O-RAN.WG1.Use-Cases-Analysis-Report-v05.00

Technical Report

O-RAN Working Group 1
Use Cases Analysis Report

O-RAN ALLIANCE e.V.
Buschkauler Weg 27, 53347 Alfter, Germany
Register of Associations, Bonn VR 11238
VAT ID DE321720189

Revision History

Date	Revision	Author	Description
2019.10.17	01.00	Arda Akman (Netsia)	v01.00 publication
2020.03.11	02.00	Arda Akman (Netsia)	v02.00 publication
2020.07.16	03.00	Arda Akman (Netsia)	v03.00 publication
2020.11.02	04.00	Arda Akman (Netsia)	v04.00 publication
2021.02.28	05.00.01	Arda Akman (Juniper Networks)	Initial version of v5.0 - Removal of “Revision History” for v4.0 changes Document version number update to v05.00.01
2021.02.28	05.00.02	Arda Akman (Juniper Networks)	Addition of CR: - ATT-2020.10.07-WG1-CR-0001-Use Case Analysis Report BBU Pooling-v03
2021.03.07	05.00.03	Arda Akman (Juniper Networks)	Addition of CR: - SAM.AO-2021.03.04-WG1-CR-0001-Naming-Change for Massive_MIMO_SU_MU-MIMO_Optimization
2021.03.07	05.00.04	Arda Akman (Juniper Networks)	Addition of CR: - NOK-2021.03.03-ORAN-CR-0001-UseCaseAnalysisReport_mMIMO_BF_Optimization
2021.03.13	05.00	Arda Akman (Juniper Networks)	Updates per WG1 review process Spec renamed to v05.00 for publication

Contents

Chapter 1 Introduction

1.1 Scope

This Technical Report has been produced by O-RAN Alliance.

The contents of the present document are subject to continuing work within O-RAN WG1 Use Case Task Group and may change following formal O-RAN approval. In the event that O-RAN Alliance decides to modify the contents of the present document, it will be re-released by O-RAN Alliance with an identifying change of release date and an increase in version number as follows:

Release x.y.z

where:

x the first digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc. (the initial approved document will have x=01).

y the second digit is incremented when editorial only changes have been incorporated in the document.

z the third digit included only in working versions of the document indicating incremental changes during the editing process.

The current document describes potential O-RAN use cases as defined by O-RAN WG1 UCTG (Use Case Task Group). The use cases are described at a very high level, emphasizing how the use is enabled by O-RAN architecture along with basic input data expectations and resulting actions. These high level use cases are prioritized within O-RAN, and selected use cases are further detailed in O-RAN WG1 UCTG and relevant O-RAN WGs to define the requirements for O-RAN components and their interfaces.

1.2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document in Release 16.

[1] 3GPP TR 21.905: “Vocabulary for 3GPP Specifications”

[2] 3GPP TS 22.261: “Service requirements for the 5G system; Stage 1”, Release 16, December 2019

[3] 3GPP TS 23.285: “Architecture enhancements for V2X services”, Release 16, June 2019

[4] 3GPP TS 23.501: “System Architecture for the 5G System (5GS); Stage 2”, Release 16, December 2019

[5] 3GPP TS 28.530: “Management and orchestration; Concepts, use cases and requirements”, Release 16, September 2019

[6] 3GPP TS 28.541: “5G Network Resource Model (NRM); Stage 2 and stage 3”, Release 16, January 2020

[7] 3GPP TS 28.552: “Management and orchestration; 5G performance measurements”, Release 16, September 2019

[8] 3GPP TS 36.300: “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 16)”, Release 16, January 2021

[9] 3GPP TS 36.902: “Evolved Universal Terrestrial Radio Access Network E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (Release 9)”, Release 9, April 2011

[10] 3GPP TS 37.340: “NR; Multi-connectivity; Overall description; Stage-2”, Release 16, April 2020

[11] 3GPP TS 38.305: “NG Radio Access Network (NG-RAN); Stage 2 functional specification of User Equipment (UE) positioning in NG-RAN”, Release 16, April 2020

[12] 3GPP TS 38.321: “Medium Access Control (MAC) protocol specification”, Release 16, September 2019

[13] 3GPP TS 38.331: “NR; Radio Resource Control (RRC) protocol specification (Release 16)”, Release 15, January 2021

