RCR Tech
NEWS. INSIGHT. IMPACT.

RETHINKING THE RAN

Building an intelligent, programmable radio access network

By Sean Kinney, Consulting Analyst, RCRTech

Sponsored by

NOKIA

CONTENTS

RCRTech Takeaways

Treat openness as architectural optionality, not a vendor-count KPI. Open RAN's durable value is the ability to introduce new radios, applications, compute and algorithms without rebuilding around one supplier. Operators should evaluate openness by whether it preserves programmability, interoperability and future technology choice, not by how many vendors appear in a deployment.

AI-RAN must prove itself against spectrum economics first. The clearest near-term business case is more capacity from existing spectrum through AI-enhanced radio algorithms. Accelerated compute is a means, not the objective. Operators should assess AI-RAN against measurable gains in spectral efficiency, power, price and performance, then expand compute only where those gains justify added complexity.

Use infrastructure refresh cycles to change the equipment and the operating model. TELUS shows that brownfield transformation is more viable when architectural change aligns with hardware replacement. Operators should use major refresh decisions to introduce virtualization, SMO, RICs, multivendor interoperability and automation together to create the programmable foundation required for increasingly AI-driven operations.

Delegate autonomy according to evidence, risk and intent. AI can automate fault detection, root-cause analysis and closed-loop network changes, but production autonomy should be graduated rather than absolute. Operators need guardrails around identity, access, conflict resolution and change management, while retaining human oversight where blast radius is high. AI should execute policy and intent, not determine them.

Build toward an intelligence fabric, with orchestration as the control plane. AI-native 6G will coordinate radio workloads, inference, distributed compute, telemetry, customer context and policy across centralized and edge infrastructure. The differentiator will be deciding dynamically what runs where and why. Operators should treat orchestration, data architecture and lifecycle management as foundational 6G investments today.

INTRODUCTION

The radio access network (RAN) is entering a period of architectural convergence. Open RAN, cloud-native infrastructure and artificial intelligence (AI) have largely developed as distinct industry movements, each with its own technical objectives and commercial narratives. Increasingly, however, they are becoming interdependent components of a broader transition toward a more programmable, adaptive and software-driven RAN.

Open RAN established much of the architectural foundation. By disaggregating hardware and software, opening interfaces and introducing new control and application layers, it expanded the ability of operators to mix suppliers, introduce third-party innovation and evolve the network through software. Cloud RAN extends that model by virtualizing and containerizing baseband functions, shifting more of the RAN toward common compute infrastructure and enabling functions to scale, update and automate with greater flexibility.

AI adds a new dimension. Machine learning (ML) is not new to mobile networks; self-organizing network functions and optimization algorithms have incorporated increasingly sophisticated intelligence for years. What is changing is the scope, location and potential impact of that intelligence. AI is moving deeper into RAN operations and, increasingly, into the radio stack itself, where new algorithms promise gains in spectral efficiency, capacity and customer experience. At the same time, operators are applying AI to network operations, root-cause analysis, intent-based optimization and closed-loop automation.

The convergence of these technologies is charting a course toward new types of intelligent outcomes. For operators, the questions are increasingly practical: can AI extract more value from finite spectrum assets? Can cloud-native architectures accelerate innovation without increasing operational complexity? Can openness preserve supplier optionality while supporting the tight integration required for high-performance AI? And how much authority can operators safely delegate to increasingly autonomous systems?

The answers will not be uniform. Brownfield constraints, network architectures, compute requirements and business priorities vary considerably by operator and deployment. But a common direction is emerging in that the RAN is evolving from relatively static communications infrastructure into a programmable platform capable of observing network conditions, interpreting intent and adapting resources around network, business and customer outcomes.

This report examines how that transition is taking shape and what the convergence of openness, cloud and AI means for the architecture and economics of the RAN.

FROM SON TO AI-RAN — WHEN CLOUD, DISAGGREGATION AND INTELLIGENCE CONVERGE

The emergence of AI-RAN can look like a break with the past as new accelerated computing architectures, new application models and a new ambition to distribute AI throughout mobile network infrastructure come to the fore. But viewed against the longer history of RAN automation, the transition is more evolutionary than abrupt. Self-organizing networks (SON), virtualization, disaggregation and cloud-native architectures have progressively created the control, software and compute foundations on which AI-RAN is now being built.

