THE EVOLUTION OF OPEN RAN

A 5G Americas White Paper

Feb. 2023

Contents

Executive Summary

This paper provides an update to 5G Americas 2020 white paper *Transition Toward Open and Interoperable Networks* [1] on the recent advancements in Open Radio Access Network (RAN) standards and the evolution in Open RAN trials and deployments.

More specifically, it focuses on:

- Interoperability offered by Open RAN systems – Dealing with open interfaces between different network functions, so as to achieve multi-vendor interoperability and coexistence.
  - » Related recent enhancements in the Open RAN standardization process, such as the O-RAN Alliance and the Telecom Infra Project
- Scalability in Open RAN systems – Dealing with cloudification of O-RAN
  - » Related recent enhancements in virtualization and cloudification of O-RAN network functions, and cloud-native RAN deployments
  - » Perspectives from brownfield and greenfield operators
- Performance offered by Open RAN systems:
  - » The role of artificial intelligence (AI)/machine learning (ML), reinforcement learning and analytics in realizing Open RAN use cases
  - » Enhancements in the standardization of RAN Intelligence Controller (RIC) functions.

1. Open RAN and Goals

An introduction to Open RAN is provided in the 5G Americas white paper, *Transition Toward Open and Interoperable Networks*. [1]

1.1 Principles of Open RAN

In brief, the Open RAN systems adopt the principles of:

- Openness: The interfaces between different functions or logical nodes in O-RAN architecture are open interfaces in order to achieve multi-vendor interoperability and coexistence across the functions.
- Virtualization: The network function implementations in O-RAN architecture are migrated from vendor-proprietary hardware to commercial-off-the-shelf cloud platforms running on whitebox hardware.
- Intelligence: The control-plane (C-plane), user-plane (U-plane) and management-plane (M-plane) functionalities of the RAN functions are subject to optimization by third-party solutions deployed in a new centralized controller function, called the RIC, that performs closed-loop control of the RAN functions over open interfaces. These solutions leverage data-driven analytics and advanced AI and ML techniques to efficiently learn intricate inter-dependencies and complex cross-layer interactions between parameters across the layers of the RAN protocol stack towards optimizing radio resource management (RRM) decisions at finer user equipment (UE)-level granularities, which cannot be captured by traditional RRM heuristics.
- Programmability: The objective targets for optimization are programmatically configured and adapted using AI/ML-driven declarative policies, based on continuous monitoring of network and UE performance. Furthermore, the ML models for training and inference are updated using life cycle management to adapt to dynamics in the network, load and traffic conditions.

1.2 Ecosystem Survey and Implications

This section talks about standardization efforts for rapid evolution of O-RAN systems.

1.2.1 2.2.1 O-RAN Alliance

A primary description of the O-RAN Alliance is given in [1]. It is responsible for defining the standardization of the Open RAN systems by the O-RAN Alliance.

1.2.1.1 O-RAN Alliance Work Group Structure

Work within the O-RAN Alliance is split and streamlined into several different work groups, a summary of which is provided here:

WG1 – Use Cases and Overall Architecture Workgroup

This working group [n2 – 4, 27] is responsible for defining the O-RAN architecture and identifying O-RAN use cases, and architecture-specific task groups, such as the ones responsible for defining the slicing architecture in the context of O-RAN.

WG2 – Non-Real-Time RIC and A1 Interface Workgroup

This working group [5 – 10, 28] is responsible for defining the Non-RT RIC architecture, the A1 interface between the Non-RT RIC and the Near-RT RIC, and the R1 interface between the rApp and the Non-RT RIC/Service Management and Orchestration (SMO) framework functions. It also defines the interface-specific application protocols, the use-case-specific policies, optimization objectives, enrichment information and Operations, Administration and Maintenance (OAM) functionalities.

WG3 – Near-Real-time RIC and E2 Interface Workgroup

This working group [11 – 16] is responsible for defining the Near-RT RIC architecture, the E2 interface between the E2 node and the Near-RT RIC, and the xApp APIs between the xApp and the Near-RT RIC framework function. It also defines the interface-specific application protocols and the use-case specific service models.

WG4 – Open Fronthaul Interfaces Workgroup

This working group [17 – 18] is responsible for defining the open fronthaul interface between the O-Distributed Unit (O-DU) and O-Radio Unit (RU) for Control, User and Synchronization (C/U/S) plane protocols, Management (M) plane protocols, and Multi-vendor IOT specifications, supporting both LTE and 5G NR systems. WG4 also standardizes the hierarchical M-plane and the hybrid M-plane models for the O-RU.

WG5 – Open F1/W1/E1/X2/Xn Interface Workgroup

This working group [72 – 75] is responsible for refining the definitions of 3GPP's F1, Xn, X2, E1 interfaces and OAM M-plane interface for supporting multi-vendor interoperability.

WG6 – Cloudification and Orchestration Workgroup

This working group [19 – 23] is responsible for defining O-Cloud infrastructure and deployment management principles of the O-Cloud infrastructure, and the OAM of

the O-Cloud infrastructure over the O2ims, O2dms and O2 interface.

WG7 – White-box Hardware Workgroup

The goal of Work Group 7 [76] is to specify and release the complete hardware reference design of a high performance, spectral and energy efficient whitebox base station.

WG8 – Stack Reference Design Workgroup

The goal of Work Group 8 [77] is to develop the software architecture, design, and release plan for the O-CU and O-DU based on O-RAN and 3GPP specifications for the NR protocol stack.

WG9 – Xhaul Transport

WG9 is focused on the transport domain [78] – consisting of transport equipment, physical media, and control/ management protocols associated with the transport network underlying the assumed Ethernet interfaces (utilized for fronthaul, mid-haul and backhaul).

WG10 – OAM for O-RAN

The WG10 is a new working group [24 – 25], created out of WG1, that focuses on OAM architecture and the OAM interface and procedures for the O-RAN network functions. Work Group 10 objectives include:

1. Develop O-RAN O1 OAM Information and data models, using industry developed baselines where available, and adding further material as needed by O-RAN's architecture and interfaces.

2. Develop O1 interface specifications to O-RAN elements consistent with the principles outlined in O-RAN whitepaper and elaborated in Architecture Description Document.

3. Develop a detailed OAM architecture consistent with the Architecture Description Document, including key management interfaces and deployment options: Develop the role of the O1 interface in Control and Management loops of O-RAN architecture, in conjunction with other Work Groups.

4. Provide coordinated definition and collection of O1 KPIs and Performance Measurement (PM); Fault Management (FM) across all WGs.

WG11 – Security Work Group

The WG11 is a new working group [26] that is responsible for specifying O-RAN security requirements and drives security requirements across each of the working groups. O-RAN is striving towards a zero-trust architecture [79] to protect against internal and external threats towards achieving a strong security posture for O-RAN implementations. The threat risks are identified in [26,67,68] due to O-RAN WG11's threat modeling process, and are attributed to the additional functions and interfaces, cloud deployments based on the O-Cloud platform defined in O-RAN architecture. O-RAN WG11 shall continue its ongoing analysis to identify additional threats, risk, and security controls.

WG11 is currently addressing ten security work items to ensure O-RAN is secure. These work items include (i) SMO Security, (ii) Non-RT-RIC Security, including rApps, (iii) Near-RT-RIC Security, including xApps, (iv) Open Fronthaul Security, (v) O-RU Centralized User Management, (vi) O-Cloud Security, (vii) Application Lifecycle Management, (viii) Certificate Management Framework, (ix) Security Logging, (x) Security Test Cases.

As the O-RAN architecture continues to evolve, WG11 will be addressing security aspects of shared O-RU, Decoupled SMO, and AI/ML. WG11 is collaborating with multiple working groups to ensure O-RAN cloud implementations, including Hybrid Cloud deployments, are secure.

The four focus groups are: (i) Open Source Focus Group (OSFG), responsible for developing open-source software of O-RAN network functions, (ii) Standard Development Focus Group (SDFG), responsible for advancing standardization of O-RAN systems, (iii) Test and Integration Focus Group (TIFG), responsible for interoperability and interface compliance, (iv) Next Generation Research Group (nGRG), responsible for research involving O-RAN and next generation telecom systems [80].

1.2.1.2 O-RAN Software Community

O-RAN Software Community (OSC) [61] is a collaboration between the O-RAN Alliance and Linux Foundation with the mission to support the creation of software for the RAN. OSC uses O-RAN specifications while leveraging other LF network projects, to address the challenges in performance, scale, and 3GPP alignment, and to enable rapid development and deployment of O-RAN-based systems:

- near-real-time RAN intelligent controller (Near-RT RIC) – with utility xApps,
- non-real-time RAN intelligent controller (Non-RT RIC) – with utility rApps,
- SMO
- cloudification and virtualization platforms, open central unit (O-CU),
- O-DU, and
- test and integration effort to provide a working reference implementation.

New features in the recent OSC releases (Release-F) include:

- Near-RT RIC xApps: Key Performance Indicator (KPI) monitoring, RAN control, Quality of Experience (QoE) predictor xApps involving E2SM-KPM and E2SM-RC service models for developing traffic steering, Quality-of-Service (QoS)/QoE and anomaly detection use-case functionalities.
- Near-RT RIC platform: Latest O-RAN WG3-standardized E2AP procedure implementations (such as E2 Node Configuration Update) and xApp API procedure (such as, REST APIs involving the Subscription Manager platform function) implementations.
- Non-RT RIC platform: Implementations of services related to the Service Management and Exposure function, and the Data Management and exposure function of the Non-RT RIC platform, and the associated procedures involving data fetch and data delivery.
- OAM O1 interface: Implementation of YANG models based on O-RAN WG10-defined Network resource model for configuration management (CM) over O1 and the associated O1-CM procedures, and automated test cases validating the end-to-end message flows over O1 and open fronthaul M-plane interfaces, implementation of topology generator and reader, and providing abstract topology for rApps involving the O-RAN interfaces across O-RAN NFs.
- O-DU: Implementation of HARQ framework support, scheduler enhancement, inter-DU handover support, idle mode paging, E2AP support on O-DU, massive MIMO and ultra-reliable low-latency communication (URLLC) functionalities.
- SMO: Implementation of O1 VES interface for alarms and PM counters, and O2 interface for Virtual Network Function (VNF)/CNF deployment and infrastructure management.

1.2.1.3 O-RAN Testing and Integration Centers

O-RAN Testing and Integration Centers were developed by the TIFG [60]. towards facilitating O-RAN community interface conformance and interoperability testing and to drive the ecosystem towards O-RAN compliant solutions. OTICs are being deployed globally initially across Asia, Europe and North America, and organize annual plugfests, where companies demonstrate key O-RAN capabilities and use cases in the areas of openness and intelligence.

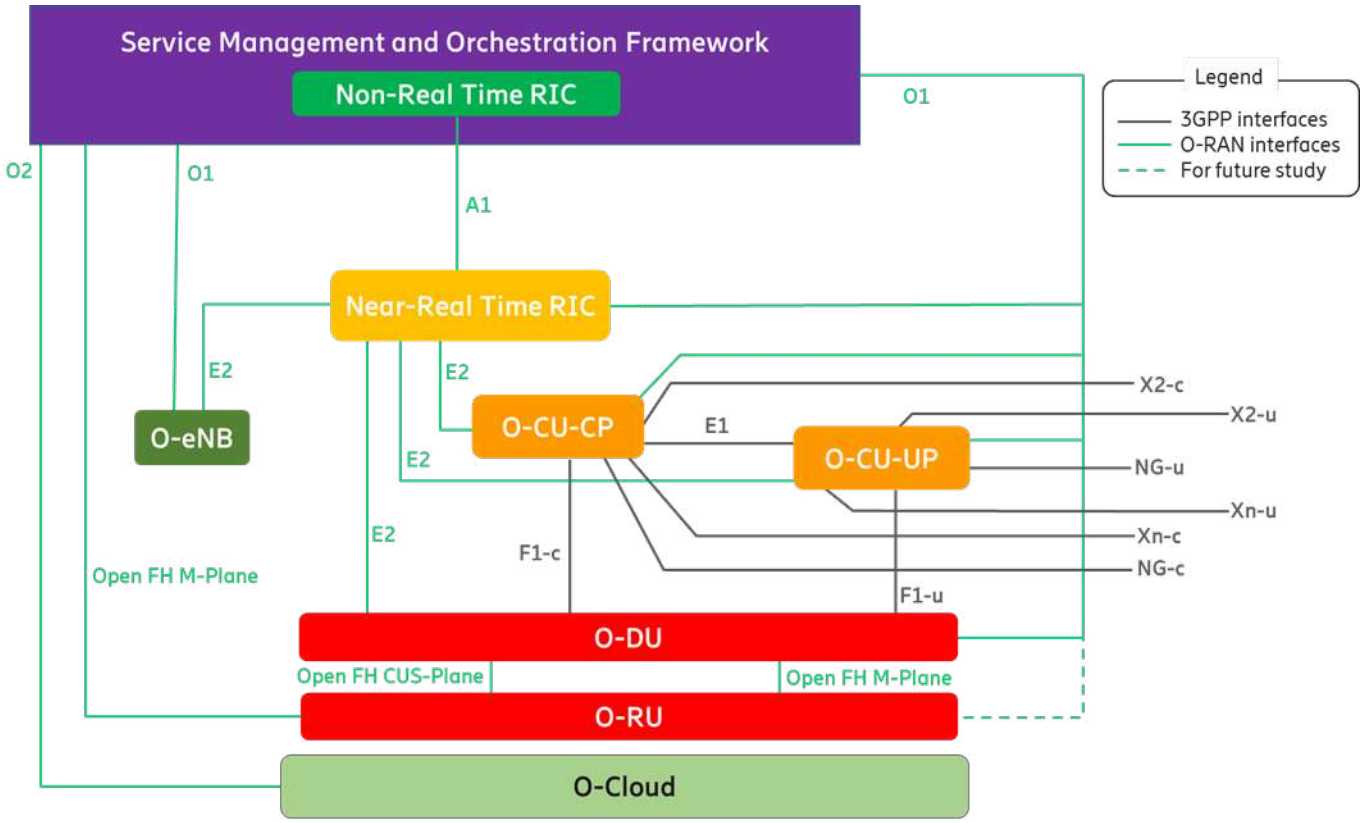
1.2.1.4 O-RAN Alliance Architecture

The interfaces between the different network functions are open and standardized, to achieve multi-vendor interoperability. Interfaces defined as open interfaces in O-RAN Alliance, which is the standardization body for Open RAN systems [2], include:

- 3GPP-defined interfaces between disaggregated RAN Network Functions, defined as open interfaces, in O-RAN [2, 72 – 75]:
  - F1-C between CU-CP and DU
  - F1-U between CU-User Plane (UP) and DU
  - E1 between CU-CP and CU-UP
  - X2-C between two eNBs and/or between an eNB-O-CU-CP pair for control-plane signaling
  - X2-U between two eNBs and/or between an eNB-O-CU-UP pair for user-plane signaling
  - Uu interface between the UE and the RAN
- New O-RAN-defined open interfaces between O-RAN NFs [2, 5 – 10, 11 – 14]:
  - E2 between Near-RT RIC and CU-CP, Near-RT RIC and CU-UP, Near-RT RIC and DU, Near-RT RIC and eNB
  - O1 between SMO and the CU-CP, SMO and CU-UP, SMO and DU, SMO and Near-RT RIC, SMO and eNB
  - Open fronthaul M-plane interface between SMO and O-RU
  - O2 between SMO and O-Cloud platform
  - A1 interface between Non-RT RIC and Near-RT RIC
  - Open fronthaul Control User Synchronization (CUS)/ M-plane interface between O-DU and O-RU
  - Y1 interface for exposure of analytics information from Near-RT RIC to authorized consumers.
- O-RAN-defined APIs in O-RAN NFs [2, 9, 15, 16]:
  - xApp APIs between xApps and Near-RT RIC platform functions
  - R1 APIs between rApps and Non-RT RIC/SMO framework functions
  - SMOS APIs between SMOSs (currently being defined)
- Open hardware that uses standard processors (e.g., x86, ARM CPUs and GPUs) allowing software from different sources to run on them that uses standardized racks, chassis, power distribution, and cabling such as those from open19.org, Open Compute Project (OCP), etc. that has an open standard coherent accelerator processor interface.
- Open software that is commercially viable to meet high performing KPI requirements that support real-time system needs that leverages adjacent software communities such as Open Networking Automation Platform (ONAP) and other open approaches to utilize existing solutions to speed time to market.

- The RRM decisions of the C-plane and U-plane functionalities across the layers of the RAN protocol stack for individual UEs are split in a fine-grained manner between the E2 node and the Near-RT RIC function for UE-level call processing and UE-level data transfer.
- In addition, the RRM decisions of the M-plane functionalities pertaining to individual cells, managed elements and managed functions are split between the E2 nodes and the Near-RT RIC / Non-RT RIC functions. The RIC functions make configuration recommendations for cells and network elements.
- The KPI optimization objectives for RRM are configured for individual UEs or UE groups or slices or cells in a programmable manner from Non-RT RIC to Near-RT RIC.

Figure 1-1 :Logical O-RAN Architecture [2]



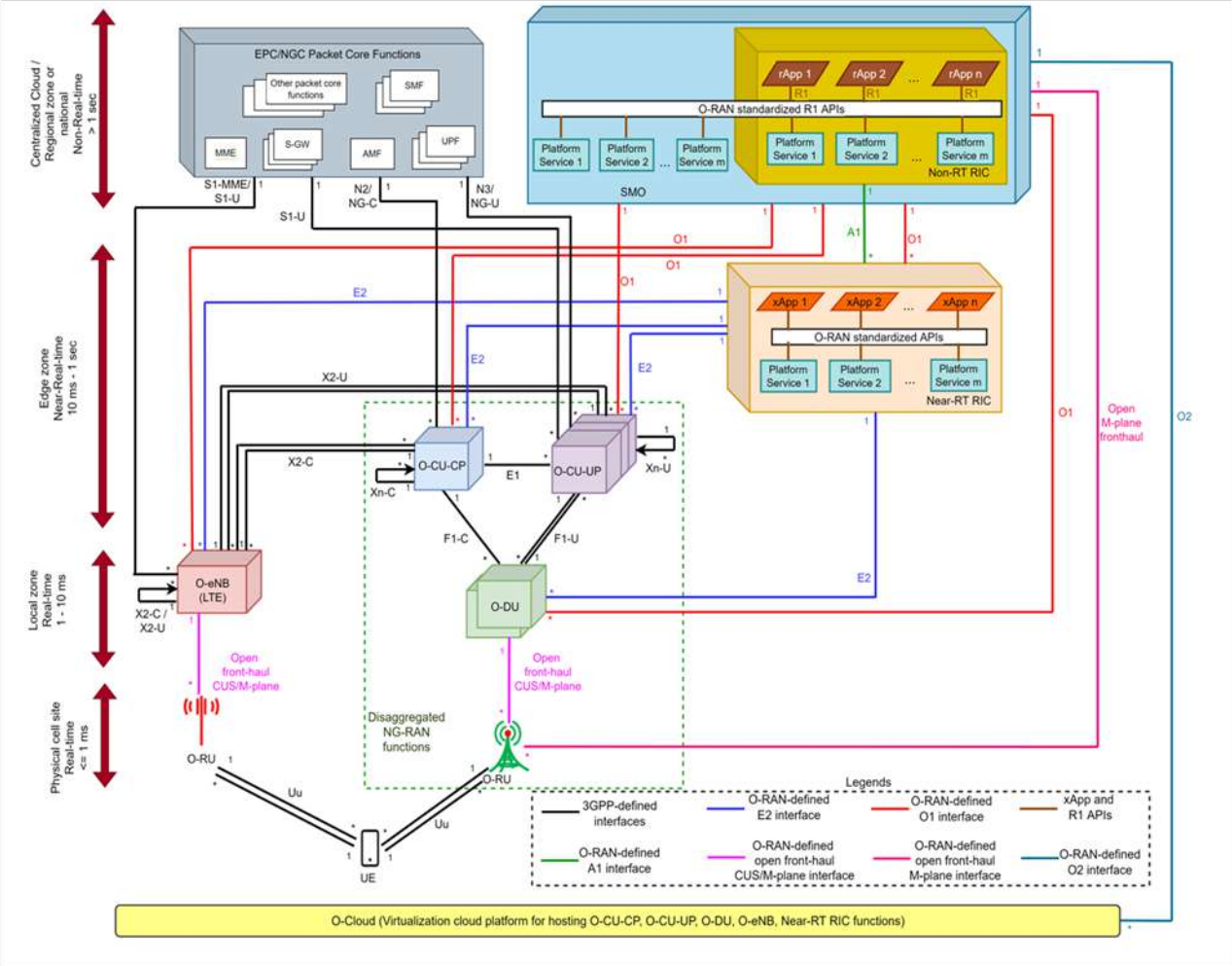
3rd party extensible applications (such as xApps in the Near-RT RIC and rApps in the Non-RT RIC) performing RAN optimization for O-RAN use cases leveraging AI/ML and deep learning tools and advanced analytics for fine-grained UE-level intelligence and optimization for individual UEs, bearers, cells, etc.

It is to be noted that the O-RAN-compliant RAN network functions, which support open and interoperable interfaces of Xn, X2, F1, E1, Uu, open fronthaul CUS/M-plane, E2, O1, etc. are prefixed by an "O-" and are referred to as O-CU-CP, O-CU-UP, O-DU, O-RU and O-eNB. The O-RAN architecture, as standardized by the O-RAN Alliance, [2] is shown in Figure 1-1 : Logical O-RAN Architecture. An informal holistic view of the O-RAN architecture with all the pertinent interfaces and network functions, along with the packet core functions, is shown in Figure 1-2.

The O-RAN architecture supports a Lower Layer Split (LLS) of the PHY layer functionalities, which was studied by 3GPP in [81], but which has not, at the time of writing, been specified by 3GPP.

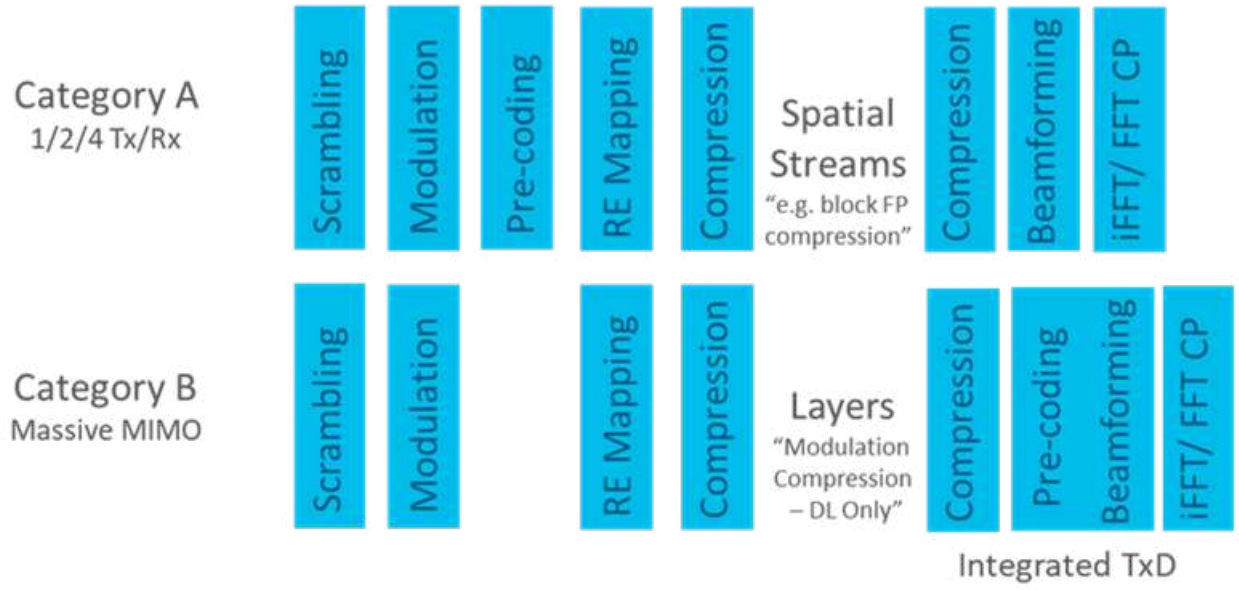
In particular, the earlier split PHY analysis (called Split 7) highlighted different variations of the Split 7 depending on what functions are located above the split point compared to those located below, with alternatives being referred to as Split 7-1, 7-2 and 7-3. O-RAN Alliance has adopted the previous conclusions of the xRAN Forum to standardize on a specific decomposition of functions between the O-RU and O-DU [17 – 18], termed 7-2x, as shown in Figure 1-3.

Figure 1-2: An informal view of the O-RAN network functions



This split can also be configured to operate in two distinct modes, termed Category A and Category B (shown in Figure 2-3). When operating in “Category A” mode of operation, the pre-coding and resource element mapping operate in the O-DU, resulting in the fronthaul interface being used to transport different spatial streams. Conversely, when operating in “Category B” mode of operation, the pre-coding functions are moved below the split, allowing the fronthaul interface to transport MIMO layers. In such a configuration, “modulation compression” can be used in the DL to effectively send only the bits equivalent to the constellation points, resulting in the bandwidth approaching that of alternative 7-3 splits. Using such an approach, a converged fronthaul interface can be used to support a variety of use cases, such as outdoor massive MIMO.

