdia Materials") are the copyrighted property of TechTarget, Inc. and its subsidiaries or affiliates (together "Informa TechTarget") or its third-party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice, and Informa TechTarget does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an "as-is" and "as-available" basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa TechTarget and its affiliates, officers, directors, employees, agents, and third-party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa TechTarget will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.