

Design and Optimization of AI-RAN Architectures for Intelligent 6G Communication Systems

Prof. S. D. Lavange¹, Prof. Y.P. Sushir²,

^{1,2} Assistant Professor, Department of Electrical Engineering, Padm. Dr. V. B. Kolte College of Engineering, MH, India

DOI: 10.5281/zenodo.19291641

ABSTRACT

Sixth-generation (6G) wireless networks are anticipated to be built upon a set of fundamental architectural principles such as pervasive intelligence, decentralization, seamless interoperability, and end-to-end digitalization. As wireless systems continue to evolve toward higher levels of complexity and service diversity, the integration of advanced artificial intelligence (AI) and machine learning (ML) techniques into the core of communication architectures has emerged as a critical requirement for 6G and future network generations. Embedding intelligence natively within wireless infrastructures is expected to enable networks that are self-learning, self-optimizing, and capable of autonomous decision-making in highly dynamic environments. In addition, the paper provides a comprehensive review of the evolution of RAN architectures, tracing their development from traditional centralized designs to more flexible, virtualized, and open architectures. It further highlights the key technological enablers that are expected to play a pivotal role in realizing AI-RAN, including digital twins (DTs) for real-time network modeling and simulation, intelligent reflecting surfaces (IRSs) for programmable wireless environments, large-scale generative AI (GenAI) models for advanced decision-making and automation, and blockchain (BC) technologies for secure, transparent, and decentralized network management. Furthermore, the paper discusses the major challenges associated with the deployment of AI-RAN from both technical and regulatory perspectives. These challenges include issues related to data availability and quality, model interpretability, computational complexity, energy efficiency, security, privacy, and compliance with evolving regulatory frameworks. Finally, the paper outlines promising future research directions for AI-RAN, emphasizing the integration of emerging technologies such as integrated sensing and communication (ISAC) and agentic AI systems, which have the potential to further enhance network intelligence, autonomy, and resilience in next-generation wireless networks.

Keywords: - 6G, AI-driven radio access networks, integrated sensing and communication, agentic AI, blockchain

1. INTRODUCTION

The evolution toward sixth-generation (6G) wireless networks marks the beginning of a new era in global connectivity. Beyond merely enhancing peak data rates and spectral efficiency, 6G is envisioned to enable a wide range of advanced and immersive services, including holographic communications [1], the Internet of Senses (IoS) [2], and numerous other applications that were previously regarded as speculative or futuristic. However, the realization of these ambitious use cases demands far more than incremental upgrades to existing network infrastructures. The unprecedented scale, heterogeneity, and dynamic behavior of 6G systems necessitate a fundamental rethinking of network design principles. In this context, key concepts such as self-constructing, self-adaptive, and self-healing network infrastructures emerge as central pillars of 6G architecture. These principles significantly challenge the conventional understanding of Radio Access Networks (RANs) and elevate Artificial Intelligence (AI) from a supporting tool to a core architectural element in next-generation RAN design. AI is expected to enable networks that can perceive their environment, learn from operational data, and autonomously optimize their performance in real time.

Historically, from first-generation (1G) systems through second-generation (2G) networks and the early stages of third-generation (3G) deployments, RANs primarily relied on rule-based mechanisms and relatively simple algorithms. These systems were typically implemented within Distributed RAN (D-RAN) architectures and focused on basic functionalities such as elementary traffic estimation and static network management. As wireless technologies progressed, the later phases of 3G and subsequent generations witnessed the gradual adoption of Machine Learning (ML) and Deep Learning (DL) techniques, alongside the introduction of more advanced RAN architectures including Centralized RAN (C-RAN), Virtualized RAN (vRAN), and Open RAN

(O-RAN). These developments enabled more sophisticated network functions such as dynamic load balancing, interference mitigation, fault detection, and anomaly identification, while also improving scalability, openness, and interoperability among multi-vendor ecosystems. Despite these advancements, prior-generation networks remained largely reactive and lacked the level of autonomy, predictive intelligence, and holistic end-to-end orchestration necessary to support the stringent and diverse requirements of future services. In contrast, 6G places a strong emphasis on pervasive and large-scale intelligence, driven by advanced AI models, with the ultimate goal of realizing AI-native networks that operate with minimal human intervention. Achieving this vision requires intelligent, real-time orchestration and autonomous decision-making directly at the RAN level, thereby enabling seamless end-to-end automation across the entire network stack.

