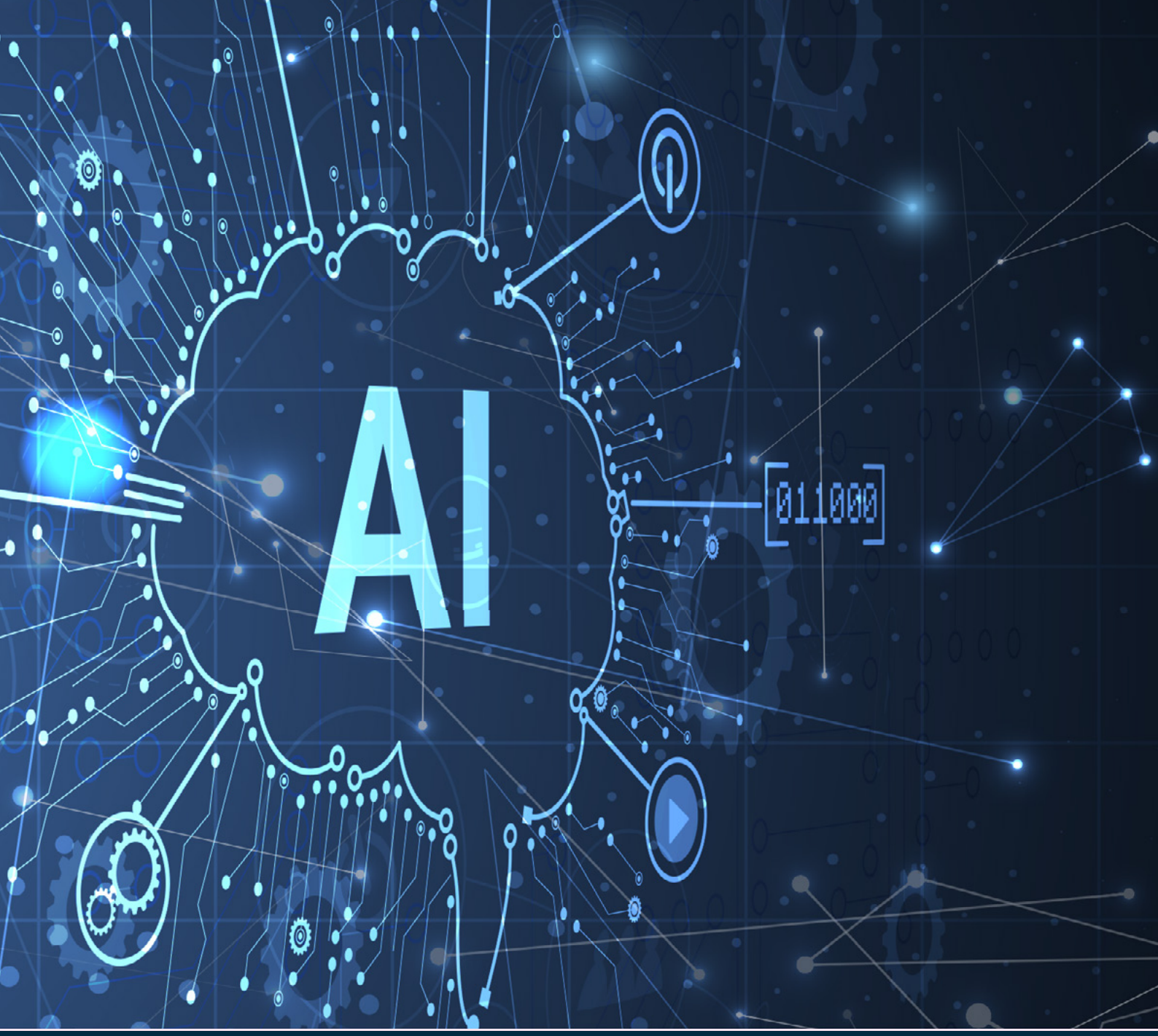


WHITEPAPER

ZTE Native AI RAN Composer

Published by
MOBILE WORLD LIVE

In partnership with
ZTE

Abstract

Future telecommunication generation e.g. 6G will follow a Native Artificial intelligence (AI) architecture, based on the necessity of dealing with the sheer amount of connectivity and complexity that 6G will bring to the table. Native AI will deliver many benefits to mobile networks, including efficient energy use, increased sustainability, new services and O&M, and a better user experience.