Figure 1-3 : O-RAN Split 7-2x modes of operation [17]



1.2.1.5 Recent Advancements in O-RAN Standards

This section discusses some key recent advancements in the O-RAN standards over the past couple of years.

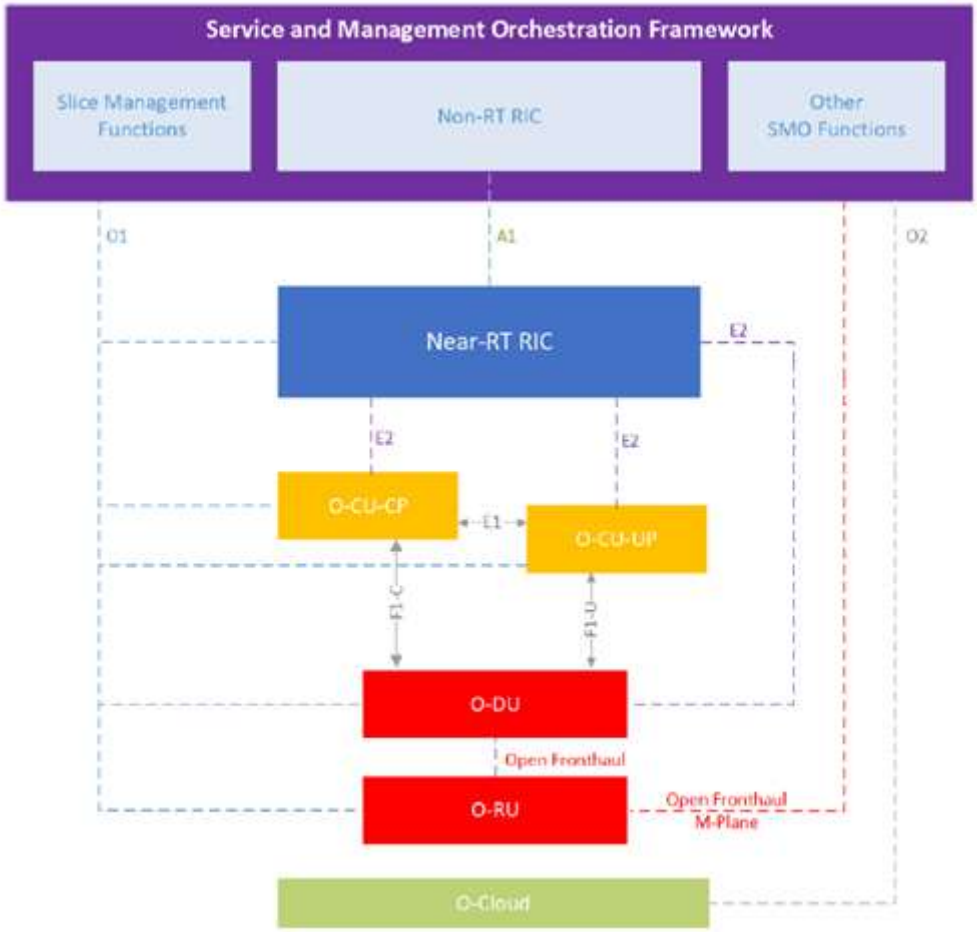
1. Services-based architecture for the SMO [27]: The decoupled SMO architecture in O-RAN WG1 presents the SMO as a set of SMO functions offering SMO services, which are a standardized cohesive set of management, orchestration and automation capabilities. The SMO functions, including the Non-RT RIC, play the role of service producers and service consumers, offering and consuming services over the SMO services-based interface. This enables interoperability and interfacing between multi-vendor solutions for the SMO functions.

2. Services-based architecture for the Non-RT RIC and R1 interface [9,10]: The Non-RT RIC architecture in O-RAN WG2 presents a services-based architecture, where the Non-RT RIC/SMO framework offers a bundle of services that shall be produced by the Non-RT RIC/SMO framework functions (acting as service producers) and that shall be consumed by the rApps (acting as service consumers) over the R1 services-based interface. rApps, as extensible applications from 3rd parties responsible for generating declarative policies and KPI targets, setting RRM objectives for RAN functionalities to the Near-RT RIC, and recommending network element configurations for OAM operations over R1, shall inter-operate with the Non-RT RIC/ SMO framework functions over the R1 services-based interface, being standardized by O-RAN WG2.

3. Service models and type definitions for E2 and A1 interface: O-RAN WG3 has standardized new service models, such as E2 Service Model – RAN Control (E2SM-RC), E2 Service Model – Key Performance Monitoring (E2SM-KPM), E2 Service Model – Cell Configuration and Control (E2SM-CCC) [13 – 14, 82 – 83] for implementing Near-RT RIC services over the E2 interface towards the realization of O-RAN use cases such as traffic steering, QoS, network slicing, massive MIMO, etc. [3]. Similarly, O-RAN WG2 has also standardized new type definitions for generating policies towards realization of traffic steering, QoS, network slicing and massive MIMO O-RAN use cases [7].

4. Cloudification and orchestration over the O2 interface: The O2 is an open logical interface within the O-RAN architecture for communication between the SMO and O-Cloud for management of O-Cloud infrastructure and the deployment life cycle management of O-RAN cloudified network functions that run on O-Cloud [19, 20]. The functions to be performed over the O2 interface include: (i) O-Cloud Infrastructure Resource Management for managing O-Cloud infrastructure and platform resources [21], (ii) O-Cloud Deployment Resource Management and Orchestration for managing the O-Cloud deployment [22], (iii) O-Cloud OAM for Fault, Configuration, Accounting, Performance, Security (FCAPS) management of the O-Cloud infrastructure instance [19 – 20].

Figure 1-4 : Slice management functionality in the SMO [4]



5. Hierarchical and hybrid M-plane for the open fronthaul interface: O-RAN WG4 has standardized two options for the OAM management of O-RU [18]. In the hierarchical M-plane, the OAM management for the O-RU happens via the O-DU. The SMO uses O1 interface for exercising the O-RU-related FCAPS operations on the O-DU, which later uses the M-plane fronthaul interface with the O-RU. In the hybrid M-plane, The OAM management for the O-RU is exercised from the SMO directly via the open fronthaul M-plane interface to the O-DU.

6. Near-RT RIC APIs: O-RAN WG3 has recently standardized APIs [15, 16] to enable interoperability and integration between 3rd party xApps, which are responsible for fine-grained RRM of C-plane, U-plane and M-plane functionalities of the O-RAN network functions over the E2 interface at near-real-time granularities using low-latency control loops, and the Near-RT RIC platform functions.

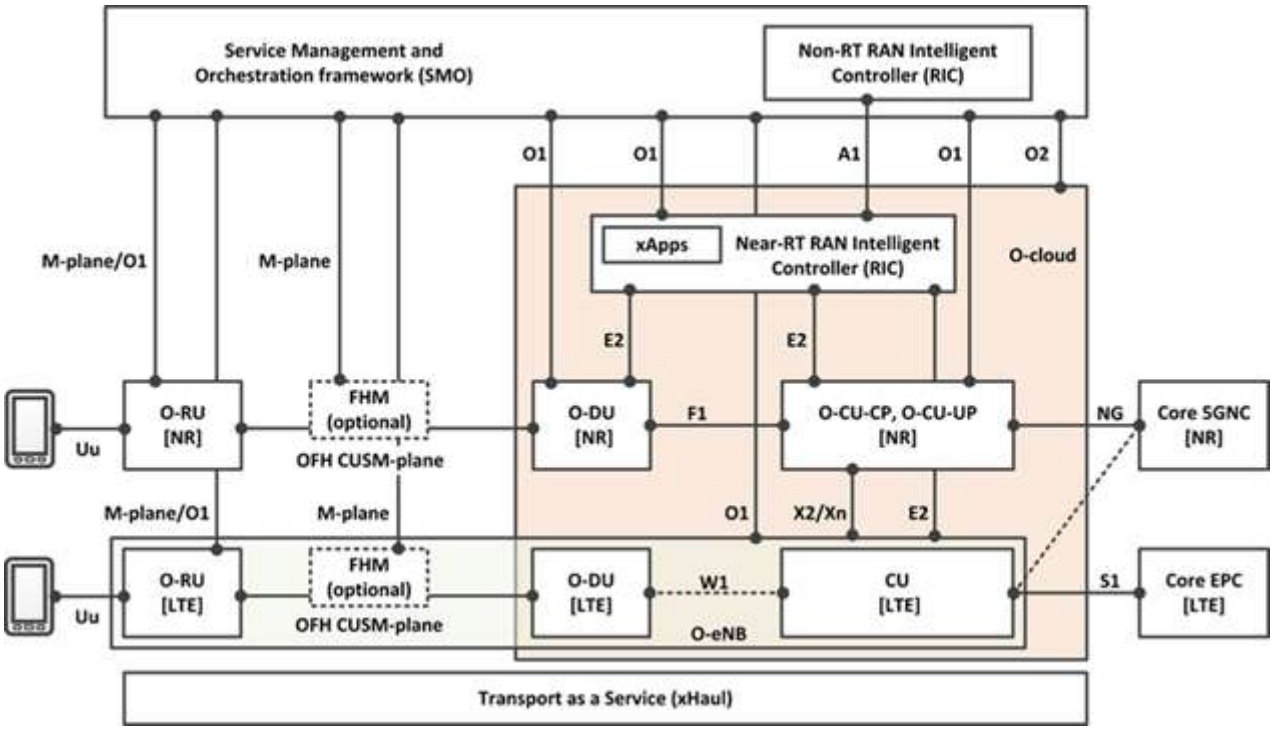
7. New interface for RAN analytics information exposure: Fine-grained RAN analytics exposed by the Near-RT RIC platform over the new Y1 services-based interface [2, 15] for consumers such as packet core functions or application servers or edge servers have a wide variety of use cases, such as the following:

- Predicted UE-level RAN throughput, exposed by the RIC, when consumed by an application server such as a video streaming server, shall enable the server towards proactively and intelligently optimizing the video bitrate resolution for the streaming video content to the UE that minimizes buffering, stalling, etc., thereby subsequently improving the average QoE of the video streaming UEs.
- Predicted UE-level RAN latency, exposed by the RIC, when consumed by an application server such as a 360-degree virtual reality (VR) streaming server, shall enable the server towards proactively and intelligently optimizing the IP packet size for the VR content that minimizes screen freezing, screen blackouts etc., thereby subsequently improving the average QoE of the VR UEs.
- Relevant details are subsequently discussed in Section 2.

8. Network slice subnet management functionality in SMO: The RAN-specific network slice subnet management functionality that produces network slice subnet management services (NSSMS) for the Network Slice Subnet Instances (NSSIs) in the RAN domain is realized in the SMO, as figured in Figure 1-4 [4, 58]. The NSSMS producer produces CM, PM and other FCAPS services for the RAN-specific NSSIs [58].

1.2.1.6 O-RAN Plugfests

Figure 1-5: O-RAN architecture and interfaces reference diagram for O-RAN PlugFest 2021 [60]



Most recently, the O-RAN Alliance successfully conducted its O-RAN Global PlugFest 2021 [60] to demonstrate the functionality and the multi-vendor interoperability of O-RAN based network equipment. The plugfest demonstrated the strength of the O-RAN ecosystem and its global drive towards open and intelligent RAN systems. The O-RAN PlugFest expanded from 4 to 7 global venues, with 94 participating companies. Many of the companies contributed to multiple venues, bringing PlugFest to a total of 144 active corporate participants compared to 70 at the 2020 PlugFest. The O-RAN reference architecture towards testing the integration of O-RAN systems and their interfaces from a compliance standpoint is shown in Figure 1-5. The key highlights of the O-RAN PlugFest 2021 are:

- The plugfest proved advanced maturity of the Open Fronthaul (OFH) implementations. Interoperability has been achieved between many vendors in different network setups, base station classes, OFH profiles and RU/CU-DU product combinations.
- The plugfest presented several demonstrations of advanced use cases utilizing the O-RAN Near-Real-Time Radio Intelligent Controller (Near-RT RIC) and Non-Real-Time RIC (Non-RT RIC), like automated network outage detection and recovery, and latency assurance for end-to-end network slicing. A lot of effort also went into testing individual RIC interfaces, application protocols, and related xApps and rApps.
- Several venues across Asia, South Asia, Europe, North America, etc. successfully tested O-Cloud products and multi-vendor virtualized RAN integrations.
- Specific tests dealt with the O-RAN infrastructure security, and several O-RAN end-to-end functionality tests passed against production core network elements, while many other utilized simulators.
- The venues proved the readiness of advanced test equipment and simulation of different parts of the network.

In Asia, the plugfest took place at four venues:

- Plugfest in Japan: It showcased multi-vendor Interoperability Testing with Fujitsu, NEC, Nokia, Altiostar and Samsung products using Open Fronthaul, both SA and NSA setup, on Sub6 GHz NR TDD, for Open Fronthaul M-plane and

CUS-plane, call processing and performance evaluation involving multi-vendor O-CU/O-DU and O-RU systems.

- Plugfest in Korea, where the host was LG Uplus, and the participants included Altiostar, Intel, NEC, Keysight technologies, etc. It showcased multi-vendor interoperability evaluation involving the open fronthaul 7.2x interface.
- Plugfest in Taiwan, where the host was Chungwa telecom: Similar to the plugfest to Korea, the plugfest in Taiwan also showcased multi-vendor interoperability evaluations involving the open fronthaul 7.2x interface.
- Plugfest in India, where the host was Airtel and the participants included Mavenir, VMware, Intel, STL, ASOCS, Capgemini, etc. - It showcased demonstration of Near-RT RIC traffic steering use case with E2 service models, APIs and O-RAN AI/ML life cycle management by Mavenir, VMware, Capgemini, TCS, Viavi etc., multi-vendor interoperability evaluation over the open fronthaul M-plane interface involving STL, Keysight, etc.

In Europe, the plugfest took place at two venues:

- Plugfest in Russia: The host was Skoltech and the participants included Foxconn, Keysight and Xilinx, and the demonstrations included testing of multi-vendor interoperability between the O-DU and O-RU NFs over the open fronthaul M-plane interface, multi-vendor interoperability between O-CU and O-DU NFs with commercial 5G SA packet core integration
- Joint European O-RAN and TIP Plugfest: The hosts included British Telecom (BT), Deutsche Telekom, Orange, Telefónica, TIM and Vodafone, and the participants included Mavenir, VMware, Radisys, Dell, Intel, Juniper, NEC, etc. The plugfest demonstrated Near-RT RIC and its interfacing with O-CU/O-DU involving the E2 interface and the interfacing between the xApps and the Near-RT RIC platform functions involving the xApp APIs, the interfacing between the SMO/Non-RT RIC and the O-CU/O-DU functions involving the O1 interface, the interoperability and interfacing between the Non-RT RIC and Near-RT RIC involving the A1 interface, demonstration of RAN slice service level agreement (SLA) assurance and mobile load-balancing xApps, etc.

In North America, the plugfest took place in the USA. The plugfest was jointly hosted by AT&T and Verizon, and the demonstrations included O-Cloud infrastructure behavior in latency sensitive applications, demonstration of successful E2AP and E2SM-KPM service model involving the RIC, RAN slice assurance xApp and AI-enabled management of multi-vendor RAN with O-RU pooling and multi-vendor slices. The participants included Juniper, Mavenir, NEC, Radisys, Intel, Viavi, etc.

1.2.2 Telecom Infra Project (TIP)

TIP was formed in 2016 as an organization that is focused on collaboration and the development of new technologies for building and deploying global telecom network infrastructure to enable access for everyone in the world [62].

There are over 500 members which include operators, suppliers, developers, integrators, and other entities. The TIP board of directors is composed of individuals from the founding tech and telecom companies. Member companies host TIP community labs, and TIP hosts an annual TIP Summit.

Within TIP, there are project groups working on different network concepts. Below is a list of project groups dedicated to the area of Open RAN platforms:

1. OpenRAN 5G NR

- The goal of the OpenRAN 5G NR Project Group [64] is to collaboratively design an open interfaced, multi-vendor interoperable, disaggregated whitebox platform for a 5G NR access point that is easy to configure, scale and deploy. The solution includes a 5G NR compatible baseband unit; antenna and radio, and the provisioning elements. The focus of this TIP project group is on use cases for outdoor macrocells and small cells as well as indoor small cells.

2. TIP OpenCellular

- The goal of this project group [84] is to provide pervasive connectivity, especially to underserved areas, with tools to build and operate sustainable cellular infrastructure using open-source technologies and an open ecosystem. The aims are to achieve its mission by providing an open-source platform to build, deploy, and operate complete (E2E) cellular networks. The OpenCellular platform has been deployed by multiple service providers in various African and Latin American countries. Africa Mobile Networks, as part of MTN and Orange, have installed sites covering over 300,000 people in sub-Saharan African countries.

3. PlugFest / Test and Integration Project Group

- The TIP PlugFest group, now known as the Test and Integration Project Group [85], was launched in 2019. The mission of the project group is to define and accelerate the development of test materials, test plans and other documents for testing compliance and interoperability of OpenRAN systems. In 2020, TIP had the O-RAN joint plugfest for the first time to showcase multi-vendor interoperability tests of O-RAN solutions. The testing occurred in the labs hosted by Deutsche Telekom (DT) and TIM and on behalf of BT, Orange, Telefónica, DTAG, and TIM. The test was focused on Open Fronthaul functionality and compliance, E2E performance and O1 management interoperability.

4. OpenRAN

- This main objective of the group is the development of fully programmable RAN solutions based on

General-Purpose Processing Platforms (GPPP), disaggregated software and open interfaces. It complements existing TIP projects and focuses on disaggregation of virtualized RAN solutions into different components and ensuring each individual component can be efficiently deployed on GPP platforms, in terms of (i) reference framework/ architecture for implementation of the eNB stack on GPPPs, (ii) reference (and optimized) implementation of the basic building blocks and algorithms, both as software libraries and FPGA register-transfer levels (RTL), (iii) hardware abstraction layer, including APIs, to abstract from application vendors the underlying hardware platform capabilities, (iv) defined KPIs and traffic model as part of the reference implementation, (v) orchestration framework to manage and provide operational capabilities, (vi) carrier-grade lab proof-of-concept evaluation of multi-vendor open solutions. The TIPOpenRAN project group has multiple subgroups [62]. Component subgroups include (i) the RU subgroup – with the goal to develop and build an RU whitebox hardware based on open and disaggregated architecture with single band, multi-band and mMIMO RU deployment options, (ii) DU/CU subgroup – to enable and accelerate development and deployment of whitebox DU/CU with open disaggregated architecture enabling multi-vendor interoperability, (iii) RIA subgroup – to enable an ecosystem that leverages AI/ML and data science technology to improve RAN performance for use cases such as RAN coverage, capacity, interference mitigation, massive MIMO, energy savings, etc., (iv) ROMA subgroup – that defines technical specifications for OpenRAN lifecycle management, automation and orchestration, (v) the Outdoor and indoor subgroups - to address the challenges of large-scale, outdoor OpenRAN deployment and to enable the development of open interface, cost effective indoor small cell whitebox systems for indoor coverage respectively, by defining requirements and aggregating technology solutions.

5. TIP System Integration and Site Optimization

- This group [85] focuses on cost analysis for site engineering (site selection and setup), connectivity systems (wireless backhaul, satellite link and efficient antenna technologies), automated maintenance and optimization, system integration and business/revenue model (network infrastructure sharing, revenue-sharing model).

1.2.3 3GPP

As seen in Section 2.2.1.4 in Figure 2-1 and Figure 2-2, Open RAN systems require open interfaces between the elements of a disaggregated RAN: namely the CU-CP (O-CU-CP in O-RAN architecture), CU-UP (O-CU-UP in O-RAN architecture), DU (O-DU in O-RAN architecture) and the O-RU. 3GPP has the following specifications for the disaggregated RAN architecture and interfaces which form the basis of O-RAN architecture and interface specifications:

- 3GPP TS 38.401 [37] that talks about the split or disaggregated RAN architecture, which is mandated in O-RAN.
- 3GPP TS 38.470 [31] for F1 general architecture and principles, and TS 38.473 [32] that talks about the F1 interface application protocol for the control-plane signaling between the CU-CP and the DU.
- 3GPP TS 37.480 [29] for E1 general architecture and principles, and TS 37.483 [30] that talks about the E1 interface application protocol for signaling between the CU-CP and the CU-UP.
- 3GPP TS 38.420 [33] for Xn general architecture and principles, and TS 38.423 [34] that talks about the Xn interface application protocol for signaling between two CU-CPs.
- 3GPP TS 36.420 [35] for X2 general architecture and principles, and TS 36.423 [36] that talks about the X2 interface application protocol for signaling between an eNB and a gNB CU-CP.
- 3GPP TS 38.425 [86] for user-plane interface specifications.

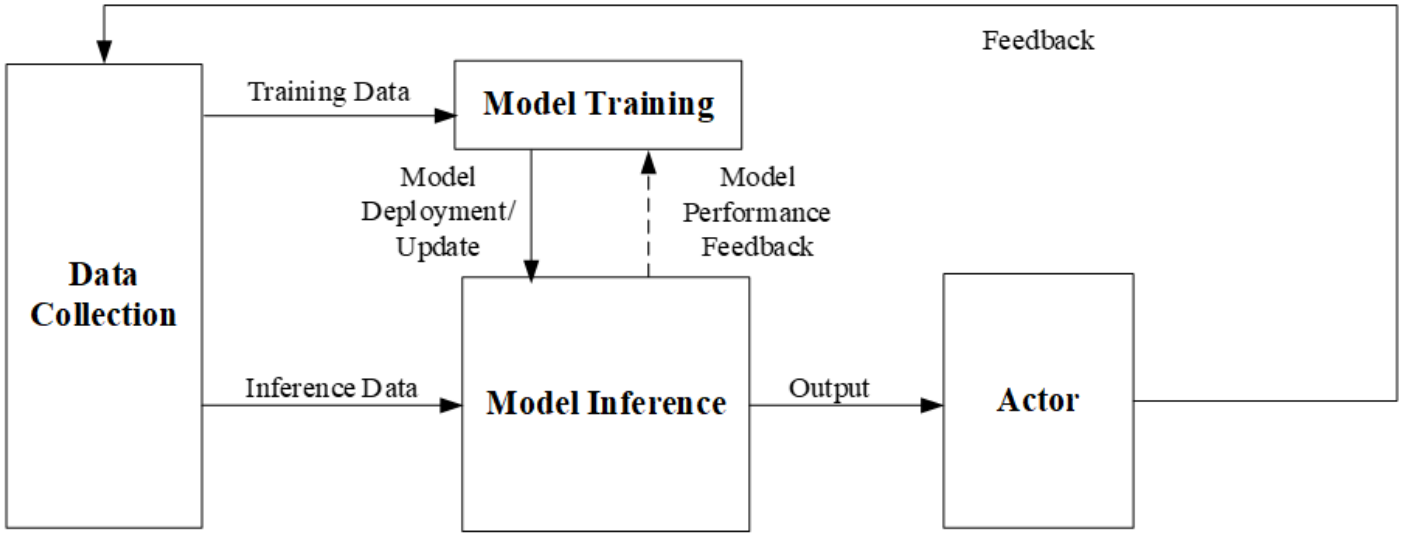
Moreover, due to recent advancements in AI/ML, 3GPP initiated multiple study and work items to incorporate AI/ML technology in cellular systems, including the introduction of Network Data Analytic Function (NWDAF) [87, 89] in the core network and Management Data Analytic Function (MDAF) [88] for OAM, and the latest Rel-17/18 work on applying AI/ML for RAN. The Rel-17 study on enhancement for Data Collection for NR and EN-DC [90] considered the basic AI/ML functional framework shown in Figure 2-6, and focused on input/output signaling requirements for three target use cases: network energy saving, load-balancing and mobility optimization. The study also investigated the message flows for two types of training deployment: training done by OAM and training done locally at RAN.

Following the Rel-17 study, there are more AI/ML for RAN activities in the ongoing Rel-18 effort. A work item on AI/ML for NG-RAN [41] will focus on updating 3GPP spec for AI/ML support. Rel-18 also works on enhancement in RAN data collection for SON (Self-Organizing Networks)/ MDT (Minimization of Drive Tests) [42] and QoE [43] that will benefit RAN AI algorithm development. In addition to previous focus on applying AI/ML to improve higher layer RAN control and management, there will be a new study in Rel-18 on AI/ML for NR air interface [91].