2. EVOLUTION OF RANS: REVISITING ARCHITECTURAL GROWTH

At the heart of wireless systems lies the RAN, the bridge that links User Equipment (UE) with the Core Network (CN). A typical RAN comprises two major components: the Remote Radio Unit (RRU) and the Baseband Unit (BBU). The RRU contains antennas that transmit and receive radio signals, while the BBU orchestrates operations.¹ Though early RAN architectures were adequate for the limited scope of wireless systems prior to the Long Term Evolution (LTE) networks, use cases such as Ultra-Reliable Low-Latency Communication (URLLC), Massive Machine-Type Communication (mMTC), and Enhanced Mobile Broadband (eMBB) has necessitated more advanced RAN solutions [3].

The D-RAN architecture has been a widely used approach, especially from 1G to 3G. In this architecture, the RRU and BBU are co-located at each cell site, facilitating direct on-site processing of radio signals and baseband functions. Each cell site processes the radio signals on the distributed RRU to directly process the radio signals on site, while simultaneously providing the baseband functions on site with the co-located BBU. This setup allows for immediate user data and control signal handling, which provides advantages in terms of reduced latency. With the development of mobile networks, the limitations of D-RAN architecture have become more apparent. Driven by challenges, inefficient resource utilization at all dispersed sites and the inability to apply complex coordination technologies across multiple cell sites, the focus has shifted towards more centralized RAN architectures.

C-RAN offers advantages over D-RAN by centralizing baseband processing into BBU pools, improving throughput, resource efficiency, and power consumption. This centralization reduces both capital and operational costs for service providers. It also enables advanced interference coordination techniques such as Coordinated Multipoint Transmission (CoMP), improving uplink performance by 50%–100% [4]. Due to these benefits, the C-RAN architecture has been mainly deployed in 4G LTE era and beyond. C-RAN utilizes a fronthaul connection (e.g., Common Public Radio Interface (CPRI)) to interconnect the centralized BBU pool and RRUs. However, a significant drawback of this architecture is that the interface between the BBU pool and the RRUs remains proprietary.

The traditional D-RAN architecture, where all RAN functions are integrated into proprietary hardware, faces limitations in supporting technologies and services required for 5G. vRAN addresses these challenges by virtualizing BBU functions on Commercial Off-The-Shelf (COTS) hardware, enhancing scalability, flexibility, and cost efficiency. It marks the beginning of hardware-software disaggregation in the RAN, eliminating the need for frequent hardware upgrades when introducing new standards or features. This enables flexible scaling and low operational costs. However, like C-RAN, vRAN retains proprietary interfaces, limiting interoperability between vendors and resulting in vendor lock-in. With technological advancements, O-RAN, introduced by the O-RAN alliance, has become a promising software-centric architecture that provides network flexibility, cost efficiency, open interfaces, Software-Defined Networking (SDN), multi-vendor capability, Network Function Virtualization (NFV), energy efficiency, and potential advantages for AI integration into RAN [5]. Beyond its role as an architectural framework, O-RAN acts as a critical enabler for the next evolutionary stage of RAN due to its innate openness, interoperability, and software-defined capabilities.

