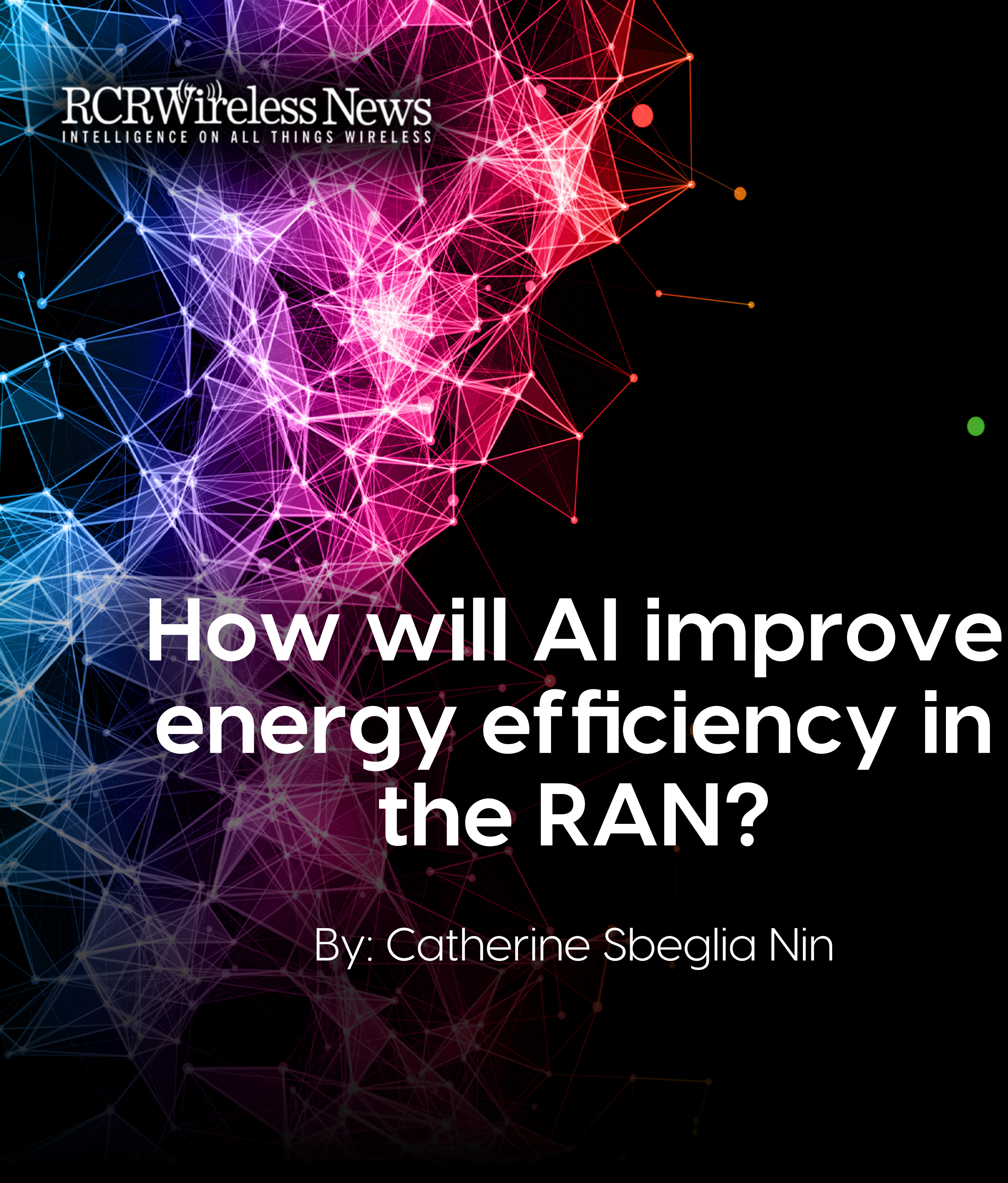
RCRWireless News
INTELLIGENCE ON ALL THINGS WIRELESS



How will AI improve energy efficiency in the RAN?