[14] 3GPP TR 38.802: "Study on new radio access technology Physical layer aspects", Release 14, September 2017

[15] 3GPP TR 38.889: “Study on NR-based access to unlicensed spectrum”, Release 16, December 2018

[16] 3GPP TSG-RAN WG3 Meeting #101-Bis

[17] 3GPP TSG-RAN WG3 Meeting #104

[18] ETSI EN 302 637-2: “Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service”, Release 1, November 2010

[19] ETSI EN 302 637-3: “Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service”, Release 1, November 2014

[20] GSMA Future Networks: “Infrastructure Sharing: An Overview”, June 2019

[21] GSMA NG.116: “Generic Network Slice Template”, Version 2.0, October 2019

1.3 Definitions and Abbreviations

1.3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

A1: Interface between non-RT RIC and Near-RT RIC to enable policy-driven guidance of Near-RT RIC applications/functions, and support AI/ML workflow.

A1 policy: Type of declarative policies expressed using formal statements that enable the non-RT RIC function in the SMO to guide the near-RT RIC function, and hence the RAN, towards better fulfilment of the RAN intent.

A1 Enrichment information: Information utilized by near-RT RIC that is collected or derived at SMO/non-RT RIC either from non-network data sources or from network functions themselves.

E2: Interface connecting the Near-RT RIC and one or more O-CU-CPs, one or more O-CU-UPs, and one or more O-DUs.

E2 Node: a logical node terminating E2 interface. In this version of the specification, O-RAN nodes terminating E2 interface are:

- for NR access: O-CU-CP, O-CU-UP, O-DU or any combination;
- for E-UTRA access: O-eNB.

FCAPS: Fault, Configuration, Accounting, Performance, Security.

Intents: A declarative policy to steer or guide the behavior of RAN functions, allowing the RAN function to calculate the optimal result to achieve stated objective.

Near-RT RIC: O-RAN near-real-time RAN Intelligent Controller: a logical function that enables real-time control and optimization of RAN elements and resources via fine-grained data collection and actions over E2 interface.

Non-RT RIC: O-RAN non-real-time RAN Intelligent Controller: a logical function 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/features in Near-RT RIC.

O-CU: O-RAN Central Unit: a logical node hosting O-CU-CP and O-CU-UP

O-CU-CP: O-RAN Central Unit – Control Plane: a logical node hosting the RRC and the control plane part of the PDCP protocol.

O-CU-UP: O-RAN Central Unit – User Plane: a logical node hosting the user plane part of the PDCP protocol and the SDAP protocol.

O-DU: O-RAN Distributed Unit: a logical node hosting RLC/MAC/High-PHY layers based on a lower layer functional split.

O-RU: O-RAN Radio Unit: a logical node hosting Low-PHY layer and RF processing based on a lower layer functional split. This is similar to 3GPP's "TRP" or "RRH" but more specific in including the Low-PHY layer (FFT/iFFT, PRACH extraction).

O1: Interface between management entities (NMS/EMS/MANO) and O-RAN managed elements, for operation and management, by which FCAPS management, Software management, File management shall be achieved.

RAN: Generally referred as Radio Access Network. In terms of this document, any component below Near-RT RIC per O-RAN architecture, including O-CU/O-DU/O-RU.

1.3.2 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

AI/ML	Artificial Intelligence/Machine Learning
eNB	eNodeB (applies to LTE)
gNB	gNodeB (applies to NR)
KPI	Key Performance Indicator
KQI	Key Quality Indicator
MIMO	Multiple Input, Multiple Output

PRB	Physical Resource Block
QoE	Quality of Experience
RIC	O-RAN RAN Intelligent Controller
SINR	Signal-to-Interference-plus-Noise Ratio
UAV	Unmanned Aerial Vehicle
V2X	Vehicle to Everything
SMO	Service Management and Orchestration

Chapter 2 Objective

This document provides O-RAN WG1 high level use case descriptions. Any use case defined in O-RAN is expected to be documented in this report. If the use case is to be studied further, it will be covered in O-RAN WG1 detailed use case specification next, and then in relevant WGs. It should be noted that not all of the use cases presented here are currently supported by O-RAN specifications and these use cases will be addressed in future O-RAN work.