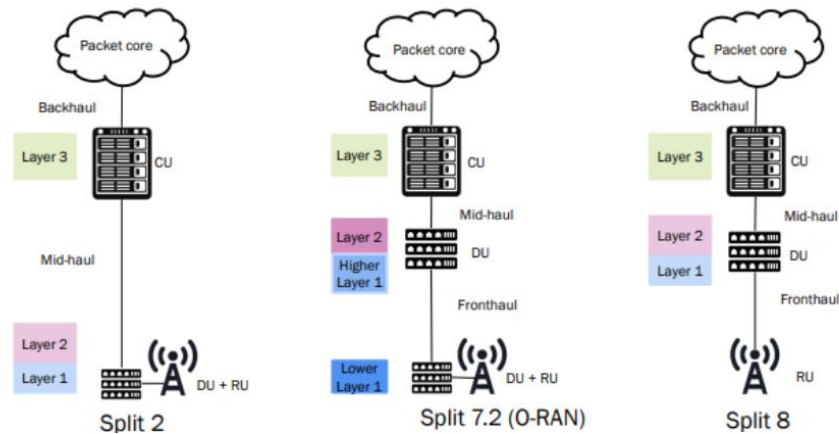


Fig - 1: Transition D-Ran to O-Ran

3. KEY TECHNOLOGY ENABLERS

6G is expected to extend existing use cases into Ubiquitous Mobile Ultra-Broadband (uMUB), Ultra-High Data Density (uHDD), and Ultra-High-Speed with Low-Latency Communications (uHSLLC). Apart from already revealed features like openness and intelligence, AI-RAN has the capacity to forecast traffic patterns in the network, dynamically regulate the spectrum, and minimize the necessity of over-provisioning. Furthermore, AI-RAN rectifies network issues and vulnerabilities, thereby improving overall performance and ensuring a future-proof network. Ultimately, AI-RAN aspires to enable networks that are self-adaptive, self-optimizing, and self-sustaining, driven by intelligent and autonomous decision-making processes. This section provides a neutral view of such enhancements driven by the key enabling technologies of AI-RAN. A Digital Twin (DT) is a virtual representation of the components and dynamics of a physical system, and DTs can be broadly categorized into three types: monitoring DTs, simulation DTs, and operational DTs [6]. Monitoring DTs support monitoring of the status of a physical system, whereas simulation DTs use various simulation tools to provide future insights. One of the important abilities of AI-RAN is to be self-healing and zero-touch, hence autonomously identify faults and perform live-action responses to ensure self-maintainability. DTs can facilitate automated network maintenance and help reduce management overheads. Another important aspect is that DTs can facilitate AI model/agent training by synthesizing data that represents various network scenarios. Rarely experienced network scenarios can be artificially simulated in a controlled environment to observe network behaviors and assist in the data collection process. By doing so, it removes the burden of collecting large amounts of datasets with controllable distributions to ensure strict Quality of Service (QoS) requirements. Finally, DTs can also improve context awareness by processing multi-sensory data to identify contextual elements related to the environment, time, and situation.

The emergence of GenAI marks another inflection point in realizing the vision of future wireless generations, where networks have the capability to adjust, optimize, and reconfigure their functionalities and parameters based on network conditions. This stands in contrast to most of the currently prevailing AI solutions, which are designed for dedicated problem-solving and function merely as add-on features in wireless networks. These are models trained on vast amounts of data to generate new content such as text, images, videos, and more. More specifically, Large Language Models (LLMs) [7], the foundation model for natural language understanding, and Large Action Models (LAMs) [8], the foundation of AI agents, are the major enablers. Given below are several ways large GenAI models facilitate the development of AI-RAN.

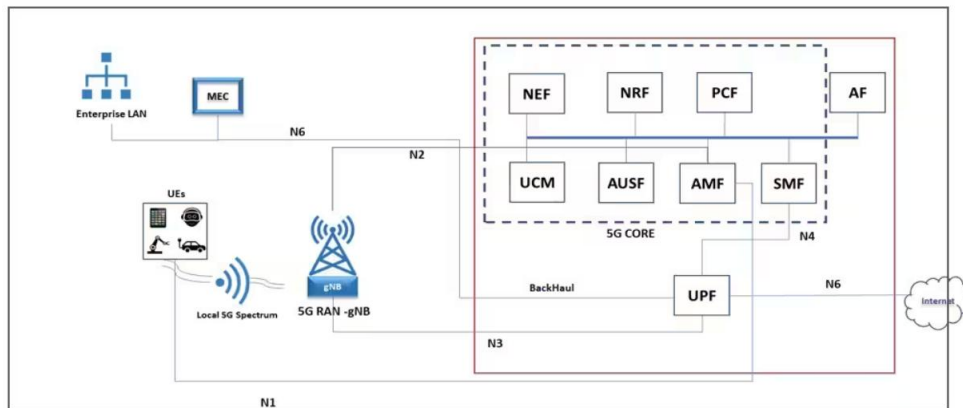


Fig - 2:Ran-Lam

4. PERSPECTIVE ARCHITECTURE

We propose an architecture that orchestrates intelligent and adaptive RAN functionalities for 6G networks., it incorporates Orchestration RAN (OrchestRAN) with AI-enhanced optimization and dynamic resource allocation for both Real-Time (RT) and Near-Real-Time (near-RT) applications. Requests from network operators are collected by the Request Collector module, covering functionalities such as network slicing, scheduling, and beamforming, along with associated location (e.g., RAN Intelligent Controller (RIC), Central Unit (CU), and Distributed Unit (DU)) and time constraints (e.g., latency requirements for different service classes). The module aggregates and forwards these to the OrchestRAN framework for processing.