By: Catherine Sbeglia Nin

Sponsored by
NOKIA

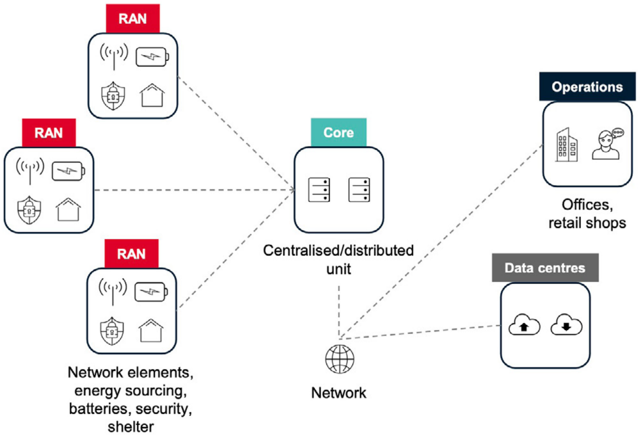
(Image courtesy of 123RF)

INTRODUCTION

The radio access network (RAN) consumes, by far, the majority of the power needed to operate mobile networks. In fact, GSMA Intelligence found that 87% of an operator's energy consumption comes from the RAN, which it said includes BTS, Node B, eNodeB and gNodeB energy usage, all associated infrastructure energy usage such as from air-conditioning, inverters and rectifiers, as well as energy usage from repeaters and consumption associated with backhaul transport.

Such findings make the RAN a clear place for telcos to start as they begin their massive energy efficiency targets against the pressing need to reduce operational costs while working towards long-term climate action goals without sacrificing network performance. And beyond cyclical hardware upgrades that deliver incremental energy efficiency gains, the ability to collect, parse and action RAN data to right-size network resource provisioning is a job well-suited to artificial intelligence (AI) and machine learning (ML). AI-enabled RAN, or AI-RAN, can understand when to scale down resource allocation to meet actual demand rather than maintaining peak load. And from there, the opportunities to infuse AI into the RAN for both internal-facing efficiencies and monetization of advanced services are quite material.

Where mobile operators use energy in their network


RAN
RAN
RAN
Network elements, energy sourcing, batteries, security, shelter
Core
Centralised/distributed unit
Network
Operations
Offices, retail shops
Data centres

Source: GSMA Intelligence

(Image courtesy of 123RF)

THE AI-RAN ALLIANCE AIMS TO 'WEAVE' AI INTO THE RAN FABRIC

Officially launched in February 2024, the AI-RAN Alliance is a group of technology and telecom leaders focused on the integration of AI into cellular technology to further advance RAN technology and mobile networks. The AI-RAN concept refers to the deployment of graphics processing unit (GPU)-based infrastructure that can run both wireless and AI workloads concurrently, turning networks from single-purpose to multi-purpose infrastructures, increasing end-to-end efficiency and enabling telcos to turn their cell sites into revenue sources.

According to Alex Jinsung Choi, chair of the AI-RAN Alliance and principal fellow of SoftBank Corp.'s Research Institute of Advanced Technology, the main mission of the group is to "weave AI right into the fabric of the radio access network. "We are not just about making networks faster and more flexible, we want to transform them into self-organization, self-optimizing, self-managing systems that can handle real-time changes and anticipate maintenance and efficiently manage resources," he told *RCR Wireless News*.

He continued: "AI is perfectly poised to revolutionize the RAN. In the AI-RAN Alliance, we are seeing a major shift towards integrating AI directly into RAN operations on shared computing platforms including GPUs and NPUs [neural processing units]… We are looking at networks that can configure, optimize and repair themselves with minimal human intervention. This could lead to revolutionary communications technologies that change how we interact with the digital world."

Unlike traditional Standard Development Organizations (SDOs), such as the 3GPP and the O-RAN Alliance, that focus on standardizing network interfaces, the AI-RAN Alliance revolves around "real-world" AI applications. "The role of SDOs is very important and their output will be utilized by the AI RAN Alliance member companies," Choi clarified. "However, SDOs, based on my experience, tend to work relatively slowly and the AI RAN Alliance wants to create a fast-moving organization to bring the rapidly evolving AI and machine learning technologies to RAN and reinvent the RAN from a more software perspective, a more AI perspective, a more AI accelerated computing perspective."