1.2.4 Open RAN Policy Coalition and U.S. Government Initiatives on Open RAN

The Creating Helpful Incentives to Produce Semiconductors and Science (CHIPS) Act of 2022 [92] was signed into law on August 9, 2022, to boost U.S. competitiveness,

Figure 1-6: Functional Framework for RAN Intelligence [41, 90, 91]



innovation and national security. The law aims to catalyze investments in domestic semiconductor manufacturing capacity. It also seeks to jump-start R&D and commercialization of leading-edge technologies, such as quantum computing, AI, clean energy, and nanotechnology, and create new regional high-tech hubs and a bigger, more inclusive science, technology, engineering, and math (STEM) workforce. The CHIPS and Science act includes $1.5 billion USD for promoting and deploying wireless technologies that use open and interoperable radio access networks and towards boosting U.S. leadership in wireless technologies and their supply chains.

During the U.S President Joe Biden's speech on 15th July 2022 in Jeddah, during his state visit to Saudi Arabia, the President remarked [48], "*…we concluded several new arrangements to better position our nations for the coming decades. Saudi Arabia will invest in new U.S.-led technology to develop and secure reliable 5G and 6G networks, both here and in the future, in developing countries to coordinate with the Partnership for Global Initiative — the Global Infrastructure and Investment, which I put together at the G7. This new technology solution for 5G, called Open RAN, will outcompete other platforms…*".

Other related U.S. Congress and Government initiatives concerning Open RAN are discussed in [44 – 47].

1.3 Architectural Considerations

This section discusses the key novel architecture principles of O-RAN, hierarchical and hybrid M-plane, disaggregation and interoperability, RAN virtualization etc.

1.3.1 Disaggregation and functional-split

The O-RAN architecture [2], based on the principles of disaggregation in RAN, is split into near-real-time (Near-RT) functions, and the real-time functions. The Near-RT functions include the O-CU-CP (responsible for RRC and Packet Data Converge Protocol [PDCP]-C) and O-CU-UP (responsible for SDAP and PDCP-U) which operate at a granularity of 10 ms to 1 second, while the real-time O-RAN functions include the O-DU (responsible for the Radio Link Control [RLC], Medium Access Control [MAC] and upper-PHY) and the O-RU (responsible for lower-PHY) that operate at a granularity of units of milli-seconds (TTIs). The O-eNB function with E-UTRA Radio Access Technology are not considered for disaggregation, though they support the LLS between the O-eNB baseband and the O-RU. The O-CU-CP, O-CU-UP, O-DU and O-eNB are referred to as the E2 nodes, since they support the E2 interface. The RAN optimization decisions for fine-grained C-plane and U-plane UE-level RRM functionalities are split with the Near-RT RIC over the O-RAN-defined E2 interface. The RRM decisions for the individual C-plane and U-plane RAN functionalities are exercised by 3rd party extensible applications, known as xApps, that are deployed in the Near-RT RIC. The E2 Service Model (E2SM) describes the functions in the E2 Node, which may be controlled by the Near-RT RIC and the related procedures, thus defining a function-specific RRM split between the E2 node and the Near-RT RIC.

They describe a set of services exposed by the E2 node that shall be subsequently used by the Near-RT RIC and the hosted xApps. These services provide the Near-RT RIC with access to messages and measurements exposed from the E2 node (such as cell configuration information, supported slices, PLMN identity, network measurements, UE Context Information, etc.), that enable control of the E2 node from the Near-RT RIC. Multiple E2SMs have been defined in O-RAN WG3 such as E2SM-RAN Control (E2SM-RC), E2SM-Key Performance Monitoring (E2SM-KPM), E2SM-Network Interface (E2SM-NI), E2SM-Cell Configuration and Control (E2SM-CCC) [11 – 14, 82 – 83]. The RAN optimization decisions of the E2 nodes for relatively coarser M-plane cell-level and network element-level RRM functionalities are split with the Non-RT RIC and SMO O-RAN NFs.

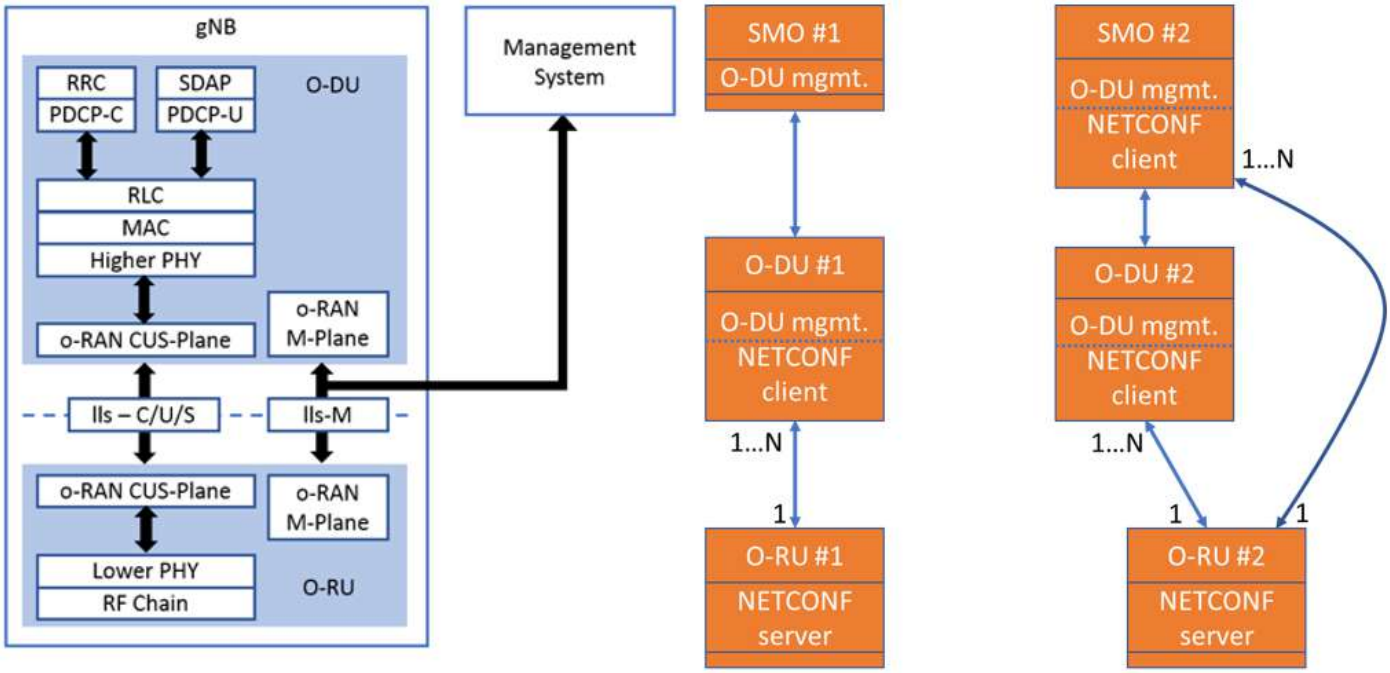
The RRM decisions for the individual M-plane RAN functionalities are exercised by 3rd party extensible applications, known as rApps, that are deployed in the Non-RT RIC [5 – 10]. In addition, the rApps are also responsible for RRM management of high-level declarative policies, generated at UE-level, UE group-level, QoS flow-level, etc., that are then sent to the Near-RT RIC over the A1 interface [5 – 10]. The Near-RT RIC xApp further enforces the A1 policies for exercising the RRM decisions for the corresponding RAN functionalities over E2.

1.3.2 Hybrid vs Hierarchical M-plane

The O-RAN fronthaul M-plane is a protocol that runs in parallel to CUS planes, with endpoints IP connectivity between the O-RU and the elements managing it (O-DU, SMO, or so called “O-RU Controllers”). The M-plane is end-to-end encrypted through Secure Shell (SSH) and/or TLS, and management instructions are based on NETCONF.

The M-plane provides functionalities related to the lifecycle of the O-RU. To begin with, it manages the startup installation procedures for commissioning, during which the O-RU establishes the management with the DU and/ or SMO based on the pre-defined IP address. Moreover, it enables software management, CM for initialization and configuration of operating parameters, performance and FM, and file management for uploads to the O-RU controller i.e. either DU and/or SMO. Among others, M-plane also manages the registration of RU as PNF, support the updates of beamforming vectors (antenna calibrations, beam-weights), and carries power efficiency commands to enable O-RU power-saving techniques.

Figure 1-7: Hierarchical vs Hybrid M-plane architectures [18]



M-Plane Architecture

The M-plane interface can be implemented hierarchical or in a hybrid manner, as shown in Figure 2-7.

In the hierarchical model [18], the O-RU is managed by one or more O-DUs, (e.g., for transport connectivity redundancy). The benefit of this model is that O-RU only has to interact with O-DU, which means that the SMO does not need to involve managing the O-RUs. This also reduces the SMO processing load. Moreover, it eliminates the need to support NETCONF on the SMO.

In the hybrid model [18], there are simultaneous logical connections from the O-RU to the SMO in addition to the logical interface between O-RU and O-DU, possibly using the same physical connections. The functions for managing the O-RU can be shared between the O-RU controllers. For example, software management can be located in the SMO framework, and performance management and fault reporting may be managed by the network management system. The advantage of this model is that the SMO can manage the O-RU, which in a way simplifies the multi-vendor network integration. However, this comes with the requirement that SMO has to support NETCONF, and the SMO processing requirements increase with the increase of the number of simultaneous sessions.

Typically, the configuration for the O-RU is performed initially as well as during operation, for which the following functions are used via the NETCONF protocol. NETCONF/YANG is used as the network element management protocol and data modeling language. Use of the standardized framework and common modeling language simplifies integration between O-DU and O-RU, natively supports a hybrid architecture which enables multiple clients to subscribe and receive information originating at the NETCONF server in the O-RU and eliminates dependency on different O-RU vendors implementation for seamless multi-vendor network integration.

1.3.3 RAN Cloudification and Virtualization

RAN virtualization involves virtualization of the CU and DU functions [20]. The key decision points deal with selection of the right commercial off-the-shelf (COTS) server hardware, the right virtualization approach and cloud operating system; and the right hardware acceleration approach in the case of compute-heavy scenarios. There is also a key consideration to reduce power consumption and increase the energy efficiency when using COTS platforms. Many use cases with low-capacity demands can run well on pure CPU platforms, but as bandwidths increase and advanced antenna systems are deployed, current x86 cores struggle to keep up with the performance demands and in the case of compute-heavy scenarios the right hardware acceleration approach. Also, each new 3GPP release has brought enhanced capabilities including supporting more spectrum, additional frequency bands, advanced features as well as air interface enhancements in performance and efficiency. A consequence of this continuous evolution is that the processing requirement for the network functions also increases. In non-Massive MIMO and low-capacity use case scenarios, RAN workloads can run on a general-purpose computing architecture (e.g., x86)—but for full 5G capabilities (for instance, mid-band capacity and Massive MIMO layers, and stringent latency demands), more processing power and hence hardware acceleration will be needed. Hardware acceleration approaches are relevant in:

- Acceleration of traffic in input / output path (e.g., virtual Centralized Unit User Plane [vCU-UP])
- Acceleration of individual functions in the L1 pipeline (for a virtual Distributed Unit [vDU])

The overall system integration, management, orchestration, and assurance are significant considerations in the virtualization journey. To enable scalable SMO across 5G RAN, open programmability of RAN is an important consideration for virtualized as well as embedded platforms. Open programmable interfaces provide a way for the SMO layer to manage different platforms and VNF and CNF workloads in a consistent and scalable fashion.

Virtualization of CU includes virtualizing the CU-CP and CU-UP. CU-CP and CU-UP can be virtualized on a COTS server. CU-UP is more demanding than CU-CP in terms of capacity and I/O throughput. CU-UP comes with high throughput user-plane traffic and handles flow control over the baseband user-plane interface (F1-U) interface. Depending on the server capabilities and workload demands, acceleration of traffic in the input/output path may be required for CU-UP workloads.

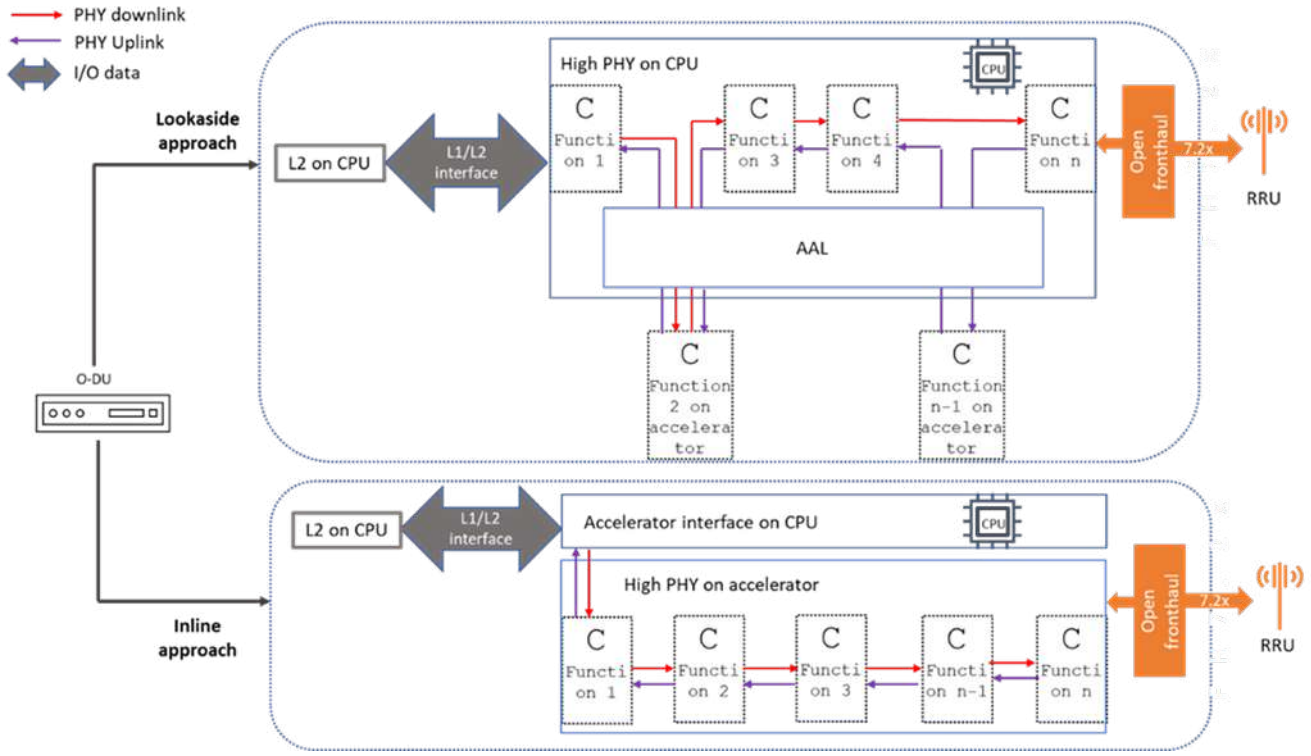
The choice of virtualization environment/Cloud OS for CU and DU is an important strategic decision. The trajectory of deployment architectures is towards a microservices-based, Kubernetes-orchestrated CNF environment [19, 20]. Containerizing the network functions and managing them in cloud-native fashion facilitates agile development, testing and deployment of services using CI/CD principles in addition to allowing greater scale, flexibility and manageability using cloud tools. Cloud-native implementation also makes the RAN functions flexible to be realized over multiple cloud environments: Private, Public, Hybrid and so on.

Virtualizing CU and DU starts with the selection of the hardware platform and the virtualization environment or Cloud OS. The hardware platform is in general a COTS server (e.g., Intel X.86 based server platform) - with NIC and hardware accelerators, where needed. Hardware acceleration will be needed for the compute-heavy functions in 5G NR.

There are two approaches in L1 acceleration, namely, the look-aside acceleration approach and inline acceleration approach as seen in Figure 2-8. Considering the downlink (DL) case, look-aside acceleration approach supports dataflow from the CPU to the accelerator and back to the CPU before being sent to the fronthaul interface. Inline acceleration supports data flow from the CPU to the accelerator and directly from the accelerator to the fronthaul interface, instead of being sent back to the CPU.

Figure 1-8: Look-aside vs Inline Acceleration

With the look-aside approach, selective functions are accelerated. In contrast, in the case of inline approach, a part of or



the entire L1 pipeline can be offloaded to the accelerator. Both approaches can be applied depending on the system vendor implementation and operator cloud infrastructure for specific deployment scenarios, as appropriate.

Accelerator Adaptation Layer API (AAL API): A key aspect of AAL API [23] work in O-RAN Alliance is to minimize fragmentation and maximize harmonization between different proposals leveraging look-aside and inline acceleration approaches. AAL API for forward error correction (FEC) profile has been completed in O-RAN. The community, however, is debating if the scope of an inline accelerator should replace the entire L1 of the O-DU network function. Definition of High-PHY profile is under active discussion within O-RAN community to achieve consensus. A potential next step in this process is to do an impact analysis of different proposals on the O-RAN architecture. For full-stack RAN virtualization, the DU is connected to the radio via a packet interface known as enhanced Common Public Radio Interface (eCPRI). There are multiple ways to divide functions between the DU and the radio; in standards discussions these are referred to as “lower layer split” (LLS) options. One possible alternative specified by the O-RAN Alliance is referred to as the 7-2x split [17]; other functional splits are also being considered.

Managing distributed vRAN workloads between far edge, edge/regional and hyperscale data center hubs is a resource as well as a service orchestration challenge. Workloads may be required to span different cloud environments as demanded by KPI requirements on capacity, scale, resiliency, latency etc. – this applies to O-DU, O-CU, Near-RT RIC and Non-RT RIC/SMO virtualization deployment scenarios. To manage infrastructure resources and the deployment life cycle for vRAN NFs and apps in the cloud using multi-vendor orchestration functions from the SMO over the new O-RAN-defined O2 interface. O2 interface procedures help optimally orchestrate the computational and storage resources for the O-RAN functions in the cloud resource pools.

The O2 [19] is an open logical interface within the O-RAN architecture for communication between the SMO and O-Cloud for management of O-Cloud infrastructure and the deployment life cycle management of O-RAN cloudified network functions that run on O-Cloud. The interface is defined in an extensible way that enables new information or functions to be added without necessarily changing the protocol or procedures. This interface enables a multi-vendor environment and is independent of specific implementations of SMO and O-Cloud.

Figure 1-9: O-Cloud infrastructure inventory [19, 20]

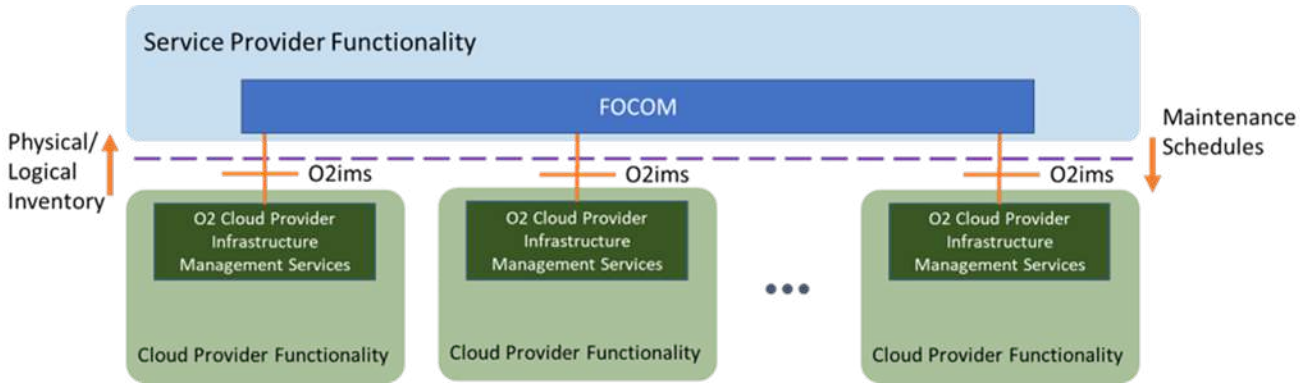
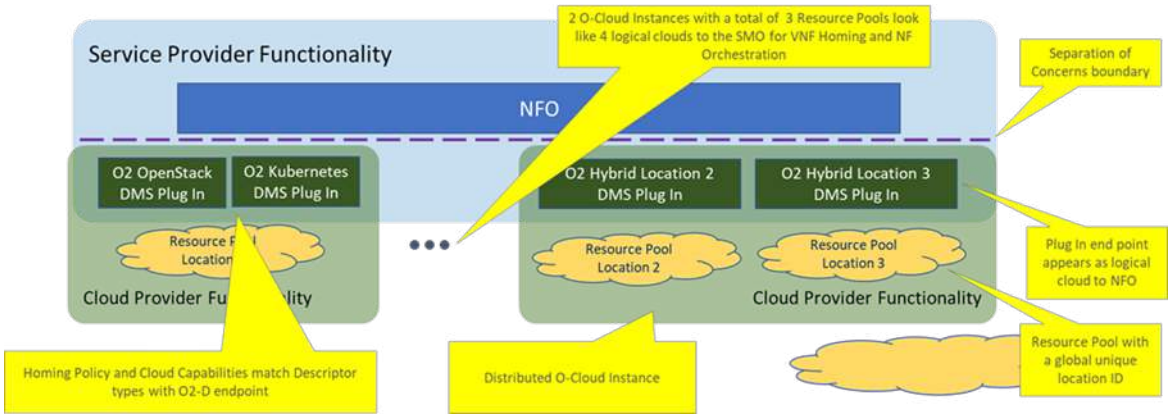


Figure 1-10 : Logical clouds [19, 20]



The O-Cloud consists of multiple Deployment Management Services (DMS) [22], which are the logical services provided by the O-Cloud for managing the life cycle of deployments using cloud resources. Each DMS can manage leased resources from multiple resource pools and span multiple locations. The O-Cloud itself must have one or more DMS available within its distributed footprint. These could be based on virtual technologies (Kubernetes/Docker, Open Stack, etc.) used, and/or O-Cloud pool locations. Each DMS endpoint provides an O2dms interface and is inventoried by the SMO as a logical cloud. The logical cloud is used by the SMO in order to select the O-Cloud to be used for a deployment during the cloud selection process. The Infrastructure Management Services (IMS) [21] are logical services provisioned by the O-Cloud, providing the interface to orchestrate O-Cloud life cycle processes with the network functions it may host along with other operational procedures. There is a single IMS for O-Cloud that manages all resources of DMSes and the resources that are not allocated to any DMS in the O-Cloud. The functions to be performed over the O2 interface include (i) O-Cloud Infrastructure Resource Management,

(ii) Managing abstracted resources and deployment, and (iii) OAM of the O-Cloud infrastructure. The O-Cloud infrastructure inventory and the logical clouds where the managed functions are deployed are shown in Figure 2-9 and Figure 2-10, respectively.

O-RAN clouds are described as a distributed cloud composed of O-Cloud pools, where each pool is a collection of O-Cloud Nodes, which are computational resource designators. The cloud is divided into the following three planes, namely, the management-plane, the control-plane, and the deployment plane. The SMO shall be able to correlate managed element telemetry to infrastructure and deployment telemetry to aggregate problems to a root cause. The O-Cloud shall be able to make all Configuration Data and any external changes to it available to the SMO. O-Cloud telemetry shall minimally consist of Fault, Performance [56, 57], and Configuration Data [58, 59]. The SMO shall be able to correlate a managed element to its deployment components. The O-Cloud shall be able to report telemetry of deployment resources relative to those identified in the deployment descriptor. The O-Cloud shall be able to report Infrastructure telemetry and identify the deployments using the resource. O-Cloud shall provide the collection and reporting of performance information of O-Cloud resources. O-Cloud shall support the capability to notify about the availability of performance information. O-Cloud shall expose the type of performance information that can be collected for the allocated O-Cloud resource(s). O-Cloud shall expose the type of O-Cloud resource, for which the performance information can be collected. O-Cloud shall provide the collection of fault information for O-Cloud resources. O-Cloud shall support providing notification of fault information related to O-Cloud resources.

O-Cloud Provisioning shall provide Query of O-Cloud Capacity. O-Cloud Provisioning shall provide Query of O-Cloud Availability. O-Cloud shall provide addition of software Images of O-RAN Cloudified Network Function to O-Cloud repository. O-Cloud shall provide Delete Software Images of O-RAN Cloudified Network Function from O-Cloud repository. O-Cloud shall provide Update Software Images of O-RAN Cloudified Network Function to O-Cloud repository. O-Cloud shall provide Query Software Images of O-RAN Cloudified Network Function from O-Cloud repository. O-Cloud shall provide Software Image properties information of O-RAN Cloudified Network Function, such as *softwareImageId*, vendor and vision. O-Cloud life cycle management will provide the (i) deploy, (ii) registration, and (iii) scale capabilities. The objective of deployment is to provide automated provisioning of the O-Cloud infrastructure, while the objective of registration is to register an O-Cloud towards making it available for deployments. Scaling capability is used to scale functional behavior and resources of O-RAN-cloudified network functions to support the required RAN services. O-Cloud supports Deploying an O-RAN O-RAN Cloudified NF instance. O-Cloud supports Terminating an O-RAN Cloudified NF instance. O-Cloud supports Horizontal Scaling (in and out) of an O-RAN Cloudified NF instance. O-Cloud supports Healing of an O-RAN Cloudified Network Function instance. O-Cloud supports Querying information about an O-RAN Cloudified NF instance. O-Cloud supports Querying status of LCM operations. O-Cloud supports upgrading of any or all components of an O-RAN Cloudified NF instance.