Within OrchestRAN, pre-trained AI models are stored in an ML/AI catalog. This includes Reinforcement Learning (RL) for resource allocation, FL for distributed intelligence, DL optimizers, Graph Neural Networks (GNNs) for topology-aware management, and transformer-based models for traffic forecasting. The Infrastructure Abstraction module provides a high-level view of physical infrastructure, divided into five logical groups: near-RT RICs, non-RT RICs, CUs, DUs, and RUs. Each group contains a variable number of nodes deployed across the network with virtualization capabilities. Upon receiving requests, the Orchestration Engine selects suitable AI models from the ML/AI catalog based on service requirements, available computational resources, and current network conditions. These models are embedded into docker containers as dApps, xApps, or rApps, and dispatched to selected nodes. The engine also generates orchestration policies that define data collection for model training and feedback loops for continuous learning. These are converted into executable O-RAN applications using E2, A1, and O1 interfaces. Intelligent RAN Controllers manage AI-driven control. The non-RT RIC oversees slicing policies and cross-slice optimization, while the near-RT RIC enhances scheduling and beam tracking, manages interfaces, and supports intelligent handovers.

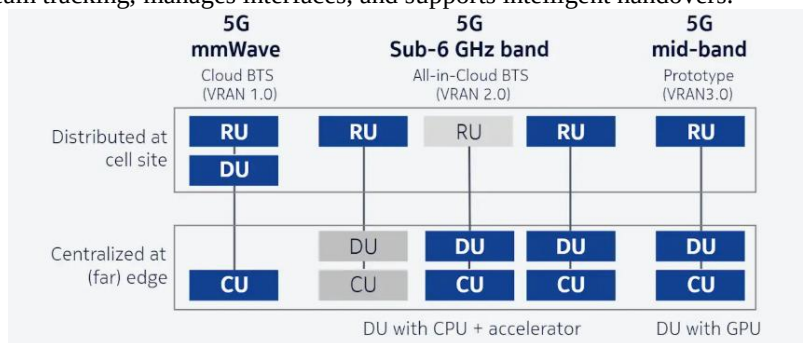


Fig - 3:Ran-Architecture

To evaluate performance, we have developed a simulation scenario in Python. The simulation considers an OrchestRAN-based distributed architecture designed to optimize beamforming and scheduling decisions using multi-agent RL. The scenario includes 100 operator requests per time slot, with 2 non-RT RICs, 5 near-RT RICs, 3 CUs, 8 DUs, and 25 RUs at 28 GHz with 400 MHz bandwidth. The model follows the 3rd Generation Partnership Project (3GPP) TR 38.901 Urban Macro (UMa) path loss model and includes ray-tracing-based fading for both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS) conditions. The performance of the proposed framework is analyzed in comparison with traditional scheduling algorithms, including Round Robin, Proportional Fair, and Max-Min Fairness, across different performance metrics such as latency and spectral efficiency. As illustrated in Fig. 4(a), OrchestRAN demonstrates a significantly lower latency compared to other scheduling techniques. The Round Robin algorithm exhibits the highest latency variance, whereas Proportional

Fair and Max-Min Fairness provide moderate latency improvements. The reduced latency in OrchestRAN can be attributed to its intelligent decision-making process with AI models based on service requirements. It dynamically allocates resources based on real-time network conditions, and hence minimizing delays. Further, as illustrated in Fig. 4(b), OrchestRAN consistently outperforms traditional scheduling methods, achieving a spectral efficiency improvement of up to 20% compared to Round Robin and 10-15% compared to Proportional Fair and Max-Min Fairness. This superior performance is due to adaptive resource allocation and RL-based optimization, which maximize spectrum utilization while maintaining QoS. Overall, these results validate the effectiveness of the AI-RAN framework based on OrchestRAN, highlighting its capability to enhance network efficiency by reducing latency and improving spectral efficiency under varying network loads. After typing the right username and password, we will get access to the web server

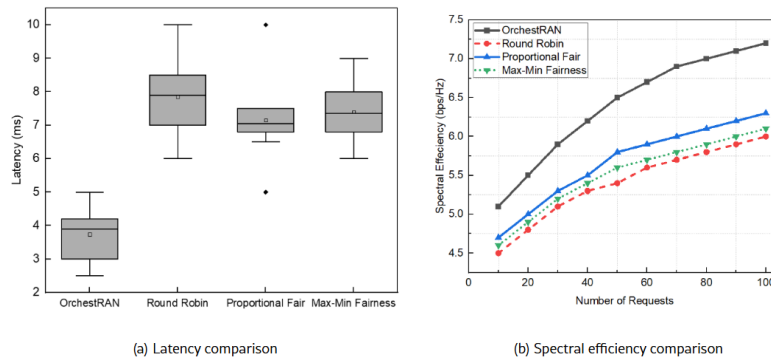


Fig - 4 LoginPerformance comparison of the proposed AI-RAN architecture based on OrchestRAN with traditional scheduling algorithms.