The AI RAN Alliance has three distinct working groups:

AI-for-RAN: This group is looking at how AI can enhance the performance of the radio access network, diving into how AI can improve efficiency, boost capacity and achieve key performance targets.

AI-on-RAN: This group explores using RAN to enhance the capabilities of AI applications. This group also addresses the challenges associated with running AI applications directly on radio access networks, making sure networks can handle growing demands without compromising things like latency and security.

AI-and-RAN: This final group is chaired by Nokia and according to the vendor's VP and Head of Cloud RAN Aji Ed, it looks at how to leverage the synergies of infrastructure and AI for different workloads. This means using the same infrastructure to run both RAN workloads and AI workloads simultaneously, with the idea that doing so will open up new revenue streams for telcos by enabling them to host various AI applications on the same platforms that run network functions.

Image courtesy of: AI-RAN Alliance

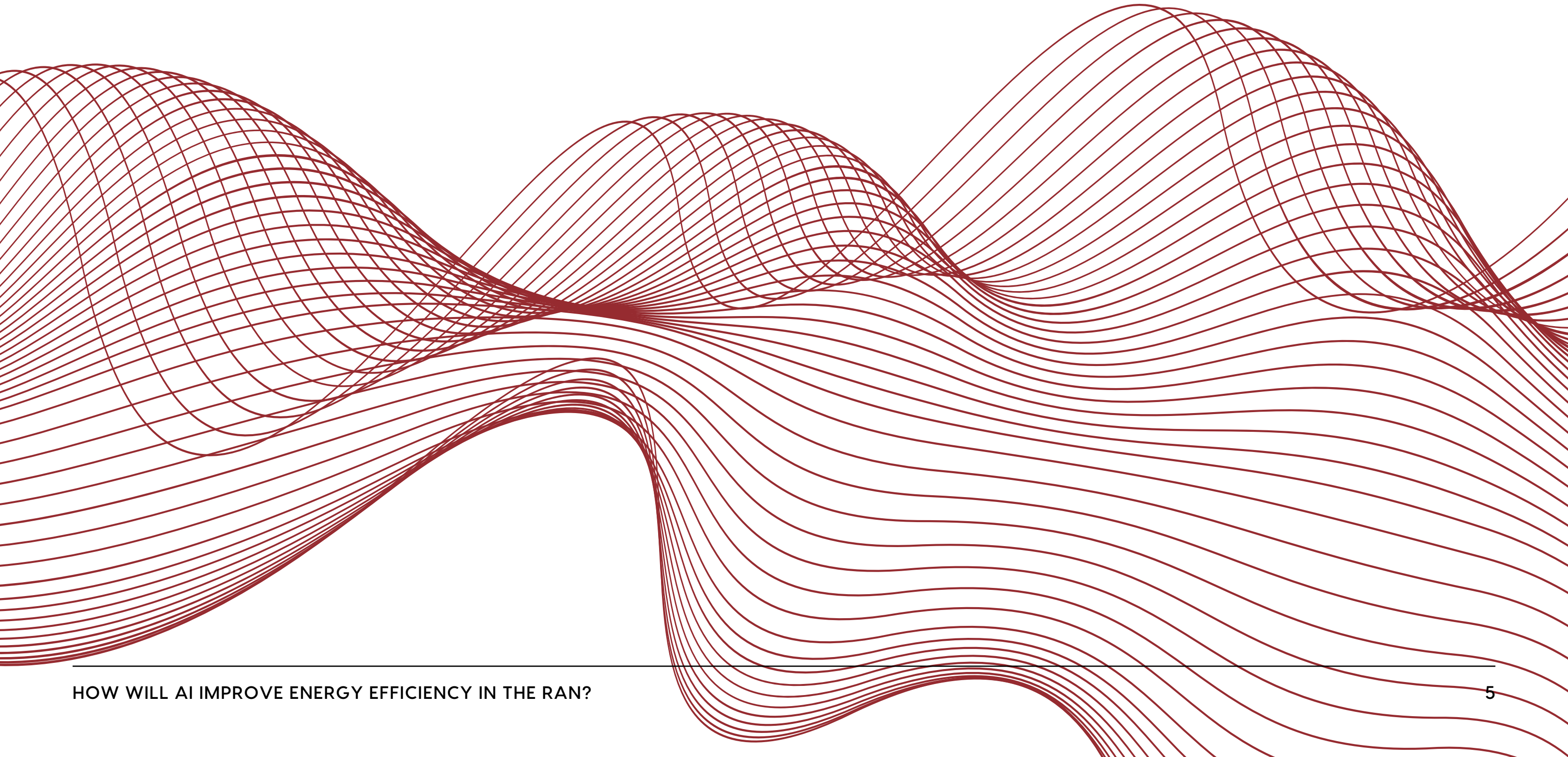
All three of these groups, in their own way, are after more RAN efficiency, and therefore, are an important piece to the larger sustainable network conversation. “The main objective includes using AI/ML to enhance network efficiency, optimizing resource allocation and providing a platform for member companies to integrate … By integrating AI into RAN, we want to ensure the network not only improves [in] cost performance and effectiveness but also creates new business models,” summarized Choi.