1.3.4 Services-based Architecture for the RIC Functions

O-RAN has adopted a services-based architecture for the Near-RT RIC [15], Non-RT RIC [10] and SMO [27] functions, as shown in Figure 2-11 and Figure 2-12, for the services-based SMO/Non-RT RIC architecture and the Near-RT RIC architecture, respectively. The functions of the Non-RT RIC, SMO and Near-RT RIC produce services that expose a set of capabilities over a services-based interface to the 3rd party rApps and xApps acting as service consumers. The functions register the services produced by them and their capabilities with the services registry, and the 3rd party xApps and rApps discover the services to be consumed by them from the services registry over the xApp APIs defined by O-RAN WG3, as well as rApp APIs defined by O-RAN WG2.

1.4 Operator Trials and Deployments

Interest in Open RAN deployments has been steadily growing over the last couple of years and you will find operators around the world have started Open RAN trials and deployments in some capacity. Figure 2-13 illustrates some of the publicly announced milestones with deployment of various forms of vRAN and Open RAN combinations.

It should be noted that as the standards bodies and the alliances have formed and shaped the technology roadmaps, there is a wide variation in the implementation of the deployments seen to date. Some of the early cases, such as the Rakuten deployment of a 4G vRAN network in Japan, pre-dates the O-RAN Alliance driven specifications, while still embracing several of the underlying principles such as software – hardware disaggregation, moving towards multi-vendor RAN and moving towards a more cloud-native application environment.

Figure 1-11: Services-based architecture for the decoupled SMO and Non-RT RIC [27]

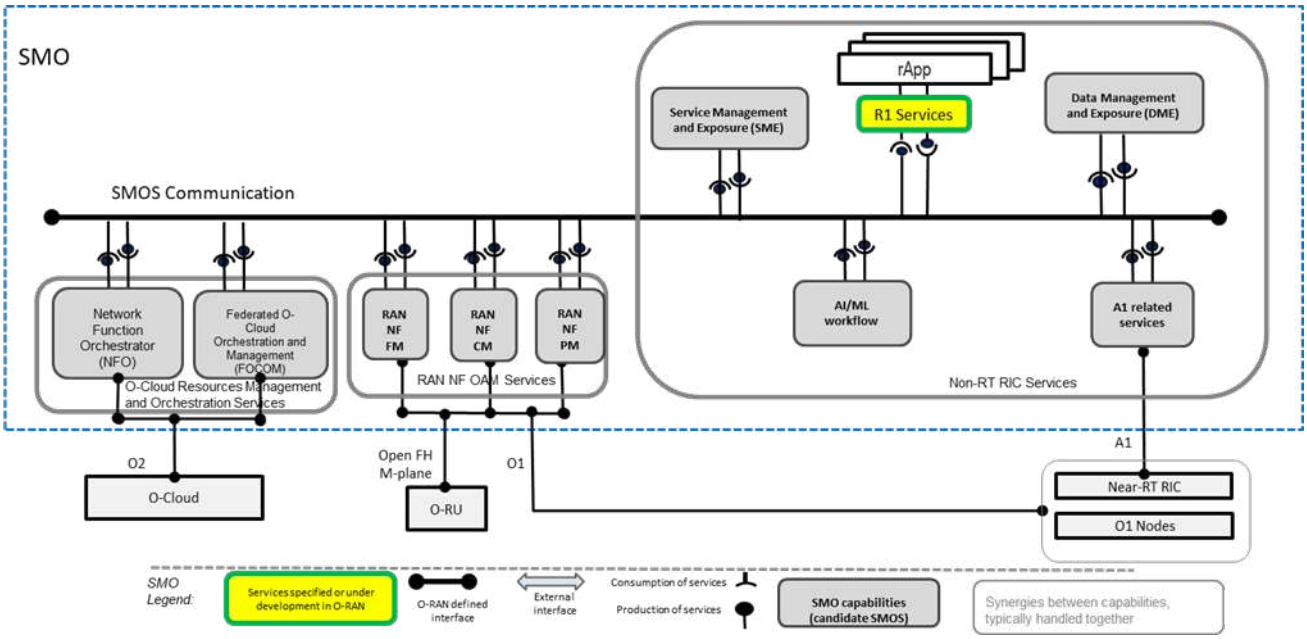


Figure 1-12: Near-RT RIC architecture [15]

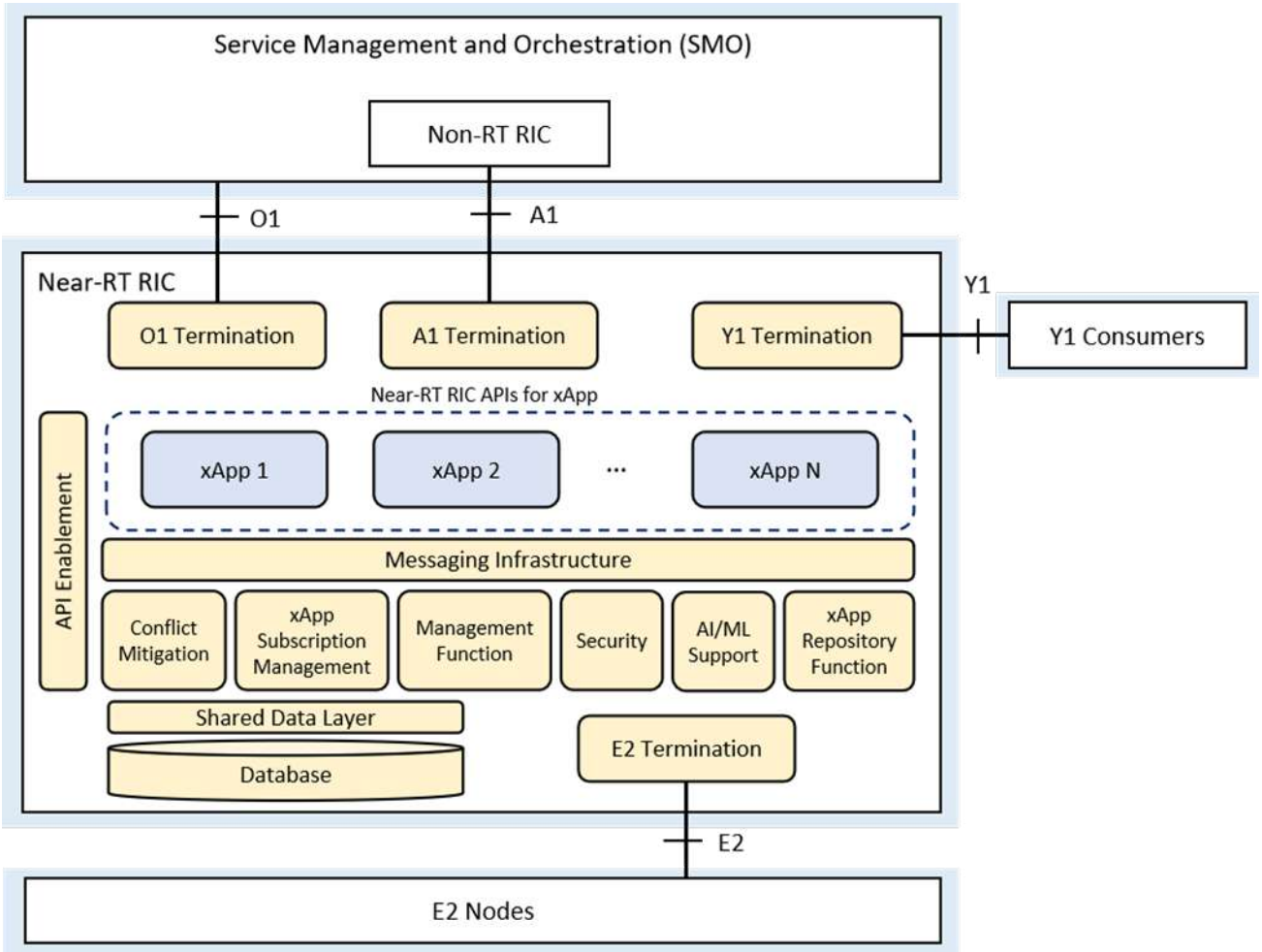
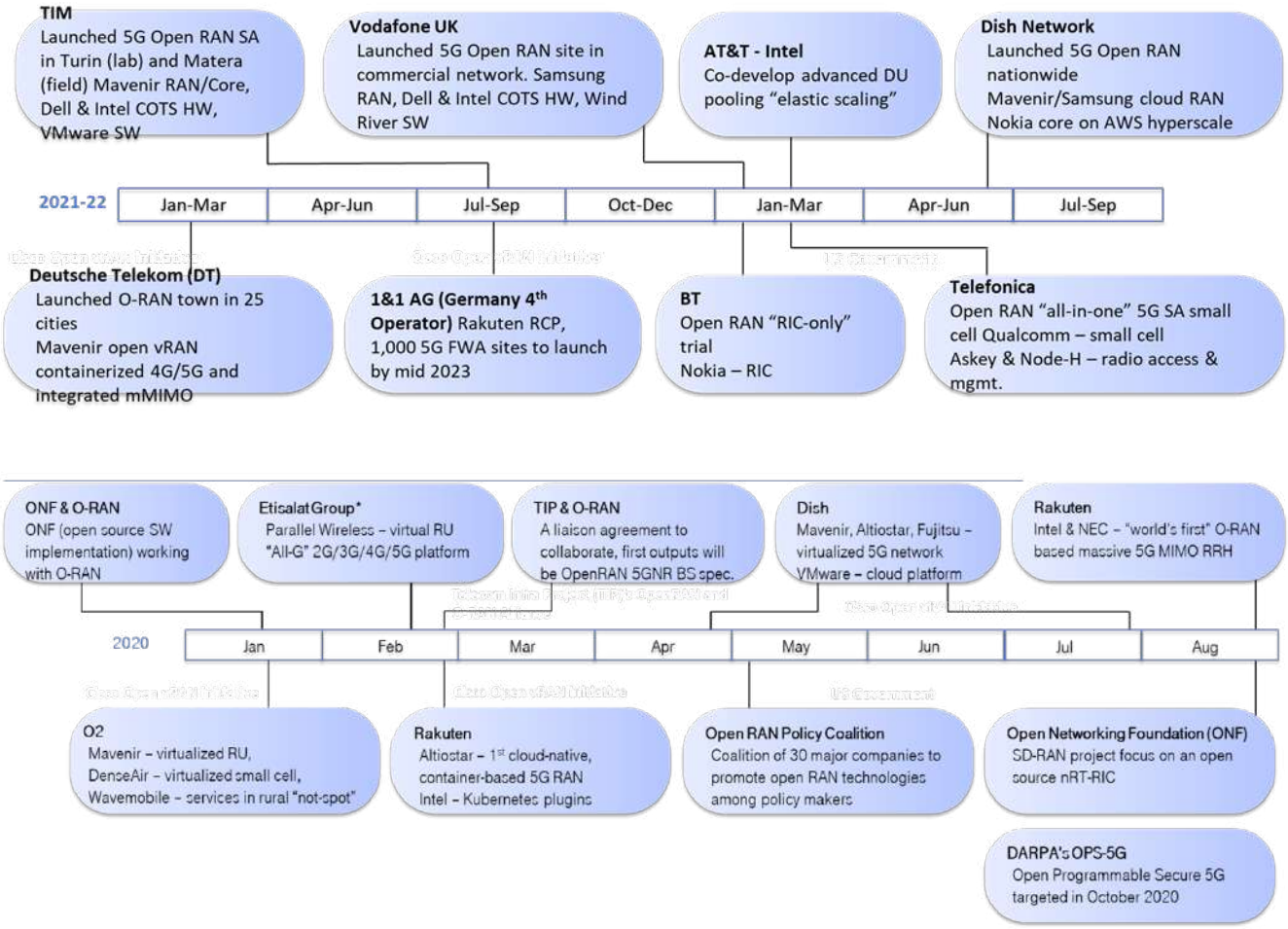


Figure 1-13 : Illustration of publicly announced milestones with deployment of various forms of vRAN and Open RAN combinations



Below are brief descriptions of some of the deployments:

British Telecommunications (BT): In January 2022, BT announced a new Open RAN trial with Nokia's RIC in the city of Hull, U.K. [49, 93]. RIC was installed across a number of sites to optimize network performance for customers of its EE mobile network. The initial phase is focused on the near real-time RIC and deployment in outdoor urban areas.

Deutsche Telekom (DT): In February 2021, DT announced the successful launch of an "O-RAN Town" in Neubrandenburg, Germany in June 2020 [50, 94]. The O-RAN Town is a multi-vendor Open RAN network that delivers O-RAN based 4G and 5G services across up to 25 cities. The town is powered up by Mavenir open vRAN, cloud-native, fully containerized 4G/5G baseband, integrated massive MIMO (mMIMO) active antenna units (AAU conforming to O-RAN Cat B specifications) using generic COTS platforms based on the latest Intel CPU and FEC acceleration technologies. Mavenir confirmed the solution has already been integrated into Telekom Germany's live network, the first live multi-vendor mMIMO deployment (4G and 5G n78) using fully standardized open fronthaul 7-2 Category B split between mMIMO RU and O-DU in Europe.

Telecom Italia (TIM): In September 2021, TIM introduced the first 5G Open RAN standalone connection on 3.7 GHz, in collaboration with Mavenir radio and core functionality, Dell and Intel for infrastructure and VMware Telco Cloud Platform for the end-to-end network function virtualization and automation software in the TIM Innovation Lab in Turin and in the field in Matera [51, 95]. TIM has already launched 4G Open RAN in Faenza in May 2021 with Mavenir RAN, MTI 4G RUs, Dell, Intel and VMware.

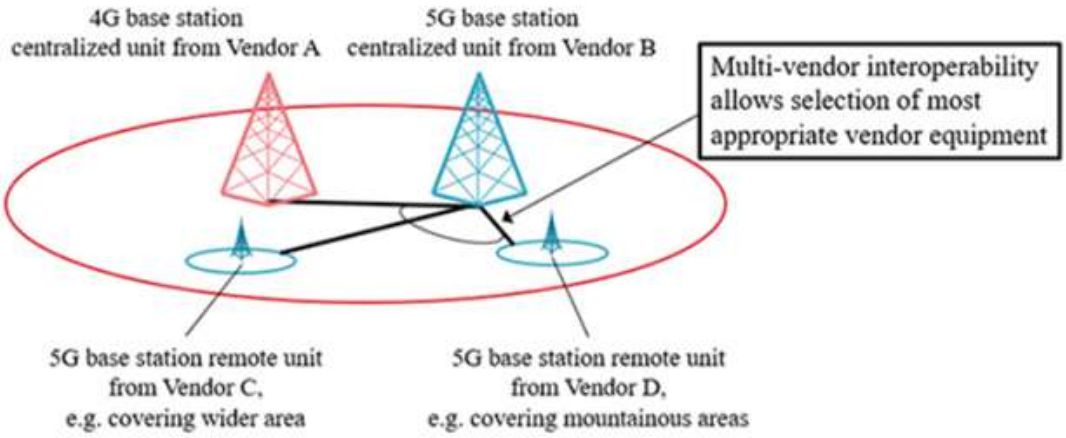
Rakuten Mobile: Rakuten has built a fully virtualized, end-to-end, cloud-native mobile network. The innovative network is fully virtualized from RAN to core and adopts 5G system architecture. Rakuten nurtures an open ecosystem through engaging

with industry leaders in crafting solutions. Rakuten Mobile is using equipment, software and services from Intel, Cisco, Nokia, Qualcomm, Altiostar, NEC, Mavenir, and Airspan [96]. The network is also cloud-native using COTS servers. Open RAN coverage has been deployed with 275,000 cells.

Vodafone: In January 2022, Vodafone U.K. turned on its first 5G Open RAN site in commercial network [52, 97]. This is a rip and replace as part of the U.K.'s 2020 ban on Huawei that Vodafone has relied on. Vodafone plans to swap-out about 2,500 sites, mainly in rural areas in the western area of U.K., to install Open RAN equipment using Samsung's RAN software/radios, Dell and Intel's servers and Wind River Studio's software management platform. Vendors that supply equipment to Vodafone are Parallel Wireless, Mavenir and U.K. based Lime Microsystem for Open CrowdCell. Vodafone has until end of 2027 to strip all Huawei products out of its network. Vodafone's antenna suppliers are Samsung and NEC which are developing Open RAN compliant equipment expecting in the summer.

Telefónica: On March 18, 2020, Telefónica announced that it would deploy Open RAN trials for 4G LTE and 5G in U.K., Germany, Spain and Brazil. Telefónica is embracing the O-RAN Alliance open interface standard and has reached agreements with Altiostar, Gigatera Communications, Intel, Supermicro and Xilinx to develop and deploy Open RAN trials in its network. Telefónica has built a network under the name Internet para Todos in Peru which covers around 800,000 people and 650 sites. About half of these sites are Open RAN sites using Parallel Wireless products. In February 2022, Telefónica validated an Open RAN "all-in-one" 5G SA small cell using Qualcomm's FSM100 RAN platform supporting both sub-6GHz and millimeter-wave spectrum at its Technology & Automation Lab [53, 98]. Askey (RAN manufacturer) and Node-H (RAN software) provided radio access, security, and management software. Telefónica aims to service the new small cell for enterprise and private network deployments.

Figure 1-14 : NTT DoCoMo 4G and 5G multi-vendor interoperability [100]



Dish: In February 2020, Dish announced its plans to build a new virtualized and open 5G network. On June 14, 2022, Dish announced that it is offering 5G broadband service to more than 20% of the U.S. population [99]. Dish's multi-vendor Open RAN 5G SA network is based on Mavenir and Samsung cloud RAN cloud-native networks (O-DU, O-CU), Fujitsu/MTI radios (O-RU) with open fronthaul CUS/M-plane interface, Nokia IMS/core on Amazon Web Services (AWS) cloud infrastructure (with plan for multi-public cloud infrastructure in the future) and VMware telco cloud management towards automation. The Dish 5G Architecture in AWS cloud is shown in Figure 2-15 [54]. The architecture shows the deployment of O-RUs in the physical cell sites, the O-DUs in Kubernetes Grid clusters in the Amazon AWS Local Data Centers (LDC) in the local zone, the O-CU-CPs and O-CU-UPs in the Amazon AWS Edge Data Centers (EDC) in local zone. The core network user-plane function (UPF) for data is deployed in the UPF in the EDCs, whereas the AMF, SMF and UPF for voice are deployed in the Amazon AWS Regional Data Centers (RDC) in the regional zone. The Near-RT RIC shall also be deployed in the EDC, co-located with the O-CU-CPs and O-CU-UPs, and the Non-RT RIC/SMO functions shall be deployed in the RDC, along with the packet core functions. The OSS/BSS and IMS functions are deployed in the Amazon AWS National Data Center (NDC).

NTT DoCoMo: In Sept. 2019, NTT DoCoMo announced that it successfully worked with Fujitsu, NEC and Nokia on multi-vendor interoperability for its 4G and 5G base station using O-RAN Alliance specifications [100]. DoCoMo will deploy this in its pre-commercial 5G network. NTT DoCoMo has adopted O-RAN fronthaul specifications to connect remote radio units with centralized baseband units, and the O-RAN X2 profile specification to connect between 4G base stations and 5G base stations from different vendors, shown in Figure 2-14 [100]. By the end of 2021, NTT DoCoMo has been building more than 10,000 base stations based on 5G Open RAN with Nokia CU/DU and Fujitsu RU and 20,000 more by March 2022. The DoCoMo's 5G deployment uses the actual O-RAN fronthaul interface between the baseband and the radio-- yielding throughputs of up to 4.2 Gbps (with carrier aggregation). DoCoMo has converted about 10M of its 82M subscribers to 5G Open RAN services. DoCoMo also made a 5G deal with Samsung in March 2021 and recently completed trials of a 5G standalone baseband unit with NEC.

AT&T: In 2019 and 2020, AT&T and Nokia conducted trials involving the Near-RT RAN Intelligent Controller in New York and New Jersey 5G mmwave and 5G FR1 sites, irrespectively, involving a 5G NSA network, where the UEs were connected in EN-DC mode [70, 101]. The Near-RT RIC was connected to 5G gNBs over E2, by means of which X2 messages between eNB and gNB were traced to the Near-RT RIC using E2SM-NI. 3 xApps were deployed in the Near-RT RIC, which included (i) a 5G measurement campaign xApp [102] that computed fine-grained UE-specific, cell-specific and node-specific measurements based on the UE-associated and non-UE-associated X2AP messages exposed over E2, (ii) Automated Neighbor Relations and admission control xApps were demonstrated by exposing the X2AP messages, containing the *MeasurementReport* RRC message container, over E2 to the Near-RT RIC. AT&T has highlighted that open architectures based on RIC and SMO would enable the operators to use their wealth of RAN operational data and customer insights to customize their networks using data science and AI technologies to their particular customer base, geography, and spectrum position [103].

Crown Castle USA: As a neutral host provider, Crown Castle has trialed and is delivering an Open RAN solution for Rudin Management Company in Manhattan, New York [104]. Rudin owns several buildings (e.g., 345 Park Ave has 44 floors), has high-value tenants, and is advancing smart building management. Rudin's drivers for the solution were to maintain control of their data (local data security), support their building management system to support eco-friendly initiatives, and support private tenant use cases. Operators are interested in improving in building coverage at Rudin locations. The flexible Open RAN platform enabled Crown Castle to deploy an LTE (i.e., not NSA) network today, which is software-upgradable to 5G SA in concert with Rudin or operator demand.

Triangle Communications: Triangle Communications, a U.S.-based telecommunications service provider, that provides mobile broadband telecommunication services to residents of Montana, announced that it replaced Huawei's equipment with fully cloud-native Open RAN equipment, including O-CUs, O-DUs and O-RUs, with converged packet core, as part of FCC's "rip and replace" program. The Open RAN equipment deliverable was completed ahead of FCC's "rip and replace" program funding [105].

1.5 Operational Considerations and Integration Challenges

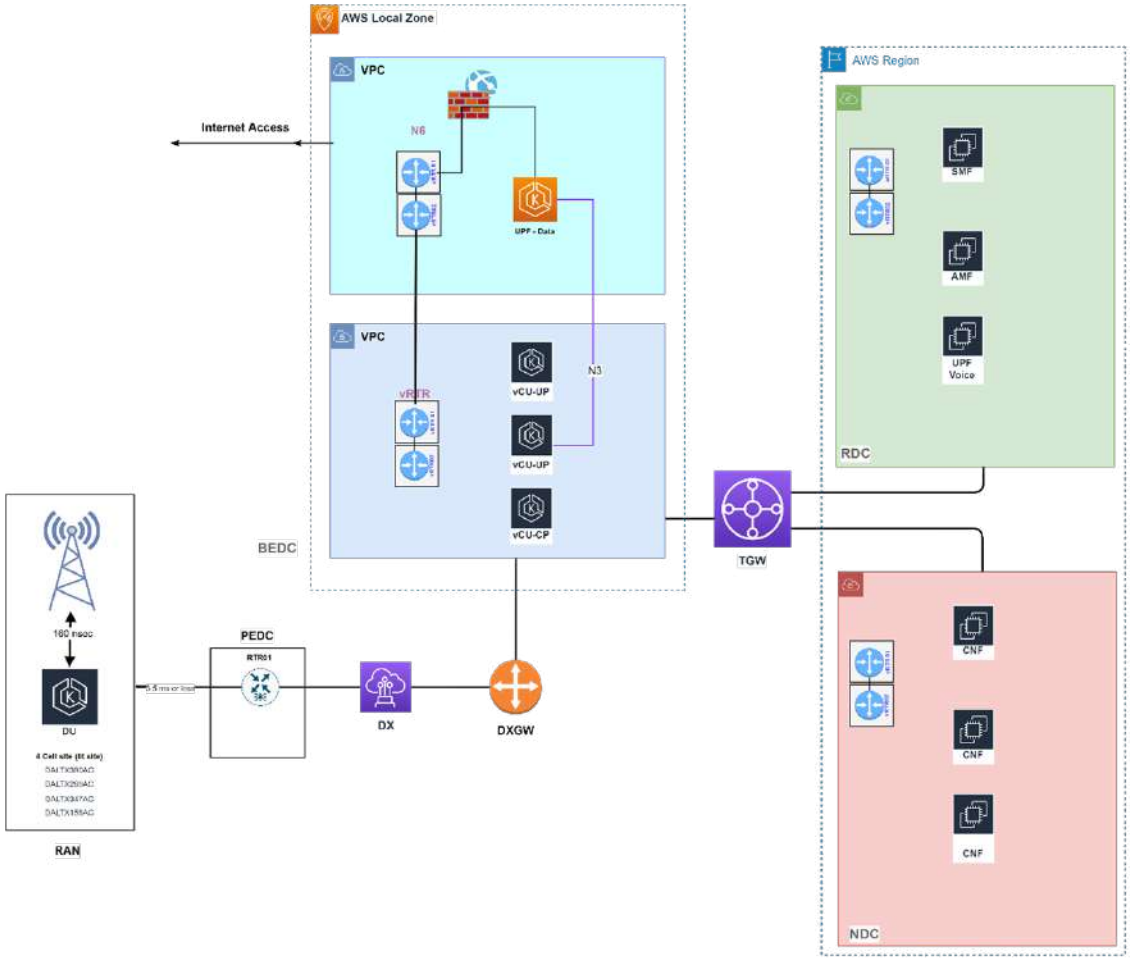
For an operator to move from a proprietary vendor deployment approach to an Open RAN model, a number of operational considerations and integration challenges could be considered:

1.5.1 Brownfield Operators

A brownfield operator with plans to introduce Open RAN needs to have the following considerations and integration challenges:

- Open interfaces and RIC integration: A key challenge, perceived by the operators, is to upgrade the current LTE macro eNBs to support (i) open X2 C-plane and U-plane interface with multi-vendor O-CU-CPs and O-CU-UPs, (ii) O-RAN-standardized E2 interface with Near-RT RIC. With a significant portion of subscriber base on LTE, a brownfield operator, upon adopting open-X2 interface with a multi-vendor O-CU-CP and O-CU-UP NFs, could have challenges in testing key call processing network interface procedures, related to connected mode mobility, load-balancing, connectivity, dual connectivity, over the open X2 interface and evaluating the network performance. The operator faces similar challenges upon supporting the E2 interface for the LTE eNBs and splitting the RRM of the key UE-associated low-latency call processing decisions for the above procedures (connectivity, dual connectivity, mobility, load-balancing) with the Near-RT RIC.
- O-RU integration: For a brownfield operator, the integration of macro sites with 5G implemented as Open RAN Option 7.x split with the underlying legacy LTE could be challenging with respect to network design and implementation. 5G features parity to traditional RANs from basic features such as uplink

Figure 1-15: Dish 5G Architecture in AWS cloud [54]



pre-scheduling (to help reduce round trip latency), uplink closed-loop power control (to maintain signal to noise ratio and minimize interference), to advanced features such as CSI-RS based mMIMO, 4CC CA (F, T, F+T combinations), to customized radios from low band + mid-band + high band RU combinations. It doesn't seem to be feasible to address these functionalities, especially below the L2 layer, purely by software. RU vendor selection should be carefully considered. For example, inter-band carrier aggregation may have challenges in multi-vendor O-RUs due to their L1 implementation difference in the O-RU [108].

