Sponsored by SoftBank

AI in the Mobile RAN

A Transformational Opportunity for Telcos

RESEARCH BRIEF

Table of Contents

RESEARCH BRIEF

AI in the Mobile RAN

A Transformational Opportunity for Telcos

Executive Summary

Artificial intelligence (AI) and Generative AI (GenAI) are changing mobile radio access networks (RANs), driving a transformation that could redefine the telecom industry. By optimizing network performance, automating operations, and unlocking new revenue streams, AI is evolving into a fundamental requirement for telcos to thrive in the 5G-Advanced and 6G era.

This research brief outlines the evolution of AI in RAN, from early self-organizing networks (SONs) to the emergence of AI-native architectures. It explores the architectural shifts needed to support AI, highlights key AI use cases, and examines the market dynamics and challenges associated with AI adoption. While AI offers transformative potential, operators should establish clear KPIs and metrics to measure AI implementation success. These should tie directly to business outcomes like reduced OPEX, improved customer experience scores, and new revenue streams.

Key takeaways:

- AI is not only about operational efficiency; it's about creating a sustainable, scalable, and intelligent RAN capable of supporting next-generation services and edge computing.
- The transition to AI-native RAN requires a shift in mindset, embracing open architectures and investing in AI-optimized infrastructure.
- Telcos have another opportunity to leverage their edge infrastructure to deliver AI services, differentiating themselves from hyperscalers and generating new revenue.

Call to action: Telco leaders must act decisively, developing a AI-native RAN roadmap and investing in the necessary skills, partnerships, and infrastructure today to benefit from an AI-powered future of mobile networks.

Research Briefs are independent content created by analysts working for AvidThink LLC. These reports are made possible through the sponsorship of our commercial supporters. Sponsors do not have any editorial control over the report content, and the views represented herein are solely those of AvidThink LLC. For more information about report sponsorship, please reach out to us at research@avidthink.com.

About AvidThink

AvidThink is a research and analysis firm focused on providing cutting-edge insights into the latest in infrastructure technologies. Formerly SDxCentral's research group, AvidThink launched as an independent company in October 2018. AvidThink's coverage includes 5G infrastructure, enterprise networks, private wireless, edge computing, SD-WAN, SASE, SSE, ZTNA, AI/cloud infrastructure, and infrastructure security. Our clients range from Fortune 500 enterprises and hyperscalers to tier-1 communications service providers, fast-growing unicorns, and innovative startups. AvidThink's research has been quoted by Forbes, the Wall Street Journal, Light Reading, Fierce Networks, Mobile World Live, and other major publications. Visit AvidThink at avidthink.com.

The Evolution of AI in RAN: From SON to AI-Native Networks

The journey of AI and machine learning (ML) in RAN is marked by an evolution from basic automation to the ambitious vision of AI-native networks. AI and RAN isn't a new concept — early efforts focused on self-organizing networks (SONs), which leveraged rule-based algorithms and limited ML to automate tasks like neighbor configuration and load balancing. While SONs from incumbent vendors like Nokia and Ericsson delivered measurable gains in efficiency, they were limited in uptake by vendor-specific implementations and a lack of open interfaces.

With the advent of 5G and the anticipation of 6G, the industry is shifting towards AI-native networks. This involves integrating AI into every layer of the RAN, from the physical layer to the service management layer, and across domains — radio, transport, cloud — for end-to-end optimization. AI-native RANs are designed with AI at their core, enabling functions like radio resource management and interference mitigation to be driven by sophisticated AI algorithms.

This transition is characterized by the increasing generalization and scalability of AI/ML models. Unlike static SON algorithms, AI-native approaches leverage vast datasets and continuous learning to adapt to dynamic network conditions. Current 5G networks are witnessing the integration of AI models for improved channel estimation, beam-forming enhancements, and mmWave planning, showcasing the early benefits of AI-native features.

The latest frontier in AI-native networks is the incorporation of Generative AI (GenAI) and Foundation Models (FMs), including Large Language Models (LLMs), into RAN operations. Unlike traditional ML models that perform specific tasks, these foundation models can be applied to multiple use cases through fine-tuning or context augmentation and other prompting techniques, allowing for faster deployment of AI capabilities across different network functions. These advanced AI systems enable automated network configuration (AIOps), dynamic code generation for automation tasks, and intelligent analysis of network data.

Looking ahead, the vision encompasses autonomous AI agents capable of self-managing network resources, optimizing for multiple objectives, and proactively healing the network through predictive maintenance. Furthermore, AI-centric RANs, equipped with GPUs and AI accelerators, could support distributed computing resources for edge AI applications, transforming the RAN into an edge computing fabric.

The transition to AI-native architectures not only promises to reduce operational costs but also opens up new revenue streams through edge AI services. Early adopters of AI in RAN are already reaping the benefits of operational efficiency, with service innovation expected to follow.

Early adopters of AI in RAN are already reaping the benefits of operational efficiency, with service innovation expected to follow.

Rethinking Telecom Network Architecture for AI-Native RAN

The realization of AI-driven RANs necessitates a significant reimagining of traditional network architectures. Legacy RANs, characterized by their closed, hardware-centric nature, present a barrier to the seamless integration of AI capabilities. To overcome this, the industry is embracing RAN disaggregation and openness, epitomized by the Open RAN (O-RAN) initiative. For operators still heavily invested in legacy equipment, a hybrid approach with well-defined interfaces between AI-enabled components and traditional infrastructure offers a pragmatic migration path that avoids disruptive 'rip and replace' scenarios.

Challenges of Today's RAN Architecture:

- **Rigidity:** Traditional RAN architectures are rigid and node-based, lacking the scalability and flexibility required for dynamic AI-driven optimization. They were not designed to accommodate the rapid evolution and diverse requirements of AI algorithms. This inflexibility limits the ability to inject new AI capabilities and adapt to changing network conditions.
- **Closed and Proprietary:** The closed nature of legacy systems, often reliant on proprietary hardware and software, hinders the integration of AI solutions from diverse vendors. This lack of openness can stifle innovation and lead to vendor lock-in, limiting operators' choices and potentially hindering the adoption of best-of-breed AI solutions.