“The soul of [the Alliance] is to actually get things done, not be bogged down by standards and so on but actually show how we can deliver the best-in-class solutions and deliver the blueprints so the larger community can benefit,” stated Nvidia’s General Manager of AI, 5G and telecoms Soma Velayutham. He added that for Nvidia, its participation in the AI-RAN Alliance is inspired by the desire to help telcos realize what it sees as a big opportunity in 6G for them to actively participate in the AI economy. “So can telcos actually transform themselves from a connectivity provider to a connectivity plus AI fabric provider?” he asked. “That’s what excites Nvidia most.”

The AI-RAN Alliance’s founding members are Softbank, Nokia, Nvidia, Arm, DeepSig, Ericsson, Microsoft, Northeastern University, Samsung Electronics, the University of Tokyo and T-Mobile US.

SOMA VELAYUTHAM,
General Manager of AI,
5G and Telecoms
Nvidia

“Can telcos actually transform themselves from a connectivity provider to a connectivity plus AI fabric provider? That’s what excites Nvidia most.”



Triangulates

Explore the technology trends shaping the future of your business

- A series of digital events brought to you by Nokia spotlighting the latest tech trends.
- Discover how mobile network operators can gain a competitive edge.
- Learn how the latest technologies enable new services for enterprises.
- Deep dive into one business-critical topic with three dynamic speakers at each event.

Visit the Triangulates website to view all the episodes.

(Image courtesy of 123RF)

'ZERO TRAFFIC, ZERO ENERGY' – USING AI FOR DYNAMIC NETWORK MANAGEMENT

Another priority for telcos, according to Choi, is the reduction of operation expenses (OpEx). "And energy consumption is one of the biggest factors of OpEx. AI will play a big role in making RAN operations more sustainable. Through predictive analytics, AI can predict the most energy-efficient times and places to operate network components; it can switch systems into low-power mode during times of low usages, such as overnight to conserve energy," he said.

By identifying lower network load and only scaling up resources when necessary, AI and ML can fine-tune energy usage based on real-time demand. "This kind of dynamic scaling can help ensure that network performance targets are always met while significantly cutting down unnecessary energy usage," said Choi. "It's not just about using less energy; it's about using the right amount of energy at the right time."

One such example is smart radio frequency (RF) channel management in which Data Quality Monitoring (DQM) — a kind of reinforcement learning model — is leveraged to provide dynamic RF channel reconfiguration, leading to material energy savings. "Dynamically, we can change channels, cells and even signal levels on and off based on network traffic," explained Choi, adding that this approach involves offloading users to neighboring cells before switching off certain cells to ensure seamless connectivity.

Nokia's approach to energy efficiency has been self-described as "extreme." For instance, the vendor takes this idea of spinning network resources up and down to reflect real-world traffic to the max with its Deep Sleep cell switch-off mode. This solution, announced in February, leverages software in the vendor's AirScale Habrok Massive MIMO radio units to reduce energy consumption by up to 97% compared to a cell on air but without traffic. The company also has a 'Zero traffic, zero energy' solution that switches off all radio resources when there is zero traffic.