Stéphane Téral, founder and chief analyst at Téral Research, traces that evolution back to a longstanding objective for automated systems: "To create a system that minimizes human interaction to the lowest level possible." Within mobile networks, that ambition became increasingly formalized as SON evolved alongside LTE. 3GPP Release 8 divided SON into distributed, centralized and hybrid approaches, placing intelligence at different points in the network and operations stack. By 2015, according to Téral, ML algorithms were being blended into SON algorithms, extending optimization beyond purely rules-based automation.

That history puts the current AI-RAN cycle into context. Machine learning is not being introduced into a previously unintelligent RAN. Rather, AI represents the latest stage in a progression toward increasingly adaptive network control.

Open RAN expanded the architectural surface available for that intelligence. Much of the early Open RAN narrative centered on disaggregation, open interfaces, multivendor interoperability and shifting network economics through greater supplier choice. Progress against those ambitions has been uneven, but Téral rejects the characterization that Open RAN failed because large-scale multivendor deployments have taken longer to materialize than initially anticipated. "No, it did not [fail]. It gives you flexibility," he said.

The consequence is that openness does not require an operator to populate every layer of the RAN with equipment from a different supplier. It creates the option to introduce different suppliers and technologies where doing so delivers value. Téral points to interoperability between radio units and baseband infrastructure as one example, but the more significant long-term contribution may be programmability. Functions that previously existed as SON modules are evolving into applications associated with the RAN Intelligent Controller (RIC), including rApps and xApps, broadening the ecosystem that can develop intelligence for the network.

"Open RAN architectures rely on cloud-native principles to run disaggregated functions…That's where there is convergence."

— Stéphane Téral, founder and chief analyst at Téral Research

Open RAN, in that sense, is an architectural enabler because it opens interfaces, separates functions and creates new points where software can be introduced. Cloud RAN extends the same trajectory by changing how those functions are deployed and operated. Téral describes the progression sequentially. "You start by disaggregating software from hardware. Then you want to run functions on virtual machines. At some point, you containerize and then you move to the cloud."

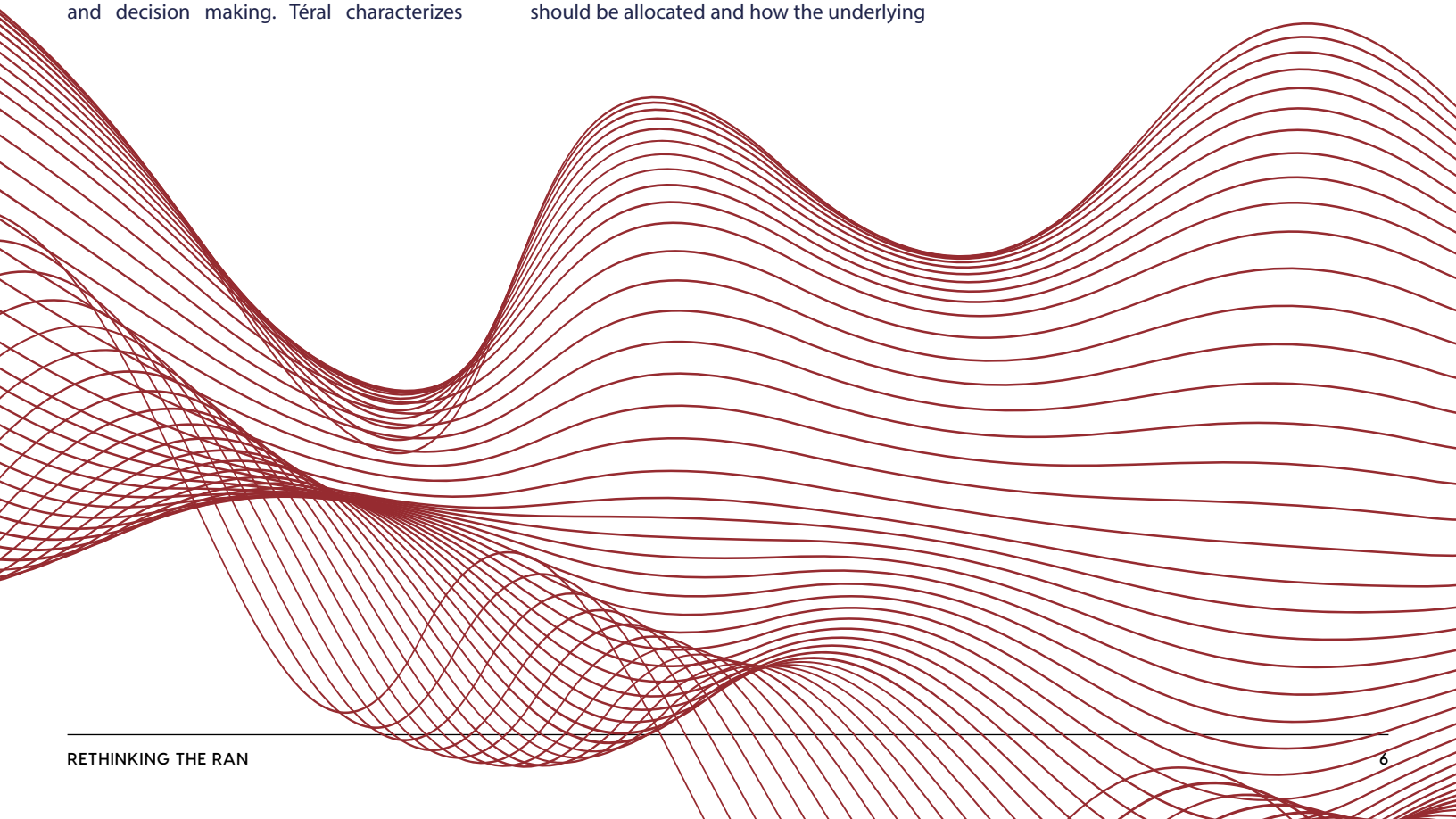
The relationship between Open RAN and Cloud RAN is complementary rather than equivalent. A RAN can be cloudified without being fully multivendor, just as an open architecture does not automatically imply a particular cloud implementation. But the two approaches increasingly reinforce one another. "Open RAN architectures rely on cloud-native principles to run disaggregated functions. So you know that's where there is convergence," Téral said.