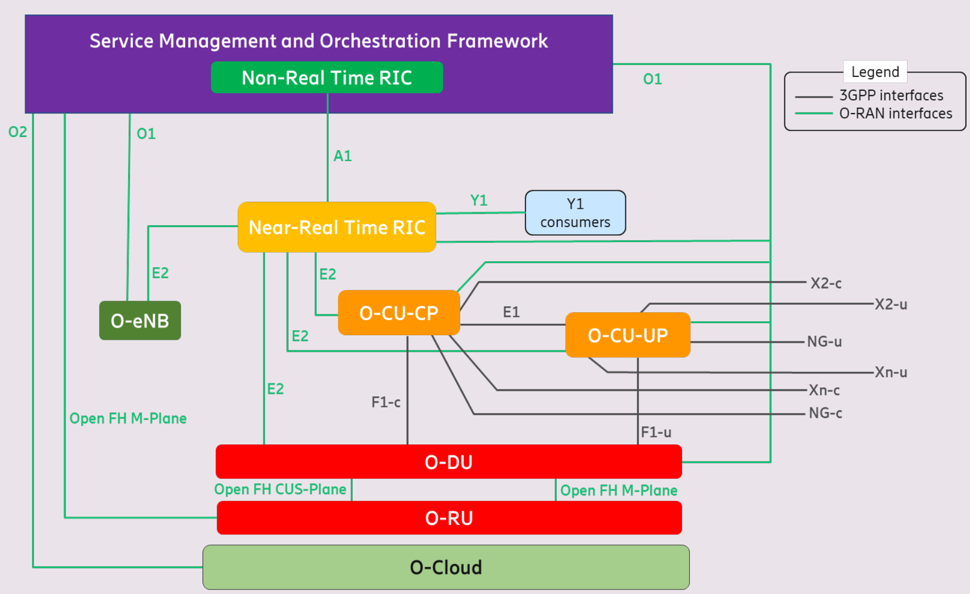
Adopting a Flexible and Open Architecture:

- **Lessons from 5G Core:** The transition of the 5G core network to a service-based architecture (SBA) provides valuable insights for RAN transformation. The SBA's modularity and use of open interfaces have enabled greater flexibility and innovation in the core network, paving the way for a similar transformation in the RAN.
- **Overcoming Vendor Resistance:** Embracing open architectures may require overcoming resistance from traditional vendors accustomed to closed, proprietary systems. These vendors may be hesitant to relinquish control and open up their systems to third-party solutions. However, the potential benefits of increased innovation and competition are likely to outweigh the challenges, and many of today's RAN incumbents are signaling a willingness to open up their architectures.
- **Feasibility of Cloud-Native RAN:** While progress towards a fully cloud-native RAN is underway, challenges remain in ensuring real-time performance and reliability. The stringent latency requirements of some RAN functions may necessitate careful consideration of hardware and software optimization in a cloud-native environment.

Cloudification & AI-driven Network Functions:

- **IT Cloud as a Roadmap:** The IT industry's journey towards cloud adoption offers a roadmap for RAN cloudification. The lessons learned in terms of scalability, agility, and cost-efficiency can be applied to the RAN, enabling operators to leverage cloud technologies for greater flexibility and efficiency.
- **Decoupling RAN Components:** Decoupling RAN components (software, hardware, and operations) allows for greater flexibility and the integration of AI-optimized solutions. This disaggregation enables operators to choose best-in-class components from different vendors and integrate AI solutions seamlessly into the RAN.
- **Shifting from ASICs:** Moving from proprietary ASICs to AI-optimized general-purpose hardware, potentially with specialized accelerators, enables greater efficiency and adaptability. This shift allows for the utilization of hardware specifically designed for AI workloads, potentially improving performance and reducing power consumption.

O-RAN LOGICAL ARCHITECTURE



Source: O-RAN Alliance

avidthink.com

The Role of Open RAN (O-RAN)

The O-RAN architecture, as shepherded by the O-RAN Alliance, with its open interfaces and focus on disaggregation, provides a platform for AI-driven modules to be deployed flexibly. The cornerstone of the new open RAN architecture is the RAN Intelligent Controller (RIC) — a platform introduced to host third-party control applications (xApps and rApps) and intelligence. The RIC architecture represents a break from previous architectures — separating intelligence from execution allows for a third-party "app store" model for network optimization algorithms. This could encourage innovation in a traditionally closed ecosystem, though AvidThink recommends learning from the SDN experience in the mid 2010s which didn't achieve its goals of an SDN app marketplace enabled by SDN Controllers. The RIC is split into two logical components to handle different timescales: a Non-Real-Time RIC (for policy, analytics and optimization tasks >1s) and Near-Real-Time RIC (control loops of 10 ms to 1s).

Together, these RIC platforms bring artificial intelligence to the wireless access network by leveraging advanced analytics and machine learning to make data-driven decisions. The RIC facilitates sophisticated optimization algorithms — often powered by AI/ML — to observe network data and instruct RAN nodes via standardized interfaces. For example, a Non-RT RIC rApp might analyze long-term cell performance trends and send optimization policies, while a Near-RT RIC xApp might dynamically adjust handover parameters or scheduling in almost real-time based on AI predictions.

Crucially, the RIC and O-RAN architecture are designed to eliminate vendor lock-in using open interfaces (like A1, E2, O1) and data models so that operators can mix and match AI applications from different developers.

This open ecosystem fosters innovation, allowing startups, academia, and operators to contribute AI algorithms, mitigating vendor lock-in and accelerating the pace of AI adoption. To support AI at scale, the RAN architecture needs to incorporate MLOps capabilities, encompassing data collection, model training, and inference.

Leading operators are even rethinking network topology to accommodate AI. SoftBank's proposed "AI-and-RAN" unified infrastructure envisions common hardware clusters handling both RAN baseband processing and AI workloads in tandem, optimizing resource usage and potentially reducing costs. This convergence of connectivity and computing infrastructure could lead to significant efficiencies and unlock new possibilities for edge AI services.

Key AI Use Cases in RAN

AI is being applied across a wide range of RAN functions, from operations and management to optimization and control, and even new service creation. This section explores some of the most impactful AI use cases in the RAN.

AI for RAN Operations (AIOps)

Operating modern mobile networks, with their tens of thousands of cells, diverse spectrum bands, and dynamic traffic patterns, has become incredibly complex. AI for IT operations (AIOps) in telecom leverages AI/ML to automate network management, fault handling, and performance optimization, with the ultimate goal of achieving a "zero-touch" network. In such a network, routine operations and even complex troubleshooting would be handled automatically with minimal human intervention. Operators like Vodafone have explicitly set "Zero Touch" targets[1] — aiming to prevent the majority of network faults before they impact customers by using predictive AI models and automation.