Additionally, ReefShark System-on-Chip (SoC), Nokia's SoC technology dynamically adjusts internal resources based on traffic demands, contributing to a 15% reduction in energy consumption. Energy efficiency of the radio network software solutions can be further enhanced with the vendor's MantaRay Energy solution, which automates and optimizes the configuration of the energy-saving software features.

It was this last solution that Ed highlighted when speaking with *RCR Wireless News*. The Nokia MantaRay SON solution, he shared, taps AI and ML to autonomously identify and address network inefficiencies by automatically switching off underutilized cell sites. "This can optimize the RAN energy based on the real traffic pattern," he said, adding that Nokia also uses digital planning services like digital twins to further improve efficiencies. "It's all [to] improve the energy efficiency for the entire site overall."

While several telcos around the world have signed on to Nokia's "extreme" approach to energy savings — including Stc in July and Chunghwa Telecom in October —

David Soldani, the SVP of next generation advanced research at Rakuten Mobile told *RCR Wireless News* in a separate conversation that switching cell sites on and off invites a "tremendous" amount of risk, and so, telcos want AI to deliver more "liquidity" when it comes to energy consumption by using it to predict — at various network levels — where such an action should be taken.

"For example, [Rakuten is] using AI to determine if certain CPUs in the hardware can be switched off … that's where AI contributes, down to the hardware. And if the platform supports it, we have the ability to go a level up to scale horizontally or vertically the resources you provide to your nodes … You can measure your energy and then move your workloads to provide less or more resources, vertically to the nodes or horizontally to the cluster, so that you have an optimal way of consuming energy," he said.

ALEX JINSUNG CHOI,
Chair, AI-RAN Alliance and Principal Fellow, SoftBank Corp.'s Research Institute of Advanced Technology

"It's not just about using less energy; it's about using the right amount of energy at the right time."

(Source: SoftBank)

SOFTBANK AND NVIDIA PILOT FIRST AI-RAN NETWORK

As part of an extensive AI tie-up between Nvidia and SoftBank, the Japanese telco successfully piloted the world's first combined AI and 5G telecom network using the Nvidia AI Aerial accelerated computing platform, an end-to-end solution based on a full-stack, virtual 5G RAN software integrated with 5G core. The field tests took place in Fujisawa City in Japan's Kanagawa prefecture.

Velayutham said that SoftBank's AI-RAN network validated all three of the AI-RAN Alliance's previously described concepts — AI-for-RAN, AI-on-RAN and AI-and-RAN — by showcasing RAN workloads co-existing with AI workloads, as well as how AI can optimize the RAN. The network also demonstrated AI applications like robotics running over a RAN network and a self-driving car using AI talking over the 5G network.

And to bring this news back to the topic of energy efficiency in the RAN: Nvidia's Senior Director for Telco Marketing Kanika Atri wrote in a blog that the live field test revealed not only that GPU-enabled RAN systems are feasible, but they are also "significantly better in energy efficiency and economic profitability."

Three AI-RAN workload distribution models

But as Velayutham pointed out, energy efficiency is only part of the equation: "It's not just about having efficiency — which is very important — but it's also about utilizing the asset to the maximum," said Velayutham. "Telecom networks are designed for peak and when you design something for peak, that actually means that most of the time, it's underutilized. How do you create more utilization? Having an orthogonal workload in the daytime …

using it for RAN when it's highly utilized and, in the nighttime, you might actually use it for AI workloads."

Therefore, AI and RAN multi-tenancy and orchestration — the ability to run and manage RAN and AI workloads concurrently — is one of the key principles of AI-RAN technology. Multi-tenancy can refer to dividing network resources based on time of day or on the amount of compute.

In the SoftBank AI-RAN trial, the pair stated that concurrent AI and RAN processing was successfully demonstrated between RAN and AI workloads, with the goal of maximizing capacity utilization. Nvidia claimed that AI-RAN enables telcos to achieve almost 100% utilization compared to 33% capacity utilization for typical RAN-only networks — an increase of up to 3x — while implementing dynamic orchestration and prioritization policies to accommodate peak RAN loads.