AI builds on that convergence by adding increasingly sophisticated optimization and decision making. Téral characterizes the current phase primarily around AI used to improve the RAN, including spectral-efficiency gains delivered through software, ML and increasingly accelerated compute. From there, the architecture can progress toward AI and RAN workloads running on shared infrastructure, with a longer-term possibility of using distributed network infrastructure to support AI applications closer to users and devices.

That evolution is not synonymous with putting GPUs at every cell site. Téral emphasized that operators retain choices around the type and location of compute, just as openness created choices elsewhere in the RAN architecture. The appropriate architecture will depend on workload, geography, latency requirements and economics.

It will also depend on orchestration. Sharing distributed infrastructure between communications and AI workloads introduces a new resource-management problem of determining when and where workloads should execute, how capacity should be allocated and how the underlying complexity can remain manageable for operators. Téral pointed to ongoing industry research around orchestrating shared AI and RAN infrastructure and cautioned against treating that problem as solved. "The bottom line is this is a work in progress," he said. "People are looking at the orchestrator as the key component of managing everything and managing the complexity."

The arc from SON through Open RAN and Cloud RAN to AI-RAN is not a series of independent technology cycles. Each expands what can be controlled in software and how dynamically the network can respond. Open interfaces create programmability; cloud-native architectures create deployment and operational flexibility; AI increases the sophistication of the decisions that software can make. Their convergence creates the foundation for a RAN that is not simply more open or more virtualized, but increasingly capable of continuously adapting its resources around network conditions, operator objectives and customer needs.



AI-RAN IS POISED TO DELIVER MATERIAL SPECTRAL EFFICIENCY GAINS AS 5G EVOLVES

For all the ambition around AI-RAN — distributed inference, physical AI, shared compute infrastructure and ultimately AI-native 6G — the near-term business case begins with the much more familiar operator problem of extracting more capacity from costly, finite spectrum.

Nokia is putting that proposition at the center of its commercial AI-RAN strategy. In July 2026, the company introduced what it calls the industry's first commercial AI-RAN platform, built on Nokia's AI-native anyRAN software and NVIDIA's Aerial AI-RAN accelerated computing platform. Nokia said the platform has already demonstrated more than 20% spectral efficiency gains, with a roadmap targeting 50% gains in 2027 and more than 100% by 2028. Pilot deployments are planned for the end of 2026, followed by commercial availability in 2027.

That roadmap gets directly at the economics of spectrum. Operators have spent billions acquiring licenses, making incremental improvements in bits per second per hertz without requiring additional spectrum purchases or network densification. "We bring twice the network capacity, twice the spectral efficiency compared to what we have today, and we are not going to stop there," Aji Ed, Nokia vice president and head of AI-RAN and Cloud RAN, explained to RCRTech.