Telecom providers and standards bodies like the TM Forum have defined levels of network autonomy (L1 through L5, akin to self-driving car autonomy levels). Today, the majority of operators are around Level 2—3 (partial or conditional autonomy) with automated processes but human oversight on changes. The goal in the next few years is to reach Level 4 ("high" autonomous network) for many operations. For example, China Mobile is on track for L4 automation in 2025, and Huawei anticipates that by 2025, a majority of their operator customers will achieve Level 3 autonomy, with Level 4 becoming more prevalent[2]. Greenfield operators provide a proof-point: Rakuten Mobile in Japan built a cloud-native network and runs over 300,000 cells with ~250 operational staff, an efficiency that requires extreme automation and AI-driven processes[3].

[1] N. Doran, "Vodafone Global NOC goes cloud-native, readying for GenAI," TelcoTitans.com, Jul. 31, 2024. https://www.telcotitans.com/vodafonewatch/vodafone-global-noc-goes-cloud-native-readying-for-genai/8171.article

[2] "What Do Level 4 Autonomous Driving Networks Mean for Operators?," Huawei BLOG, Mar. 22, 2023. https://blog.huawei.com/2023/03/22/what-do-level-4-autonomous-driving-networks-mean-for-operators

[3] "AI | Automation | Telecom," Rakuten.com, 2023. https://symphony.rakuten.com/blog/how-we-did-it-rakuten-mobile-taps-power-of-industrialization-for-lean-scalable-network

Key capabilities in AI-driven RAN operations include:

- **Automated Fault Management:** AI systems can predict and detect faults or outages, triggering self-healing actions like rebooting a failing base station or rerouting traffic. This proactive approach aims to resolve issues before users are impacted. Vodafone, for example, set an original goal to use AI to prevent 50% of network faults before they impact customers.
- **Performance Optimization:** AI can continuously tune network parameters such as antenna tilt, power levels, and channel allocations to adapt to changing conditions like busy hour loads or special events. This includes load balancing across cells and interference management without the need for manual drive testing.
- **Predictive Maintenance:** Machine learning models analyze sensor data from RAN hardware, such as temperatures and error rates, to predict equipment failures in advance. This allows for just-in-time maintenance scheduling, reducing unplanned downtime and optimizing maintenance costs.
- **Closed-Loop Automation:** Combining monitoring, analysis, decision-making, and action, AI systems form a closed loop that observes the network, decides on optimizations, executes them, and then repeats the cycle. This is fundamental to zero-touch networks and is facilitated by the RIC and orchestration software.

GenAI can be applied to automate tasks like analyzing network logs, generating configuration scripts, or even interfacing with engineers in natural language. While still nascent, the potential of GenAI to further automate and simplify network management is significant.

The business case for AIOps in RAN is compelling. Networking vendors like Ericsson have warned that without increased automation, operators' operational expenses (OPEX) could double over the next five years just to manage the growing network complexity[4]. AI helps tame this complexity, simplifying operations and preventing OPEX growth. The benefits are both operational and financial: fewer outages, leading to higher customer satisfaction; faster rollout of new network features; and significantly reduced labor costs per cell site managed. As networks evolve to 5G-Advanced and 6G, with ultra-dense small cells and network slicing, AIOps will shift from a competitive advantage to a necessity.

Generative AI (GenAI) in AIOps:

An emerging trend is the use of GenAI, including large language models (LLMs), in network operations. GenAI can be applied to automate tasks like analyzing network logs, generating configuration scripts, or even interfacing with engineers in natural language. While still nascent, the potential of GenAI to further automate and simplify network management is significant.

AI for RAN Processing and Control

Beyond operations, AI is being integrated into the RAN's control and processing functions, enabling more intelligent and adaptive network behavior.

- **AI-driven O-RAN RIC:** The O-RAN RIC, with its xApps and rApps, provides a platform for deploying AI-powered control loops. These applications can leverage AI to optimize RAN functions in real-time or near real-time, adapting to dynamic conditions and improving network performance. For example, an AI-powered xApp could dynamically adjust handover parameters or scheduling based on real-time conditions.
- **Intelligent Traffic Management & QoS Optimization:** AI can power real-time traffic steering, optimizing network resource allocation based on user demand, application requirements, and network conditions. This ensures efficient use of resources and maintains quality of service (QoS). AI can also be used for dynamic resource allocation in 5G, where features like Massive MIMO and scheduling offer many degrees of freedom.

[4] S. Dux, "RAN automation becomes an operator opex imperative - Mobile Europe," Mobile Europe, Sep. 09, 2024. https://www.mobileeurope.co.uk/ran-automation-becomes-an-operator-opex-imperative/

- **AI for Energy Efficiency and Sustainability:** AI plays a crucial role in reducing RAN energy consumption, aligning with operators' carbon reduction goals. AI algorithms can intelligently power down network elements during periods of low activity, optimizing energy usage without compromising performance. Ericsson states that the RAN accounts for over 80% of mobile operators' energy use, making AI-driven energy efficiency crucial. Trials have shown significant gains: intelligent RAN energy-saving features have cut energy use by 15—25% in tests[5]. However, while AI-driven energy efficiency in RAN is impressive, these gains could be offset by the additional power requirements of AI accelerators themselves - careful system-level optimization is needed.
- **AI for Physical Layer Signal Processing:** AI is even being applied to the physical layer, with use cases like AI-driven channel estimation for improved uplink performance. This involves using AI to enhance the accuracy and efficiency of channel estimation, which is a critical process in wireless communication. This demonstrates the potential for AI to enhance even the most fundamental aspects of RAN functionality.

Real-world case studies illustrate the gains from AI-driven RAN control :

- **China Mobile & ZTE** (User Experience): In a joint solution, ZTE and China Mobile used AI algorithms to improve user experience over the RAN. They demonstrated a more than 15% improvement in the performance of live streaming, HD video and VIP user guarantees, leading to a better experience and driving upsell opportunities[6].
- **SoftBank** (Wireless Performance Enhancement): SoftBank's verification testing of three AI-for-RAN use cases[7] demonstrated significant performance improvements including 20% higher uplink throughput in poor network conditions through AI-enhanced channel interpolation, 13% improved downlink throughput for high-speed (80 km/h) devices using AI-driven SRS prediction, and 8% average throughput gains with AI-powered MAC scheduling for optimized MU-MIMO pairing—all achieved without deploying additional base stations, potentially reducing infrastructure investment needs.
- **Ericsson** (Intelligent RAN Automation): Ericsson's portfolio of Intelligent RAN Automation solutions reportedly can yield around 15% spectral efficiency gain on average[8]. This translates to more capacity from existing spectrum — a direct boost to network performance.
- **Nokia** (AI SON in live network): As mentioned, a Nokia SON deployment in a live network helped the operator detect and resolve RAN anomalies much faster than before, preventing minor issues from escalating. Faster troubleshooting improves overall network quality metrics (e.g., call drop rate, throughput). At Vodafone Idea, Nokia's MantaRay SON applies up to 700,000 autonomous changes a day to optimize the RAN for resiliency and performance[9].