Chapter 3 Use Cases

3.1 Use case 1: Context-Based Dynamic HO Management for V2X

3.1.1 Background Information

V2X communication allows for numerous potential benefits such as increasing the overall road safety, reducing emissions, and saving time. Part of the V2X architecture is the V2X UE (SIM + device attached to vehicle) which communicates with the V2X Application Server (V2X AS). The exchanged information comprises Cooperative Awareness Messages (CAMs) from UE to V2X AS [18], radio cell IDs, connection IDs, and basic radio measurements (RSRP, RSPQ etc.).

3.1.2 Motivation

As vehicles traverse along a highway, due to their high speed and the heterogeneous natural environment V2X UE-s are handed over frequently, at times in a suboptimal way, which may cause handover (HO) anomalies: e.g., short stay, ping-pong, and remote cell. Such suboptimal HO sequences substantially impair the functionality of V2X applications. Since HO sequences are mainly determined by the Neighbour Relation Tables (NRTs), maintained by the xNBs, there is hardly room for UE-level customization.

This UC aims to present a method to avoid and/or resolve problematic HO scenarios by using past navigation and radio statistics in order to customize HO sequences on a UE level. To this end, the AI/ML functionality that is enabled by the Near-RT RIC is employed.

3.1.3 Proposed Solution

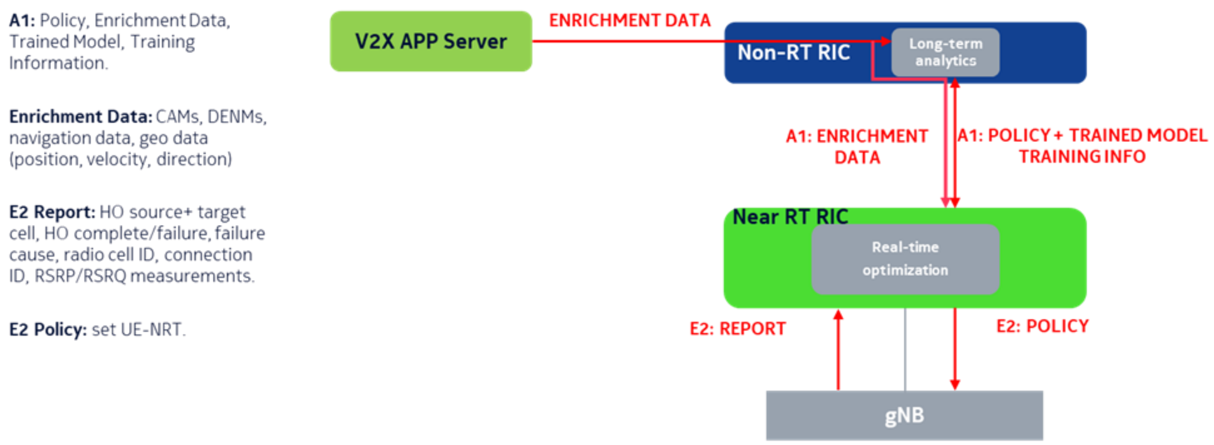


Figure 3.1.3- 1: Dynamic Handover Management for V2X use case

In order to prevent anomalous HO sequences causing performance degradation for the V2Xs, an xApp can be built with two main functionalities. Applying long-term analytics by using AI/ML algorithms is the first function expected from the xApp. As it is suggested by O-RAN framework, non-real-time intelligent management of RAN functions are deployed in the Non-RT RIC. V2X AS maintains the UE-based HO events and mobility data which are shared with Non-RT RIC over O1 interface. Hence, Non-RT RIC can identify causes of HO anomalies and discover optimal HO sequences. Finally, a data base is maintained to keep track of resolutions to anomalies based on these feedbacks. The other function of the xApp is performing real-time optimization which is offered by trained ML model on the Near-RT RIC. Near-RT RIC will monitor UE specific real-time mobility context based on V2X data. Deployed ML model is used to detect/predict unexpected HO events and generate desired HO sequence which can eliminate/avoid HO anomaly. Finally, Near-RT RIC is expected to create and update UE-specific NRTs to improve performance of the V2X users.

The xNB is assumed to host, besides the default NRT, also UE-specific NRTs (UE-NRTs) for V2X (and potentially other types of) users. If the UE-NRT for a specific V2X user exists, it is used. If not, the default NRT remains valid.