As such, the field network also provided the opportunity to compare the multiple workload distribution models that are emerging for AI-RAN — RAN-only, RAN-heavy or AI-heavy. These labels refer to how much of the server is dedicated to RAN vs AI workloads at any given time, which again, can be adjusted dynamically depending on traffic.

In the AI-heavy scenario, Nvidia used a one-third RAN and two-third AI workload distribution and claimed that for every dollar of CapEx investment in accelerated AI-RAN infrastructure, telcos can generate 5x the revenue over five years, with the overall investment delivering a 219% profit margin, considering all CapEx and OpEx costs.

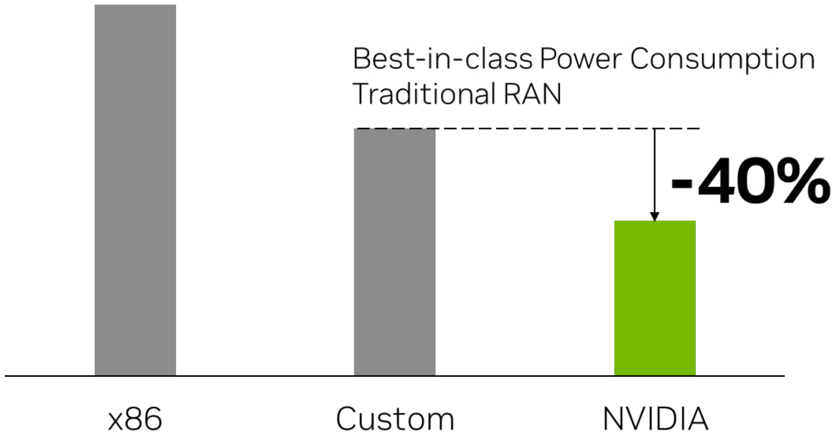
In the RAN-heavy scenario, Nvidia used two-thirds RAN and one-third AI workload distribution, which showed revenue divided by CapEx for Nvidia-accelerated AI-RAN is 2x, with a 33% profit margin over five years.

Finally, in the RAN-only scenario, Nvidia concluded that using the Aerial RAN Computer-1 is more cost efficient than custom RAN-only solutions.

"From these scenarios, it is evident that AI-RAN is highly profitable as compared to RAN-only solutions, in both AI-heavy and RAN-heavy modes. In essence, AI-RAN transforms traditional RAN from a cost center to a profit center. The profitability per server improves with higher AI use. Even in RAN-only, AI-RAN infrastructure is more cost-efficient than custom RAN-only options," wrote Atri.

And when it comes to power performance specifically, in the 100% RAN-only mode, power performance in Watt/Gbps of the GB200-NVL2 server — which resides inside Nvidia's AI Aerial accelerated computing platform — achieved 40% less power consumption than existing RAN-only systems and 60% less power consumption than commercial-off-the-shelf (COTS) x86-based vRAN, as well as similar efficiencies across distributed-RAN and centralized-RAN configurations.

Here seems like a good place to insert a crucial point — We're not talking about generative AI (gen AI), and that distinction matters, particularly when discussing energy and power. "There is a key difference between generative AI solutions, which dominate attention and dominate investment in energy-intensive GPU clusters," stated Choi. "The energy saving applications of AI/ML … for the radio access network, which is typically non-gen AI types of AI models … the prevailing trend shows that any additional energy requirements to implement operationalized AI/ML model[s] is significantly overweighted by the energy savings AI enables."



(Source: Nvidia)

(Image courtesy of 123RF)

PREPARING FOR THE FUTURE – FROM CLOUD-NATIVE TO AI-NATIVE

The vision for 6G is that AI will be end-to-end and touch every aspect of the network, from design and planning to optimization; however, in a previous conversation with *RCR Wireless News*, McKinsey and Company Senior Partner Tomás Lajous acknowledged that being cloud-native is a requirement for being AI-native." The notion of having a cloud-native telco, I think, is part and parcel to having an AI-native telco," Lajous said. "If we were to start a telecom company from scratch today, what would be the best way to put it together? And that's where we landed on the best way to put it together is by having AI at the core. And that means having AI assist essentially every decision and operating model, and a culture that embraces AI in order to do so, all the way from marketing and call centers to the network."