Furthermore, we are seeing ecosystem collaboration here: at MWC Barcelona 2024, the AI-RAN Alliance was launched with founding members including Arm, Nokia, Ericsson, Samsung, NVIDIA, Microsoft, as well as operators like SoftBank and T-Mobile joining forces. This alliance's goal is to accelerate integration of AI into RAN and promote use cases where the RAN not only is aided by AI, but also acts as a platform for AI inferencing. The presence of cloud players and chipset firms alongside traditional RAN vendors underscores the cross-industry importance of AI-driven RAN optimization. The formation of the AI-RAN Alliance with hyperscalers, chip companies and operators suggest that even traditional competitors recognize the need for ecosystem collaboration - no single entity can drive this transformation alone.

---

[5] E. Fersman, J. Pettersson, A. Höglund, E. Sanders, and Lackis Eleftheriadis, "Intelligent sustainability: the role of AI in energy consumption, management and new revenues," ericsson.com, Jun. 20, 2023. https://www.ericsson.com/en/blog/2023/6/intelligent-sustainability-ai-energy-consumption

[6] "Boosting network monetisation with RAN-native AI," GSMA Foundry, Nov. 14, 2024. https://www.gsma.com/get-involved/gsma-foundry/gsma_resources/ran-native-ai-network-monetiztion-china-mobile-zte/

[7] "SoftBank Corp. Demonstrates RAN Performance Enhancement with AI Technology" SoftBank, 2025. https://www.softbank.jp/en/corp/news/press/sbkk/2025/20250303_06/

[8] K. Hill, "Ericsson Q&A: Navigating network data challenges and AI/ML," RCR Wireless News, Sep. 22, 2023. https://www.rcrwireless.com/20230922/featured/ericsson-qa-navigating-network-data-challenges-and-ai-ml

[9] "Vodafone Idea Adopts Nokia's AI-powered MantaRay SON Solution to drive automation and improve customer experience," Itvoice.in, Dec. 12, 2024. https://www.itvoice.in/vodafone-idea-adopts-nokias-ai-powered-mantaray-son-solution-to-drive-automation-and-improve-customer-experience

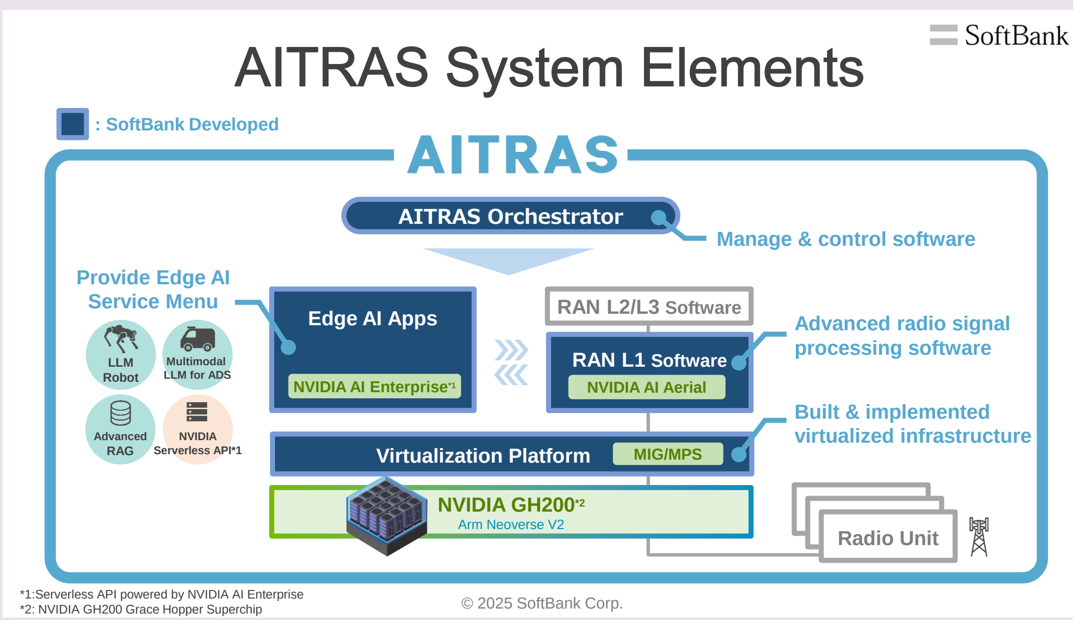
AI in Edge RAN: A New Monetization Frontier

A transformative concept gaining traction is leveraging the RAN edge for AI services, turning cellular sites and edge data centers into points of presence for AI computing. This allows telecom operators to offer AI inference services at the edge, unlocking new revenue streams beyond connectivity. Unsurprisingly, NVIDIA is championing this concept, hoping to drive more GPU sales into an early but promising use case.

In 2021 NVIDIA published an "AI-on-5G" white paper arguing that virtualized RAN (vRAN) deployments could turn cell sites into AI data centers. By colocating GPU-based servers at 5G base station sites (or centralized RAN hubs), operators can simultaneously run the RAN software and serve AI inference workloads for customers. In late 2022, NVIDIA reinforced this vision in a developer blog post[10], stating that AI would have significant impact on the RAN and highlighting four ways: energy savings, mobility optimization, load balancing, and enabling cloud RAN. Notably, that last point — Cloud RAN — ties directly into making the RAN a general compute platform, which could then host other applications.

This idea is not just theory. NTT DOCOMO deployed the world's first GPU-accelerated 5G Open RAN in its commercial network (with partners NVIDIA, Fujitsu, and Wind River), using GPUs to handle baseband processing[11]. This validated that general-purpose accelerators can run RAN workloads and left the door open to use those same GPUs for AI tasks. SoftBank is one of the early adopters explicitly pursuing this model: it announced experiments and plans involving distributed data centers that host both telco and generative AI functions at sites through their converged AI-RAN solution, AITRAS, which has the capability of dynamically allocating resources using its orchestrator function. SoftBank's research suggests that an "AI-heavy" mode of operation (where extra compute at sites is used for AI services) can generate revenue — their joint analysis with NVIDIA showed an example AI-RAN infrastructure yielding a 219% ROI over 5 years, with 5× return on CAPEX by monetizing AI workloads on the network[12]. In essence, SoftBank envisions RAN sites not just carrying voice and data, but also becoming AI hubs that customers pay to use, turning the RAN into a profit generator instead of a cost center.