The input samples for handover profiling can come from the V2X AS and the xNB:

- CAMs, navigation data, geo data (position, velocity, direction),
- Radio cell ID, connection ID, basic radio measurements (RSRP, RSPQ etc.),
- Handover event (source and target cell IDs),
- Handover success / error codes.

The long-term analytics function and the real-time optimization function can be thought of as ML training and ML inference, respectively.

3.1.4 Benefits of O-RAN Architecture

The long-term optimization function can jointly process UE level radio and navigation data. Both the analysis and the prediction functions can employ a range of AI/ML methods, whose availability is facilitated by the O-RAN architecture.

3.1.5 Notes / FFS

An option for associating the data coming from the gNB and that coming from the V2X AS can be achieved by the ECGI+C-RNTI [3][8]. The ECGI uniquely identifies the cell, while the C-RNTI uniquely identifies a UE within a gNB. The ECGI+C-RNTI pair constitutes a globally unique UE identifier.

3.2 Use case 2: Flight Path Based Dynamic UAV Radio Resource Allocation

3.2.1 Background Information

This use case provides the background, motivation, and requirements for the support the use case of flight path based dynamic UAV Radio Resource Allocation, allowing operators to adjust radio resource allocation policies through the O-RAN architecture, reducing unnecessary handover and improving radio resource utilization.

3.2.2 Motivation

The field trials' results show that the coverage for low altitude is good and can provide various services for terrestrial UEs with good performance. However, since the site along the flight is mainly for terrestrial UEs, the altitude of the UAV is always not within the main lobe of the ground station antenna. And the side lobes give rise to the phenomenon of scattered cell associations particularly noticeable in the sky. The cell association pattern on the ground is ideally contiguous area where the best cell is most often the one closest to the UE. As the UE move up in height, the antenna side lobes start to be visible, and the best cell may no longer be the closest one. The cell association pattern in this particular scenario becomes fragmented especially at the height of 300m and above. Hence, at higher altitudes, several challenges that lead to a different radio environment are:

- LOS propagation/uplink interference
- Poor KPI caused by antenna side lobes for base stations
- Sudden drop in signal strength

These challenges directly impact on the mobility performance of the drone and the service experience of the user. Hence, we would like to support the use case of flight path based dynamic UAV Radio Resource Allocation to resolve the above issues.

3.2.3 Proposed Solution

To manage resource allocation of UAVs based on Flight Path, Non-RT RIC retrieves necessary Aerial Vehicles related metrics from network level measurement reports. Moreover, flight path information of Aerial Vehicle, climate information, flight forbidden/limitation area information and Space Load information are received from an application, e.g. UTM (Unmanned Traffic Management) for constructing/training relevant AI/ML model. For example, this could be UL/DL interference from/to Aerial vehicles, the detection of Aerial Vehicle UEs, and available radio resource (e.g. frequency, cell, beam, BWP, numerology) prediction. The Near-RT RIC supports deployment and execution of AI/ML model from Non-RT RIC. Based on this, Near-RT RIC can perform the radio resource allocation for UAVs considering the radio channel condition, flight path information and other application information with on-demand coverage. Moreover, Aerial Vehicles performance reports should be sent to Non-RT RIC for evaluation and optimization of the ML model.

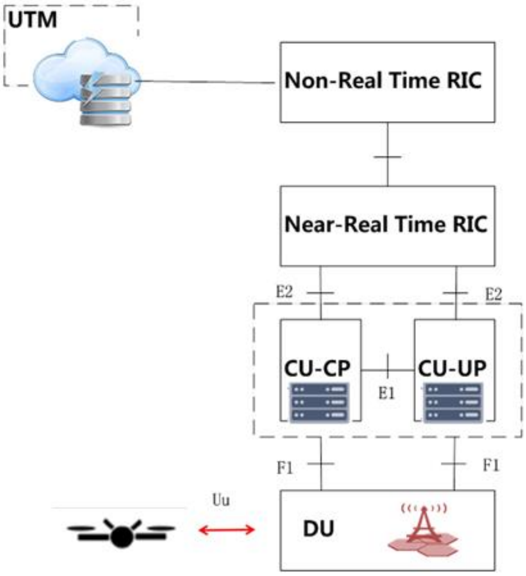


Figure 3.2.3-1: Use case of flight path based dynamic UAV Radio Resource Allocation

3.2.4 Benefits of O-RAN Architecture

An effective functional module, retrieving the application information, performing machine learning and training is not provided by current gNB/eNB architecture. Though, the O-RAN architecture components can collect both the acquired application information and radio environment information, and execute AI/ML models based on received information. The flight path based dynamic UAV Radio Resource Allocation mechanism can be supported by the RIC function modules, i.e. Non-RT RIC and Near-RT RIC. Therefore, we provide the description of O-RAN support use case for flight path based dynamic UAV Radio Resource.