The path to those gains is not a single AI model applied uniformly across a network. Radio environments differ by spectrum band, duplexing scheme, antenna configuration, traffic profile and location. Ed pointed to AI-based multi-user MIMO pairing, channel estimation and link-adaptation algorithms as examples of capabilities that can be applied differently across TDD massive MIMO, FDD and uplink- or downlink-dominated environments. Gains can also vary considerably by network condition, with congested sites and cell-edge scenarios presenting particularly valuable optimization targets.

"So it's not going to be one size fits all," Ed said. Instead, different algorithms and features will address different deployment conditions. Nokia's architectural response is similarly heterogeneous. The new platform uses a common software-defined architecture across three deployment paths. Existing AirScale customers can add a GPU-powered capacity plug-in unit designed to introduce AI-accelerated processing while preserving installed infrastructure. A standalone GPU-powered AI-RAN node provides more compute headroom and can operate independently, in clusters or alongside AirScale. For cloud-native deployments, Nokia is also supporting GPU-powered COTS server configurations through ecosystem partners. All three are

"This is extremely important for our customers because... this is the scarcest resource they have today. Any additional benefit that you get on top, it's extremely valuable."

— Aji Ed, Vice President, Head of AI-RAN and Cloud RAN, Nokia

based on Nokia anyRAN software, and the platform supports 4G, 5G and future 6G evolution.

That deployment flexibility reflects a broader point about AI-RAN: accelerated computing is not itself the outcome. "It is not about we are bringing GPU. It's about bringing the right compute at the right place, and with the right configuration. That's the most important thing," Ed said.

This is a salient point because AI-RAN has to compete against highly optimized purpose-built RAN architectures on power consumption, cost and performance. In discussing the economics of accelerated compute, Ed reduced the test to three variables — "the power, the price, and the performance." The price and power profile must remain comparable, he argued, while performance must improve.

Nokia's bet on GPU acceleration is also part of a much larger strategic alignment with NVIDIA. In October 2025, the companies announced a partnership spanning AI-RAN and AI data center networking, alongside a $1 billion NVIDIA equity investment in Nokia. The transaction was completed in November 2025, and Nokia said the capital would help accelerate development of its 5G and 6G RAN software on NVIDIA architecture and advance its broader AI and cloud networking strategy. The companies are also collaborating around Nokia switching, telemetry and optical technologies for NVIDIA AI infrastructure.

The strategic logic extends beyond using GPUs to make the RAN perform better. Nokia and NVIDIA are working toward infrastructure capable of running radio and AI workloads on the same accelerated computing platform. T-Mobile US has already piloted NVIDIA AI-RAN infrastructure with Nokia anyRAN software in the United States, and in March 2026 the companies showed how cell sites and mobile switching offices could support distributed edge AI workloads while continuing to provide 5G connectivity. The work includes physical AI applications spanning computer vision, industrial safety and utility inspection.

This creates a potential progression in the economics of AI-RAN. Initially, accelerated computing has to earn its place by improving the performance of the radio network itself. Over time, the same infrastructure could provide additional utilization and monetization opportunities by hosting AI inference and other distributed workloads. The software model is central to that proposition. Nokia is introducing AI-RAN with a subscription model through which operators can continually activate new algorithms, optimization capabilities and performance improvements rather than tying gains exclusively to hardware replacement cycles. Nokia describes the shift as moving innovation from the cadence of specialized silicon toward the cadence of software.

Ed framed the value in cumulative terms, describing "software that creates the value that compounds over a period of time with the new algorithms, new machine learning models, new complex models."

Openness becomes more consequential as intelligence moves deeper into the radio stack. Nokia's platform is designed to be O-RAN compliant and introduces support for dApps, distributed applications operating much closer to the baseband and at sub-millisecond timescales. But opening deeper control surfaces also raises the stakes around stability and accountability. "It has to be tightly governed, tightly controlled," Ed said, emphasizing reliability, value creation and governance as prerequisites for introducing third-party functionality at this level.

The commercial test for AI-RAN is straightforward even if the architecture behind is not. Operators do not need accelerated compute simply because GPUs can run RAN workloads. They need it when software and AI can produce enough additional capacity, efficiency and functionality to change network economics. Nokia is positioning spectral efficiency as the first proof point, and a software-defined, accelerated RAN as the platform on which additional value can compound.

OPEN RAN IN THE BROWNFIELD — THE TELUS TRANSFORMATION TURNS ARCHITECTURE INTO AN OPERATING MODEL

For TELUS, Open RAN was not introduced into a clean-sheet network. It became part of a broader brownfield transformation in which hardware lifecycle, regulatory requirements, vendor strategy and a desire to fundamentally change the network operating model converged.