- Disaggregation and cloudification of network functions: It can also be a challenge to the operator to disaggregate the RAN NFs, virtualize them and migrate them as CNFs to a cloud platform. The requirement of computational and storage spaces in cloud platforms, and the management of infrastructure and deployment services for cloud platforms would need to be considered. More typically, the NFs must be realized using worker nodes that include virtual machine (VM) instances of a cloud platform, and a set of CNFs are managed in a Kubernetes cluster, and the functionalities pertaining to the NFs are implemented as micro-services running on containerized pods. The Kubernetes clusters are deployed in data centers offered by the cloud provider, and it poses challenges to a brownfield operator with legacy cell-site deployments to evaluate the migration of NFs to cloud platforms in data center deployments, especially for widespread and pervasive LTE deployments, most of which will likely undergo cell-site upgrades to 5G. Most importantly, the brownfield operator needs to evaluate the choice of data centers and cloud platforms for the TTI-level operations involving the O-DU network function, which may pose challenges. Moreover, choosing the cloud provider and cloud infrastructure resources for the user-plane functions such as the O-CU-UP for handling/serving high-volume user-plane data traffic also creates challenges for the operator.
- Hardware performance: Hardware performance of x86 is perhaps closing the gap of the monolithic and ASIC based systems with the next generation silicon like Intel 3rd generation XEON processors with ready libraries like OPENESS and Open Vino has made it possible to achieve almost the same type of performance as monolithic RAN systems. System integration requires complete vertical stack validation i.e., end-to-end working solution including radio frequency (RF)/ radio and hardware and not only for the cloud certification. The system integrator must have rich tools and capabilities on automation, data and AI. Smart power management/distribution of servers in each rack to ensure reliable operation and services including distributed power detection and prediction of COTS hardware to avoid potential power overload

causing DU/CU server failure in local data center and/or edge data center. Addressing RAN KPIs and counters improvement to support RAN benchmarking, together with cloud and automation parts is challenging for a brownfield operator.

1.5.2 Greenfield Operators

A greenfield operator with plans to introduce Open RAN could consider the following:

- Evolution of the O-RAN interface specifications: The maturity and readiness of the O-RAN specifications, involving the O-RAN interfaces and APIs, service/policy/data models, etc. are a key challenge to the adoption of Open RAN systems by greenfield operators, starting with O-RAN-based deployment of 5G systems. Moreover, with greenfield operators adopting cloud-native technologies for implementation of O-RAN NFs and their related RAN functionalities, there could arise compatibility issues with existing 3GPP technology-specific solution sets in terms of protocol and payload, which may not be cloud-native. As example, cloud-native implementations of OAM functionalities will involve using Open CI/CD techniques and they may need to co-exist and be integrated with existing Non-O-RAN OAM tools.
- Realization of O-RAN NFs: O-RAN NFs such as the O-DUs deal with TTI-level operations, such as functionalities involving the scheduler that does allocation of frequency-time resources (bandwidth parts, BWPs, TTIs), MCS selection, HARQ retransmission management, etc. With the real-time TTI-level granularity for functionalities involving the O-DU, the cloudification of O-DU NFs with baseband pooling and their deployment in local zone cloud data centers, as opposed to the physical cell sites much closer to the radio units and the UEs, raise challenges in terms of meeting the low-latency requirements and conforming to performance requirements.
- Challenges with disaggregated O-RAN NFs and integration: The disaggregation model will require a new approach to security and trust practices for the operator to cover multi-layer, multi-player security test process. This will be impacted by Cloud related trust considerations (Private Vs Public Vs Hybrid Cloud) as well.
- Cloud footprint and scaling: With RAN optimizations involving the RIC functions that leverage sophisticated AI/ML and reinforcement learning techniques, a significant amount of effort shall be spent on training models offline and updating them online, and deploying the trained/updated models in the inference engine. Offline AI/ML model training, typically done in the Non-Real-Time RIC, requires huge computation and storage spaces; especially for more sophisticated deep learning and deep reinforcement learning (RL) models. Such computation for training ML models necessitates the usage of GPU. And as the network gets larger, there would be more cells and higher number of UEs, resulting in increased transactions per second. This further increases the computational footprint in terms of number of vCores, pod instances etc. required to train optimal ML models towards achieving higher performance. There is thus a cost factor for the greenfield operator due to this increased footprint of computational and storage resources in the cloud, which may dilute the performance benefits resulting from harnessing deep AI/ML and RL models.
- Organizational preparedness: Finally, Open RAN implementations require the Operational team to adapt and get skilled in new Cloud RAN-specific life cycle management paradigms and new maintenance & troubleshooting practices.

1.5.3 Shifting Operator Role and realizable TCO savings

In addition to the role of an integrator that the operator has to play (or rely on outsourcing that activity and still oversee everything with less direct involvement but with full responsibility), the other area of possible focus is the need for more immediate consideration given towards the upskilling of network and field operations teams to run a variety of services in a complex Open RAN network, with critical implications for day to day performance variations or security issues. Operators will need training and hands-on experience in every functional block or system component, dealing with a set of known vendors and potentially with implementations from unknown open-source contributors.

The Open RAN move towards standard COTS and/or white label hardware is expected to drive significant cost savings and supply-chain simplicity with hardware replacements and inventory management, which is a very desirable outcome for most network operations teams. On the other hand, with more vendors to deal with, the relationship value (measured in payments) is lower for each vendor compared to a fully sourced single vendor revenue model. While potential benefits from the lower-cost lure of Open RAN may offset some of that, the tradeoffs will likely vary case by case. For instance, some operators and vendors are concerned that the use of a system integrator will potentially come at a steep cost and that it could be a risk to the business given the likely need for complicated business models around support agreements with the component/functionality suppliers resulting in lengthy resolution processes.

1.5.4 Performance Considerations

The premise of Open RAN includes leveraging the skillsets of a broad community of designers, engineers, developers etc. As the Open RAN ecosystem is designing novel architectures for next generation technologies, the use cases themselves are evolving and requirements are being investigated. These dynamic aspects are a big challenge for proprietary RANs and Open RAN systems too. It is to be

seen whether the architectural aspects (e.g. RIC and use of AI/ML) are robust enough to meet the challenging requirements of the upcoming use cases, and if the overall implementation flexibility and resulting performance with the open community based design will be better relative to proprietary RAN systems where vendors provide their special sauce to improve spectral efficiency, manage interference and increase system throughput using components and designs they have full control over. Lack of specialized proprietary implementations of highly advanced functionalities (e.g., digital beamforming, MU-MIMO etc.) could limit relative performance and flexibility in the near term.

1.6 Advantages and Challenges with Open RAN Architectures

Having discussed the principles of Open RAN architecture and standards, this section summarizes the advantages and challenges associated with adopting Open RAN architectures.

The key benefits of Open RAN architectures include (i) avoiding vendor lock-in with open interfaces that enable multi-vendor interoperability, (ii) cost reduction with the adoption of COTS platforms and open whitebox hardware while minimizing the usage of vendor-proprietary hardware, (iii) enhancing smartness of the network so as to offer performance and experience guarantees to the end-user and enterprise customers, towards the realization of use cases by leveraging the RIC that does optimization of C-plane, U-plane and M-plane functionalities using sophisticated AI/ML tools, fine-grained intelligence and programmable policies, (iv) spur innovation by fostering a competitive ecosystem of 3rd party solutions that inter-operate with each other using open and standard interfaces and APIs towards optimizing the network and achieving performance/experience guarantees, (v) collaboration towards development of open hardware and software, (vi) flexibility in deployment of network functions, such as an open choice in deploying an O-DU either at a customer premise closer to the physical site or a nearby local/edge data center, and facilitating aggregation from multiple DUs from the local data center at the CUs deployed in the edge data center, etc.

While these advantages have certainly motivated greater interest towards adopting and trialing with Open RAN systems by operators (both greenfield and brownfield), there are also challenges which need to be carefully considered and resolved towards a successful pervasive adoption and deployment of production-grade Open RAN systems. These challenges include:

- Open RAN TCO benefits for the operator are yet to be clearly proven, (ii) the necessity to have high bandwidth and low-latency for the open fronthaul M-plane interface between the O-DU and O-RU for exercising TTI-level operations, iii) In addition to fronthaul, centralization of virtualized CU workloads requires additional transport capacity planning for mid-haul (iv) requirements for new hardware (such as COTS, accelerators, open and whitebox hardware) and software (such as cloud-native and virtualization software) in the context of telco systems, (v) interoperability between multi-vendor NFs over open and O-RAN-defined interfaces and the development and operational complexity involved, especially with brownfield operators, in upgrading and interoperating with their legacy deployments, vi) Centralized CU and DU deployments present large failure domains and hence local redundancy and geo-redundancy for service resiliency must be carefully planned, adding to the cost and complexity. (vii) complexity in intelligent automation involving the RIC and third-party xApps/rApps, arising from multiple xApps/rApps trying to access the same resource and making potentially conflicting changes in the same resource that could result in system instability, (viii) reliability and availability, dealing with the availability of necessary compute and storage from cloud providers in cloud infrastructure resources, while minimizing the loss in services offered by the network functions, (ix) security and trust issues, Open RAN disaggregation presents an expanded threat surface and enhanced security risks arising from the addition of new interfaces and new disaggregated network functions, which, if not subject to secure management, can result in increased vulnerability; whereas the trust issues between multiple vendors could impact the coexistence of their respective network functions that must integrate and inter-operate over open interfaces.
- Open RAN-specific security has become an important area of consideration for U.S. government such that it has summarized the considerations in a recent paper on Open RAN security considerations. In terms of Zero-Trust Architecture (ZTA) considerations [79], there is also an increasing awareness that cloud deployments must assess internal threats, in addition to external threats, to ensure the migration of 5G critical infrastructure to the cloud is secure. This is a new paradigm for securing the RAN requiring the pursuit of a ZTA framework to protect the network from internal and external threats. U.S. Department of Homeland Security (DHS) Cybersecurity and Infrastructure Security Agency (CISA) advises for 5G critical infrastructure deployments in the cloud to “assume the adversary is already inside the network”. A white paper detailing security considerations in Open RAN are detailed in [67]. The Enduring Security Framework (ESF) [67 – 68] chartered by the Department of Defense, DHS, Office of the Director of National Intelligence, and the IT, Communications and Defense Industrial Base Sector Coordinating Councils aims to address risks that threaten U.S. critical infrastructure has come up with key guidance to build a ZTA complaint 5G cloud infrastructure. The O-RAN Alliance’s WG11 for security is evolving security specifications to address these risks [26].

2. O-RAN use-case realization using AI/ML

For mobile networks to evolve from a design that offers best-effort services to a design that offers performance and user experience guarantees, intelligence needs to be an integral component of the network. O-RAN WG1 Use Case Task Group [3] has defined a set of use cases, such as traffic steering, QoS-based resource optimization, QoE optimization, RAN Slicing Service Level Assurance, massive MIMO beamforming optimization, Dynamic Spectrum Sharing etc., along with the corresponding use-case requirements. Realization of these use cases require meeting certain performance guarantees and service assurances that mandate the usage of AI/ML tools. This section introduces the concepts of application of AI and ML to Open RAN networks, identifies architecture requirements with specific use cases, discusses the deployment of scalable and practical AI/ML models towards the realization of Open RAN use cases in operational 5G and beyond 5G networks.

2.1 The Role of AI/ML in 5G and Beyond 5G RAN

Enhancing RAN performance with the use of AI and ML has many potential benefits and considerations. 5G networks enable operators to provide a vastly expanded range of services across a diverse set of technologies and spectrum. The flexibility and richness of 5G could make it more complex to optimize and manage, with a wider range of performance KPIs parameters to optimize. 5G telecommunication services focus on the following categories of use cases [106]:

- Enhanced Mobile Broadband (eMBB) – for bandwidth-intensive HD 4K to 8K video/VR streaming, immersive AR/VR, etc.
- URLLC – for intelligent transportation services, factory automation, remote telesurgery, real-time drone surveillance etc.
- Massive IoT (MIoT) – for wearables, etc., requiring high coverage to support network densification.
- High Performance Machine Type Communication (HMTC) – for mission critical communication, requiring ultra-high reliability and high availability/coverage of the network.
- V2X (Vehicle to Everything) – for intelligent transportation services, connected vehicles, autopilots and self-driving cars, etc.

Beyond 5G networks will likely focus on provisioning newer use cases like metaverse, telepresence, etc., that deliver an altogether new experience to mobile UEs [107]. Enterprises expect mobile network operators to deliver such applications to UEs with QoS assurances in network performance that enrich the end-user's QoE. Towards this end, mobile network operators are slicing their network resources for serving these use cases. Network operators are dealing with complexities of deploying and managing 5G services, while maintaining previous generations of wireless networks, and some operators are already on track to introduce Open RAN networks. The traditional human-intensive means of deploying, optimizing and operating radio access networks may not be able to achieve the level of optimization needed for provisioning 5G services, due to the heterogeneity and diversity of services and use cases that shall be provided by 5G systems.

Traditional RRM solutions, largely based on heuristics, also do not sufficiently account for intricacies resulting from rapidly-changing wireless network dynamics, and are not adequately optimized to handle user-customized optimization decisions for key RAN functionalities (such as connected and idle mode mobility, radio bearer admission, radio resource control and spectrum allocation, multi-RAT dual connectivity, carrier aggregation, dynamic spectrum sharing, etc.) [3] pertaining to evolving use cases and slicing requirements for 5G and beyond. This necessitates the need to have more data-driven, AI/ML-based solutions that can learn intricate inter-dependencies between RAN parameters, arising from complex interactions across the layers of the RAN protocol stack due to RRM decisions, and quantify their impact on individual UEs and collectively on the entire network.

Taking the example of the traffic steering feature to control the mobility of UEs in RAN, the handover optimization is an age-old problem in cellular RAN, solutions which have been widely discussed and implemented. However, the requirements and deployment scenarios keep changing with evolving radio access technologies, newer use cases and slice requirements that traditional handover procedures and optimization techniques are not primed to handle. To illustrate this further, even as handover processing has been featured in 3GPP specifications since the 2G days, 3GPP standards for 5G, as recent as Release 16 (2020), have introduced a new handover feature, called Dual Active Protocol Stack (DAPS) handover (which enables the UE to stay connected to the same serving cell, even after receiving the handover command from the O-CU-CP hosting the cell up until the UE establishes a successful RACH to the target cell, thereby avoiding interruption in connectivity and data transfer) [32, 34], for processing the handover of URLLC

UEs towards meeting their stringent latency requirements control-plane and user-plane requirements. Likewise, traditional RRM or legacy SON solutions for handover, largely based on heuristics involving signaling measurement and load thresholds for cells, are not primed to handle optimal UE-centric handover decisions for serving new use cases and slicing requirements.

AI-enabled solutions manage the scale of complexity with advanced capabilities in auto-configuration, self-driving and self-healing of networks that use new learning-based technologies to automate operational network functions and reduce OPEX [107]. This new "intelligent" RAN should be able to sense its environmental and application context, as well as interpret and act on the contextual information in real-time extremely efficiently. Furthermore, device and resource control functionality should be able to take advantage of the de-coupling of the UP and CP in Open RAN to offer efficient and optimized closed-loop network management capabilities using advanced analytics and data-driven approaches, including advanced AI/ML-enabled applications close to the edge of the RAN networks.

The key benefits of Open RAN with respect to AI/ML-based optimization and automation are:

- Use of interoperable open interfaces to perform data collection, and configuration changes for these tasks.
- Use of open APIs to implement algorithm clusters (such as rApps and xApps in RICs) to allow multiple solutions to be tried and tested for best results for the same use case.
- Allow operators to take control of their networks and innovate at their own pace.
- Inherent ability in O-RAN to offer efficient, optimized RRM for decisions concerning use cases such as load-balancing, mobility management, multi-connection control, QoS management, network energy savings, slicing, massive MIMO etc. through closed-loop control of C-plane, U-plane and M-plane functionalities at finer granularities towards enhancing network performance and user experience.

All variants of Open RAN architecture (such as O-RAN, TIP Radio Intelligence and Automation, etc.) target achieving these goals by embedding intelligence, at component and network levels, to enable dynamic RRM and optimize network-wide efficiency. In O-RAN Alliance, "Intelligent RAN" is a key stated objective. In the O-RAN Alliance's reference architecture, the introduction of the hierarchical non-Real-Time (non-RT) and near-Real-Time (near-RT) RIC with the A1 and E2 interfaces is aimed at enabling an entirely new ecosystem of intelligent features and applications residing close to the edge of the RAN network to fulfill the above stated goals. This chapter focuses on the O-RAN architecture from the O-RAN Alliance towards the utility and applicability of AI/ML techniques for efficient network operations.

2.2 AI/ML Functionality in O-RAN Architecture

As discussed earlier, the Near-RT RIC and the Non-RT RIC are two network functions in the O-RAN architecture that are dedicated for building AI/ML models towards equipping the underlying O-RAN NFs with intelligence and subsequently optimizing the RAN-related functionalities.

2.2.1 Analytics and AI/ML framework functions in the Near-RT RIC

Referring to Figure 1-12, The Near-RT RIC architecture includes the platform functions concerning analytics and AI/ML [15]:

1. The AI/ML support function is a platform function for AI/ML training. The services offered by the function include:

- Data pipeline: The AI/ML data pipeline service for the AI/ML support function in Near-RT RIC offers data ingestion and preparation for xApps. The input to the AI/ML data pipeline may include E2 node data collected over E2 interface, enrichment information over A1 interface, information from xApps, and data retrieved from the Near-RT RIC database through the messaging infrastructure. The output of the AI/ML data pipeline may be provided to the AI/ML training capability in Near-RT RIC.
- Training: The AI/ML training service for the AI/ML support function in Near-RT RIC offers training of xApps within Near-RT RIC. The AI/ML training provides generic and use-case-independent capabilities to AI/ML-based applications that may be useful to multiple O-RAN use cases, such as traffic steering, QoS optimization, QoE enhancement, slicing, MU-MIMO, RAN sharing, etc.

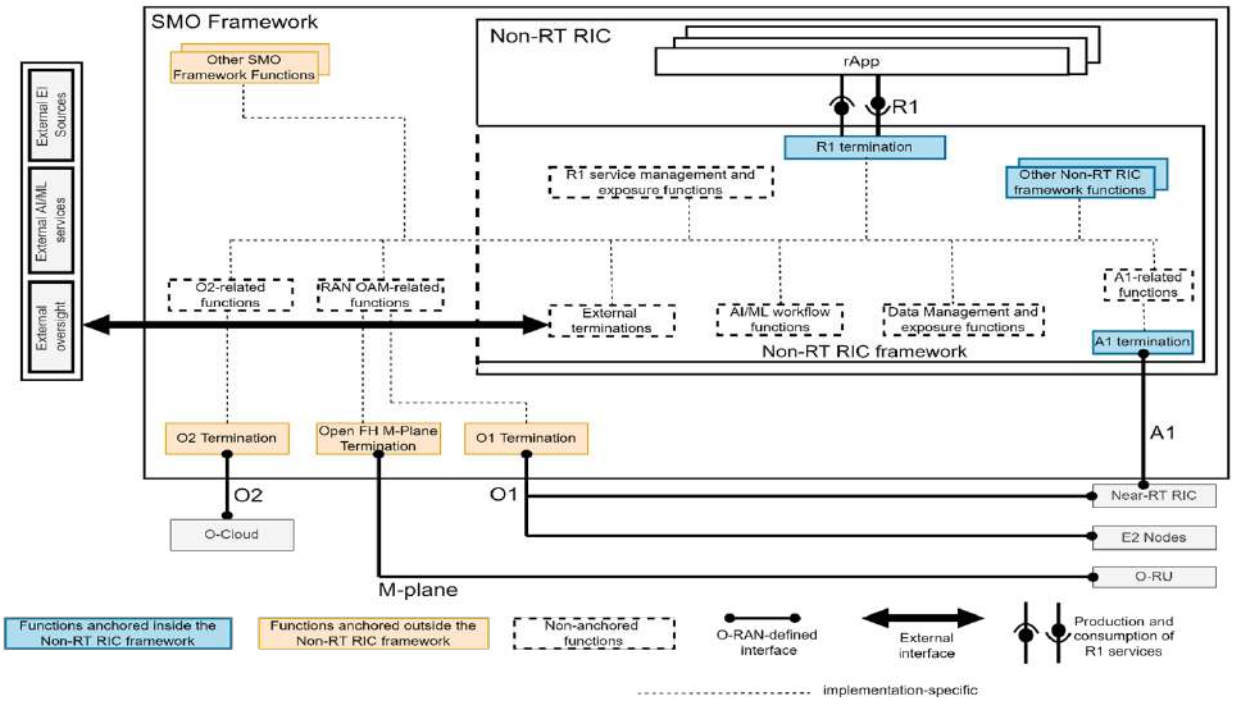
API messaging infrastructure for AI/ML services: O-RAN WG3 standardizes APIs and interfaces that enable the xApps to communicate with the AI/ML support functions in the Near-RT RIC platform towards training AI/ML models and subsequent deployment of updated models in the xApps [15, 16]. These APIs are standardized so as to enable multi-vendor interoperability between 3rd party xApps and the Near-RT RIC platform that offers AI/ML services.

2. The Y1 termination function: The Y1 is a new interface between the Near-RT RIC and Y1 consumers [15, 16]. This interface enables RAN analytics information exposure from the Near-RT RIC. Y1 termination is a function which

terminates the Y1 interface from Y1 consumer. Y1 termination communicates with Y1 consumers via Y1 interface and exposes RAN analytics information service(s) from Near-RT RIC. Y1 interface allows the Y1 consumers to subscribe to or request the RAN analytics information service(s) provided by Near-RT RIC.