3.3 Use case 3: Radio Resource Allocation for UAV Application Scenario

3.3.1 Background Information

As shown in Figure 3.3.1-1, this scenario refers to a Rotor UAV flying at low altitude and low speed and carrying cameras, sensors and other devices mounted. The Operation Terminal uses the 5.8 GHz frequency band remote control the UAV for border/forest inspection, high voltage/base station inspection, field mapping, pollution sampling, and HD live broadcast.

At the same time, the UAV control mobile station and the UAV anti-weapon combination provide the UAV control, fight against illegal UAV and other services to ensure low-altitude safety in special areas.

The UAV Operation terminal, the anti-UAV weapon, and the UAV control mobile station are connected with the UAV Control Vehicle using 5G network. UAV Control Vehicle deploys network equipment, including O-CU,O-DU, the Non-RT RIC, the Near RT RIC function modules and Application Server (In this use case, it is an Edge computing Service Platform) to provide reliable network services through 5G networks.

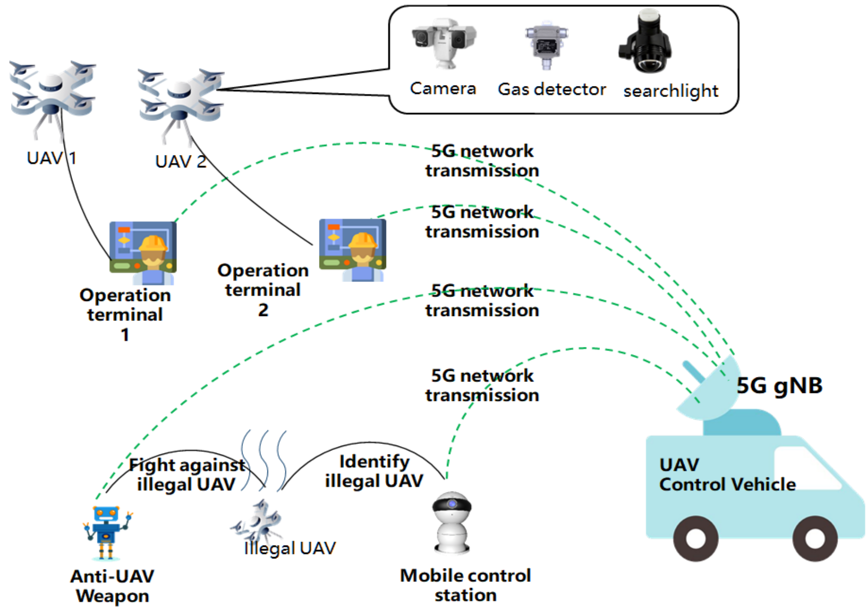


Figure 3.3.1-1: UAV Control Vehicle Application scenario

3.3.2 Motivation

UAV terminal access is a new scenario of 5G networks. It has a large amount of high real-time data that is more suitable for local processing. For the existing radio resource allocation strategy, there is no solution for such uplink high-rate data transmission.

