3rd URSI AT-AP-RASC, Gran Canaria, 29 May – 3 June 2022

Intelligent and QoE-aware Open Radio Access Networks

Georgios Kougioumtzidis* (1), Vladimir Poulkov(1), Zaharias D. Zaharis(2), and Pavlos I. Lazaridis(3)
(1) Technical University of Sofia, Sofia, 1000, Bulgaria, email: gkougioumtzidis@tu-sofia.bg; vkp@tu-sofia.bg
(2) Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece, email: zaharis@auth.gr
(3) University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom, email: p.lazaridis@hud.ac.uk

Abstract

The open radio access network (O-RAN) concept refers to the architectural design of next generation RAN that is built on the principles of openness and intelligence. This paper presents an overview of the concept of O-RAN, by analyzing its architecture and examining its main building blocks. Moreover, it highlights the significance of quality of experience (QoE) for the envisioned future wireless networks, and analyzes the importance of integrating QoE-awareness in O-RAN's design. Furthermore, it provides an analysis of the methodology of embedding artificial intelligence models in O-RAN's architecture with the form of xApps.

1. Introduction

Upcoming beyond 5G (B5G) and 6G mobile networks should support ever-increasing mobile traffic and the advent of innovative and heterogeneous usage scenarios, including telesurgery, vehicles-to-everything (V2X), industry 4.0, and holographic telepresence communications. To meet these challenges, the architectural design of mobile networks must be enriched to be able to provide programmable, virtualized, flexible, intelligent, and energy efficient connectivity services.

The open radio access network (O-RAN) concept as introduced by the O-RAN alliance, can be considered as a game-changer technology, showing strong capabilities to turn mobile RAN domain towards customized solutions to meet the aforementioned diversified service requirements of use cases of future wireless networks (FWNs). O-RAN alliance is a consortium shaped by the fusion of the cloud-RAN (C-RAN) alliance and the xRAN forum. Its goal is to evolve current RAN topologies and deploy virtualized and fully interoperable next generation RAN (NG-RAN) based on two fundamental pillars: openness and intelligence [1].

The cardinal principles on which O-RAN is founded, include virtualized RAN elements, white-box hardware, open-source software, standardized interoperable interfaces, specified application programming interfaces (APIs), and use of commercial-off-the-shelf (COTS) hardware components. Accordingly, the benefits of adopting O-RAN-based solutions, include the efficient use of network resources, optimized network operation, simpler and faster migration to new technologies, lower capital expenditure (CAPEX) and operational expenditure (OPEX), higher revenues and creation of new markets and business models.

A key feature in the design of the inherently user-centric FWNs, and consequently, of O-RAN, will be the quality of experience (QoE). The concept of QoE within the context of mobile networks refers to assessing the quality of network services and operations based on the end-users' perspective. QoE is a very broad and interdisciplinary metric, as an extensive number of factors belonging to various fields influence the quality of a communication service as it is perceived by end-users. To address the extremely complex challenge of embodying QoE-awareness into O-RAN's operation, it is important to embed appropriate artificial intelligence (AI) and machine learning (ML) methods into its design.

In this paper, we study the concept of O-RAN and analyze its architecture and main building blocks, and highlight the advantages provided by its deployment. Moreover, we discuss the role and importance of QoE in FWNs, and underline the necessity of incorporating QoE-awareness in O-RAN's design framework. Finally, we analyze the methodology of embedding AI/ML solutions into O-RAN's architecture, and provide a practical guide for xApps development.

2. O-RAN Architecture

O-RAN is intended to support NG-RAN implementations based on the concepts of openness and intelligence. It offers well-designated and standardized interfaces to provide an open, interoperable ecosystem that supports and complements 3GPP and other industry standardization groups. The O-RAN architecture encompasses the RAN elements, the operation support system (OSS), the radio engineering systems, and well-defined interfaces as shown in Fig. 1. The central unit (CU), distributed unit (DU) and radio unit (RU) are respectively in charge of the functionality of radio resource control (RRC), radio link control (RLC)/media access control (MAC), and physical

(PHY) layer of the radio protocol stack. The antenna and radio frequency (RF) processing units, as well as the lower-level PHY (PHY-low) layer are embedded in O-RAN RU (O-RU), which is responsible for establishing the link between the PHY layer and the user equipment (UE). The O-RAN DU (O-DU) is in charge of the higher-level PHY (PHY-high) layer operations, including channel modulation and coding/decoding, and MAC and RLC layer. O-RU and O-DU are connected via the open fronthaul interface. The O-RAN CU (O-CU) oversees the RRC, packet data convergence protocol (PDCP), and service data adaptation process (SDAP) layer of the radio protocol stack. O-CU is also known as multi-radio access technology (multi-RAT), as it supports the operation of 4G long-term evolution (LTE) and 5G new radio (NR) O-DUs simultaneously. For the enablement of multi-RAT O-CU functionality, the network function virtualization infrastructure (NFVI) technology is exploited. The O-CU is connected with the O-DU via the F1 interface, which is responsible for application-level data exchange regarding resource coordination, cross-link interference mitigation, flow control management, and control trace sessions [2].