Canada announced in 2022 that operators would be prohibited from deploying new Huawei and ZTE 5G equipment and would have to remove existing 5G equipment and managed services from those suppliers by June 28, 2024; existing 4G equipment and services face a December 31, 2027 removal deadline. TELUS had already broadened its supplier base, selecting Ericsson and Nokia for 5G in June 2020 and subsequently naming Samsung as another 5G infrastructure partner. TELUS went beyond just swapping equipment; instead, it used the refresh cycle to change the RAN architecture itself.

"We were in our hardware refresh cycle, and we were anyway swapping our network," Sushil Rawat, director of RAN strategy at TELUS, said. "So it helps when you're riding an ongoing hardware refresh cycle and just change the architecture because then the technology itself does not cost more to do that truck rolls and make those changes in the network."

That distinction helps explain both the pace of TELUS' deployment and one of the broader realities of brownfield Open RAN economics. Operators rarely evaluate a new architecture independently of the residual value of infrastructure already in the field. Replacing equipment early can undermine otherwise attractive TCO calculations. Aligning architectural change with an existing refresh cycle changes the equation. "It's out of necessity. It's not just out of choice of a technology that you make," Rawat said.

TELUS formally expanded its work with Samsung from greenfield deployments into brownfield virtualized and Open RAN in 2024. Samsung supplies its vRAN software, Open RAN-compliant radios and SMO capabilities, while the architecture supports third-party radio integration. Wind River provides cloud infrastructure, and HPE supplies ProLiant servers using Intel Xeon processors with Intel vRAN Boost for the distributed unit infrastructure.

The resulting architecture is deliberately multivendor. Rawat described servers from standard COTS suppliers, a cloud layer independent of the DU software, Samsung and third-party radios, and an automation environment that combines Samsung components with other software developed or integrated by TELUS. The RIC follows the same principle in that Samsung provides the platform, while applications can come from Samsung or third parties.

TELUS is intentionally treating interoperability as something that has to be demonstrated operationally rather than inferred from standards compliance alone. "We have at least kept this principle where we will have openness in the architecture, and not only on paper, but also exercise having interoperability with third party," Rawat said.

The deployment has moved quickly. As of July 2026, Rawat said approximately 25% of the TELUS network was running on Open RAN, with plans to reach about 40% by the end of 2026 and 50% by the end of 2027. TELUS has publicly set a target of moving 100% of its network to Open RAN by the end of 2029.

But the larger significance of the transformation is that virtualization and

openness are changing how TELUS can automate and optimize the network. Rawat said the original TCO analysis included the SMO, RIC architecture, tool consolidation and simplification of the automation platform — much more than just the cost of radios, servers and software.

That operating model is now extending to AI. In 2025, TELUS and Samsung announced deployment of an AI-powered RIC designed to work across both legacy hardware-based infrastructure and the operator's open virtualized RAN. The platform supports Samsung and third-party applications for functions including anomaly detection, energy management and load balancing. Rawat sees current AI-for-RAN capabilities as a natural extension of this automation journey. Smaller agents can take on repeatable operational tasks first, while more complex and less deterministic activities require additional training, infrastructure and controls.

"I would call it an extension of automation, which is more self-defined," he said. "Now, with the help of AI, I think it is becoming more reusable and easier to configure an AI agent to develop certain use cases." For TELUS, however, the value of AI-RAN ultimately has to be measured in outcomes. One is operational efficiency; another is the ability to increase the performance of the radio stack itself.

"When it comes to my perspective [on] AI-RAN, [it] is how does AI increase the efficiency of the RAN stack that is deployed, right? How do we deliver more data? How do we deliver more data using the same spectrum asset that you have?" Rawat said.

That pragmatic framing also informs TELUS' view of accelerated compute. Rawat said he did not currently see a requirement to deploy GPUs at cell sites within the next 12 months, although TELUS is already using accelerated compute centrally for model training, anomaly detection and related workloads. His position is not dogmatic; he said that if a new scheduler or algorithm delivers performance unavailable with existing architecture, "There might be an opportunity to do that."

The harder problem is allowing AI to move from analysis into action. Rawat described production deployment as fundamentally different from demonstrating an AI use case in a lab. Once an agent can modify a live network, identity management, access controls, conflict management and established change-management processes become essential. "You can build a use case. You can demonstrate it in the lab, quick and easy," he said. "Taking it to production, scaling it for day-to-day operation, this is a very important aspect of it."