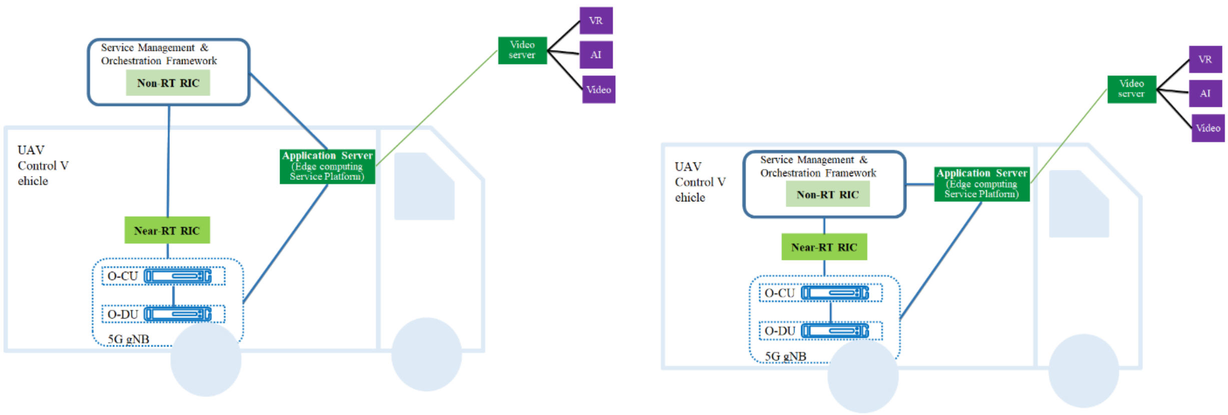
In this use case, UAV Control Vehicle is required to provide reliable network services through 5G networks. The data of the UAV terminal interacting with the network includes control data and application data, and the control data includes navigation commands, configuration changes, flight status data reporting, etc. Control data requires low latency and low bandwidth requirements. The application data includes 4K high-definition video data, which has obvious uplink and downlink service asymmetry, and the uplink has high requirements on network bandwidth.

The UAV Control Vehicle deploys edge computing services that application services and content are placed on the edge instead of core network so that local processing of video and control information can be managed. At the same time, real-time data services can be provided to third-party applications through a video server.

3.3.3 Proposed Solution

As shown in Figure 3.3.3-1, this use case involves two options of network architecture. Option A is that O-CU, O-DU and Near-RT RIC are deployed on the Control Vehicle, Non-RT RIC and core network are deployed on the central cloud. The Control Vehicle is connected to core network and Non-RT RIC via fiber optics. Option B is a private network, all modules, including the gNB, Near-RT RIC, Non-RT RIC and the necessary core network function modules, are deployed in the Control Vehicle.

In both deployment options, radio resource allocation for UAV application use case requires some of the basic functionalities of the O-RAN components. Non-RT RIC shall retrieve UE-level radio resource requirements from Application Server to generate UE based radio resource allocation policies. In this scenario, Near-RT RIC shall support execution and interpretation of retrieved policies to create configuration parameters to be applied on the E2 nodes. In addition to UE-specific radio resource adjustment with respect to received parameters, E2 nodes shall provide information about UE registration or UE status change to Near-RT RIC so that it can be transferred to Application Server. As it is stated in previous section, UAV terminals require low latency and high throughput in the uplink direction. For this reason, Application Server is needed to receive user plane data from UAV terminals.



Option A Option B

Figure 3.3.3-1: Network architecture for UAV Control Vehicle Application scenario

3.3.4 Benefits of O-RAN Architecture

The 5G network supports real-time high-definition video transmission and remote low-latency control of UAV, and finally provides various industry services such as inspection, security, surveying and mapping.

In the UAV Control Vehicle Application scenario, there are a small amount of control data interaction requirements between the terminal and the network interaction, as well as the large bandwidth requirements for uploading HD video.

The service asymmetry raises new requirements for resource allocation of the gNB. At the same time, the existing network operation and maintenance management platform (OSS system) can only optimize the parameters of a specific group of UEs, but not for individual users. Therefore, under the O-RAN architecture, the radio resource requirements for different terminals are sent to the gNB for execution by means of the Near-RT RIC function module.

The UAV control vehicle has flexible layout features. In this use case, the application service and content are deployed on the edge computing platform instead of the core network; the Non-RT RIC function module is used to schedule radio resources instead of the core network's QoS mechanism. In this way, the load and overhead of the core network can be reduced, and the forwarding and processing time in data transmission can be reduced, and the delay can be reduced.

3.4 Use case 4: QoE Optimization

3.4.1 Background Information

The highly demanding 5G native applications like Cloud VR are both bandwidth consuming and latency sensitive. However, for such traffic-intensive and highly interactive applications, current semi-static QoS framework cannot efficiently satisfy diversified QoE requirements especially taking into account potentially significant fluctuation of radio transmission capability. It is expected that QoE estimation/prediction from application level can help deal with such uncertainty and improve the efficiency of radio resources, and eventually improve user experience.

