



AI-RAN Architecture Overview and Component Definitions

Version 1.2

ai-ran.org

 **AI-RAN**TM
ALLIANCE

Contents

1. Introduction and Core Concept	3
1.1. Dual Functions of AI-RAN.....	3
2. Architecture Diagram and Structure.....	3
2.1. The AI-RAN Architecture	4
3. Detailed Component Definitions	4
3.1. AI-RAN Infrastructure	4
3.2. Workloads (AI-and-RAN).....	5
3.3. Management, Control and Orchestration.....	5
3.4. Core Network	7
3.5. UE.....	7

1. Introduction and Core Concept

The **AI-RAN Architecture** represents a functional framework for integrating **Artificial Intelligence (AI)** directly into the **Radio Access Network (RAN)**, designed to create a more intelligent, efficient, and autonomous network platform. The high-level components and the architecture of AI-RAN is described in this document.

1.1. Dual Functions of AI-RAN

The architecture supports two primary functions, enabling both network optimization and new service enablement:

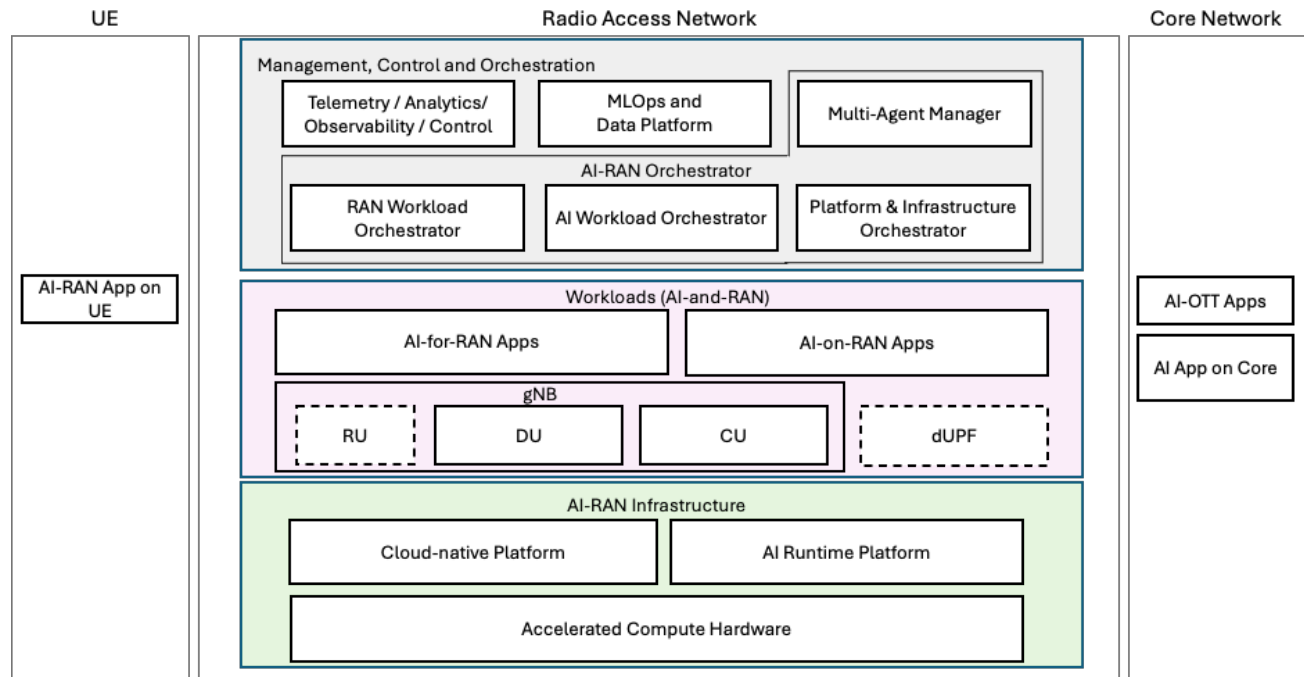
- **AI-for-RAN:** The application of AI/Machine Learning (ML) models specifically to **optimize the radio access network's performance**. This includes resource management, signal quality improvements, traffic prediction, energy consumption optimization and predictive maintenance.
- **AI-on-RAN:** Utilizing the accelerated computing RAN infrastructure to **host and accelerate new AI, Generative AI (GenAI), and Extended Reality (XR) applications** at the network edge, leveraging proximity for edge services over 5G and future 6G networks.

2. Architecture Diagram and Structure

The AI-RAN architecture is broadly categorized into three major functional domains, which run in coordination on the physical infrastructure:

Domain	Key Components	Description
AI-RAN Infrastructure	Accelerated Compute Hardware, Cloud-native Platform	The foundational layer providing the hardware and platform software for operating environment.
Workloads (AI-and-RAN)	gNB (DU, CU, RU), AI-for-RAN Apps, AI-on-RAN Apps, dUPF	The user plane and control plane functions for both network access and AI services.
Management, Control and Orchestration	AI-RAN Orchestrator, Telemetry / Analytics, MLOps & Data Platform, Multi-Agent Manager	The control plane responsible for coordinating and optimizing all workloads.

2.1. The AI-RAN Architecture



3. Detailed Component Definitions

3.1. AI-RAN Infrastructure

This layer provides the physical and virtualized foundation for hosting all RAN and AI workloads.