It is not necessary to wait for 6G to see how Native AI can improve mobile networks. 5G will inevitably benefit from AI to live up to its maximum potential, and Native AI can also bring similar benefits to 5G as with 6G networks.

New solutions like ZTE's Native AI-based RAN Composer enables an end-user-focused experience via service-aware resource allocation for 5G networks, which can boost traffic, increase network and energy efficiency, and enable intent-driven user experience optimization.

Contents

The need for Native AI in Telco networks

In today's world, mobile networks have become an integral part of our daily lives, supporting a multitude of applications and services. However, the ever-increasing demand for data rates, connectivity, and seamless user experiences poses significant challenges for mobile network operators. To address these challenges, it requires a transformation of traditional network management strategy. This article will discuss the challenges faced by modern mobile networks, why traditional RAN management approaches fall short, and the need for a dynamic, AI-driven solution.

5G is designed around three main concepts: Enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (URLLC) and Massive Machine-Type Communications (mMTC) – in other words, ultra-fast data speeds, ultra-low latency/high reliability and the ability to support millions of IoT connections. This means that in addition to consumer services, 5G can and does support many enterprise applications and services, especially in vertical industry sectors such as manufacturing. Last year in China alone, the number of connected IoT devices outstripped the number of human connections.

The mobile industry has already set its sights on developing 5G-Advance and 6G. While the exact technical details of how 6G will work are still being hashed out, one thing that's certain is that AI will be a crucial component. The 6G architecture is already being designed Native AI end to end, from the user equipment (UE) to the RAN and even the transport system, giving 6G mobile networks the ability to calculate the optimal service between any given end points. This could even give rise to a new quality evaluation mechanism, QoAIS (quality of AI service), as services increasingly rely on AI to deliver the best customer experience.

While this highlights the important role that Native AI will play in 6G networks, the same is actually also true for 5G networks. The difference, of course, is that 5G RAN architectures weren't designed with Native AI in mind. But 5G RANs can still make use of Native AI – in fact, they'll have to as 5G adoption grows.

We don't have to wait for 6G to leverage Native AI for next-gen mobile services. 5G can leverage Native AI solutions today to boost traffic, guarantee B2B SLAs, improve O&M efficiency, and enable intent-driven experiences.

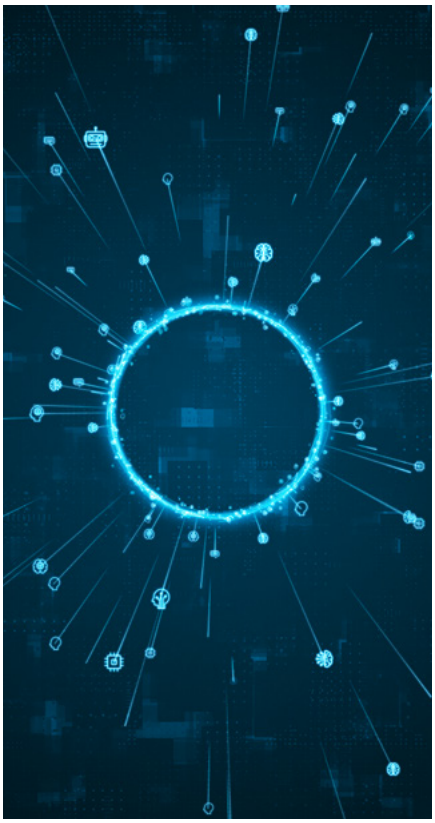
The industry's first Native AI

ZTE's RAN Composer, the industry's first Native AI engine for 5G RANs, elevates radio resources management to a new level, extending AI from network management system to base station.