3.4.2 Motivation

The main objective is to ensure QoE optimization be supported within the O-RAN architecture and its open interfaces. Multi-dimensional data, e.g., user traffic data, QoE measurements, network measurement report, can be acquired and processed via ML algorithms to support traffic recognition, QoE prediction, QoS enforcement decisions. ML models can be trained offline and model inference will be executed in a real-time manner. Focus should be on a general solution that would support any specific QoE use case (e.g. Cloud VR, video, etc.).

3.4.3 Proposed Solution

Traditional technologies involves manual configuration of RAN parameters for the congested cells to improve QoE of the users. However, O-RAN architecture facilitates QoE optimization in real-time with proactive closed-loop network optimization. Limited radio resources are utilized in a way that congested cells are dynamically detected and resources allocated to critical users.

The proposed solution consists of O-RAN components, Non-RT RIC, Near-RT RIC and E2 Nodes, empowered with Machine learning algorithms. The main objective of Non-RT RIC is constructing AI/ML model that is trained with data retrieved from SMO, network level measurements and policies to be sent Near-RT RIC for managing RAN parameters. The ML model will be deployed in the Near-RT RIC to assist QoE optimization such as making predictions on application/traffic types, QoE and available bandwidth. To achieve all these functions, E2 Nodes shall provide the PMs with required granularity to SMO over O1. Also, RRM behaviour updates shall be allowed by E2 Nodes through E2 to support QoS enforcement.

3.4.4 Benefits of O-RAN Architecture

The proposed solution to support a real-time QoE optimization for any specific use e.g. Cloud VR, video requires features provided by O-RAN architecture. Retrieving of measurement metrics, AI/ML training and executing the AI/ML model in near-real time are main properties offered by the O-RAN architecture.

3.5 Use case 5: Traffic Steering

3.5.1 Background Information

5G systems will support many different combinations of access technologies namely; LTE (licensed band), NR (licensed band), NR-U (unlicensed band), Wi-Fi (unlicensed band), (see TR 38.889 [15]). Several different multi-access deployment scenarios are possible with 5GC to support wide variety of applications and satisfy the spectrum requirements of different service providers;

- Carrier aggregation between licensed band NR (Primary Cell) and NR-U (Secondary Cell)
- Dual connectivity between licensed band NR (Primary Cell) and NR-U (Secondary Cell)
- Dual connectivity between licensed band LTE (Primary Cell) and NR-U (Secondary Cell)
- Dual connectivity between licensed band NR (Primary Cell) and Wi-Fi (Secondary Cell)

Note: The scenario of dual connectivity between NR and Wi-Fi is for future study

The traffic steering use case allows, using the A1 interface, flexibly configure the desired optimization policies and utilize the appropriate performance criteria to proactively manage user traffic across different access technologies. A1 interface can also provide the enrichment information, e.g., radio fingerprint information based on the data analytics of the historical RAN data

3.5.2 Motivation

Imbalances in the traffic load across cells of different access technologies, or variances in their available bandwidth and QoS attributes, may give rise to situations with suboptimal spectrum utilization and negative impact on the user experience. The 3GPP Self-Organizing Network (SON) function Mobility Load Balancing (MLB) is the legacy approach to improve the user experience by balancing the load through optimization of the handover triggers and handover decisions using load information shared between neighbouring cells. Normally all user are treated equally in this respect. In addition, the statistical characteristics of the radio network information have not been fully exploited and utilized to enhance the network and user experience performance.

3.5.3 Proposed Solution

3.5.3.1 Policy based traffic management optimization

Impairments of the user experience due to local load imbalances across cells, or due to variances in their available bandwidth offered, are example scenarios addressed by traffic management policies over the A1 interface. The traffic management policy allows allocating cells in order of priority to individual users, for the control plane and user plane respectively. A traffic management policy could be issued for any reason, e.g. reasons not known to the RAN.

The Non-RT RIC monitors the user experience by UE level performance measurements and the resource utilization on cell level. The Non-RT RIC assesses the observed performance vs. the expected service level requirements. If the requirements are breached the Non-RT RIC locates the cell where the breach is detected and assesses the local load or bandwidth conditions in the associated cells. It may desire to relocate one or more users to other cells e.g. in order to increase their available radio resources, offer increased bandwidth, desire to off-load a certain cell to improve conditions for the remaining users, or for any other reason.