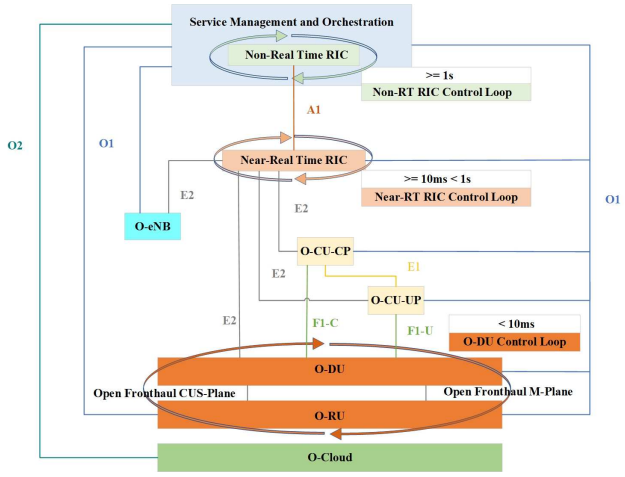

Figure 1. O-RAN logical architecture and control loops [11].

2.1 Service Management and Orchestration

The service management and orchestration (SMO) unifies a multitude of management and orchestration (MANO) services, including 3GPP next generation core (NGC) network management and end-to-end (E2E) network slice management. The main duties of SMO include the following: 1) supply of the interface of fault, configuration, accounting, performance, and security (FCAPS) for the network functions (NFs); 2) large-timescale RAN optimization; and 3) open cloud (O-Cloud) MANO through the O2 interface, including resource discovery, scaling, FCAPS, software management, and create, read, update, and delete (CRUD) O-Cloud resources [3].

2.2 Non-Real Time RAN Intelligent Controller

The non-real time RAN intelligent controller (non-RT RIC) is positioned in SMO. The control loop-based functions of non-RT RIC (reaction time >1s) are typically utilized for non-RT management, service and policy administration, RAN analytics, and ML model training. Through the A1 interface, non-RT RIC receives information from near-real time RAN intelligent controller (near-RT RIC), regarding traffic patterns and user mobility behavior, and exchanges policy update data to optimize RAN operation. Simultaneously, it receives feedback through the O1 interface with regard to network condition. Non-RT RIC applications (rApps) aim to improve the control and optimization of RAN elements and resources. The gathered data may be utilized by ML models to enhance the decision process [4].

2.3 Near-Real Time RAN Intelligent Controller

The near-RT RIC is connected with CU and DU via the E2 interface, implementing radio resource management (RRM) functionalities with integrated intelligence. Near-RT RIC utilizes the 7-2x gNB split, and executes operationally demanding control loop-based functions (10ms $\leq$ reaction time < 1s), including per-UE controlled load balancing, resource block (RB) management, and interference detection and mitigation. Near-RT RIC supports the development of applications called xApps, which utilize 2 near-RT RIC databases, one including information about UEs (UE-NIB), and one including information about RAN nodes (R-NIB). It includes mechanisms for toning down conflicts originated by xApps requests, and utilizes the O1 interface for xApps orchestration. It also includes a functions library for supporting AI-based operations. O-RAN architecture supports the deployment of RAN elements in cloud environments, facilitating SMO functionalities. The O2 interface is used to support cloud-related management functions, including virtual resource management [5].

2.4 Advantages of O-RAN Architecture

The advantages provided by O-RAN architecture can be summarized in the following [6]: 1) reduction of network's CAPEX and OPEX, as open interfaces, open source software and hardware reference designs, and O-RAN's native cloud functionality minimize CAPEX, and RAN automation minimizes OPEX by integrating intelligence in RAN design and adopting new learning-based techniques to greatly automate operational network services; 2) increased network efficiency and operation, as RAN automation allows for continuous monitoring of network performance and resources, as well as for provision of effective, optimal radio resource management through closed-loop control; and 3) network agility, as O-RAN's native cloud infrastructure can introduce new features through simple software upgrades.