5. CONCLUSIONS

This study has highlighted the potential of integrating AI into RAN architectures, thereby enabling intelligent, self-optimizing, and fully automated networks for 6G and beyond. The proposed AI-RAN framework demonstrates improved network performance and adaptability. However, challenges such as explainability, scalability, data protection, and compliance must be addressed. Further research is needed, particularly in security protocols, standardized AI agent architectures, and sustainable AI operations, to fully realize AI-RAN.

6. REFERENCES

- [1]. T. Gong, P. Gavrilidis, R. Ji, C. Huang, G. C. Alexandropoulos, L. Wei, Z. Zhang, M. Debbah, H. V. Poor, and C. Yuen, "Holographic MIMO Communications: Theoretical Foundations, Enabling Technologies, and Future Directions," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 196–257, 2024, doi: 10.1109/COMST.2023.3309529.
- [2]. N. Sehadi, L. Bariah, W. Hamidouche, H. Hellaoui, R. Jantti, and M. Debbah, "Generative AI for Immersive Communication: The Next Frontier in Internet-of-Senses Through 6G," *IEEE Communications Magazine*, vol. TBD, no. TBD, pp. 1–13, 2024, doi: 10.1109/MCOM.001.2400199.
- [3]. J. L. Frauendorf and É. Almeida de Souza, "The Evolution of RAN (Radio Access Network), D-RAN, C-RAN, V-RAN, and O-RAN," in *The Architectural and Technological Revolution of 5G*, Springer, Cham, 2023. DOI: https://doi.org/10.1007/978-3-031-10650-7_10.
- [4]. I. Chih-Lin, J. Huang, R. Duan, C. Cui, J. X. Jiang, and L. Li, "Recent Progress on C-RAN Centralization and Cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014, doi: 10.1109/ACCESS.2014.2351411.
- [5]. N. A. Kaim Khani and S. Schmid, "AI-RAN in 6G Networks: State-of-the-Art and Challenges," *IEEE Open Journal of the Communications Society*, vol. PP, pp. 1–1, Jan. 2023, doi: 10.1109/OJCOMS.2023.3343069.
- [6]. L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-Twin-Enabled 6G: Vision, Architectural Trends, and Future Directions," *IEEE Communications Magazine*, vol. 60, no. 1, pp. 74–80, 2022, doi: 10.1109/MCOM.001.21143.
- [7]. H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin, C. Chen, H. Wu, D. Yuan, L. Jiang, D. Wu, X. Liu, C. Zhang, X. Wang, and J. Liu, "Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities," *IEEE Communications Surveys & Tutorials*, vol. TBD, no. TBD, pp. 1–1, 2024, doi: 10.1109/COMST.2024.3465447.
- [8]. Bohora, B., Maharjan, S., and Shrestha, B. R; "IoT Based Smart Home Using Blynk Framework". *Zerone Scholar*, . (2016). 1(1), 26-30.
- [9]. J. Zhang, T. Lan, M. Zhu, Z. Liu, T. Hoang, S. Kokane, W. Yao, J. Tan, A. Prabhakar, H. Chen, Z. Liu, Y. Feng, T. Awalganekar, R. Murthy, E. Hu, Z. Chen, R. Xu, J. C. Niebles, S. Heinecke, H. Wang, S. Savarese,

and C. Xiong, “xLAM: A Family of Large Action Models to Empower AI Agent Systems,” arXiv Preprint, arXiv:2409.03215, 2024. [Online]. Available: <https://arxiv.org/abs/2409.03215>.

[7]. Reddy, P. S. N., Reddy, K. T. K., Reddy, P. A. K., Ramaiah, G. K., & Kishor, S. N. “An IoT based home automation using android application.”; International IEEE Conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), October, 2016, pp. 285-290

[8]. J. Guo, L. Gao, N. Li, S. Yang, J. Zhu, X. She, J. Wang, and P. Chen, “System-Level Simulation and Performance Evaluation for 6G Ultra Massive MIMO,” in Proc. 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), 2023, pp. 1–6, doi: 10.1109/VTC2023-Spring57618.2023.10199530.

[9]. Singh, S., Saha, D., Khaware, P., Das, S., Raj, D., Das, S., & Nandi, C. S.; “Home Automation and Internet of Things”. International Advanced Research Journal in Science, Engineering and Technology, 2016, 3(6)..

[10]. C. Sandeepa, E. Zeydan, T. Samarasinghe, and M. Liyanage, “Federated Learning for 6G Networks: Navigating Privacy Benefits and Challenges,” IEEE Open Journal of the Communications Society, vol. 6, pp. 90–129, 2025, doi: 10.1109/OJCOMS.2024.3513832.