AI-RAN Alliance
Vision and Mission White Paper

ai-ran.org

AI-RAN™ ALLIANCE

Contents

1. Introduction

The Radio Access Network (RAN) plays a pivotal role in our increasingly connected world. Over the course of five generations of technological innovation, RAN has vastly expanded the boundaries of connectivity, enabling the seamless communication experiences we rely on today. Despite these advances, RAN's full potential remains largely untapped. Now, with the power of Artificial Intelligence (AI), we stand on the brink of a new era for RAN—one that is more adaptive, intelligent, performant, efficient, and versatile.

Traditional RAN architectures provide a solid foundation, offering robust performance that meets the needs of modern telecommunication. However, AI presents an opportunity to elevate RAN further, enabling not only higher levels of efficiency and automation but also potential for new business opportunities. Integration of AI with RAN allows for RAN-related performance improvements, e.g., by intelligently optimizing the use of radio resources, dynamically adapting to shifting traffic and spectrum conditions, and improving the overall efficiency while delivering superior user experiences. It can also allow for RAN infrastructure to be used for AI application workloads not necessarily linked to RAN-specific operations. This potentially unlocks new business opportunities for the CSPs and the infrastructure providers as it diversifies the ways they can monetize the deployed networks and valuable spectrum investments.

The evolution of AI with RAN is closely tied to advances in compute architectures ranging from edge processing for low-latency AI inferences to cloud-based resources for AI model training, each playing a crucial role in the application of AI technologies with RAN. Cloudification and programmability further enable this transformation, allowing for higher levels of automation, more efficient operations and agility to create new services.

Although early explorations of AI-RAN concepts and applications are promising, collaborative effort across the ecosystem to validate them must accelerate. Thus, there is a clear need for further development of AI-RAN on real-world scenarios to identify where AI integration can be truly beneficial. Moreover, efforts to identify and address challenges related to the introduction of AI-RAN should be undertaken by the telecom industry to ensure the promised benefits are realizable.

1.1. Why we need AI-RAN Alliance

To address these objectives, the AI-RAN Alliance was formed, bringing together industry leaders, research institutions, and academia to collaboratively explore and develop solutions and accelerate the realization of an AI-native RAN. By pooling expertise and resources, the alliance builds an ecosystem to establish a roadmap for implementing AI-RAN at scale, focusing on practical use cases that maximize benefits. This partnership seeks to create an AI-native RAN that not only meets the demands of the present but also anticipates and adapts to the needs of the future, ensuring a more connected and efficient world for all.

1.2. Founding Members

The AI-RAN Alliance was founded by a coalition of industry and academic leaders, including network operators, equipment manufacturers, and research institutions. The eleven founding members of the AI-RAN Alliance are:

- Arm
- DeepSig
- Ericsson
- Microsoft
- Nokia
- Northeastern University
- NVIDIA
- Samsung
- SoftBank
- T-Mobile
- The University of Tokyo

This diverse group of founding members, together with other members, share a vision of an AI-native future for RAN, where networks are more intelligent, sustainable, and adaptable. By collaborating on research and innovation, the alliance aims to accelerate the development of AI based solutions that will shape next-generation connectivity, unlocking new business models and transforming industries through intelligent, AI-native RAN systems.

2. Vision and Mission

Our vision is to transform the telecommunications industry by integrating AI with RAN, driving innovation, increasing efficiency, and unlocking new economic opportunities for the telecom ecosystem – further growing the value of RAN by embracing AI across the stack.

The AI-RAN Alliance brings together industry leaders and academic institutions to advance mobile network performance and capabilities through state-of-the-art AI innovation. Our mission is to maximize RAN asset utilization, generate new revenue streams, and drive the telecom industry's seamless and profitable transition toward 6G.

3. Strategic Goals and Objectives

The objective of the AI-RAN Alliance is to explore various applications and coexistence of AI with RAN, enhancing the performance of RAN and unlocking new capabilities and business opportunities.

This section outlines the alliance's strategic development objectives, describes how we foster cross-industry partnerships, and explains how we address industry standards.