3. QoE-awareness in O-RAN

The envisioned 5G usage scenarios of enhanced mobile broadband (eMBB) and ultra-reliable low latency communications (uRLLC) are, respectively particularly bandwidth consuming and latency sensitive. Current demanding interactive applications such as online video streaming, multiplayer gaming and connected vehicles, are typically served in a best effort manner, without application-oriented optimization. To make these applications meet varied QoE requirements, it is necessary to develop a more holistic approach in the network operation. Regarding this, QoE prediction can provide awareness towards the QoE levels in the application layer, as well as the application-oriented parameters of the communication link, thus enhancing the effective utilization of radio resources [7].

In a multi-tenancy environment, the continuous service monitoring is critical to guarantee that tenant's QoE meets the agreed levels. When a tenant is allocated to a specific network, the network configuration guarantees the quality of service (QoS) for delivering the service requested by the tenant and ensured by specific service level agreements (SLAs). It is also important to monitor the E2E service flow to secure that traffic flow is in proportion to the network capacity, and that the level of QoE remain immutable. Consequently, a QoE and service flow monitoring framework would guarantee that the tenant's QoE remains within the agreed levels, and efficiently allocate radio resources, enabling the network operator to serve a larger number of users [8].

Application-oriented QoE prediction and real-time QoE-driven proactive closed-loop network optimization can aid in preserving QoE. The radio resource allocation should be steered where they are most urgently required, thus avoiding QoE degradation. Such an approach, would optimize QoE whilst effectively exploiting the radio resources. AI/ML models can be developed to optimize QoE, by taking advantage of the software-defined RIC and open interfaces of O-RAN. These models can gather and process multi-dimensional data to support operations such as traffic management and QoE prediction, and enforce close-loop QoS decisions [9]. The development of an automated policy control for QoS/QoE behavior, permits network operators to fine tune the behavior in real-time. The appropriate network configuration and the behavior modification in response to the tenant's QoE requests, can be achieved through the dictation of policy and control from the non-RT RIC and near-RT RIC respectively, in the form of rApps and xApps [7].

4. Embedding AI/ML in O-RAN with xApps

The utilization of AI and ML techniques is expected to be a catalyst in the design of FWNs, by providing solutions to highly complex problems at a wide range of levels, including PHY layer, MAC layer, RRM, and RAN OSS [10]. AI/ML assisted solutions in O-RAN fall into the three types of control loops as shown in Fig. 1 [11]: 1) loop 1 controls the transmission time interval (TTI) level scheduling and operates within the TTI time scale (<10 ms); 2) loop 2 is positioned in the near-RT RIC, operates within the range of 10-500 ms, and performs resource optimization; and 3) loop 3 is positioned in the non-RT RIC, operates within a time scale greater than 500 ms, and manages policies and orchestration. AI/ML related operations can be distributed into the three loops and run in parallel. According to the particular usage scenario, the ML model training location depends on the computational complexity, the data availability, the response time requirements, and the type of ML model.

The native programmability and openness of O-RAN architecture allows the NG-RAN design to exploit the concept of continuous integration and continuous delivery (CI/CD). Such an approach aims at continuous software upgrades, by targeting the testing automation and the reliable code integration and modularity, while the network remains operational. Moreover, complementary to virtual machines, software containerization provides the ability to create isolated spaces within the same operating system for advanced automation. Dedicated RAN operations in the form of xApps can be modularized, separated, upgraded and delivered by various software developers. These xApps can operate as separate entities that interact via the open standardized interfaces, and can be effortlessly uninstalled, upgraded, or replaced [12].

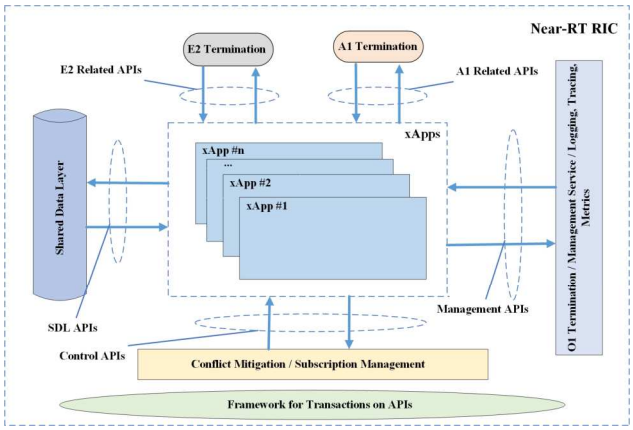

Figure 2. Embedding xApps in near-RT RIC [13].

The near-RT RIC can host multiple xApps running in parallel on APIs, as shown in Fig. 2. The xApps can connect through the A1 and E2 interfaces and manage RAN operations using ML models. The outputs of xApps are coordinated by utilizing conflict mitigation and subscription management. The operation of xApps requires coordination for timing demands, and monitoring for proper configuration, balanced resource processing, and message handling. The utilization of xApps permits the non-RT RIC to adopt a modular approach, as each xApp contains a different ML model. The xApps can either

gather data for further processing or apply changes in network function [13].