TELUS' brownfield transformation illustrates a broader evolution of Open RAN. The immediate task was replacing and modernizing network infrastructure. The architectural opportunity was to make that infrastructure virtualized, interoperable and programmable. The next phase is using those characteristics to introduce increasingly intelligent automation while preserving the operational controls required of a production carrier network.

"There might be use cases in the future where it will require more time-sensitive processing and needs to be closer to the application or endpoint...When it comes to the applications today, I think they require more centralized GPU resources."

— Sushil Rawat, Director of RAN Strategy, TELUS

T-MOBILE US PUTS AI TO WORK BY JOINING AUTOMATION WITH CUSTOMER OUTCOMES

For T-Mobile US, the transition toward an AI-native network begins with a change in what the network is being optimized to accomplish. Traditional KPIs around availability, capacity, alarms and network-element performance remain essential to operations, but they do not necessarily describe what a customer is experiencing. A network can appear healthy while an application is slow, a video stalls or a user cannot upload a photo.

T-Mo EVP and Chief Network Officer Ankur Kapoor explained the operator's approach to determining the value AI brings: "If it's not delivering a differentiated customer experience, it's only operational benefits." Instead of measuring success only by whether network infrastructure moves from red to green, T-Mobile US is incorporating application responsiveness, video performance, customer-care contacts, time to resolve customer pain points and Net Promoter Score into its view of network performance. The shift is from optimizing network elements toward optimizing outcomes experienced by individual users.

That philosophy is already influencing how much operational authority T-Mobile US delegates to software. Kapoor described the traditional network operations center as three layers: Tier 1 identifies problems, Tier 2 diagnoses them and determines an appropriate response, and Tier 3 executes changes on the production network. T-Mobile US has fully automated the Tier 1 function, while extensive automation performs root-cause analysis at Tier 2. At Tier 3, the level of human involvement depends on the potential consequences of an action.

Dynamic CX provides an example of where T-Mobile US has moved further into closed-loop operations. Introduced commercially in June 2026, the AI-powered capability automatically adapts network parameters as demand changes during major events. T-Mobile US subsequently reported that Dynamic CX executed more than 11,000 automated network optimizations across 12 major fan events during the summer soccer tournament.

"We make thousands and thousands of parameter changes, antenna changes, azimuth changes, and it's all closed loop, right?" Kapoor said. "Everything in between happens in a fully autonomous fashion because we are looking for not the network nodes and how the network nodes are performing. We're looking for are the customers getting that experience or not?"

The willingness to close the loop is not universal. Kapoor said T-Mobile US retains humans when potential changes carry greater risk, particularly in the core. Before

"Our vision was no automation is good enough unless it's really impacting customer experience."

— Ankur Kapoor, Executive Vice President and Chief Network Officer, T-Mobile US

human oversight is removed, the company tests whether automated decisions produce outcomes equivalent to or better than those a human operator would have produced. The result is less a binary transition from manual to autonomous operations than a graduated model in which software earns authority according to the use case, evidence and potential blast radius.

Intent-based networking adds another layer. Multiple objectives can compete for the same finite resources: customer experience, network efficiency, regulatory obligations, security requirements and public-safety services. T-Mo's work on conflict management in intent-based networks received a TM Forum Moonshot Attendees' Choice award in 2026 for using AI to identify and reconcile competing network intents. The key distinction, according to Kapoor, is between executing policy and determining it. "What AI does is actually executes on intent. It's not what governs. The business principles are the governance factor in that."

He used T-Priority, the company's service for first responders, to illustrate the point. Business and policy rules establish the priority; intelligent network systems determine dynamically how much capacity a particular activity requires and allocate resources accordingly. A first responder making a voice call requires a different resource profile than one operating a drone and transmitting imagery. Intent provides the objective and hierarchy, and AI continuously determines how network resources should satisfy them.

None of that is possible without a knowledge architecture capable of connecting network telemetry with application and customer context. T-Mobile US builds that knowledge base using measurements from devices and actual network usage rather than relying exclusively on synthetic testing. "This is real people in real places doing real things," Kapoor said. "We analyze that in real time. We look at what applications are being used, and we tune the network to deliver on what the customers are trying to do."

The same customer-outcome philosophy is increasingly extending beyond network optimization into network-native AI services. T-Mobile US launched its Live Translation beta in 2026, embedding real-time translation into the network rather than requiring a dedicated application or specific handset. The service was introduced as part of a broader network AI platform designed to support additional intelligent services over time.