But achieving this, he continued, relies on a "very deep technical architecture that goes with it. And the best way to do [that] is by bringing in the cloud." In this way and in this context, cloud is "required" for AI.

By embracing cloud-native, Blue Planet VP's Kailem Anderson told *RCR Wireless News* more recently, 5G is striving to be more elastic, flexible and portable. "The most important part of 5G that plays forward in 6G is that it's dynamic, which means that whether it's for a consumer or a business service, the customer expects an on-demand experience," he said, adding that this required the entire software stack that supports 5G to be "reimagined." This, he continued, brought upon the "rapid acceleration and transformation of the software stack," introducing new technologies like AI-driven automation. As such, the concepts of AI and cloud-native — as well as network disaggregation — get "clumped together" for Anderson because it's all about taking a data-driven approach to enable the delivery of real-time customer experiences.

AI-RAN Alliance's Choi agreed with Anderson, commenting, "The shift towards cloud-native architecture and disaggregation of network functions over the last several years in 5G has laid a robust foundation for the advent of 6G."

Choi added, as well, that the flexibility and modularity of cloud-native and disaggregated systems allows telcos to integrate AI features "seamlessly, ensuring that the networks are not just more powerful but also responsive to market dynamics."

Cloud-native, but also software-defined and centralized

Another principle of 5G that adds to its dynamic nature is the use of software-defined networking (SDN), which allows data to move easily between distributed locations. This, of course, is key for cloud applications. The establishment of SDN in 5G, according to Ronnie Vasishta, who is responsible for the telecom business, strategy and products at Nvidia, has made it so that 6G "all of a sudden doesn't seem like that big of a challenge."

He explained the transition from 5G to 6G further: "If you have software definition, as you're going from 5G to 5G-Advanced and you're including ultra-reliable low-latency communication [and] precision positioning… all of those are software upgrades." A software-defined infrastructure will also present telcos with the opportunity to leverage that hardware for other things, said Vasishta. "So you can also put hardware on the base station, in a mobile switching office, you can put it in a centralized cloud — same software, same hardware, it just scales it."

But let's dig into this centralized radio access network (cRAN) configuration a bit more. While Nokia's Ed confirmed that RAN centralization, when compared to a distributed deployment, has a host of benefits related to energy efficiency, including the ability to pool resources and enhance hardware utilization by centralizing virtual functions, he also said the idea that AI RAN requires a centralized architecture is a misconception. "That's not necessarily the case. It must be possible to deploy AI RAN in a very small dRAN [distributed RAN] configuration or a very large cRAN configuration," he said.

And that's because not every telco and not every network is ready to centralize, primarily due to challenges related to a lack of necessary transport infrastructure. "If you look at the reality, not all of the networks around the world are ready for centralization today… At the same, that's the journey that we need to take, but we need to be realistic in terms of how we take the journey forward because we need to be able to support the current network architecture as well with AI RAN approach, which means it would be a hybrid model where we would need to have the right solutions and right business