SOFTBANK AITRAS



Source: SoftBank

[10] "How AI-Enabled Functionality Is Transforming 5G RAN | NVIDIA Technical Blog," NVIDIA Technical Blog, Dec. 05, 2022. https://developer.nvidia.com/blog/how-ai-enabled-functionality-is-transforming-5g-ran/

[11] S. McDowell, "NVIDIA Powers First GPU-Accelerated 5G Network," Forbes, Oct. 09, 2023. https://www.forbes.com/sites/stevemcdowell/2023/10/09/nvidia-powers-first-gpu-accelerated-5g-network/

[12] "Challenges in the Telecommunications Industry and AI-RAN's Return Expansion Mechanism," SoftBank, 2025. https://www.softbank.jp/en/corp/technology/research/story-event/073.

Telcos have several advantages in this edge AI arena:

- **Proximity and Low Latency:** Their distributed infrastructure provides proximity to end-users and data sources, crucial for real-time AI applications. This allows for much lower latency compared to sending data back to a centralized cloud.
- **Data Sovereignty and Trust:** Telcos are often seen as trusted local entities, addressing concerns about data privacy and security. Many countries and enterprises prefer to keep sensitive data within their borders, and telcos can provide that assurance.
- **Existing Customer Base:** Operators can leverage their existing relationships with enterprise clients to upsell AI services. For example, a telco could offer a manufacturing company not just a private 5G network but also an AI application hosted on the edge of that network for quality inspection via video analytics.
- **Partnerships with Hyperscalers:** Collaborations with hyperscalers can bring cloud AI services to the telco edge, although competition with hyperscalers also exists. These partnerships can provide access to advanced AI software and expertise.

Successful edge AI monetization requires not just technical capabilities but also clear service definitions, pricing models, and go-to-market strategies that align with enterprise needs. Operators should start with specific industry verticals where they have existing relationships rather than pursuing generic edge AI offerings.

While there is optimism around edge AI monetization, there is also skepticism due to the recent experiences with mobile edge computing (MEC). However, the current demand for AI computing is far greater than previous edge application demand, suggesting a more promising future for edge AI services, assuming current AI/GenAI trends hold.

Several operators have launched AI inferencing services from regional data centers to edge sites, demonstrating the viability of this concept. Examples include:

- **Singtel:** Partnered to launch an AI lab on 5G, offering AI services to enterprises in Singapore[13].
- **Swisscom:** Integrating edge AI into its offerings[14].
- **Telenor:** Experimenting with AI solutions for industry in its 5G networks[15].
- **Indosat Ooredoo:** Bringing AI apps into its Indonesian network[16].
- **SoftBank:** A front-runner in AI and RAN convergence, conducting trials with enterprises for edge AI over 5G[17].

Successful edge AI monetization requires not just technical capabilities but also clear service definitions, pricing models, and go-to-market strategies that align with enterprise needs.

By hosting AI at the RAN edge, telcos can tap into enterprise digital transformation budgets and even consumer services (like cloud gaming with AI, immersive media). It's a natural extension of the "telco to tech-co" aspiration: many operators are indeed aiming to become more like technology companies with AI at their core strategy, rather than just connectivity utilities. Large operators (Deutsche Telekom, Orange, etc.) have launched AI initiatives and even "AI factories" to develop and productize AI solutions across their business[18]. The coming years will reveal how much of this translates into tangible business, but the groundwork is being laid now.

[13] "Case Study: NVIDIA AI Enterprise for Singtel 5G Services," NVIDIA, 2025. https://www.nvidia.com/en-us/case-studies/5g-enterprise-services/

[14] "Swisscom cooperates with NVIDIA," Swisscom.ch, Jan. 16, 2024. https://www.swisscom.ch/en/about/news/2024/01/cooperation-with-nvidia.html

[15] N. G. Eriksson, "Telenor and Partners Unveil Cutting-Edge 5G Solutions Lab in Gothenburg," Telenor IoT, Mar. 03, 2025. https://iot.telenor.com/press-release/telenor-and-partners-unveil-cutting-edge-5g-solutions-lab-in-gothenburg/

[16] J. O'Halloran, "Indosat Ooredoo Hutchison claims AI-RAN first for Southeast Asia," ComputerWeekly.com, 2025.

[17] "SoftBank deploys AI-RAN solution for Edge applications at Nvidia's Santa Clara HQ," Datacenterdynamics.com, Mar. 19, 2025.

[18] S. Creasy, I. Ferrero, Tomás Lajous, V. Trigo, and B. Vieira, "How generative AI could revitalize profitability for telcos," McKinsey & Company, Feb. 21, 2024. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/how-generative-ai-could-revitalize-profitability-for-telcos

It will take ingenuity and significant effort on the part of telcos to succeed. They will need to invest in high-performance edge hardware, cultivate developer ecosystems, and ensure strong backend integration with core cloud and enterprise systems. They also need to differentiate their offerings from hyperscalers and other GPU-as-a-Service "neoclouds" and address concerns about reliability and QoS for both RAN and non-RAN workloads.

AI-Optimized RAN Computing — The Changing Role of GPUs

The increasing use of AI in RAN is driving a shift in the underlying computing infrastructure. Traditional base stations, reliant on specialized processors like DSPs and FPGAs, are being augmented or even replaced by more versatile and AI-optimized hardware. This section explores the evolving landscape of RAN computing, with a focus on the role of GPUs and other accelerators.