AI-RAN extends that model from centralized platforms toward distributed infrastructure. T-Mobile US announced an AI-RAN Innovation Center with NVIDIA, Nokia and Ericsson in Bellevue, Washington, in 2024, with the objective of bringing accelerated computing, RAN workloads and AI applications onto common infrastructure. By March 2026, T-Mobile had become the first U.S. operator to pilot NVIDIA AI-RAN infrastructure with Nokia anyRAN software, demonstrating how cell sites and mobile switching offices could simultaneously support 5G connectivity and distributed AI workloads. Physical AI trials span vision-based city operations, utility inspection, industrial safety and other applications that require intelligence closer to devices and physical environments.

Kapoor connected these efforts directly. Dynamic CX and SON systems provide centralized intelligence and automation. The core becomes a platform for services such as Live Translation. AI-RAN creates a way to distribute more compute and intelligence toward cell sites. "What we're trying to do with AI-Ran is we are trying to bring that intelligence into the cell sites," he said. "Cell sites…all the way from 1G to 5G, they've just moved bits and bytes, right? They have not moved tokens. They have not been intelligent, right?"

T-Mobile US is already testing radio and AI workloads on common cell-site infrastructure. The longer-term significance is not simply that AI can run closer to the user. Distributed intelligence can process local data, return what is relevant to platforms such as Dynamic CX, and support increasingly adaptive network decisions while also creating infrastructure for new AI applications. "I'm not going to say that every tower is going to become a robot, but every tower is going to become an actual intelligent fabric, intelligent fabric," Kapoor said. "So it's all very interlinked."

That interconnection is the larger point. T-Mobile US' agentic operations, intent-based networking, Dynamic CX, network-native AI services and AI-RAN work are being developed in individual domains today, but they point toward a common operating model. The network increasingly observes what users and applications are trying to accomplish, interprets those requirements against policy and available resources, and takes progressively more autonomous action to deliver the intended experience.

"Our vision isn't just to automate the network," Kapoor said. The objective is "really delivering on those customer experiences," enabling new services and eliminating customer pain points.

ON THE PATH TO 6G, THE RAN IS BECOMING AN INTELLIGENCE FABRIC

The convergence of openness, cloud and AI is changing what the radio access network is capable of becoming. Individually, none of these developments represents a clean break with the past. Open RAN extends decades of disaggregation and software-defined networking. Cloud RAN moves network functions onto increasingly flexible compute infrastructure. Machine learning builds on a long history of SON and closed-loop optimization. What is different now is that these trajectories are beginning to intersect inside a common architecture.

The current state remains pragmatic. Open interfaces provide operators with greater architectural optionality and new points of programmability, but brownfield economics continue to determine the pace of adoption. Cloud-native architectures increase software velocity, but they do not eliminate the need for purpose-built performance. AI is already producing value in automation, root-cause analysis, intent-based optimization and increasingly sophisticated radio algorithms, yet operators remain appropriately cautious about delegating control over deterministic production networks to probabilistic systems.

The near-term measure of progress will be outcomes. Nokia is positioning spectral efficiency as the major near-term AI-RAN business case as delivered by applying new algorithms and accelerated compute to produce more capacity from existing spectrum. TELUS demonstrates how an open, virtualized brownfield RAN can become a platform for continuously expanding automation. T-Mobile US is moving the optimization target further outward from network-element health toward application behavior, customer intent and differentiated experience. Together, these approaches suggest that AI-RAN will not be defined by a particular accelerator, interface or deployment topology. It will be defined by the ability to continuously match network resources to network, business and customer objectives.

This also points toward the architectural challenge ahead. As AI moves deeper into the radio stack and AI workloads begin sharing infrastructure with RAN workloads, orchestration becomes increasingly important. Operators will need to coordinate heterogeneous compute, determine where inference should execute, manage application and model lifecycles, and enforce governance across distributed infrastructure. The RAN will have to balance radio performance with AI workload demand while preserving reliability, security and predictable service behavior.

This is where the current 5G/5G-Advanced transition begins to intersect with the longer-term vision for AI-native 6G. Rather than adding intelligence as another software layer on top of the network, 6G is increasingly being conceived around intelligence embedded throughout the architecture in radio optimization, orchestration, service delivery and potentially the physical infrastructure itself.

The result should not be viewed as a network with AI but as an intelligence fabric — a distributed system in which connectivity, compute, data and policy can be coordinated dynamically. Cell sites that historically moved bits can also become locations for processing and inference; centralized platforms can interpret broader network and customer context; and software can continuously decide what intelligence belongs where. T-Mobile US' vision of towers becoming an "intelligent fabric" captures the direction of travel.