3.1. Development

The AI-RAN Alliance three key development areas are:

- **AI-for-RAN**: focusing on the integration of AI into RAN to significantly enhance RAN performance, such as improving spectral and operational efficiency, optimizing radio resource management, and enabling predictive maintenance.
- **AI-and-RAN**: seeks to create a shared infrastructure between RAN and AI workloads, enabling concurrent resource utilization on this converged computer-and-communications infrastructure. The goal is to increase the overall utilization of such platform infrastructure supporting the RAN workloads and open new monetization opportunities for running non-RAN AI applications on the same infrastructure.
- **AI-on-RAN**: aims to enable new RAN services to enhance AI applications running at the network edge, to be able to offer new consumer and enterprise services and applications from the edge of the network.

AI-RAN promises to offer the telecommunications industry several benefits, including:

- **Increased Efficiency:** AI solutions to optimize network operations, leading to higher spectral efficiency, reduced power consumption and new intelligent spectrum coordination schemes.
- **Cost Reduction:** AI technologies help lower operational costs by automating processes and improving resource management.
- **Enhanced Network Performance:** AI integration leads to more reliable and higher-quality network performance, improving user experience.
- **New Revenue Streams:** AI-RAN capabilities create opportunities for new services and applications, unlocking additional revenue sources.
- **Future-proofing Infrastructure:** Scaling up AI capabilities future-proves existing infrastructure for 6G and beyond, protecting investments and facilitating smooth transitions to next-generation technologies.
- **Knowledge Sharing Among Ecosystem Members:** AI-RAN Alliance fosters a collaborative environment where knowledge sharing among operators, vendors, academia and other stakeholders accelerates innovation and the development of standardized solutions.
- **Strategic Path Towards Next-Generation RAN:** AI-RAN Alliance provides a roadmap for evolving RAN architectures, ensuring a strategic transition to the next generation of networks through advanced capabilities such as intent-based networking and AI-native operations.

3.2. Collaboration

Integrating AI with RAN requires close collaboration among telecom vendors and operators, AI and cloud technology providers, hardware and software specialists, and research institutions. The AI-RAN Alliance is committed to building an effective ecosystem for its members. This includes establishing cross-industry working groups to facilitate collaboration and providing shared labs equipped with simulators and real-time systems, access to comprehensive data sets, and robust test, benchmarking tools, and AI-RAN proof of concepts. The Alliance will produce whitepapers, blueprints, and evaluation and benchmarking results to further foster industry-wide collaboration.

3.3. Standards

The AI-RAN Alliance is not a standard development organization (SDO) and as an independent body, the Alliance is not mandated to provide feedback to any SDOs. However, while formal standardization is outside its scope, the AI-RAN Alliance's outputs are designed to facilitate AI integration into RAN systems, supporting industry efforts where appropriate. Development of guidelines and reference materials, such as blueprints, can potentially be used by standard-setting bodies like 3GPP and O-RAN Alliance.

4. Organizational Structure

As highlighted above, the AI-RAN Alliance comprises of diverse members, united with the objective of advancing the implementation of AI with RAN and leveraging AI as a tool to substantially enhance RAN efficiencies and ecosystem benefits. The structure of the alliance is designed to promote collaboration, ensure efficient decision-making, and drive innovation in line with strategic goals.

4.1. Membership

The AI-RAN Alliance is comprised of Executive and General Members. Executive Members form the core leadership of the alliance, seated on the Board of Directors, and guide the strategic direction of the organization. General Members represent organizations committed to advancing and contributing to the alliance's objectives. All members participate actively in working groups and the alliance's technical initiatives.

Each member plays a distinct role in contributing to the advancement of AI-RAN technologies, with specific responsibilities in research, development, testing, and commercialization.

4.2. Governance

The AI-RAN Alliance is governed by its Board of Directors, which is composed of the founding Executive Members. The Board is responsible for the strategic oversight of the alliance, shaping its long-term vision, and approving key decisions such as the formation of new working groups. All activities of the alliance are guided by a set of foundational

by-laws, and working procedures which define membership categories, governance protocols, and operational procedures.

While the Board of Directors is responsible for overseeing the decision-making in the alliance, the technical and marketing development is guided by the Technical Steering Committee (TSC) and Marketing Steering Committee (MSC), respectively.

Operating within the strategic direction set by the Board of Directors, the TSC owns the alliance's technology roadmap and workplan, and is the principal forum for the discussion and management of the work products. The TSC ensures that the technical roadmap is aligned the objective of the alliance and is responsible for managing resources for Working Groups (e.g., labs) and supporting the setup of infrastructure needed for testing and implementation. In addition, the TSC is responsible for initiating, reviewing, and responding to technical liaisons from other organizations, as well as providing and reviewing technical content for the alliance's marketing efforts, including whitepapers, blueprints, and demos.