ASICs and GPU-based RAN Computing — a Discussion

- **GPUs in RAN:** GPUs, with their parallel processing capabilities, have emerged as a flexible platform for both running virtualized RAN (vRAN) software and AI workloads. They excel at the matrix operations commonly found in deep learning algorithms. NTT DOCOMO's deployment of a GPU-accelerated 5G Open RAN demonstrates the feasibility of using GPUs for baseband processing[19]. Furthermore, platforms like NVIDIA's Aerial SDK allow RAN functions and AI inference to be performed on the same GPU, offering a compelling proposition for operators looking to consolidate hardware. The transition from proprietary ASICs to general-purpose hardware with AI acceleration is similar to what happened with IT servers decades ago - this commoditization could dramatically reshape vendor dynamics and pricing models.
- **ASICs and FPGAs:** While GPUs offer flexibility, some vendors are developing custom AI accelerators in the form of ASICs. The key tradeoff is between the adaptability of GPUs (which can run many different AI models and be repurposed as needs change) versus the power efficiency and performance optimization of purpose-built ASICs (which excel at specific workloads but lack versatility). These ASICs can be highly power-efficient but may lack the versatility of GPUs and often run vendor-proprietary models. FPGAs, with their reconfigurable nature, also play a role in accelerating specific AI tasks or offloading parts of Layer 1 processing. The industry is actively evaluating the optimal mix of GPUs, FPGAs, and ASICs for AI-optimized RAN deployments.

It will take ingenuity and significant effort on the part of telcos to succeed. Telcos will need to invest in high-performance edge hardware, cultivate developer ecosystems, and ensure strong backend integration with core cloud and enterprise systems. Telcos also need to differentiate their offerings from hyperscalers and other GPU-as-a-Service "neoclouds" and address concerns about reliability and QoS for both RAN and non-RAN workloads.

- **Hybrid Models:** The future likely involves hybrid models, where AI workloads are integrated with traditional RAN processing. This could involve using GPUs for both vRAN and AI tasks or offloading specific AI functions to dedicated accelerators while maintaining core RAN processing on specialized hardware.

One notable device in this space is NVIDIA's ARA (AI RAN Accelerator) or ARC-1[20]. This is a packaged computer that includes all baseband hardware required to run a 5G base station with AI. By employing these accelerators, RAN can run complex AI models that were previously impossible in real-time. NVIDIA suggests that the same AI accelerators used in RAN can simultaneously handle generative AI functions (image processing, voice recognition, etc.) at the edge.

[19] "Enabling the World's First GPU-Accelerated 5G Open RAN for NTT DOCOMO with NVIDIA Aerial," NVIDIA Technical Blog, Sep. 26, 2023. https://developer.nvidia.com/blog/enabling-the-worlds-first-gpu-accelerated-5g-open-ran-for-ntt-docomo-with-nvidia-aerial/

[20] "Nvidia shows what open RAN could really look like with ARC-1," Lightreading.com, 2024. https://www.lightreading.com/open-ran/nvidia-shows-what-open-ran-could-really-look-like-with-arc-1

At the same time, adding GPUs at cell sites raises practical issues: (1) these sites were not originally designed as data centers. Upgrades may be needed for power supply and cooling. Some recent telecom equipment cabinets now include liquid cooling options for high-density compute, or novel designs to vent heat efficiently. However, telcos will need to replace outdated systems with edge-specific modern hardware and beef up power/cooling to host AI at the edge. (2) RAN workloads are critical and time-sensitive, any other hosted functions will need to be appropriately segregated and scheduled to not impact RAN performance. These operational challenges are important to consider as telcos plan AI-native RAN deployments.

Evolving Radio & Baseband Hardware for AI

- **RF SoCs embedding AI accelerators:** The integration of AI accelerators into RF SoCs is another trend to watch. This enables AI-driven optimization at the radio level, potentially leading to more intelligent and adaptive radio functions.
- **AI-driven spectrum optimization and adaptive radio functions:** AI can play a crucial role in optimizing spectrum usage and enabling adaptive radio functionalities. This includes dynamic spectrum sharing, where AI algorithms can intelligently allocate spectrum resources based on real-time demand and interference patterns.

Cloud vs. Distributed AI Compute

- **Centralized vs. decentralized AI models in RAN computing:** The deployment of AI models in the RAN can be centralized, with AI processing occurring in a central location or edge data center, or decentralized, with AI models deployed closer to the network edge, even at individual base stations. The choice depends on factors such as latency requirements, data privacy concerns, and the nature of the AI application.
- **The rise of federated learning and RAN opportunities:** Federated learning, where AI models are trained on decentralized datasets without sharing raw data, offers a promising approach for RAN. This allows telcos to train models across their network without centralizing sensitive data - this addresses both privacy and bandwidth constraints.

The integration of AI accelerators, whether GPUs or specialized ASICs, is crucial for enabling AI-native RAN. These accelerators empower the RAN to run complex AI models in real-time, unlocking new possibilities for optimization and control. The evolving landscape of RAN computing, with its mix of centralized and distributed AI processing, highlights the need for flexible and adaptable infrastructure to support the diverse requirements of AI-driven RAN.

In wrapping up this section, it's worth noting the competitive dynamic: traditional RAN silicon vendors (like Qualcomm, Ericsson's ASIC division, Nokia's ReefShark partnership, etc.) are now competing or collaborating with GPU vendors (NVIDIA, AMD) to provide the brains of future base stations. The presence of general-purpose compute in RAN also invites software innovation, potentially lowering barriers to entry in the RAN market.

The AI Revolution in RAN Operations & Management

AI is not just transforming the RAN itself but changing how networks are operated and managed. This section explores the impact of AI on RAN operations and management, paving the way for more efficient, automated, and intelligent networks.

Zero-touch operations represent not just cost savings but a fundamental cultural shift for telcos - the workforce implications (retraining, reorganization) may be more challenging than the technical implementation.

AI-powered Zero-Touch Network Operations

The vision of zero-touch network operations, where networks self-manage and self-optimize with minimal human intervention, could become a reality thanks to recent advancements in AI.

- **Real-time network adaptation using AI:** AI and ML models enables networks to adapt to changing conditions in real-time, dynamically adjusting parameters and resources to maintain optimal performance. This includes responding to traffic fluctuations, optimizing for different service levels, and even predicting and preventing potential issues.
- **Closed-loop automation for predictive maintenance:** AI-powered closed-loop automation systems can monitor network health, predict potential failures, and trigger automated responses, such as rerouting traffic or scheduling maintenance, to ensure continuous operation.

GenAI & Large Language Models in RAN Management

Generative AI and Large Language Models (LLMs) are poised to play an increasingly important role in RAN management, offering new capabilities for automation and intelligent decision-making.

- **AI for automated configuration & policy enforcement:** AI can automate network configuration tasks, ensuring consistent and accurate implementation of policies across the network. This reduces manual effort, minimizes errors, and enables faster deployment of new services and features.
- **Intelligent troubleshooting with instant AI-driven code generation:** LLMs with advanced reasoning can be used to analyze network logs, identify anomalies, and even generate code for automated troubleshooting and remediation. This accelerates the resolution of network issues and reduces downtime.