The Native AI engine, performing inside the 5G nodes, leverages algorithms, network data and base station computing power that results in multi-dimensional awareness of the whole network, including UE capability, service characteristics, and network serving capability. The engine enables drilling from the cell level to the RF fingerprint level. The cell can be divided into small grids, each representing the wireless characteristics of everyone and everything connected to the cell.

Based on this foundation, the Native AI engine conducts traffic-pattern analysis to determine the application or service the customer is using, together with the RF fingerprint knowledge base, this enables the network to allocate different radio resources to support each service ideally for every user.

In addition, functions like "Quality on Demand" utilizing GSMA's Open Gateway API can be enabled. For example, in B2B2C cases, e.g. Pay-TV, the Pay-TV service provider can enable high value customers for premium services to use ensured bandwidth services to guarantee the quality of the movie or if the operator sees that TikTok is a major traffic generator, ZTE's Native AI engine can enable the TikTok service provider to schedule the capability from the Open Gateway API to improve application performance.



Key benefits of Native AI

Better network efficiency

There are several key benefits that Native AI delivers to 5G, starting with better network efficiency. To illustrate this, let's take a closer look at how service-aware resource allocation works.

When the data enters the 5G RAN, the Native AI engine performs real-time traffic pattern analysis to determine the type of application or service being used – for example, an AR service, a HD video or a cloud game.

The engine then looks at the available radio resources across different bands. Users will typically be on a multiband network with at least three bands (for example, 900/1800/2100 MHz). These multi-band resources constitute a resource pool from which the operator can allocate radio resources according to the service requirements they need.

The result is a very precise resource allocation, exactly fulfilling the demand of the end-user, helping to improve network efficiency.

Supreme user experience

As mentioned above, Native AI reaches all the way down to the individual grids in the RF fingerprint of each cell, enabling the network to know exactly what application or service each customer using via traffic pattern analysis. This also enables the operator to evaluate network performance as it relates to each service in use.

Traditionally, operators use KPIs for network performance evaluation, but this is regardless of the service being used on the network. So when it comes to each individual user, their experience may fluctuate above or below the expected experience. Consequently, the actual experience not consistent or guaranteed.

However, with the Native AI engine, precise radio resource allocation according to the service requirements can ensure that the network delivers optimal performance for each service for each user.

Higher energy efficiency

Native AI can also improve the energy efficiency in a couple of ways. Apart from the fact that it requires no extra hardware for the cell site, Native AI's precise resource allocation also means that no extra power is waste on unnecessary data transmission.

The traffic pattern analysis described above not only identifies which services are in use within the RF fingerprint, but also detects co-coverage across cells and bands. For example, let's say you have two layers of co-coverage – one at 1800 MHz, and the other at 1800 MHz and 2100 MHz. If you need to wake up a cell at 2100 MHz, traditionally you have to wake up the whole band across all the cells covering that user.

With Native AI, you can wake up the minimum amount of co-coverage needed based on the service to be supported. For example, if you have co-coverage at cells 4, 5, 6 and 7, you can wake up the 2100 band only in Cell 4 if that's all you need to support the service that customer is using.

To an extent, 5G can do this quite granularly without Native AI, but adding Native AI to the mix can extend that granularity further to realize even more energy savings. Figures will vary based on specific deployment scenarios, of course, but generally speaking, without Native AI, 5G can already realize 20% energy savings compared to 4G. Native AI can boost that to 25-28%. The overall electricity saving will be 52,000,000 kWh yearly for 7,200 sites, which is enough to provide electricity for a community of over 70,000 people for a year.

ZTE's exploration of Native AI

Service-aware resource allocation is just one of a number of possible use cases for Native AI in 5G. ZTE has been exploring various use case possibilities, and so far has determined the following categories of use cases:

- Intelligent traffic pattern analysis and intelligent user steering
- Traffic prediction-based energy savings and enhancement
- CSI feedback
- Smart O&M (equipment risk identification, KPI abnormality detection, network diagnosis, etc.)