The MSC provides strategic direction and oversight for the organization's marketing initiatives. The MSC aims to enhance AI-RAN's brand visibility, engage diverse audiences, and support the organization's mission to advance AI research, analysis, and maturation. The committee guides and evaluates marketing strategies, campaigns, and activities to ensure alignment with the AI-RAN Alliance's goals.

The three Working Groups (WGs) are where the core of technical collaboration occurs, each tasked with a specific focus area. Each of the WGs operates under its respective charters, approved by the Board of Directors, that defines its scope, deliverables, and objectives. The groups are led by Chairpersons, who are responsible for driving active engagement and ensuring progress in-line with the goals established in their charters.

All WGs follow the Working Group Policies, which provide governance on group formation, operations, and performance evaluations. These policies ensure that all work within the groups is executed in alignment with the alliance's broader goals.

5. Working Groups and Roadmaps

Each working group is responsible for work items that cover various focus areas, collectively contributing to the alliance's overall goals.

5.1. Working Group 1: AI-for-RAN

WG1 focuses on the application of AI for RAN systems to improve efficiency, capacity, and performance metrics. The group's objectives include conducting literature reviews, defining use cases, and implementing proof-of-concept systems, with an emphasis on leveraging and advancing AI/Machine Learning (AI/ML) efforts in industry initiatives, potentially also standards.

The WG1 focus areas are:

- **AI-native air interface and signal processing,** where the use cases focus on improving the signal processing chain by replacing or augmenting traditional methods with AI/ML techniques for enhanced efficiency, reduced overhead, and better system performance. Some specific examples cover AI for receiver optimization, end-to-end learned air interface, Joint Communication and Sensing (JCAS) / Integrated Sensing and Communication (ISAC), Distributed MIMO (D-MIMO), Extra-Large Scale MIMO (XL-MIMO), etc.
- **Positioning and beam management,** where AI is used to improve the positioning accuracy, improve beamforming agility and performance. Examples include SRS and PRS based accurate positioning, CSI compression and prediction for beam management, etc.
- **Radio resource management and scheduling,** leveraging AI to enhance the dynamic allocation and utilization of radio and spectrum resources. This ensures the RAN can efficiently adapt to fluctuating traffic patterns and varying network conditions, ultimately contributing to optimized resource use and improved overall performance.
- **Energy and spectrum efficiency,** where the focus is on reducing the energy consumption, thus the operational cost of RAN components, and optimizing spectrum usage. Some specific use cases include cell energy saving and spectrum sensing/sharing.
- **Network optimization and anomaly detection,** where AI is leveraged to detect and predict network anomalies, as well as to reduce operational risks and enhance

network resilience. Specific examples include RAN digital twin and site-specific optimization, anomaly detection, fault prediction and mitigation, air-interface assurance, semantic communications, etc.

While AI has demonstrated promising performance across various domains, further integration of AI with RAN presents new research and engineering challenges.

The requirements on RAN in terms of availability, reliability, security and privacy are very high. RAN is also a highly distributed system with real-time functionalities operating under resource and latency constraints, imposing further requirements on AI algorithms, their implementation and their performance.

AI functionality in RAN is subject to challenging constraints. Solutions need to address scalability, security, privacy, resource efficiency, robustness, real-time processing, and more. The network usage will also evolve over time and data patterns and evolving business and use-case needs. AI-native RAN includes automated monitoring and management of ML models and pipelines to ensure they remain accurate and aligned with this evolution and variation.

5.2. Working Group 2: AI-and-RAN

Concentrating on the co-location of non-RAN AI workloads with RAN infrastructure, this group explores how both can coexist on shared platforms to maximize resource utilization and infrastructure value. In this regard, the group aims to increase platform utilization while creating monetization opportunities by running AI on shared computational resources. The group's activities include validating multi-tenancy systems and creating resource-partitioning frameworks driven by AI, for example, in edge data centers.

The focus areas are:

- **Design architecture and components of multi-tenant system,** including configuration management, security and privacy ensuring tenant isolation, automation, orchestration, and service management.
- **Lifecycle management of AI and RAN workloads,** where the management of AI and RAN workloads is addressed, ensuring seamless integration and operation of these workloads within the AI-RAN infrastructure.