The main stages, which need to be followed from any AI/ML-driven solution in the form of an xApp, include the following [14]: 1) model capability query, which is performed by the SMO in the initial execution or in an update of a model, focusing on the ML engine and availability of data; 2) model selection and training, where the ML training host starts the model training and returns the trained model to the non-RT RIC for execution; 3) ML inference host configuration, which utilizes the model description file and online data; and 4) corresponding actions using the related actors, which is based on the result of the model inference. Depending on the location of the ML inference, as well as the actors and type of actions, different interfaces (O1, A1 and E2) are employed.

5. Conclusions

The O-RAN concept will usher in a new era in communications, opening new opportunities for emerging technologies and industries. O-RAN can be seen as a catalyst for creating new business models that will cut expenses, enhance corporate efficiency, and facilitate more innovation. In this paper, we have provided an overview of the concept of O-RAN and analyzed its architecture and main building blocks. Moreover, we have studied the key role of QoE in FWNs realization, and highlighted the significance of integrating QoE-awareness in O-RAN's architecture. Finally, we have analyzed the methodology of embedding AI/ML-based optimization models in O-RAN's design, and presented a practical guide for deploying xApps.

6. Acknowledgements

This research was supported by the European Union, partially through the Horizon 2020 Marie Skłodowska-Curie Innovative Training Networks Programme "Mobility and Training for beyond 5G Ecosystems (MOTOR5G)" under grant agreement no. 861219, and partially through the Horizon 2020 Marie Skłodowska-Curie Research and Innovation Staff Exchange Programme "Research Collaboration and Mobility for Beyond 5G Future Wireless Networks (RECOMBINE)" under grant agreement no. 872857.

References

[1] O-RAN Alliance, "O-RAN: Towards an Open and Smart RAN," Oct. 2018.

[2] "Toward Next Generation Open Radio Access Network--What O-RAN Can and Cannot Do!," [Online] Available: arXiv:2111.13754, Nov. 2021.

[3] A. Garcia-Saavedra and X. Costa-Perez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," IEEE Communications Standards Magazine, pp. 1-8, Oct. 2021, doi: 10.1109/MCOMSTD.101.2000014.

[4] D. Wypiór, M. Klinkowski and I. Michalski, "Open RAN—Radio Access Network Evolution, Benefits and Market Trends," Applied Sciences, vol. 12, no. 1, Jan. 2022, doi: 10.3390/app12010408.

[5] S. Kukliński, L. Tomaszewski and R. Kołakowski, "On O-RAN, MEC, SON and Network Slicing integration," in In Proc. 2020 IEEE Globecom Workshops (GC Wkshps), Taipei, Taiwan, Dec. 7-11, 2020, doi: 10.1109/GCWkshps50303.2020.9367527.

[6] S. K. Singh, R. Singh and B. Kumbhani, "The Evolution of Radio Access Network Towards Open-RAN: Challenges and Opportunities," in In Proc. 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Seoul, Korea (South), Apr. 6-9, 2020, doi: 10.1109/WCNCW48565.2020.9124820.

[7] O-RAN Alliance, "O-RAN Minimum Viable Plan and Acceleration towards Commercialization," Jun. 2021.

[8] A. Perveen, R. Abozariba, M. Patwary and A. Aneiba, "Dynamic traffic forecasting and fuzzy-based optimized admission control in federated 5G-open RAN networks," Neural Computing and Applications, Jun. 2021, doi: 10.1007/s00521-021-06206-0.

[9] O-RAN Alliance, "O-RAN Use Cases and Deployment Scenarios," Feb. 2020.

[10] R. Ferrús, O. Sallent, J. Pérez-Romero and R. Agustí, "Applicability Domains of Machine Learning in Next Generation Radio Access Networks," in In Proc. 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, Dec. 5-7, 2019, doi: 10.1109/CSCI49370.2019.00203.

[11] O-RAN Alliance, "AI/ML workflow description and requirements," Oct. 2020.

[12] M. Dryjański, Ł. Kułacz and A. Kliks, "Toward Modular and Flexible Open RAN Implementations in 6G Networks: Traffic Steering Use Case and O-RAN xApps," Sensors, vol. 21, no. 24, Dec. 2021, doi: 10.3390/s21248173.

[13] P. H. Masur and J. H. Reed, "Artificial Intelligence in Open Radio Access Network," [Online] Available: arXiv:2104.09445, Apr. 2021.

[14] S. Niknam et al., "Intelligent O-RAN for Beyond 5G and 6G Wireless Networks," [Online] Available: arXiv:2005.08374, May 2020.