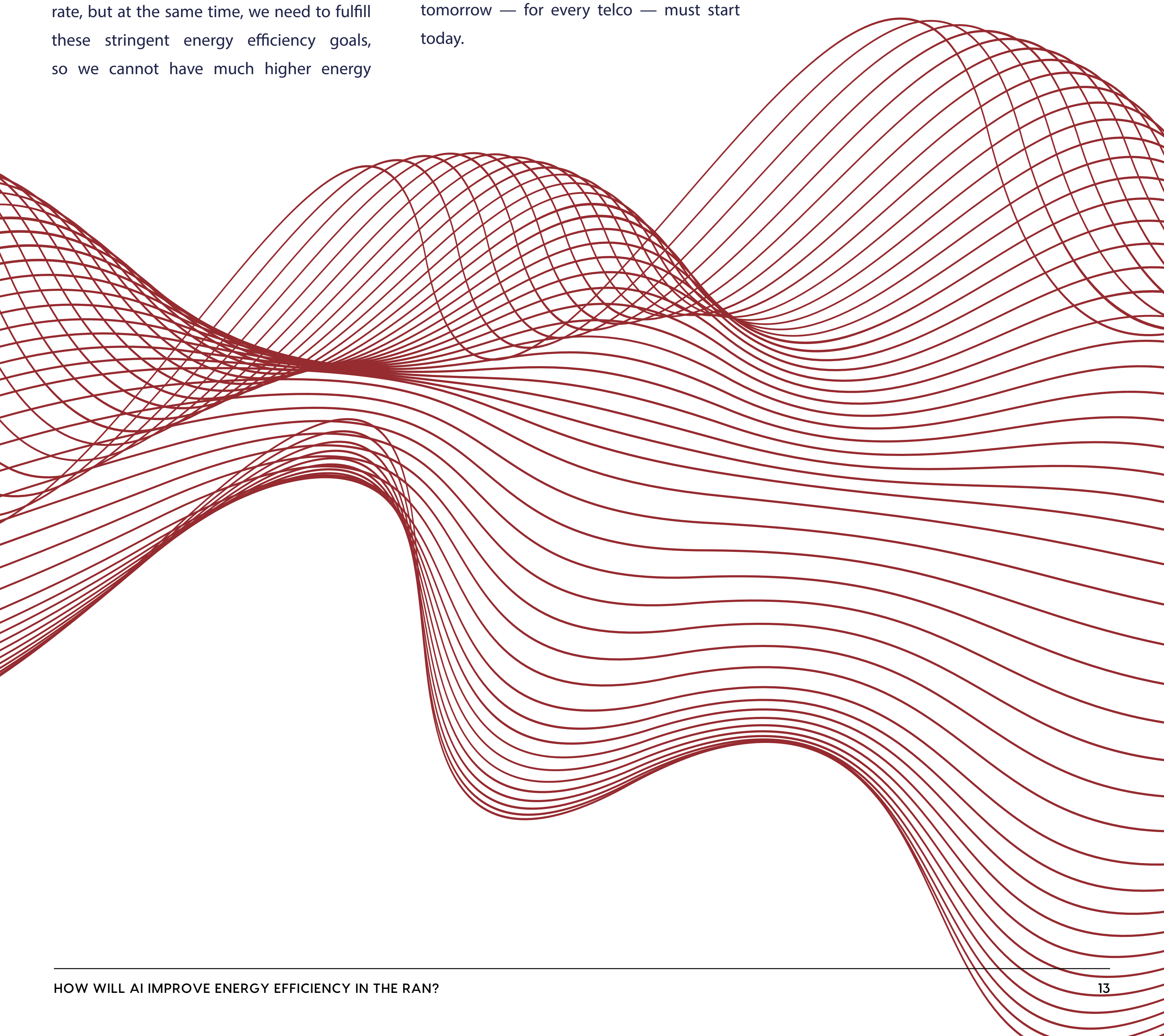
AJI ED,
VP and Head of Cloud RAN,
Nokia

"Not all of the networks around the world are ready for centralization today… At the same, that's the journey that we need to take, but we need to be realistic."

CONCLUSION

The need for RAN modernization is only becoming more and more pressing as network complexities and demand for data are set to rise. Therefore, as 5G continues to evolve, first with 5G-Advanced and then into AI-native 6G, the AI-RAN concept must be "applicable in the current, existing topology," even if the big dream is a fully autonomous, modernized and centralized RAN. "6G will need to deliver much more data at a faster rate, but at the same time, we need to fulfill these stringent energy efficiency goals, so we cannot have much higher energy consumption per traffic in 6G," cautioned Ed. "This AI-native air interface has the potential to maximize network efficiencies … so this will be a key component when we build sustainable 6G networks."

Telcos — and frankly, modern society — cannot afford to wait around for the stars to perfectly align; the journey towards the sustainable and efficient networks of tomorrow — for every telco — must start today.



Featured Companies

NOKIA

At Nokia, we create technology that helps the world act together. We put the world's people, machines and devices in sync to create a more sustainable, productive and accessible future.

Learn more