Each of these categories is determined in large part by the requirements for AI training and inference. Put simply, the feasibility of using Native AI for a given use case depends on the location of the data used to train the AI model and the location of the available computing power necessary to crunch that data – i.e. the base station or the UE.

For example, for the use case of intelligent user steering, the AI training and inference can both be conducted on the base station. This means that the base station has both the data needed to do the training and the computing power to support the training via cards inside the baseband unit at times when the base station is experiencing a low traffic load. Because the base station is the node closest to the user, the training and inference combination can be done in near-real time or real time, which makes it ideal for intelligent user steering (more on this a little later).

For the second use case category, equipment risk identification and KPI abnormality detection, training and inference can all be done on the device side in non-real time,, although transmission bandwidth also becomes a factor here. For traffic prediction-based energy savings, the training can be done on the device side while the inference is done in the base station in non-real time. And In the case of traffic pattern analysis, the inference is done in the base station but the training can actually be done offline, as it requires machine learning to process the massive amount of data. ZTE has found that using offline data to train AI is accurate enough to support traffic pattern analysis use cases.

Native AI-based RAN Composer

ZTE has channelled its Native AI research into its new RAN Composer, a Native AI-powered intelligent RAN that operates on three layers, starting of course with the physical network from which data is collected.

Above that is the intelligent service layer, which is where the data models, computing power and algorithms reside.

The computing power component enables base stations to orchestrate and share computing power resources in the network. RAN Composer can support inter-site computing power sharing that gives the baseband sufficient computing power to support more AI-based training.

As for the algorithm, ZTE is using the foundation large language model (LLM), which works together with its traditional smaller algorithm models for things such as traffic, the experience, energy efficiency, load, coverage, capacity and fault detection.

Via open APIs such as the GSMA Open Gateway APIs, this intelligent service layer is what empowers the intelligent applications mentioned earlier: supreme user experience, better energy efficiency, and higher O&M efficiency.

There are also several specific benefits that RAN Composer enables, as we will see in the next section.

RAN Composer benefits

Boosting B2C 5G traffic

RAN Composer leverages Native AI to stimulate increases in 5G traffic and boost operator revenue. Here's how:

The traffic pattern analysis feature can already recognize over 16,000 service types. From there, the RAN Composer performs intelligent parameter optimization based on historical learning, and then implements intelligent user steering, in which users are connected with the right radio resources within the cell that deliver optimal throughput and performance. This not only improves the user experience, but also boosts traffic.

ZTE has conducted a case study with China Mobile to illustrate this. For China Mobile, video comprises more than 70% of its traffic. By using RAN Composer, China Mobile was able to help users have the expected experience according to service requirements, which in turn improved user loyalty.

In this way, low video definition ratio (less than 720p) could be reduced from 21% to 16%. At the same time, the high definition video (over 720p) could be increased by more than 20%. These improvements also result in a 10% increase in total cell traffic.

Guaranteeing B2B SLAs

The RAN Composer can also help to guarantee the SLA of vertical applications. Here, the traffic pattern analysis is more focused on industrial protocol scenarios. For example, the intelligent RAN can identify an AGV robotic arm and do the precise scheduling based on historical learning and different QoS templates, with different parameter settings for different service requirements. From there, the RAN Composer performs a close-loop performance evaluation.

If the use case involves near-real-time or real time coordination between multiple AGVs, this comes with very stringent jitter requirements. For traditional networks, there is no guarantee of super low jitter. However, the RAN Composer can provide microsecond-level jitter to guarantee multiple AGV coordination. By the same token, RAN Composer can meet the 99.999% SLA requirement for applications requiring low latency and high reliability, such as cloud PLC, and guarantee sufficient bandwidth for data-intensive applications like computer vision on factory lines.