2.2.2 Analytics and AI/ML framework functions in the Non-RT RIC

Figure 2-1 : Non-RT RIC Reference architecture, as defined in O-RAN WG2 [10]



O-RAN WG2 defines the O-RAN Non-RT RIC as a logical function in the SMO that enables non-real-time control and optimization of RAN elements and resources, AI/ML workflow including model training and updates, and policy-based guidance of applications and features for the Near-RT RIC [10]. It logically terminates the A1 interface and provides policy-based guidance, enrichment information and AI/ML model management for the Near-RT RIC. The relevant requirements of the Non-RT RIC architecture, pertaining to AI/ML features, include:

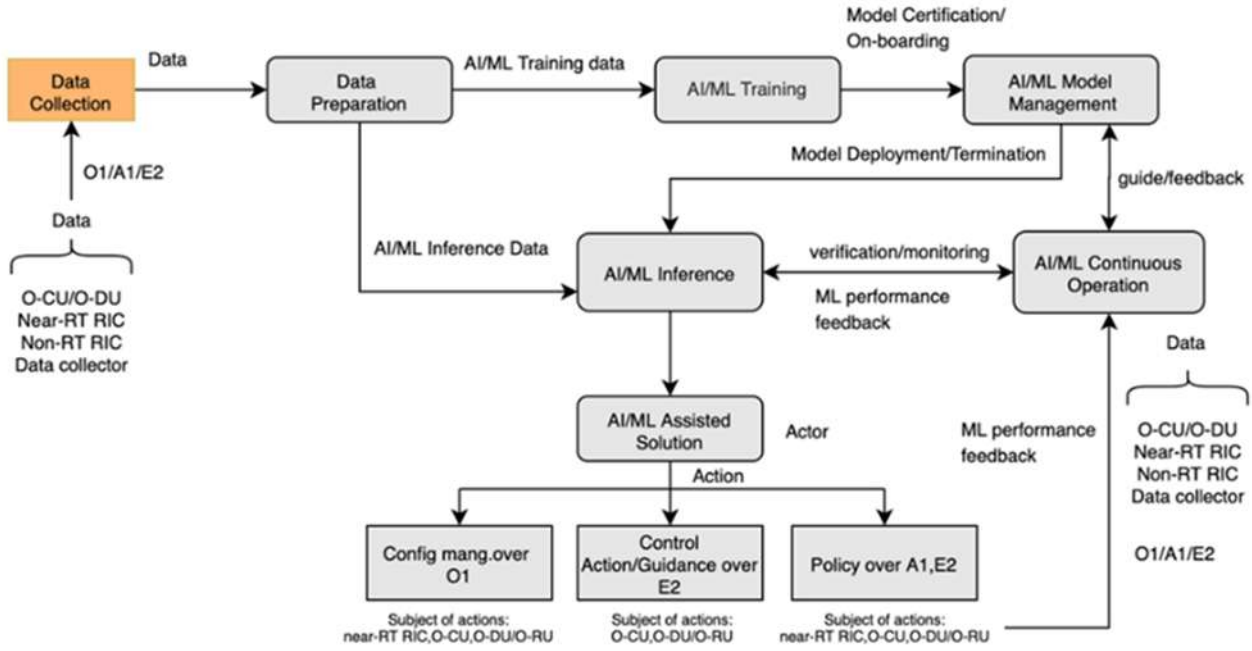
- To train AI/ML models
- To allow service consumers to store and retrieve trained AI/ML models,
- To monitor the performance of the deployed AI/ML models in runtime

Figure 2-1 shows the Non-RT RIC reference architecture and the AI/ML platform functions offering AI/ML-related functionality [10]. The AI/ML workflow function in the Non-RT RIC architecture offers AI/ML model training, version control and ML model catalog maintenance.

2.3 AI/ML Life Cycle Management in O-RAN Architecture

O-RAN Alliance WG2 provides the general framework of AI/ workflow and pipelines, which addresses the ML components within the logical functions (Non/Near-RT RIC) in the O-RAN architecture [109]. The potential mapping relationship between the ML components and network functions, interfaces defined in O-RAN are illustrated in Figure 2-2 and are subsequently. detailed.

Figure 2-2 : AI/ML workflow across the RIC and O-RAN functions [109]



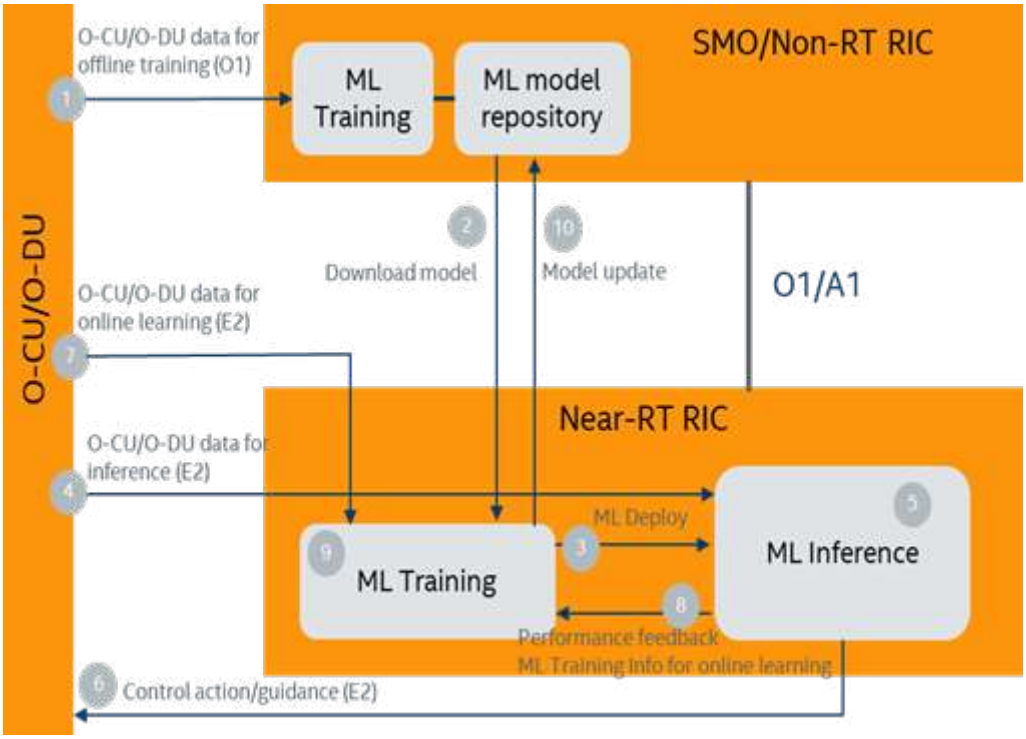
As model management, data preparation, AI/ML training, AI/ML inference and performance monitoring are implementation variability components, there are many combinations of deployment scenarios.

- Scenario 1: AI/ML Continuous Operation, AI/ML Model Management, Data Preparation, AI/ML Training, and AI/ML Inference are all in Non-RT RIC.
- Scenario 2: AI/ML Continuous Operation, Data Preparation (for training), and AI/ML Training are in Non-RT RIC while AI/ML Model Management is out of Non-RT RIC (in or out of SMO). Data Collection (for inference), Data Preparation (for inference), and AI/ML Inference are in Near-RT RIC.
- Scenario 3: AI/ML Continuous Operation and AI/ML Inference are in Non-RT RIC. Data Preparation, AI/ML Training, AI/ML Model Management are out of Non-RT RIC (in or out of 88SMO).
- Scenario 4: Non-RT RIC acts as the ML training host for offline model training while the Near-RT RIC as the ML training host for online learning and ML inference host.
- Scenario 5: Continuous Operation, Model management, Data Preparation, and ML Training host are in Non-RT RIC. O-CU and O-DU act as the ML inference host.

To explain one of these scenarios in more detail, Scenario 4 is considered (as shown in Figure 2-3), as it is one of the more widely recommended and adopted scenarios. The following steps are followed (not necessarily in a strict ordering):

- The data for offline training from the E2 nodes (O-CU-CP, O-CU-UP, O-DU, O-eNB), the Near-RT RIC and the O-RU are sent over the O1 and open M-plane fronthaul interfaces to the SMO. This data includes (but not limited to) PM data [56, 110], KPI data [57], CM data [58, 114], FM [113] – alarms and threshold crossing data [113], network interface trace data or UE-specific MDT trace data [111, 112], etc. This data is collected and stored in the SMO/Non-RT RIC framework.

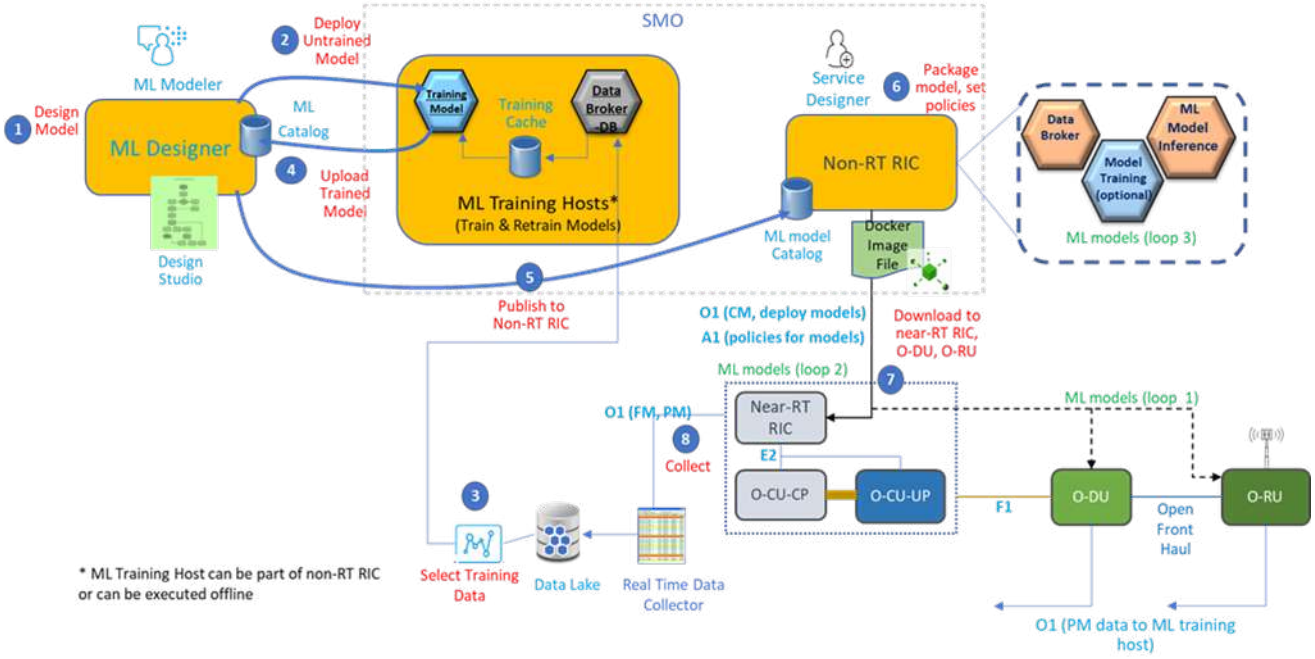
Figure 2-3 : Machine Learning deployment scenario (Scenario 4) [109]



Previously, a PM/FM/trace job control activation [59] would be initiated from the SMO to the E2 node and/or O-RU functions, either via some pre-configured application in the SMO/Non-RT RIC framework or via the rApps deployed in the Non-RT RIC.

- Upon data collection, the data is prepared, aggregated and fed to the AI/ML training host in the Non-RT RIC/SMO framework towards building the AI/ML training models, based on the choice of AI/ML algorithms (such as reinforcement learning – Q-learning/Deep Q-Network [DQN], Recurrent Neural Networks – Long Short Term Memory [LSTM], ARIMA time series prediction, deep learning, supervised learning techniques, etc. [107]). The AI/ML training host takes the aggregated data (PM/CM/FM/trace/topology, etc.) and the parameters to be optimized (target cell for HO, MU-MIMO layers, etc.) and trains the ML model based on a given choice of AI/ML algorithm towards meeting the target goal (maximize throughput, minimize latency, maximize reliability, minimize call drop, etc.) [6]. Previously, the SMO/Non-RT RIC would have received a request from the Near-RT RIC for offering AI/ML training services.
- The trained ML model is saved in the ML repository in the SMO/Non-RT RIC framework and is published to the ML catalog maintained in the SMO/Non-RT RIC framework. Once the Non-RT RIC responds to the Near-RT RIC via the A1 interface on the availability of the ML model sought by the Near-RT RIC along with details of the ML model catalog and the information on the repository where the ML model is stored, the ML model can be downloaded via the O1 interface and deployed in the inference engine of the Near-RT RIC.
- The Near-RT RIC can receive data from the E2 nodes (O-CU-CP, O-CU-UP, O-DU, O-eNB) via the E2 interface into the inference engine where the ML model is deployed. This data is based on the E2 Service Models (E2SMs) supported by the E2 nodes for the Near-RT RIC use cases and xApps.
- Based on the deployed ML model, the Near-RT RIC can perform inference based on E2 data reports received via E2AP Indication procedures. Once the Near-RT RIC makes inference, it generates control action or imperative policy guidance via E2AP control or policy procedure back to the E2 nodes via the E2 interface using the respective E2SMs.
- The deployed ML model can also be subject to online ML training updates based on the subsequent E2 data received by the Near-RT RIC from the E2 nodes, and based on continuous monitoring of network performance data and the feedback generated by the ML inference engine to the online ML training engine in the Near-RT RIC. The updated ML model is again deployed in the inference engine of the Near-RT RIC. If there is a significant update to the ML model, the Near-RT RIC uploads the updated model to the SMO/Non-RT RIC via the O1 interface and the model is updated in the ML model repository in the SMO/Non-RT RIC.

Figure 2-4 : ML life cycle management and implementation example [109]



The above steps are captured in the ML life cycle management and implementation example shown in Figure 2-4.

2.4 Interface Service Models for AI/ML-enabled XApp Design in Near-RT RIC

As discussed earlier, The E2 Service Model (E2SM) describes the functions in the E2 Node, which may be controlled by the Near-RT RIC and the related procedures, thus defining a function-specific RRM split between the E2 node and the Near-RT RIC. They describe a set of services exposed by the E2 node that shall be subsequently used by the Near-RT RIC and the hosted xApps. These services provide the Near-RT RIC with access to messages and measurements exposed from the E2 node (such as cell configuration information, supported slices, PLMN identity, network measurements, UE Context Information, etc.), that enable control of the E2 node from the Near-RT RIC. Multiple E2SMs have been defined in O-RAN WG3 such as E2SM-RAN Control (E2SM-RC), E2SM-Key Performance Monitoring (E2SM-KPM), E2SM-Network Interface (E2SM-NI), E2SM-Cell Configuration and Control (E2SM-CCC). In this section, we discuss how E2SMs are used in building ML models for xApps in the Near-RT RIC. This section discusses how E2SMs are used towards the building of AI/ML-enabled xApps in the Near-RT RIC as well as continuous network performance monitoring.

2.4.1 E2SM-KPM

Using E2SM-KPM [13], the E2 node can stream UE-level, cell-level and E2 node-level PM data across the layers of the RAN protocol stack at near-real-time granularities (ranging from 10 ms to 1 second) to the Near-RT RIC.

The PMs, standardized in O-RAN WG3, include packet delay measurements [56, 110] at PDCP, RLC, MAC/PHY layers, Radio Resource Utilization measurements, UE throughput measurements, RRC connection number/connection establishment/ re-establishment measurements, mobility management measurements (number of intra-RAT and inter-RAT handovers), transport block (TB) related measurements (number of TBs modulated with QPSK, 16QAM, 64QAM, 256QAM), CQI-related measurements (wideband and sub-band CQI), QoS flow-related measurements (number of QoS flows setup, release, modification), data radio bearer (DRB)-related measurements (number of DRBs setup, release, attempts, in-session activity time), received random access preambles per cell or synchronization signal block (SSB), distribution of RSRP values per SSB, number of UEs with active DRB transmission, packet loss rate due to over-the-air transmission losses, packet drop rate due to high PDCP traffic load, PDCP data volume measurements in terms of amount of successfully-transmitted PDCP SDU bytes, IP latency measurements due to buffering in the RLC layer caused by network congestion, UE and bearer context release measurements (average and distribution), call duration, the average or distribution of RSRP/RSRQ of UEs subject to handover with respect to the serving cell and the target cell, etc.

E2SM-KPM enables streaming these measurements from the E2 node to the Near-RT RIC at periodic intervals. It is to be noted that, while 3GPP TS 28.552 and TS 32.425 [56, 110] discuss streaming these PMs at a cell-level or an E2 node-level, E2SM-KPM additionally facilitates the reporting of these PMs at a per UE-level at near-RT periodicities.

2.4.2 E2SM-RC

Using E2SM-RC [14], the E2 node can stream or send the following information at UE-level to the Near-RT RIC at near-real-time periodicities, such as:

1. Context information

UE-specific RAN state and context information (L2 PDCP/RLC/MAC state variables, RLC buffer occupancy, etc. [117 – 119]), E2 node information (serving cell context information, neighbor cell information, etc.), UE-specific L3 RRC measurements (serving cell RRC, neighbor cell RRC measurements [115, 116]), RRC state information of the UEs [115, 116].

2. UE-specific signaling information

This includes information about the slice profile and the PDU sessions subscribed by the UEs, information about the QoS flows of the PDU sessions and their 5QI (or QCI) profile, information about the DRB and the 5QI profile, the mapping of QoS flows to DRBs, the primary serving cell for the UE and the secondary cells (in case of Carrier Aggregation), the secondary node for the UE (in case of EN-DC or MR-DC [120]), etc. [37]

3. Configuration information

This includes information pertaining to PDCP and RLC configuration for the UEs, PDCP duplication, cell selection/reselection priority for the UEs with respect to NR-specific ARFCN and EUTRA-specific EARFCN bands, MAC layer logical channel configuration, DRB split ratio, scheduler configuration information such as scheduling request periodicity, buffer status reporting periodicity, semi-persistent scheduling periodicity, discontinuous reception (DRX) cycle periodicity, number of HARQ processes and CQI configuration information, etc. [58, 59, 114]

4. Network interface or RRC messages

This includes reporting a copy of network interface messages between E2 nodes and RRC messages between the UE and the E2 node to the Near-RT RIC. [29 – 37, 115 – 116]

The Near-RT RIC can also exercise UE-level control actions back to the E2 nodes using E2SM-RC for functionalities that include (i) radio bearer control – such as controlling the QoS profile of the DRB, the mapping of QoS flows to the DRB, configuring the logical channel for the DRB, admission control for the DRB and the PDCP/RLC configuration, controlling the DRB termination, split ratio and PDCP duplication, (ii) radio resource allocation control – such as DRX parameter configuration, scheduling request periodicity configuration, semi-persistent scheduling periodicity control, grant configuration, etc. (iii) connected mode mobility control – such as choice of the optimal target cell for UE handover, conditional handover for UEs and DAPS to control the handover of URLLC UEs, (iv) radio access control – such as admission control for the UE in terms of PDU sessions, DRB configuration, RACH backboff control, access bearing control, RRC connection release control, RRC connection reject control, (v) dual connectivity control – in terms of choice of secondary node, PSCell, etc. (vi) carrier aggregation control – in terms of choice of secondary cells, (vii) idle mode mobility control – in terms of choice of cell reselection priority, (viii) measurement reporting configuration control – in terms of controlling the measurement objects, reporting objects, etc. [14]

Thus, E2SM-RC helps in exchange of fine-grained information involving the UEs, QoS flows, DRBs, PDU sessions, slices, cells, etc. at near-real-time granularities with the Near-RT RIC.

2.4.3 Other E2SMs

Other E2SMs, standardized (or being standardized) in O-RAN WG3, include E2SM-Cell Configuration and Control (E2SM-CCC) [82] and E2SM-Network Interface (E2SM-NI) [83]. E2SM-CCC enables reporting of cell-level and network element-level configuration information [58, 114] to the Near-RT RIC, and facilitates the Near-RT RIC to control the configuration parameters associated with the network elements at near-real-time periodicity. On the other hand, E2SM-NI enables tracing of UE-associated network interface messages from the E2 nodes to the Near-RT RIC, thereby enabling the Near-RT RIC gain access to fine-grained information about UE state information exchanged as part of network interface procedures between the E2 nodes.

The E2SMs thus facilitate reporting the relevant UE and RAN data as state information to ML/AI engine in the Near-RT RIC, and also facilitate optimization of control parameters by the Near-RT RIC xApps back to the E2 nodes.

2.5 Interface Data Models for AI/ML-enabled rApp Design in Non-RT RIC

The O1 and Open Fronthaul M-plane interfaces [24, 25, 18] are used for OAM, exercising RAN FCAPS functionality from the SMO over the O-RAN NFs such as O-CU-CP, O-CU-UP, O-DU, O-eNB and O-RU. The management service producer for the O-RAN NFs (such as O-CU-CP, O-CU-UP, O-DU, O-eNB, Near-RT RIC) produces management services that are exposed over O1 for consumption by the management service consumer in the SMO, which are subsequent used by the rApps in the Non-RT RIC. These services provide the rApps in the Non-RT RIC with access to messages and measurements exposed from the E2 node and the Near-RT RIC (such as cell configuration information, supported slices, PLMN identity, network measurements, UE Context Information, etc.), that enable CM of the E2 node from the rApps in the Non-RT RIC/SMO over R1 and O1 interfaces, as well as offering policy guidance and enrichment information and ML model training services over R1 and A1 to the xApps in the Near-RT RIC.

2.5.1 O1-PM and Open Fronthaul M-plane PM

Using O1-PM, the E2 node and Near-RT RIC can stream performance measurement data (largely cell-level and E2 node-level) at non-real-time granularities (in the order of > 1 sec) to the SMO.

The PMs, standardized in O-RAN WG10 [56, 110], include cell-level and E2 node-level packet delay measurements at PDCP, RLC, MAC/PHY layers, Radio Resource Utilization measurements, UE throughput measurements, RRC connection number/connection establishment/re-establishment measurements, mobility management measurements (number of intra-RAT and inter-RAT handovers), transport block related measurements (number of TBs modulated with QPSK, 16QAM, 64QAM, 256QAM), CQI-related measurements (wideband and sub-band CQI), QoS flow-related measurements (number of QoS flows setup, release, modification), DRB-related measurements (number of DRBs setup, release, attempts, in-session activity time), received random access preambles per cell or SSB, distribution of RSRP values per SSB, number of UEs with active DRB transmission, packet loss rate due to over-the-air transmission losses, packet drop rate due to high PDCP traffic load, PDCP data volume measurements in terms of amount of successfully-transmitted PDCP SDU bytes, IP latency measurements due to buffering in the RLC layer caused by network congestion, UE and bearer context release measurements (average and distribution), call duration, the average or distribution of RSRP/RSRQ of UEs subject to handover with respect to the serving cell and the target cell, etc.

O1-PM enables streaming and/or file-based reporting of these measurements from the O-RAN NFs (O-CU-CP, O-CU-UP, O-DU, O-eNB, etc.) to the SMO using PM job control configuration that describes the list of network elements and cells subject to PM streaming, measurement granularity, reporting periodicity, file-based reporting or streaming.

Similarly, the O-RU [18] also does a file-based reporting of PMs over the open fronthaul interface based on PM job control configuration from the SMO.

Though UE-level PMs are not streamed over O1-PM from the O-RAN NFs (O-CU-CP, O-CU-UP, O-DU, O-eNB), the Near-RT RIC, which receives UE-level PMs from the E2 nodes using E2SM-KPM, can further stream these PMs to the SMO based on job control configuration by the SMO, wherein the Near-RT RIC aggregates the UE-level PMs and exposes them at the configured reporting periodicity to the SMO over O1. The reason for doing so is to enable the Non-RT RIC function build offline ML training models that may also require UE-level PM data as input features, in addition to cell-level and E2 node-level data. Furthermore, the PMs related to Near-RT RIC functional procedures can also be streamed/reported to the SMO from the Near-RT RIC over O1-PM.

2.5.2 O1-Trace

Using O1-Trace [24, 25, 110, 111], the E2 nodes can stream/report cell-level call traces, UE-level MDT traces, Radio Link Failure (RLF) traces and RRC Connection Establishment Failure (RCEF) traces to the SMO via O1.

The cell-level call traces include tracing of network interface messages over X2, Xn, F1, E1, etc. involving the desired cells, whereas UE-level MDT traces involve UE-level signal quality measurements (for DL), data volume measurement for DL/UL per DRB per UE, average UE throughput measurement separately for DL/UL and per DRB per UE, packet delay measurement separately for DL/UL and per DRB per UE, etc. Similarly, RLF and RCEF traces are used for tracing failures concerning radio links and connection establishments, along with root cause diagnostic data.

Similar to PM, the tracing feature is also configured using trace job control procedures, which specifies the interfaces to be traced, the depth of tracing (minimum, medium, maximum), the tracing area scope such as the list of cells, tracking area code, trace events, etc.

2.5.3 O1-CM and Open Fronthaul M-plane CM

Using O1-CM [24, 25, 58, 114, 121, 122], the E2 nodes and the Near-RT RIC can report and notify their configuration parameters associated with the network elements over O1 to the SMO. Similarly, the O-RU can report its configuration parameters over the open fronthaul M-plane to the SMO.

These include configuration parameters concerning the E2 nodes (such as X2/Xn blacklist neighbors, X2/Xn whitelist neighbors, etc.), NR cells (such as ARFCN/frequency bands, PCI, neighbor cells), bandwidth parts (such as, sub-carrier spacing, etc.), cell sector, NR cell relation (such as, cell individual offset, whether HO is allowed towards a given neighbor cell, etc.), NR frequency relation (such as, cell selection and reselection priority, etc.), SON functionalities associated with the network elements/cells (such as, Automatic Neighbor Relations configuration, Centralized Energy Savings configuration, etc.), Radio Resource Management Policies (such as RRM policy ratio for Physical Resource Blocks [PRB] allocation for the cells/network slices/O-DUs, for RRC connected users in the O-CU-CP/ NR cells/network slices, for DRBs in the O-CU-UP), O-RU configuration such as antenna (beam tilt, antenna spacing), etc.