Operational Efficiency Gains & Business Impact

The adoption of AI in RAN operations and management leads to significant operational efficiency gains and positive business impact.

- **AI-driven cost reduction in RAN OpEx:** AI-powered automation reduces the need for manual intervention, leading to lower operational expenses. This includes reduced labor costs, optimized resource utilization, and proactive prevention of costly network outages.
- **AI's role in maximizing infrastructure utilization:** AI can optimize the utilization of network infrastructure, ensuring that resources are allocated efficiently and effectively. This leads to improved network capacity, better service quality, and a higher return on infrastructure investments.

The AI revolution in RAN operations and management is not just about doing things faster or cheaper; it's about changing how networks are run. By embracing AI, telcos can achieve much higher levels of efficiency, agility, and intelligence than previously, enabling them to meet the growing demands of a connected world and unlock new opportunities for innovation and growth.

Market Dynamics & Industry Challenges

The integration of AI in mobile RAN presents a market opportunity for the telecom industry, but it comes with its share of challenges. This section explores the market dynamics and industry challenges associated with AI adoption in RAN.

AI's Role in Bringing New Opportunities for the RAN

- **Driving efficiency amid stagnating ARPU:** AI offers a way for telcos to improve efficiency and reduce costs in the face of stagnating average revenue per user (ARPU). By optimizing network operations and resource utilization, AI can help telcos maintain profitability and competitiveness.
- **AI-driven business models to monetize RAN investments:** AI can enable new business models and revenue streams, such as offering AI inference services at the edge or providing AI-powered network optimization as a service to enterprise customers. This allows telcos to monetize their RAN investments beyond traditional connectivity services.

Challenges in AI Adoption for RAN

- **Vendor lock-in and legacy system inertia:** Existing vendor relationships and legacy systems can hinder the adoption of open, AI-native architectures. Overcoming vendor lock-in and integrating AI into legacy systems can be challenging.
- **Data privacy and security concerns:** The use of AI in RAN raises concerns about data privacy and security, especially given the sensitive nature of network data. Telcos need to address these concerns by implementing robust data governance frameworks and ensuring compliance with regulations.

- **Lack of telco AI expertise and workforce readiness:** The telecom industry faces a skills gap in AI and data science. Telcos need to invest in upskilling their workforce and attracting AI talent to successfully implement and manage AI-driven RAN solutions. Culturally, the shift to data-driven decision-making can be non-trivial in companies used to deterministic network engineering. There can be resistance or caution among staff about trusting AI systems. Change management and demonstrating AI's value through pilot successes is important to get buy-in from network operations teams.
- **Architectural innovations required:** The transition to AI-native RAN requires architectural innovations, including the adoption of open interfaces, cloud-native platforms, and distributed AI computing infrastructure. Likewise, advanced real-time resource scheduling is required to ensure that RAN workloads are not impacted by running other workloads on shared infrastructure. Strict segregation and segmentation of data capabilities will be needed to protect privacy of RAN subscriber data along with any enterprise AI workload data that's running on shared infrastructure.
- **Sophistication in orchestration and resource management:** Managing and orchestrating AI-driven RANs requires sophisticated tools and expertise. Telcos need to invest in orchestration platforms and develop the skills to manage complex AI-powered networks. Similarly, there is a lack of orchestration solutions that can package up excess GPU or AI accelerator capacity in a distributed RAN environment and make that available for use by non-RAN workloads — and manage resource availability in real-time.
- **Data collection and ecosystem challenges:** Collecting and managing the vast amounts of data required for AI training and inference can be challenging. Building a robust ecosystem of partners and vendors is also crucial for successful AI adoption. For long-term success in the ecosystem, startups will need access to a corpora of training data — which is hard to get from carriers due to privacy concerns. Organizations like the AI-RAN Alliance or the O-RAN Alliance can play a pivotal role in curating, normalizing, sanitizing, and hosting such data sets to encourage participation from a larger base of contributors.
- **Regulatory and security concerns:** Regulators are starting to look at the use of AI in critical infrastructure. There are potential regulatory risks if an AI system causes a network incident or is found to be discriminatory (even inadvertently). Security is also a concern: AI systems themselves could be targets for attack or could introduce vulnerabilities (e.g., feeding false data to an AI to manipulate its decisions — an adversarial attack). Conversely, AI can enhance security (by detecting fraud or attacks), but operators must ensure their AI operations are robust and secure.
- **Monetization and ROI Uncertainty**: While we have noted many potential revenue opportunities and cost savings, there's a question of ROI and timing. If not properly aligned with business goals, AI initiatives could fail to justify themselves. This is why many operators pick low-hanging fruit first (like energy saving, which directly cuts costs) or customer experience improvements (reducing churn) to get quick wins. From a market perspective, there is also the risk of hype vs reality — inflated expectations could lead to disappointments if results don't materialize quickly.

Examining the Landscape

- **Hyperscalers' AI ambitions in edge computing:** Hyperscalers, with their cloud expertise and AI capabilities, are increasingly targeting the edge computing market, posing a competitive challenge to telcos. That same expertise is also needed by the telcos to upskill their employees and to leverage to build their own AI assets (proprietary FMs, etc.). In addition, hyperscalers represent a potential partner for telcos who want to reach developers and enterprises to offer their RAN AI computing capacity.
- **AI-native startups disrupting traditional RAN players:** AI-native startups like Rimedo Labs and Aira Technologies are emerging with innovative solutions for RAN optimization and automation, potentially disrupting traditional RAN vendors — though they will likely partner with the incumbents than compete head-on.
- **Computing infrastructure players - CPU/GPUs, systems, software:** The computing infrastructure landscape is also evolving, with new players offering AI-optimized hardware and software solutions for RAN.

Despite these challenges, the potential benefits of AI in RAN are too significant to ignore. Telcos that proactively address these challenges and embrace AI-native strategies will be well-positioned to lead in the next generation of mobile networks.

Strategic Recommendations for Telcos

To capitalize on the transformative potential of AI in RAN, telcos need to adopt a proactive and strategic approach. This section provides key recommendations for telco leaders.