Intent-driven supreme user experience

The Native AI RAN Composer also furthers the shift towards "intent driven" experiences.

Thanks to the rise of LLMs, we are on the way to transforming from the graphic user interface to the natural user interface (NUI), which is necessary to introduce the 'intent-driven' experience. Currently, the intent should include time, location, service type and target (e.g. vMOS>=4) , it becomes possible to spot weak intents with LLM – for example, where performance needs to be improved.

The Native AI RAN Composer can do the intent translation from the system automatically. Within one minute, the operator can retrieve the four factors (time, location, service type and target), after which the system automatically generates everything, including the traffic pattern analysis dictated by the self-generation/self-optimization policy, and evaluates performance from that.

One result of this is a dramatic improvement in O&M efficiency, which we evaluate by looking at how many human resources are needed to carry out a task and how much time it takes them to complete it. Previously, it required significant time and human resources.

In this case, the current way of guaranteeing intent-driven experience requires up to six O&M people to handle awareness, analysis and configuration, which takes up to 50 minutes to do. Then you need six people to do the subsequent performance monitoring. With RAN Composer, one person can handle all of this in around one minute.

This also translates to user experience improvements. For livestreaming, for example, you can prioritize different guarantee levels, so that a user with a Level 1 guarantee can be prioritized over other users with Level 2 or Level 3 guarantees to have better throughput and lower latency.

Conclusion

5G comes with the promise of ultra-fast speeds, ultra-low latencies and massive connectivity to support the IoT, but it must deliver that promise in the context of complex multiband environments whilst also supporting the innovative new services that 5G enables for consumer and enterprise customers alike. All of this created unprecedented challenges to operators in network operation and development.

6G shows us that Native AI is a key technology to managing all of this – and that we can implement Native AI in 5G today. Extensive research from ZTE shows how – with the right model training and inference – a Native AI engine can be built into 5G RANs to support multiple use cases that improve network efficiency and lower energy consumption, as well as deliver the supreme experience that users expect from new 5G-powered services.

The proof is in ZTE's Native AI RAN Composer, which leverages algorithms, network data and base station computing power to create multi-dimensional awareness of the whole network, including UE capability, service characteristics, and the network serving capability. By radio resources management revolution based on Native AI, RAN Composer can help boost traffic, guarantee SLAs for enterprise customers, improve energy efficiency, and realize the full potential of intent driven experiences.

ZTE helps to connect the world with continuous innovation for a better future. The company provides innovative technologies and integrated solutions, its portfolio spans all series of wireless, wireline, devices and professional telecommunications services. Serving over a quarter of the global population, ZTE is dedicated to creating a digital and intelligent ecosystem and enabling connectivity and trust everywhere. ZTE is listed on both the Hong Kong and Shenzhen Stock Exchanges.

www.zte.com.cn/global

Mobile World Live is the premier destination for news, insight and intelligence for the global mobile industry. Armed with a dedicated team of experienced reporters from around the world, we are the industry's most trusted media outlet for breaking news, special features, investigative reporting, and expert analysis of today's biggest stories.

We are firmly committed to delivering accurate, quality journalism to our readers through news articles, video broadcasts, live and digital events, and more. Our engaged audience of mobile, tech and telecom professionals, including C-suite executives, business decision makers and influencers depend on the unrivalled content and analysis Mobile World Live provides to make informed business decisions every day.

Since 2016, Mobile World Live has also had a team of in-house media and marketing experts who work directly with our brand partners to produce bespoke content and deliver it to our audience in strategic yet innovative ways. Our portfolio of custom work - including whitepapers, webinars, live studio interviews, case studies, industry surveys and more – leverage the same level of industry knowledge and perspective that propels our newsroom.

Mobile World Live is published by, but editorially independent from, the GSMA, producing Show Daily publications for all GSMA events and Mobile World Live TV – the award-winning broadcast service of Mobile World Congress and home to GSMA event keynote presentations.

Find out more at www.mobileworldlive.com