2.5.4 O1-FM and Open Fronthaul M-plane FM

These services are responsible for reporting errors and events to the SMO [24, 25, 113]. The SMO performs fault supervision operations on the O-RAN functions in the underlying network elements. The associated procedures include: (i) Fault Notification procedures [24, 59] – which include notifying new alarms by the FM service producer, notifying changed alarms by the FM service producer, notifying cleared alarms by the FM service producer, notifying alarm lists that got rebuilt by the FM service producer, subscribing and unsubscribing to network events by the FM service consumer in the SMO, retrieving the alarm list by the FM service consumer in the SMO, notification of correlations by the FM service producer, retrieval of alarm count in the SMO, (ii) Fault Supervision control procedures [24, 59]– which include acknowledgment and rejection of alarms by the FM service consumer in the SMO, clearing of alarms by the FM service consumer in the SMO, notification of changed acknowledgment state in the SMO, notification of potential faulty alarm list in the SMO.

2.5.5 O2 OAM

These services are responsible for FCAPS operations involving the O-Cloud platform, such as PMs, FMs, CMs etc. [19 – 20] towards O-Cloud infrastructure [21] and deployment management [22] services and provisioning of network, computational and storage resources for the Cloudified Network Functions in the O-Cloud platform.

2.5.6 A1 type definition for A1 policies

O-RAN WG2 has defined use-case-specific type definitions for generating A1 policies from the Non-RT RIC to the Near-RT RIC, towards which the rApps in the Non-RT RIC can leverage the RAN OAM data obtained via the O1 interface towards making informed AI/ML-driven policy recommendations to the Near-RT RIC for fine-grained RRM [5 – 8]:

- QoS target: The Non-RT RIC can generate policies related to Guaranteed Flow Bitrate, Maximum Flow Bitrate, Priority Level and Packet Delay budget for the QoS flows pertaining to a UE ID, a slice ID (given by Slice/Service Type [SST] and Slice Differentiator [SD]), a cell ID (given by NR-CGI or ECGI), a group ID (given by RFSP or SPID), etc.
- QoE target: The Non-RT RIC can generate policies related to QoE score, initial buffering, session rebuffering frequency, video stall ratio for the video sessions pertaining to a UE ID, a slice ID, a cell ID and/ or a group ID. The UE-level QoE targets involve setting objectives for UL/DL throughput, UL/DL packet delay, UL PDCP SDU packet loss rate, DL RLC SDU packet loss rate, UL/DL reliability, etc.
- Traffic Steering preferences: The Non-RT RIC can generate policies related to the list of cell IDs and/ or ARFCNs to be used as candidates for primary cell handover.
- Slicing SLA recommendations: The Non-RT RIC can generate policies related to maximum number of UEs to be admitted in a cell and/or an E2 node per slice, maximum number of PDU sessions to be admitted in an E2 node per slice, Guaranteed DL/UL throughput per slice, Maximum DL/UL throughput per slice, Maximum DL/UL packet delay per slice, maximum DL PDCP SDU packet loss rate per slice, maximum UL RLC SDU packet loss rate per slice, minimum DL/UL reliability per slice, maximum DL/UL jitter per slice, DL/UL priority per slice, etc. for slices pertaining to a UE, a UE group, a cell, QoS flows, etc.
- MIMO recommendations: The Non-RT RIC can generate policies over A1 related to whether individual UEs can be configured in SU-MIMO or MU-MIMO mode. These recommendations are then sent over A1 to the Near-RT RIC, which enables the Near-RT RIC make RRM decisions regarding the choice of beamforming indices for individual UEs.

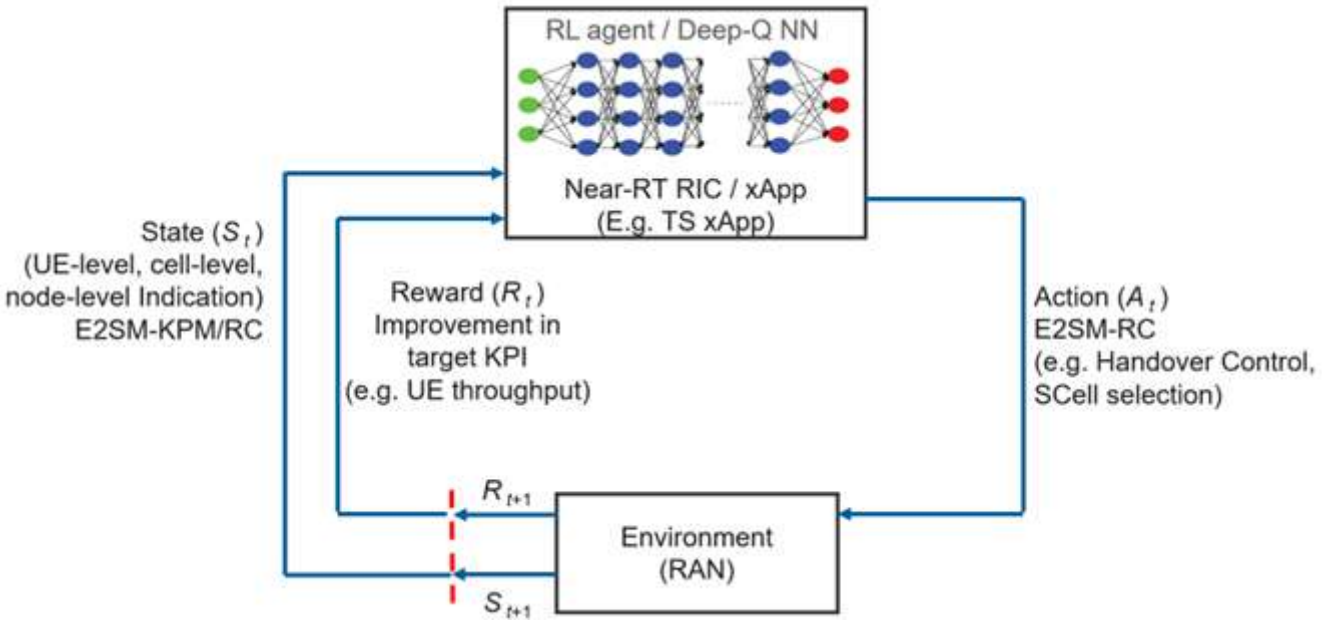
2.6 Advanced Learning Algorithms for xApp and rApp Development in the RIC

This section discussed advanced ML algorithms for xApp and rApp development in the Near-RT RIC and Non-RT RIC functions, respectively.

2.6.1 RL Algorithms in the RIC

As discussed in Section 2.1, the evolution of radio access technologies and newer use cases demand the necessity to use ML/AI algorithms that explore the complex and intricate inter-dependencies between the parameters, state, context and performance information across the layers of the RAN protocol stack for optimizing the RRM decisions of the control variables towards meeting the target KPI objective. The parameters, state, context and performance information generated from the E2 nodes as indications at UE-level/cell-level/node-level are considered as input features to the ML engine, which outputs the RRM decisions for the control parameters sent as control actions from the Near-RT RIC to the E2 nodes. The information sent via the indication messages from the E2 node to the Near-RT RIC and the control actions sent from the Near-RT RIC back to the E2 nodes are based on E2SMs [107], as discussed previously.

Figure 2-5 : Reinforcement learning engine in the Near-RT RIC based on E2SM [107].



It is to be noted that the Near-RT RIC cannot do offline ML model training, but can perform online ML model training and RL. The Non-RT RIC has the required computational and storage capacity to perform offline ML model training. So, the Near-RT RIC can request the Non-RT RIC to build an ML model with offline training, which can later be downloaded in the Near-RT RIC, where the model can be updated and deployed in the xApp that acts as an inference host. The general steps for building ML algorithms to facilitate intelligence in the RIC are as follows:

ML/AI model development using E2SM data: The E2 interface is used to send indications, containing relevant data, from the E2 node to the Near-RT RIC using E2SMs. The xApp in the Near-RT RIC shall access the data and may decide to request for ML/AI training services to the Non-RT RIC for generating an offline-trained ML model. The corresponding rApp in the Non-RT RIC shall receive this request. Once the Non-RT RIC receives the request for ML/AI training service from the Near-RT RIC, the rApp requests for data to be streamed from the Near-RT RIC. The xApp in the Near-RT RIC can use the O1 interface to send the E2SM data to the SMO at non-RT granularities. The SMO/Non-RT RIC stores the E2SM data and the rApp in the Non-RT RIC asks the Non-RT RIC/SMO framework to perform offline AI/ML training towards building an offline ML model and provides the required hyper-parameters, in the process, for the model training. The ML model is stored in a repository in the SMO/Non-RT RIC framework and the model is uploaded to the catalog. Once available, the ML model is downloaded in the Near-RT RIC over

the O1 interface, and the details are responded back to the Near-RT RIC over A1. The model is then deployed in the xApp that acts as the inference host. The ML model downloaded in the Near-RT RIC can further be subject to updates via online training based on the data sent to the Near-RT RIC from the E2 node. The updated ML model can be pushed to the SMO/Non-RT RIC, and the updated model is stored in the repository with the details updated in the catalog. The updated ML model is then deployed in the xApp as the inference host.

ML/AI-driven A1 policy: The Non-RT RIC rApps use the PM/KPI data received from the O-CU-CP, O-CU-UP and O-DU over the O1 interface for developing AI/ML-driven A1 policy and enrichment information, which are sent to the Near-RT RIC and consumed by the corresponding xApps. This A1 policy information is used by the Near-RT RIC xApps to set the Radio Resource Management objectives and targets for RRM.

Figure 2-5 illustrates how RL models can be built in the Near-RT RIC. The engine receives state information and computes rewards based on E2SMs (E2SM-KPM and E2SM-RC) that stream indication messages and generates control action to optimize RRM decisions based on E2SM-RC.

There are various categories of RL algorithms [123]:

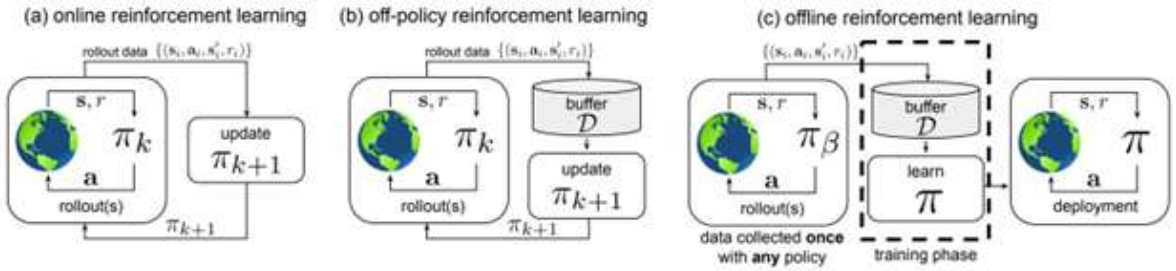
Model-free vs Model-based: Model-free RL algorithms do not model the state transition probability in the environment due to actions, but estimates the reward from state-action samples towards taking subsequent actions. Whereas model-based algorithms model the state transition probability to learn the inner-workings of the environment towards predicting the optimal control actions, accordingly.

Off-policy vs On-policy: In Off-policy RL algorithms, the target policy (the policy that the RL agent is trying to learn to determine and subsequently improve its reward value function) is different from the behavior policy (the policy used by the RL agent to generate action towards interacting with the environment). Off-policy RL agent makes use of a replay buffer which consists of data samples from the environment pertaining to all prior policies towards generating a newer/updated policy. On-policy RL algorithms use the same policy for both target and behavior.

Offline vs Online: In offline RL, a fixed training dataset of logged experiences is collected in a replay buffer based on any behavior policy, which could be potentially unknown. The RL agent is trained without any interactions with the environment, but based on this fixed offline training dataset of logged experiences. The policy is deployed online only after it is fully trained. Whereas, in online RL, the agent interacts with the environment online and a policy is updated to a newer policy based on the streaming data from the environment collected by the policy itself.

The illustrations of online RL, off-policy RL and offline RL are shown in Figure 2-6 [69]. Figure 2-6 (a) shows online reinforcement learning, where the RL agent interacts with the environment towards online exploration in order to update the target policy iterate from to . Figure 2-6 (b) shows online Off-policy RL, where the target policy is different from the behavioral policy. The RL agent employs a replay buffer D that consists of samples from various episodes pertaining to to by interacting with the environment using online exploration towards updating and determining the next target policy . Figure 2-6 (c) shows Offline RL, where the RL agent does not directly interface with the environment, but rather employs a replay buffer that stores samples pertaining to a behavioral policy , which are used by the RL agent in training the offline RL model towards determining the target policy . The trained RL model along with the target policy is then deployed in the inference engine towards exercising inferences back to the environment.

Figure 3-6 : Online RF, Off-policy RL and Offline RL [69]



Value-based RL vs policy-based RL: In value-based RL, the values of the action candidates based on the state vector are computed by the RL agent, and the action with the best value is determined; whereas, in policy-based RL, the RL agent learns the stochastic policy that maps the state vector to the action.

An RL agent can be present in the Near-RT RIC framework or in the xApp. As seen in Figure 2-5, the indication messages from the E2 nodes using the relevant E2SMs can constitute the state vector to the RL agent, and the state vector could be UE-specific. The E2 nodes and the underlying RAN constitute the environment. The KPI target and the objectives are set by the Non-RT RIC and are sent to the Near-RT RIC via the A1 interface, and the RL agent computes the reward as an improvement in the target KPI, further computed from the indication messages sent from the E2 node using E2SM. The E2SM CONTROL actions and the parameters constitute the actions taken by the RL agent in terms of optimizing the variables and decisions. Taking the example of an RL algorithm such as DQN, the Non-RT RIC can be leveraged for training the offline RL model for learning the reward as a function of the RAN data obtained from E2SM indication message content and the optimization variables controlled by E2SM CONTROL actions. Once the offline ML model is downloaded in the Near-RT RIC, the xApp can exploit the learnings of the ML model towards making inference on the control variables using E2SM, based on the incoming stream of indication data from the E2 nodes. Moreover, the downloaded ML model can also be subject to policy updates via further explorations by interacting with the environment. The RL agent can generate a random control action that gets reflected in the environment towards generating indication messages, which can be used to update the RL target policy in the RL agent. The updated model can then be subsequently deployed in the xApp, which makes further inferences based on the updated model. The updated model is also uploaded to the Non-RT RIC. And the updates to the ML can continue with further exploration, until convergence that minimizes the loss function (standard Bellman error, in the case of Q-learning) [107].

It is to be noted that RL can be applied on systems that are modeled as Markov Decision Processes (MDP) [123], where the probability of transition of the current state vector to the new state vector is dependent on the current state vector and the action taken by the RL agent towards transition to the new state vector. As an example, for the traffic steering O-RAN use case, the traffic steering xApp can make use of a RL agent that receives the UE-specific E2SM-KPM and E2SM-RC indication reports (containing UE context/state information and PMs, serving cells and serving E2 node context and PM information, UE's L3 RRC information for neighbor cells and neighbor cell context information, etc.) as the state vector from the E2 nodes (environment) to the Near-RT RIC, and the RL agent in the xApp can generate a UE-specific handover control action that optimizes the decision of the target cell for the UE towards optimizing the mobility/handover decisions that maximize a given KPI target for the UE (such as throughput/latency, etc). RL models are usually a good choice for closed-loop control systems such as the Near-RT RIC for optimizing the RRM decisions in the E2 nodes.

Thus, the Near-RT RIC leverages fine-grained UE-level intelligence, consisting of values of UE-specific PMs, KPIs, UE-specific state variables across the layers of the RAN protocol stack, UE-specific parameters exchanged across network interface procedures, information elements pertaining to entities of the individual UEs, such as DRBs, QoS flows, PDU sessions, PRB and logical channels pertaining to the UEs, etc. towards making optimized RRM decisions, down to the granularity of individual UEs, for a plethora of O-RAN use cases, as discussed below in Sec 3.6.3.

2.6.2 Other ML Algorithms in the RIC

The other ML algorithms include supervised learning, unsupervised learning, other forms of RL, etc. [109]

Supervised learning is an ML task that aims to learn a mapping function from the input to the output, given a labeled data set. Input data is called training data and has a known label or result. Supervised learning can be further grouped into Regression and Classification problems. Classification is about predicting a label whereas Regression is about predicting a quantity. Supervised learning algorithms include: (i) Regression: Linear Regression, Logistic Regression, (ii) Instance-based Algorithms: k-Nearest Neighbor (KNN), (iii) Decision Tree Algorithms: CART, (iv) Support Vector Machines: SVM, (v) Bayesian Algorithms: Naive Bayes, (vi) Ensemble Algorithms: Extreme Gradient Boosting, Bagging: Random Forest, (vii) Recurrent neural network models such as LSTM. Supervised learning algorithms in the RIC deal with prediction of RAN KPIs, network performance, QoE prediction, etc.

Unsupervised learning is an ML task that aims to learn a function to describe a hidden structure from unlabeled data. Input data is not labeled and does not have a known result. Some examples of unsupervised learning are K-means

clustering, principal component analysis (PCA) for root cause diagnostics that are useful for detecting anomalies in the network, etc.

Other reinforcement learning algorithms [123] that can be considered relevant to the RIC include multi-armed bandit learning, on-policy RL models such as State-Action-Reward-State-Action (SARSA), Proximal Policy Optimization (PPO), trust region policy optimization, etc., Actor-critic RL, and other variants of Q-learning models such as Deep Deterministic Policy Gradient (DDPG) which is useful for RL algorithms dealing with continuous action spaces, DQN for RL algorithms dealing with discrete action spaces, Monte-Carlo Tree Search for model-based RL, etc.

2.6.3 O-RAN Use Cases and XApp and RApp Functionalities

As a functionality of the O-RAN traffic steering use case [3, 124], a mobility management (a.k.a., handover) xApp [14] is responsible for making optimal decisions in terms of identifying the optimal target primary of individual UEs for handovers. The xApp is also responsible for optimizing the cell selection and reselection priority for the UEs in terms of which ARFCNs shall be prioritized to serve as primary cells. And a corresponding mobility management (handover) rApp is responsible for making policy recommendations and optimization objectives for handovers and sending them over A1 to the Near-RT RIC for enforcement of these policies towards making RRM decisions for choosing the optimal target cell.

As another functionality of the O-RAN traffic steering use case [3, 124] or as a functionality of the QoS optimization use case [3, 124], a carrier aggregation xApp [14] is responsible for choosing the optimal secondary cells based on the traffic subscribed by the UEs. And a corresponding carrier aggregation rApp is responsible for making policy recommendations on which ARFCNs could be secondary cells for the UEs and sending these policy recommendations along with optimization objectives for carrier aggregation and sending them over A1 to the Near-RT RIC for enforcing these policies towards making RRM decisions for making the optimal secondary cells.

As another functionality of the O-RAN traffic steering use case [3, 124], a dual connectivity xApp [14] is responsible for choosing the optimal secondary node for each UE. The xApp is also responsible for choosing which DRBs of the UE must be configured in dual connectivity mode. And the corresponding dual connectivity rApp is responsible for making policy recommendations on the inter-RAT B1 event measurement objects and the criteria for configuring a UE in DC mode and adding a secondary node for the UE. The rApp also recommends the 5QIs/QCIs of those DRBs that can be configured in EN-DC mode, and sends these policy recommendations and optimization objectives over A1 to the Near-RT RIC, which enforces these policies in making RRM decisions.

As another functionality of the O-RAN Quality-of-Service use case [3, 124], a QoS flow configuration xApp [14] is responsible for the QoS configuration of a DRB for a given UE, and in deciding which QoS flows subscribed by a UE need to be multiplexed to the DRB, based on the QoS profile of the 5QI flows. And the corresponding QoS flow configuration rApp is responsible for policy recommendations on which type of QoS flows and slices must be mapped to which type of DRBs, based on the 5QI of the QoS flows, and the 5QI/QCI of the DRBs and S-NSSAI of the slices. These recommendations, received by the Near-RT RIC over A1, are enforced by the Near-RT RIC towards making RRM decisions.

As a functionality of the O-RAN slicing use case [3, 124], an RRM allocation xApp [14, 82] is responsible for setting the UE- and/or cell- and/or O-DU-specific RRM policy ratio in terms of PRB allocation, cell- and/or O-CU-CP-specific RRM policy ratio in terms of number of RRC connected UEs, UE- and/or O-CU-UP-specific RRM policy ratio in terms of multiplexing slice-specific PDU sessions to DRBs, etc. The corresponding RRM allocation rApp can generate policy recommendations and set optimization objectives, as discussed in Sec 3.5.6, over A1 to the corresponding RRM allocation xApp towards enforcing RRM decisions by the xApp.

As a functionality of the O-RAN MIMO use case [3, 124], a beamforming configuration xApp [14] is responsible for setting the SU-MIMO or MU-MIMO configurations in terms of optimizing the number of MIMO layers, optimal SSB and CSI-RS beamforming weights based on the recommended transmission mode, etc., to name a few. The Near-RT RIC is also aided by AI/ML-driven declarative policies and enrichment information, generated from the Non-RT RIC via the A1 interface, based on recommendations for the MIMO use case discussed in Sec 3.5.6.

As another functionality of the O-RAN traffic steering use case [3, 125], a Frequency Layer Management rApp optimizes the automated load-balancing features of the radio network between different frequency layers. This rApp is composed of several smaller functions to provide the load

balance recommendations, and some of these functions are also shared with other centralized SON algorithms. The rApp supports load-balancing, where the goal is to improve user experience (primarily downlink throughput), resource utilization and spectral efficiency through the optimal re-distribution of users between frequency layers. This is achieved by producing deep insights into network and user characteristics using trace records, performance management and CM data. This knowledge is then used to produce and apply individually tuned load-balancing profiles using the distributed SON feature, load-based distribution at release (LBDAR). LBDAR is a RAN DSON (Distributed SON) feature that performs load-triggered distribution of UE at connection release. The rApp uses insights to tune the DSON feature and LBDAR profiles of congested cells via R1 and O1 interfaces [24, 126], thereby improving user experience (user throughput), spectral efficiency, and resource utilization.

As another functionality of the O-RAN QoE optimization use case [3, 125], a Performance Diagnostics rApp analyzes communications service providers' whole RAN to detect and classify cell issues. Identified issues are further investigated down to root cause level, enabling fast and accurate optimization of end-user performance. After the data processing and transformation is complete, the AI model detects anomaly cells and classifies coverage, handover or external interference issues, based on over a hundred network KPIs. Several dozen issue classes can be implemented in the solution. If a new issue type arises in the network, the AI software categorizes it as an "out-of-class" issue, enabling model retraining to be considered. The second step is the root cause analysis and reasoning, where other AI techniques are used to further break down a classified issue to its root cause level. It is possible to generate a knowledge graph that reveals the specific root causes that lead to an identified network issue. The rApp's output provides network engineers with actionable insights and enables faster and more effective optimization steps via R1 and O1 interfaces [24, 126].

As another functionality of the O-RAN QoE optimization use case [3, 125], a RET Optimization rApp leverages RL to enable continuous tilt optimization by learning how the performance of each cell reacts to antenna tilt changes. The ability to automatically adapt to the characteristics of each cell and the surrounding network leads to optimized radio environment and traffic distribution, thereby significantly improving the end-user experience. Benefits include:

- Adapting the optimization via R1 and O1 interfaces [24, 126] based on the characteristics of each cell and its influencing area,
- Throughput improvement (UL/DL) and DCR (dropped call rate) reduction while carrying more traffic by setting intelligent A1 policies [6,7],
- Continuous closed-loop optimization (via R1 and O1 interfaces [24, 126]) automatically maintaining the optimum settings as the network evolves and traffic distributions change.

In summary, Section 2 has vastly discussed the AI/ML framework in the RIC, which includes model management, data preparation, AI/ML training, AI/ML inference and performance monitoring, AI/ML-enabled xApp and rApp development for O-RAN use cases, ML and RL algorithms in Open RAN systems, etc. It is possible to implement many more use cases leveraging the AI/ML framework. It is however important to note that the framework itself is evolving as new proposals come in from the Open RAN ecosystem members and this will likely enable more flexibility going forward. Moreover, there are ongoing discussions in terms of evaluating the scale of ML model training and deployment in terms of computation and storage with the expansion of the network and deeper penetration of mobile UE devices in the network. In O-Cloud, where the O-RAN NFs including the RIC functions are deployed, large-scale ML training operations is supported by scaling up the number of containerized pods in the Kubernetes clusters. The increased AI/ML footprint with increase in the cloud infrastructure compute and storage resources would result in more vCPUs that would subsequently require more servers, thereby potentially driving up O-Cloud infrastructure costs for the O-RAN operator. Therefore, even as it is easy to scale the computational and storage resources of an increasingly large/dense network using O-Cloud infrastructure resources, the increased footprint may adversely impact cost efficiency. Hence, the operator needs to address the trade-off between increased accuracy and cost factor, while building AI/ML models for a large-scale network. Even as this increase in cost could be compensated in the foreseeable future by a higher revenue from an increased subscriber base, the operator may still have to make a compromise on the RIC platform's ability to generate efficient ML models with higher prediction accuracies for a larger/denser network, if cloud infrastructure scaling costs would have to be reduced. Future work would involve dimensioning benchmarks for the O-Cloud infrastructure compute and storage resources [20, 23] based on network size, subscriber base and AI/ML-based use-case requirements, and devising a cost model that would enable an operator to make informed decisions on performance – cost tradeoffs in Open RAN systems.