- **Accelerated Computing Hardware:** Hardware that is designed for efficient processing of RAN and AI/ML parallel workloads.
- **Cloud-native Platform:** The cloud-native platform hosting the workloads, consisting of:
 - **Host OS:** the operating system that has enhanced capabilities to provide real-time and deterministic low latency performance required for the RAN workloads.
 - **CaaS (Container as a Service):** Container as a Service is the core Kubernetes (<https://kubernetes.io/>) functionality that deploys, manages, and scales containerized applications. It abstracts the underlying infrastructure from the containerized applications.
 - **PaaS (Platform as a Service):** Platform as a Service is part of the Kubernetes cloud native platform that includes necessary management and observability tools from the CNCF (<https://www.cncf.io/>) projects.
 - **Infrastructure Management:** Infrastructure Management System is responsible for hardware, software and networking of the IT infrastructure.

- **Deployment Management:** This entity is responsible for "Pre-boot Execution Environment (PXE)," a network-based booting method that allows a computer to load an operating system from a remote server instead of a local hard drive or other storage device.
- **AI Runtime Platform:** A software environment optimized for the underlying accelerated hardware, responsible for the execution, loading, inference, and resource allocation of AI models.

3.2. Workloads (AI-and-RAN)

These components define the access network and user data handling.

- **gNB (Base Station):** The 3GPP-defined base station that connects User Equipment (UE) to the network. It consists of:
 - **CU (Centralized Unit):** The centralized component of the RAN workload.
 - **DU (Distributed Unit):** The distributed component of the RAN workload.
 - **RU (Radio Unit):** The physical radio component.
- **UE (User Equipment):** The end-user device, which may also host AI-RAN applications.
- **dUPF (Distributed UPF):** The distributed User Plane Function (dUPF) is a core network function performing local breakout for user data handling and direct routing of user traffic to and from the co-located AI-on-RAN workloads. In addition, MUP of SRv6 also has the same feature of dUPF. It depends on the core network design of cloud gateway.
- **AI-RAN Apps:**
 - **AI-on-RAN Apps:** The AI, GenAI, XR applications and/or workloads that run on the RAN infrastructure in parallel with the RAN workloads.
 - **AI-for-RAN Apps:** These are the AI/ML applications that run on the RAN infrastructure to improve RAN's performance. This application can be hosted on or proximal to gNB or CU or DU.

3.3. Management, Control and Orchestration

This collective layer ensures all resources and workloads are deployed, scaled, and maintained efficiently.

- **AI-RAN Orchestrator:** Consists of one or more instances of RAN Workload Orchestrator, AI Workload Orchestrator, Platform and Infrastructure Orchestrator and Multi-Agent Manager.
 - **RAN Workload Orchestrator:** This is the cloud orchestration system that manages the deployment, scaling, lifecycle, and coordination of cloud-native network functions (CNFs), such as CU, DU. It automates the creation, configuration, and operation of

containerized AI-RAN services, enabling seamless integration, policy enforcement, and efficient resource allocation on the AI-RAN infrastructure.

- **AI Workload Orchestrator:** This is the cloud orchestrator that manages the deployment, scaling, lifecycle, and coordination of AI/ML workloads.
- **Platform & Infra Orchestrator:** An orchestrator that manages Kubernetes (K8s) clusters in the AI-RAN. It's designed to streamline and automate the deployment, scaling, and lifecycle management of multiple Kubernetes clusters. It provides centralized control to provision, configure, monitor, and maintain clusters across the network ensuring consistent policies, security, and infrastructure resource allocation across the AI-RAN.
- **Multi-Agent Manager:** This is the Agentic AI manager for Operations & Management of AI-RAN and is a central software component that oversees the orchestration, deployment, and lifecycle management of autonomous AI agents within AI-RAN. It enables integration of AI agents that monitor, analyze, and optimize RAN operations. This component acts as the “control tower” for agentic AI systems in telecom environments, making AI-RAN autonomous, adaptive, and efficient by leveraging advanced AI-driven decision-making.
- **Telemetry, Monitoring, Observability & Analytics:** This component is responsible for collecting, processing, and analyzing operational data from AI-RAN applications, infrastructure, and hardware devices. It gathers metrics, logs, traces, events, and alarms from AI-RAN CNFs and platform services to provide real-time insight into system health, performance, and security. The observability framework enables proactive fault detection, performance optimization, and anomaly detection, while supporting FCAPS objectives. It also integrates with the AI/ML data platform and data warehouse to feed analytics pipelines for closed-loop automation and intelligent decision-making across the AI-RAN architecture. The Telemetry / Monitoring / Observability / Analytics function may be realized by, or include capabilities from, existing RAN management components such as the RIC, SMO, OSS, or EMS, SON, depending on the deployment scenario and vendor implementation.
- **MLOps & Data Platform:** It's the system responsible for overseeing and automating the end-to-end process of developing, deploying, monitoring, and maintaining the AI-RAN (AI-for-RAN) Apps. It manages the iterative stages of the AI-RAN Apps' lifecycle, data collection and preparation, model training and evaluation, deployment, and ongoing monitoring—to ensure model reliability, performance, and continuous improvement. It integrates tools for workflow automation, versioning, testing, and governance, enabling AI-RAN Apps' algorithm developers to efficiently scale and maintain AI/ML solutions throughout their operational lifespan.

3.4. Core Network

The core network is essential for completing the data connections with the UEs. The network functions in the core network may provide some additional data points to AI-RAN functions and applications. Two such functions are:

AI App on Core: an illustrative application that provides additional data to AI-RAN. An example function will be NWDAF.

AI-OTT Apps: Over the top AI application instances hosted in the core network that may have some associations with AI-on-RAN apps in AI-RAN.

3.5. UE

The applications that run on the UE in conjunction with the AI-for-RAN and AI-on-RAN applications running in the RAN.

AI-RAN App on UE: Applications running on the UE that have AI/ML components in the UE side as necessary. Example application for AI-for-RAN is 2-sided CSI compression, and AI-on-RAN is AI/ML inferencing model that is split between the UE and the AI-RAN.