Getting there will be incremental. The pathway runs through measurable spectral efficiency gains, increasingly autonomous operations, validated multivendor interoperability, cloud-native lifecycle management and progressively more capable orchestration. But the (next) end state is becoming clearer: an AI-native 6G network that goes beyond connecting users and devices to continuously observing, reasoning and adapting around the outcomes the network is intended to deliver.

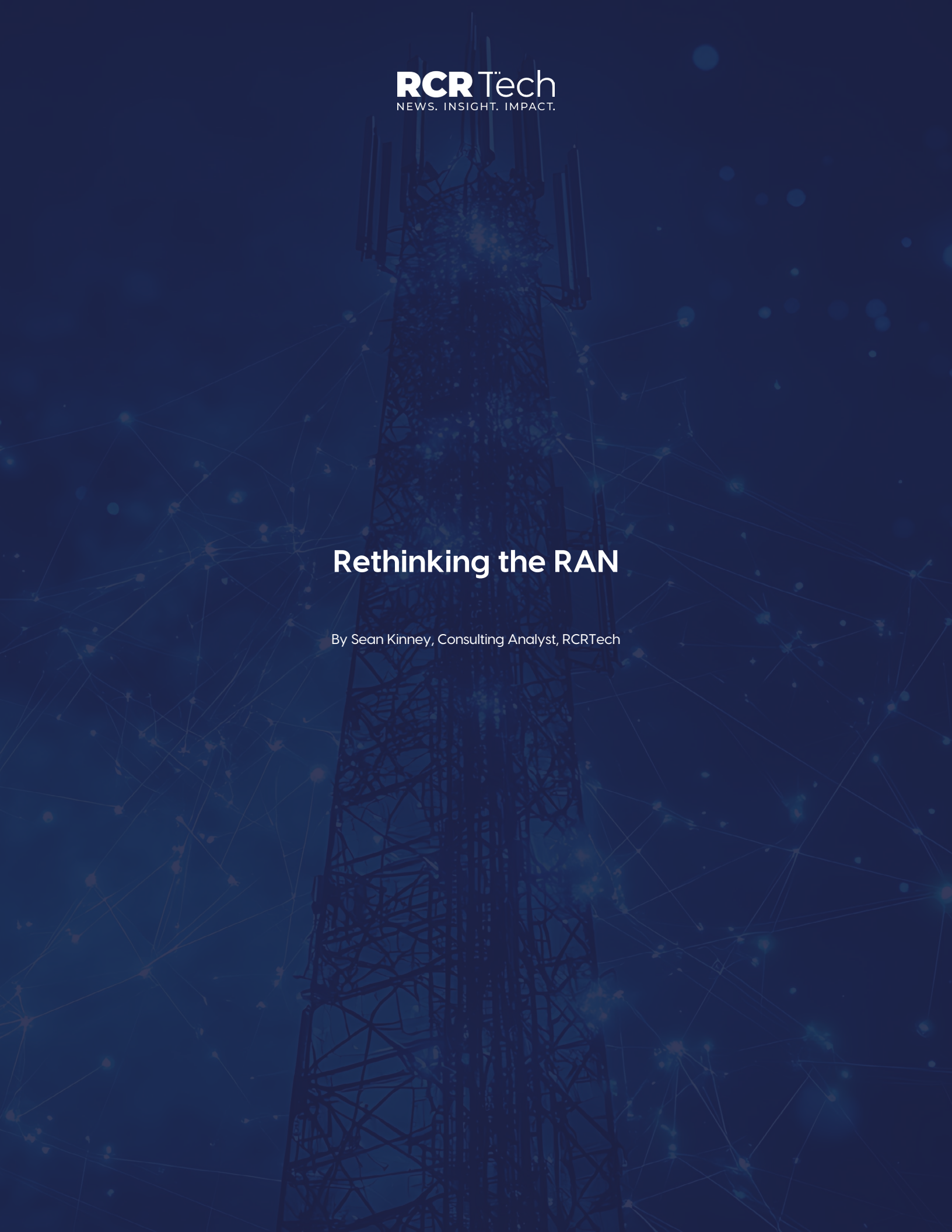
ACKNOWLEDGEMENTS

NOKIA

Nokia is a global leader in connectivity for the AI era, providing the critical network infrastructure the world relies on. Every day, we're powering our customers with advanced connectivity across fixed, mobile, and transport networks, delivering the performance and security they need to meet the demands of an AI-enabled future.

Learn more

RCR Tech
NEWS. INSIGHT. IMPACT.



Rethinking the RAN

By Sean Kinney, Consulting Analyst, RCRTech