Conclusion

As an update to the previous white paper *Transition Toward Open and Interoperable Networks* published by 5G Americas in November 2020, this white paper started with a review and recap on the principles of Open RAN systems, and provided comprehensive detailing on the current updates to the Open RAN ecosystem survey, the architectural considerations of an Open RAN system, operator trials and deployments of Open RAN systems along with operational considerations and integration challenges. The paper then detailed on the role of AI/ML in Open RAN systems towards the realization of use cases, AI/ML functionality and life cycle management in O-RAN architecture, interface models and advanced learning algorithms for AI/ML-enabled xApp and rApp design in O-RAN systems.

In particular, the white paper focused on the recent advancements in the standardization activities concerning the O-RAN Alliance involving each O-RAN working group, the evolutions in the O-RAN architecture, the contributions to the OSC and the engagements of O-RAN testing and integration center with focus on the most recent O-RAN PlugFest events. The white paper also further discussed the Telecom Infrastructure Project (TIP) related working groups along with the focus on each group, the alignments between Open RAN systems-related standardization and 3GPP, mainly in; terms of which aspects in 3GPP form the basis of O-RAN Alliance and how 3GPP standards are evolving towards addressing the key focus areas in the design of Open RAN systems, mainly concerning AI/ML. The white paper also debriefed about the Open RAN policy coalition and U.S. Government initiatives concerning Open RAN, mainly in terms of recent legislations based on Open RAN in the U.S. Congress, and statements from the executive wing of the U.S. Government on Open RAN.

The white paper further detailed the key architectural considerations in Open RAN systems, with focus on O-RAN NF disaggregation and functional-split involving O-CU-CP, O-CU-UP, O-DU and the RIC functions, hybrid and hierarchical M-plane for O-RU OAM, principles of RAN cloudification and virtualization in O-RAN, and services-based architecture involving the RIC functions, operator trials and deployments involving Open RAN systems, and operational considerations and integration challenges involving brownfield operators, greenfield operators, realizable total cost of ownership, and performance considerations. The white paper then presented the advantages and challenges in adopting Open RAN architectures towards provisioning mobile telecommunication services.

The white paper then discussed the requirements and realization of O-RAN use cases, delineating the role of AI/ML as an integral component of O-RAN architecture and the RIC NFs for provisioning 5G and beyond 5G services. The paper detailed how O-RAN architecture is inherently equipped with AI/ML functionality, and delves into the architectural aspects of analytics and AI/ML framework functions in the Near-RT RIC and Non-RT RIC functions. The paper further discussed AI/ML life cycle management in O-RAN architecture, with specific focus on the E2 service models, the O1 data models and the A1 type definitions associated with the RIC interfaces (E2, O1 and A1) for AI/ML-enabled xApp and rApp design. The paper also described advanced ML/AI and RL algorithms for developing xApps and rApps towards meeting the requirements of O-RAN use cases in terms of network performance and user experience guarantees, and subsequently realizing the associated use-case functionalities. The paper concluded the discussions on AI/ML by detailing the process and challenges related to the scaling of computational and storage resources in O-Cloud platforms for building AI/ML training models, updating them and making inference decisions based on these models. In the process, the paper threw relevant insights on the performance-cost tradeoffs.

Appendix

Acronyms

AAU: Active antenna units

AI: Artificial Intelligence

ARFCN: Absolute Radio Frequency Channel Number

ASIC: Application-Specific Integrated Circuit

AWS: Amazon Web Services

BT: British Telecom

CM: Configuration Management

CNF: Cloudified Network Function

COTS: Commercial Off-the-shelf

CP: Control-Plane

CPU: Central Processing Unit

CQI: Channel Quality Indicator

CU: Centralized Unit

CUS-Plane: Control User Synchronized-Plane

DAPS: Dual Active Protocol Stack

DARPA: Defense Advanced Research Projects Agency

DCR: Dropped call rate

DDPG: Deep Deterministic Policy Gradient

DL: Downlink

DMS: Deployment Management Services

DQN: Deep Q-Network

DRB: Data radio bearer

DRX: Discontinuous Reception

DSON: Decentralized Self-Organizing Networks

DT: Deutsche Telekom

DU: Distributed Unit

EARFCN: E-UTRA (Evolved Universal Terrestrial Radio Access) Absolute Radio Frequency Channel Number

EDC: Edge Data Centers

ESF: Enduring Security Framework

FCAPS: Fault, Configuration, Accounting, Performance, Security

FCC: Federal Communications Commission

FEC: Forward error correction

FM: Fault Management

FPGA: Field Programmable Gate Arrays

GPPP: General-Purpose Processing Platforms

GPU: Graphical Processing Unit

HARQ: Hybrid Automatic Repeat request

HO: Handover

IMS: Infrastructure Management Services

IOT: Internet of Things

IP: Internet Protocol

KPI: Key Performance Indicator

LBDAR: Load-based distribution at release

LCM: Life Cycle Management

LDC: Local Data Centers

LF: Linux Foundation

LLS: Lower Layer Split

LSTM: Long Short Term Memory

LTE: Long Term Evolution

MDAF: Management Data Analytic Function

MDP: Markov Decision Processes

MDT: Minimization of Drive Test

MIMO: Multiple Input Multiple Output

ML: Machine Learning

NDC: National Data Center

NETCONF: NETwork CONFiguration protocol

NF: Network Function

NIC: Network Interface Card

NMS: Network Management System

NR: New Radio

NSA: Non-Stand-Alone

NTIA: National Telecommunications and Information Administration

NWDAF: Network Data Analytics Function

OAM: Operations, Administration and Maintenance

OCP: Open Compute Project

OFH: Open Fronthaul

ONAP: Open Networking Automation Platform

OpenRAN: Open Radio Access Network

OPEX: Operational Expense

OSC: O-RAN Software Community

OTIC: Open RAN Testing and Integration Center

PCA: Principal component analysis

PDCP: Packet Data Convergence Protocol

PDU: Protocol Data Unit

PHY: Physical Layer

PM: Performance Measurement

PNF: Physical Network Function

PPO: Proximal Policy Optimization

PRB: Physical Resource Blocks

QCI: QoS Class Index

QoE: Quality of Experience

QoS: Quality-of-Service

RACH: Random Access Channel

RAN: Radio Access Network

RCEF: RRC Connection Establishment Failure

RDC: Regional Data Centers

RET: Remote Electrical Tilt

RFSP: Index to RAT/Frequency Selection Priority

RIA: RAN Intelligence and Automation

RIC: RAN Intelligence Controller

RL: Reinforcement learning

RLC: Radio Link Control

RLF: Radio Link Failure

ROMA: RAN Orchestration and Management Automation

RRC: Radio Resource Connection

RRM: Radio resource management

RSRP: Reference Signal Received Power

RU: Radio unit

SA: Standalone

SARSA: State-Action-Reward-State-Action

SDAP: Service Data Adaptation Protocol

SLA: Service Level Assurance

SMO: Service Management and Orchestration

SMOS: Service Management and Orchestration Services

SON: Self-Organizing Networks

SPID: Subscriber Profile Identity

SSB: Synchronization Signal Burst

SSH: Secure Shell

SST: Slice/Service Type

TB: Transport Block

TCO: Total Cost of Reduction

TIFG: Testing Integration Focus Group

TIP: Telecom Infra Project

TLS: Transport Layer Security

TTI: Transmission Time Protocol

UE: User Equipment

UP: User Plane

UPF: User Plane Function

URLLC: Ultra-Reliable Low-Latency Communication

VM: Virtual Machine

VNF: Virtual Network Function

VR: Virtual Reality

YANG: Yet Another Next Generation – data modeling language

ZTA: Zero-Trust Architecture

References

[1] 5G Americas, “Transition Toward Open and Interoperable Networks,” A 5G Americas White Paper, 2020

[2] O-RAN Working Group 1, Architecture Task Group, “O-RAN Architecture Description”

[3] O-RAN Working Group 1, Use Case Task Group, “O-RAN Use Cases Detailed Specification”

[4] O-RAN Working Group 1, Network Slicing Task Group, “O-RAN Slicing Architecture”

[5] O-RAN Working Group 2, “A1 Interface: General Architecture and Principles”

[6] O-RAN Working Group 2, “A1 Interface: Application Protocol”

[7] O-RAN Working Group 2, “A1 Interface: Type Definitions”

[8] O-RAN Working Group 2, “A1 Interface: Use Cases and Requirements”

[9] O-RAN Working Group 2, “R1 Interface: General Architecture and Principles”

[10] O-RAN Working Group 2, “Non-RT RIC Architecture”

[11] O-RAN Working Group 3, “E2 Interface: General Architecture and Principles”

[12] O-RAN Working Group 3, “E2 Interface: Application Protocol”

[13] O-RAN Working Group 3, “E2 Service Model: Key Performance Monitoring”

[14] O-RAN Working Group 3, “E2 Service Model: RAN Control”

[15] O-RAN Working Group 3, “Near-RT RIC Architecture”

[16] O-RAN Working Group 3, “Near-RT RIC APIs specification”

[17] O-RAN Working Group 4, “Control, User and Synchronization Plane Specification”

[18] O-RAN Working Group 4, “Management Plane Specification”

[19] O-RAN Working Group 6, “O2 Interface: General Architecture and Principles”

[20] O-RAN Working Group 6, “Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN”

[21] O-RAN Working Group 6, “O2ims Interface Specification”

[22] O-RAN Working Group 6, “O2dms Interface Specification”

[23] O-RAN Working Group 6, “O-RAN Acceleration Abstraction Layer: General Architecture and Principles”

[24] O-RAN Working Group 10, “O-RAN Operations and Maintenance Interface Specification”

[25] O-RAN Working Group 10, “O-RAN Operations and Maintenance Architecture”

[26] O-RAN Working Group 11, “O-RAN Security Requirements Specification”

[27] O-RAN Working Group 1, “Decoupled SMO Architecture Technical Report”

[28] O-RAN Working Group 2, “AI/ML workflow description and requirements”

[29] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 37.480: E1 General Architecture and Principles”

[30] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 37.483: E1 Application Protocol”

[31] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 38.470: F1 General Architecture and Principles”

[32] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 38.473: F1 Application Protocol”

[33] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 38.420: Xn General Architecture and Principles”

[34] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 38.423: Xn Application Protocol”

[35] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 36.423: X2 Application Protocol”

[36] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 36.423: X2 Application Protocol”

[37] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, NG-RAN, “Technical Specification 38.401: NG-RAN Architecture description”

[38] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects, Management and Orchestration, "Technical Specification 28.104: Management Data Analytics (MDA)"

[39] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects, "Technical Specification 23.288: Architecture enhancements for 5G System (5GS) to support Network Data Analytics Services"

[40] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, E-UTRA and NR, "Technical Report: Study on enhancement for Data Collection for NR and EN-DC"

[41] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network Meeting, "RP-213602 - New Work Item Document on Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN"

[42] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network Meeting, "RP-213553 - New Work Item Document on further enhancement of data collection for Self-Organizing Networks / Minimization of Drive Test tracing in NR and EN-DC"

[43] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network Meeting, "RP-213594 - New Work Item Document on enhancement of NR QoE management and optimizations for diverse services"

[44] S.3189 -- 116th United States Congress, "A bill to use proceeds from spectrum auctions to support supply chain innovation and multilateral security", Jan 2020

[45] H.R.6624 -- 116th United States Congress, "Utilizing Strategic Allied Telecommunications Act of 2020", April 2020

[46] H.R.4032 -- 117th United States Congress, "Open RAN Outreach Act", Jun 2021

[47] United States Department of Commerce, National Telecommunications and Information Administration, "NTIA Comments on Promoting the Deployment of 5G Open Radio Access Network", GN-Docket No. 21 -- 63, July 2021, available at https://www.ntia.gov/fcc-filing/2021/ntia-comments-promoting-deployment-5g-open-radio-access-networks

[48] The White House, U.S., "Remarks by President Biden on his meetings in Saudi Arabia", July 2022, available at https://www.whitehouse.gov/briefing-room/speeches-remarks/2022/07/15/remarks-by-president-biden-on-his-meetings-in-saudi-arabia

[49] Peter Cohen, "BT, Nokia trial Open RAN RIC in Hull, U.K.", RCR Wireless, Jan 2022, available at https://www.rcrwireless.com/20220126/open_ran/bt-nokia-trial-open-ran-ric-in-hull-uk

[50] Alan Weissberger, " Deutsche Telekom Achieves End-to-end Data Call on Converged Access using WWC standards", available at https://techblog.comsoc.org/category/deutsche-telecom/, Sept 2022

[51] Juan Pedro Tomas, "Telecom Italia expands O-RAN footprint in Italy", RCR Wireless, Sept 2021, available at https://www.rcrwireless.com/20210915/business/telecom-italia-expands-o-ran-footprint-italy

[52] Bevin Fletcher, "Vodafone turns on first U.K. 5G Open RAN site", Fierce Wireless, Jan 2022, available at https://www.fiercewireless.com/tech/vodafone-turns-first-uk-5g-open-ran-site

[53] Bevin Fletcher, "Telefónica validates Open RAN 5G small cell with Qualcomm", Fierce Wireless, Feb 2022, available at https://www.fiercewireless.com/tech/telefónica-validates-open-ran-5g-small-cell-qualcomm

[54] A. Latif, A. Khamas, S. Goswami, V. P. Talari, Y. Jung, "Telco Meets AWS cloud: Deploying DISH's 5G Network in AWS Cloud", Feb 2022, available at https://aws.amazon.com/blogs/industries/telco-meets-aws-cloud-deploying-dishs-5g-network-in-aws-cloud/

[55] Matt Kapko, "Mavenir CTO details Dish's 5G Open RAN Journey on AWS", SDXCentral, Sept 2021, available at https://www.sdxcentral.com/articles/news/mavenir-cto-details-dishs-5g-open-ran-journey-on-aws/2021/09/

[56] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects, Management and Orchestration, "Technical Specification 28.552: 5G Performance Measurements"

[57] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects, Management and Orchestration, "Technical Specification 28.554: 5G Key Performance Indicators"

[58] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects, Management and Orchestration, "Technical Specification 28.541: 5G Network Resource Model"

[59] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects, Management and Orchestration, "Technical Specification 28.532: Generic Management Services"

[60] C. Kan, L. Linsheng, P. Jurcik, I. Wong, “O-RAN Global Plugfest 2021 demonstrates stronger ecosystem and maturing solutions”, O-RAN Testing and Integration Focus Group (TIFG) co-chairs, available at https://www.o-ran.org/blog/o-ran-global-plugfest-2021-demonstrates-stronger-ecosystem-and-maturing-solutions

[61] O-RAN Software Community (OSC) Documentation, “Welcome to O-RAN SC F Release Documentation Home”, available at https://docs.o-ran-sc.org/en/latest/

[62] Telecom Infra Project (TIP) OpenRAN at https://telecominfraproject.com/openran/

[63] Telecom Infra Project (TIP) OpenRAN, “TIP OpenRAN Release 2 Roadmap”, available at https://telecominfraproject.com/openran/

[64] Telecom Infra Project (TIP) OpenRAN, “OpenRAN 5G NR Base Station Platform Requirements document, available at https://telecominfraproject.com/openran/

[65] Telecom Infra Project (TIP) OpenRAN, “OpenRAN: The next generation of radio access networks”, available at https://telecominfraproject.com/openran/

[66] Telecom Infra Project (TIP) OpenRAN, “TIP OpenRAN: Towards Disaggregated Mobile Networking”, available at https://telecominfraproject.com/openran/

[67] U.S. Department of Defense, “Open Radio Access Network Security Considerations”, available at https://media.defense.gov/2022/Sep/15/2003077576/-1/-1/0/ESF_OPEN_RADIO_ACCESS_NETWORK_SECURITY_CONSIDERATIONS.PDF

[68] National Security Agency, “Enduring Security Framework (ESF)”, available at https://www.nsa.gov/About/Cybersecurity-Collaboration-Center/Cybersecurity-Partnerships/ESF/

[69] S. Levine, A. Kumar, G. Tucker, J. Fu, “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems”, arXiv:2005.01643, available at https://arxiv.org/abs/2005.01643

[70] B. Fletcher, “AT&T, Nokia prove open RIC platform in live 5G network”, Fierce Wireless, Jun 2020, available at https://www.fiercewireless.com/tech/at-t-nokia-prove-open-ric-platform-live-5g-network-trial

[71] I. Morris, “NTT Docomo pitches itself as Open RAN pre-integrator”, LightReading, Oct 2022, available at https://www.lightreading.com/open-ran/ntt-docomo-pitches-itself-as-open-ran-pre-integrator/d/d-id/781376

[72] O-RAN Working Group 5, “NR C-plane profile”

[73] O-RAN Working Group 5, “NR U-plane profile”

[74] O-RAN Working Group 5, “O1 interface specification for O-CU-CP and O-CU-UP”

[75] O-RAN Working Group 5, “O1 interface specification for O-DU”

[76] O-RAN Working Group 7, “O-RAN Deployment Scenarios and Base Station Classes”

[77] O-RAN Working Group 8, “O-RAN Base Station O-DU and O-CU Software Architecture and APIs”

[78] O-RAN Working Group 9, “O-RAN Synchronization Architecture and Solution Specification”

[79] S. Rose, O. Borchert, S. Mitchell, S. Connelly, “Zero-Trust Architecture”, NIST Special Publication, Technical Report 800-207, Aug 2020.

[80] O-RAN Alliance, “http://www.o-ran.org/”

[81] 3GPP Technical Report TR 38.801, “Study on new radio access technology: Radio access architecture and interfaces”

[82] O-RAN Working Group 3, “E2 Service Model (E2SM) - Cell Configuration and Control”

[83] O-RAN Working Group 3, “E2 Service Model (E2SM) - Network Interface”

[84] Telecom Infra Project (TIP) OpenRAN, “OpenCellular” at https://telecominfraproject.com/opencellular/

[85] Telecom Infra Project (TIP) OpenRAN, “Test and Integration” at https://telecominfraproject.com/test-and-integration/

[86] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network “Technical Specification 38.425: NR user plane protocol”

[87] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects “Technical Specification 23.288: Architecture enhancements for 5G System (5GS) to support Network data analytics services”

[88] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects “Technical Specification 28.104: Management Data Analytics Function”

[89] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects “Technical Specification 29.520: Network Data Analytics Services”

[90] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network "Technical Specification 37.817: Study on enhancement for Data Collection for NR and EN-DC"
[91] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network Meeting, "RP-213599 - AI/ML for NR Air Interface"
[92] H.R.4346 -- 117th United States Congress, "Chips and Science Act", Jul 2021
[93] "BT and Nokia trial Open RAN solution in Hull, U.K., to enhance mobile broadband experience", https://newsroom.bt.com/bt-and-nokia-trial-open-ran-solution-in-hull-uk-to-enhance-mobile-broadband-experience/
[94] "Telekom switches on O-RAN Town in Neubrandenburg", https://www.telekom.com/en/media/media-information/archive/telekom-switches-on-o-ran-town-in-neubrandenburg-630566
[95] "Dell, Intel, VMware, Mavenir power TIM's Open RAN", https://www.telecomlead.com/telecom-equipment/dell-intel-vmware-mavenir-power-tims-open-ran-101596
[96] Matt Kapko, "Rakuten Mobile Dismisses Open RAN Skeptics", https://www.sdxcentral.com/articles/news/rakuten-mobile-dismisses-open-ran-skeptics/2020/03/
[97] "Vodafone switches on U.K.'s first 5G Open RAN site" at https://www.vodafone.com/news/technology/5g-open-ran-first-uk-site
[98] Juan Pedro Tomas, "Telefónica validates its new O-RAN 5G standalone small cell", https://www.rcrwireless.com/20220215/open_ran/telefónica-validates-new-o-ran-5g-standalone-small-cell
[99] M. Alleven, "Dish confirms it's offering 5G to more than 20% of U.S. population", in https://www.fiercewireless.com/5g/dish-confirms-its-offering-5g-more-20-us-population
[100] I. Scales, "NTT DoCoMo gets its O-RAN working", in https://www.telecomtv.com/content/fronthaul/ntt-docomo-gets-its-o-ran-working-36364/
[101] "Nokia and AT&T run successful trial of the RAN Intelligent Controller over commercial 5G", in https://www.nokia.com/about-us/news/releases/2020/06/18/nokia-and-att-run-successful-trial-of-the-ran-intelligent-controller-over-commercial-5g/
[102] "Measurement Campaign xApp", https://wiki.o-ran-sc.org/display/RICA/Measurement+Campaign+xApp
[103] M. Kapko, "AT&T's vRAN Test with Nokia emboldens Open RAN vision", in https://www.sdxcentral.com/articles/interview/atts-vran-test-with-nokia-emboldens-open-ran-vision/2021/03/
[104] Juan Pedro Tomas, " The Rudin Family, Crown Castle launch multi-tenant CBRS network at 345 Park Avenue", https://www.rcrwireless.com/20210308/network-infrastructure/inbuildingtech/the-rudin-family-crown-castle-launch-multi-tenant-cbrs-network-at-345-park-avenue
[105] "Triangle Communications and Mavenir Announce First Deployed Network for FCC Rip and Replace", https://www.mavenir.com/press-releases/triangle-communications-and-mavenir-announce-first-deployed-network-for-fcc-rip-and-replace/
[106] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects, "Technical Specification 23.501: System Architecture for the 5G System (5GS) Stage 2"
[107] A. Lacava, M. Polese, R. Sivaraj, R. Soundrarajan, B. S. Bhati, T. Singh, T. Zugno, F. Cuomo, T. Melodia, " Programmable and Customized Intelligence for Traffic Steering in 5G Networks Using Open RAN Architectures", Networking and Internet Architecture [cs.NI], arXiv:2209.14171 at https://arxiv.org/abs/2209.14171
[108] "Validate your O-RAN radio units" in https://www.viavisolutions.com/en-us/products/tm500-o-ru-tester
[109] O-RAN WG2 Technical Report, "AI/ML workflow description and requirements"
[110] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects, "Technical Specification 32.425: Evolved Universal Terrestrial Radio Access Network - Performance Measurement"
[111] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects, "Technical Specification 32.422: Subscriber and equipment trace; Trace control and configuration management"
[112] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects, "Technical Specification 32.423: Subscriber and equipment trace; Trace data definition and management"
[113] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects, "Technical Specification 28.545: Fault Supervision"

[114] 3rd Generation Partnership Project (3GPP) Technical Specification Group Services and Systems aspects, “Technical Specification 28.658: Evolved Universal Terrestrial Radio Access Network (E-UTRAN) - Network Resource Model”
[115] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 38.331: NR Radio Resource Control”
[116] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 36.331: E-UTRA Radio Resource Control”
[117] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 38.323: NR Packet Data Converge Protocol (PDCP) specification”
[118] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 38.322: NR Radio Link Control (RLC) specification”
[119] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 38.322: NR Medium Access Control (MAC) specification”
[120] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 37.340: Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity”
[121] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 28.622: Generic Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)”
[122] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network, “Technical Specification 28.662: Generic Radio Access Network (RAN) Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)”
[123] Barto, A. G., Sutton, R. S. (2018). Reinforcement Learning: An Introduction. United States: MIT Press.
[124] O-RAN WG3, “Near-Real-Time RAN Intelligent Controller, Use Cases and Requirements”
[125] O-RAN WG2, “Non-RT RIC and A1 Interface: Use Cases and Requirements”
[126] O-RAN WG2, “R1 Interface: Use Cases and Requirements”

Acknowledgments

5G Americas' Mission Statement: 5G Americas facilitates and advocates for the advancement and transformation of LTE, 5G and beyond throughout the Americas.

5G Americas' Board of Governors members include Airspan Networks, Antel, AT&T, Ciena, Cisco, Crown Castle, Ericsson, Intel, Liberty Latin America, Mavenir, Nokia, Qualcomm Incorporated, Samsung, Shaw Communications Inc., T-Mobile USA, Inc., Telefónica, VMware and WOM.

5G Americas would like to recognize the significant project leadership and important contributions of group leader Dr. Rajarajan Sivaraj, Director of RIC Architecture and Standards, Mavenir along with many representatives from member companies on 5G Americas' Board of Governors who participated in the development of this white paper.