- **Validating the multitenancy scenarios,** where the non-RAN AI and RAN workloads can run side by side and maintain RAN service level agreements in all scenarios.
- **Data center optimization for AI and RAN workloads,** where data center resources are optimized to maximize resource utilization, reduce operational costs, and contribute to integration goals.

5.3. Working Group 3: AI-on-RAN

Dedicated to identifying key RAN requirements, for present and future systems, to deliver and benchmark AI applications over RAN connectivity and infrastructure. The group's objectives include reviewing current AI-centric techniques, identify challenges, define use cases, develop test plans, and provide system implementation blueprints.

A critical aspect of WG3's work is the role of network-differentiated connectivity and device capabilities which are essential to enhance the Quality of Experience (QoE) for AI-powered applications.

The focus areas are:

- **AI-based multimedia applications,** where AI is employed to enhance multimedia experiences such as video analytics, augmented/virtual reality (AR/VR), immersive gaming, and next-generation multimedia applications like tactile Internet and holographic communications.
- **AI-based security and critical applications,** where AI is leveraged to strengthen security applications, such as intrusion detection, and support critical applications, such as healthcare.
- **AI-based automation and industrial applications,** that deals with the application of AI for automating industrial operations and managing autonomous vehicles, such as unmanned aerial vehicles (UAVs), drones and automated guided vehicles (AGVs).
- **GenAI/AI-based network services,** that focuses on the use of AI and GenAI to enhance network-based services such as customer service, user interactions, user positioning, etc. These AI solutions improve communication efficiency and enable seamless collaboration between users and applications.
- **Efficient AI/ML model splitting,** where an AI/ML model may be split between the device and RAN depending on several different factors such as radio link and device battery life conditions.

5.4. Roadmaps

Each working group has set its own roadmap to guide their work progress. The roadmaps outline specific results and deliverables that will be shared with the community.

The roadmaps focus on a systematic development and evaluation of AI techniques for each of the focus areas outlined in section 5.1. They start surveying existing standards and reviewing the latest AI/ML and RAN literature. The next step is to identify the key challenges, opportunities, and the most relevant use cases to be prioritized in the workstreams. The groups also establish performance benchmarks and design high-level system components for each of the selected use cases. For the evaluation, labs are built using appropriate tools as well as member capabilities, frameworks and platforms while ensuring vendor and technology neutrality in core functions to ensure a multi-vendor solution space and foster interoperability. Different implementations are evaluated through a comparative analysis which should leverage common benchmarks and enable fair comparison and highlight tradeoffs. Additionally, blueprints for system implementation, reference interoperation and evaluation frameworks are developed, and the need for standardization is assessed, including identifying relevant organizations for collaboration.

The tasks above collectively provide a systematic approach to exploring AI techniques, from initial research to potential standardization, fostering innovation, practical implementation, and realistic evaluation and validation.

To foster collaboration, expand membership, and drive industry adoption of AI-RAN solutions and blueprints, the marketing roadmap includes producing content such as whitepapers and case studies on working group initiatives and building partnerships with industry stakeholders. Additionally, the strategy involves participating in public events, such as MWC, and organizing live demos to showcase AI-RAN innovations, and joint PR to highlight member activities, while regularly providing updates from the working groups to engage the community and highlight progress and opportunities.

6. Conclusion

As the telecom industry embarks upon the transformative journey towards the integration of AI throughout the RAN and 6G, the AI-RAN Alliance offers a unique opportunity for members to lead this technology shift. Top industry leaders, academic institutions, and innovators collaborate to shape the future of RAN technologies and applications through AI integration. The AI-RAN Alliance drives impactful innovations for enhancing spectral efficiency, optimizing infrastructure, deploying cutting-edge AI services at the network edge and much more.

We invite interested parties to consider joining the AI-RAN Alliance to:

- Access a collaborative ecosystem that accelerates research and development in AI-RAN.
- Contribute to new and existing AI-RAN use cases that enhance network performance, efficiency, sustainability, and functionality.
- Work alongside global leaders to pioneer the next generation of mobile networks.
- Test and validate AI/ML concepts and ideas in industry-led testbeds, with results published in a peer-reviewed manner.

As we progress, we will soon share more detailed updates on each working group. Stay tuned for further insights into how the AI-RAN Alliance is driving AI integration in RAN use cases, unlocking new opportunities for innovation, value creation, and advancing the ever-evolving intersection of AI and RAN capabilities.