Develop an AI-Native RAN Roadmap

- **Define AI-driven transformation milestones:** Establish clear milestones and timelines for integrating AI into the RAN, starting with pilot projects and gradually scaling up to network-wide deployments.
- **Build a modular, cloud-native RAN architecture:** Embrace open architectures and cloud-native principles to enable flexibility, scalability, and the seamless integration of AI solutions.
- **Have a data modernization strategy in place:** AI feeds on data and it's important to establish a data curation and management strategy across the operator — data management/ETL processes, data lakes and warehouses, etc.

Invest in Researching GPU-Based RAN Computing/AI Accelerator Role:

- **Evaluate GPU-enabled AI-driven network performance:** Conduct thorough evaluations of GPU-based RAN solutions to assess their performance benefits and cost-effectiveness compared to traditional hardware.
- **Evaluate AI accelerators in RAN processing:** Explore the use of specialized AI accelerators, such as FPGAs and ASICs, to optimize specific RAN functions and AI workloads.

Explore Feasibility of Edge AI Inferencing for New Revenue Streams:

- **Monetize telco infrastructure for AI at the edge:** Explore using the telco's edge infrastructure to offer AI inference services to enterprise customers, creating new revenue streams.
- **Position AI services as a telco differentiator against hyperscalers:** Develop unique AI service offerings that leverage the telco's network assets and data, differentiating them from hyperscaler offerings.
- **Ensure underlying infrastructure can provide reliability and QoS for RAN while serving non-RAN workloads:** Invest in robust infrastructure that can support both RAN and AI workloads, ensuring reliable performance and QoS for all services.
- **Create ability to rapidly test hypotheses in the market and iterate:** Establish an agile approach to developing and deploying AI services, allowing for rapid testing and iteration based on market feedback.

Strengthen AI Ecosystem & Partnerships:

- **Collaborate with AI startups and academia:** Foster collaboration with AI startups and academic institutions to access cutting-edge AI research and talent.
- **Engage in industry consortia for AI-driven RAN evolution:** Actively participate in industry consortia, such as the O-RAN Alliance and the AI-RAN Alliance, to contribute to the development of AI-driven RAN standards and best practices.

Conclusion

The integration of AI into mobile RAN represents a key moment for operators. By embracing AI-native architectures and investing in AI-driven solutions, telcos can unlock new levels of efficiency, flexibility, and innovation. AI is not just about optimizing existing networks; it's about creating a foundation for the next generation of mobile services and experiences.

The transition to an AI-native RAN requires a shift in mindset, embracing open ecosystems, and fostering collaboration across the industry. Telcos must be proactive in developing their AI capabilities, investing in research and development, and building a skilled workforce.

The journey towards AI-native RAN is not without its challenges, but the potential rewards are immense. Telcos that fail to embrace AI risk being left behind, missing out on the opportunity to transform their networks and redefine their role in the digital era. Telecom leaders must act now to develop a comprehensive AI-driven RAN strategy, investing in the necessary infrastructure, skills, and partnerships to lead in the AI-powered future of mobile networks.

Appendix: AI-RAN Alliance Demos at MWC Barcelona 2025

To provide visibility into the state of AI-RAN progress in the ecosystem, we believe it may be helpful to review the demonstrations hosted by the AI-RAN Alliance at the recent Mobile World Congress in Barcelona. The AI-RAN Alliance, having grown to 75 members spanning 17 countries since its founding at MWC 2024, showcased a diverse range of cutting-edge demonstrations. These demonstrations, categorized across the Alliance's key focus areas — AI-for-RAN, AI-and-RAN, and AI-on-RAN — provide an indication of the industry's progress toward AI-native networks.

The demonstrations at MWC 2025 align with the report's findings regarding AI integration into wireless networks. Most notably, the AI-for-RAN category featured six demonstrations focused on performance enhancement, including initiatives from SoftBank and Fujitsu showing deep learning for channel estimation, Samsung's AI-based higher-order modulation techniques, and DeepSig's learned air interface with online learning. These demonstrations exemplify the capabilities of AI in enhancing RAN functionality, particularly in areas like spectral efficiency and coverage extension. The AI-and-RAN and AI-on-RAN categories showcased other important advances, with demonstrations like Keysight and Northeastern University's AI-RAN orchestration platform addressing the critical infrastructure sharing challenges, while examples of object detection and integrated sensing applications demonstrated the edge AI service opportunities that could form new monetization vectors for operators.

The multi-vendor and cross-industry participation observed at these demonstrations — including traditional vendors (Samsung, Fujitsu), silicon providers (NVIDIA, Arm), test equipment manufacturers (Keysight, VIAVI), and academic institutions — underscores our report's emphasis on ecosystem collaboration as essential for AI-RAN advancement. The Alliance's recent launch of "Data-for-AI" and "Test Methodology" initiatives further addresses the data management and benchmarking challenges identified in our strategic recommendations section. As the industry progresses toward AI-native networks, these demonstrations offer evidence that the potential benefits described throughout our report can become reality, even as challenges in architecture evolution, operational practices, and monetization strategies remain to be fully addressed.

List of AI-RAN Alliance Demonstrations at MWC Barcelona 2025

#	Category	Demonstration Title	Organizations
1	AI-for-RAN	Learned Air Interface with Online Learning	DeepSig, NVIDIA
2	AI-for-RAN	Realization of Deep Learning for UL Channel Estimation/Interpolation in Live RAN Testbed	Fujitsu, NVIDIA, SoftBank
3	AI-for-RAN	AI-based PUSCH Channel Estimation	Keysight, NVIDIA, Samsung
4	AI-for-RAN	AI/ML Optimized Higher-Order Modulations with a Neuromorphic Receiver	NVIDIA, Samsung, VIAVI
5	AI-for-RAN	AI-based 5G Beamforming for Mobility-Aware Interference Mitigation and Power Saving	Singapore University of Technology and Design (SUTD), VIAVI, Yonsei University
6	AI-for-RAN	AI-based Spectrum Sensing in the RAN	Northeastern University Open6G
7	AI-and-RAN	AI-RAN Orchestration	Keysight, Northeastern University Open6G
8	AI-on-RAN	AI-Driven Spectrum Sensing for Dynamic & Privacy-preserving AI Model Partitioning over 5G Network	Keysight, Singapore University of Technology and Design (SUTD), LITE-ON, NeuroRAN
9	AI-on-RAN	Integrated Sensing and Communications (ISAC)	Northeastern University, Tiami Networks
10	AI-on-RAN	AI-on-RAN Object Detection	Arm, Effnet AB, Phluido, Tannera

AvidThink, LLC
1900 Camden Ave
San Jose, California 95124 USA
avidthink.com

Rev A