Further in a multi-access system, apart from selecting the appropriate access technology to steer the traffic, there is a need to switch the traffic across different access technologies based on changes in radio environment and application requirements and even split the traffic across multiple access technologies to satisfy performance requirements. The different types of traffic and frequency bands in a commercial network may require appropriate bearer selection (Master Cell Group (MCG) bearer, Secondary Cell Group (SCG) bearer, Split bearer) and bearer type change for achieving load balancing, low latency and best in class throughput in a multi-access scenario with 5GC networks., (see TS 37.340 [8]).

The Non-RT RIC creates corresponding traffic management policies directed to identify UEs expressing the priority order of cells to be allocated to each one for their control planes and user planes respectively. The policies are sent over the A1 interface to the Near-RT RIC, who uses the policies when enforcing the radio resource control.

3.5.3.2 Enrichment information based traffic steering optimization

The current traffic management solutions are usually implemented by relocating users among cells, which highly depends on the measurement report (MR) from the UE feedback. The statistical characteristics of the radio network and the UE behaviors information have not been fully exploited and utilized to enhance the network and user experience performance.

The enrichment information can be utilized to optimize traffic management performance. The enrichment data could include Network Radio fingerprint information, UE trajectory information (e.g., way points in cell-level or beam-level), UE mobility profile (e.g., stationary, horizontal, vertical speed), UE service type (e.g., delay sensitivity or bit error rate sensitive), traffic pattern (e.g., average UL/DL packet size, periodicity). With the assistance of the enrichment information provided by Non-RT RIC, Near-RT RIC can efficiently derive optimized solutions.

For example Non-RT RIC can derive the radio fingerprint enrichment information via the data statistical analysis of historical measurement results. The radio fingerprint information captures the mapping relationship of the intra-frequency measurement results and the inter-frequency measurements. Then Non-RT RIC can deliver it to Near-RT RIC to help it predict the inter-frequency measurement, reduce the unnecessary inter-frequency measurement, accelerate the process of traffic optimization and improve the network system performance and user experience.

3.5.4 Benefits of O-RAN Architecture

The use case explores the opportunity with the policy based A1 interface allowing to address specific users to obtain an intent driven by e.g. UE specific service level requirements.

3.6 Use case 6: Massive MIMO Beamforming Optimization

3.6.1 Background Information

Massive MIMO is seen as one of the key technologies for 5G. Due to the multi-antenna transmission and reception, this technology can inherently to provide diversity and improve capacity by targeting high gain antenna beams towards one or multiple subscribers, thus improving the receive power levels and spatially filtering the interference from neighboring subscribers and transmission points. In addition, a spatial multiplexing operations regime can improve the network capacity by transferring multiple data streams towards/from one or different subscribers utilizing a spatial re-use of the scarce time/frequency resource blocks. Further advantages include the controlling of electromagnetic (EM) emissions or advanced network management technologies like beam shaping, beam-based load balancing, optimized beam mobility, adaptive cell coverage areas, especially in highly dense urban 3D environments and mobile subscribers. In order to further optimize networks, e.g., maximize spectral efficiency, optimize coverage, or maximize cell capacity, fully digital beamforming (BF) methods are to be employed for below 6 GHz frequency wireless telecommunications. Grid of Beams (GoB) is a BF method which aims at selectively covering regions of interest with a suitable subset of radio beams (chosen from a dictionary of possible beams). Beam-based Mobility Robustness Optimization is a BF method enhancing beam specific mobility performance e.g. by adding beam specific individual offsets. Mobility Robustness Optimization is a well-known SON concept [9][13]. First supporting measurements have been specified in LTE in Release 9 [8]

3.6.2 Motivation

The massive MIMO BF optimization use case aims at proactively and continuously improving cell and/or user-centric coverage, capacity and mobility performance in a massive MIMO deployment area by adapting beam configurations (e.g., number of beams, beam boresights, vertical beam widths, horizontal beam widths, beam black lists/white lists, beam mobility thresholds) in a multi-cell, possibly multi-vendor, deployment scenario to non-real time (e.g., 3D construction, 3D terrain topology, network, weather seasons, intra-day cell splitting/merging/shaping, traffic distribution, beam conflicts) and near-real time (e.g., moving users/hotspots, changing traffic distribution, crowd source data) changes. The high number of configuration parameters per array antenna, the amount of available measurement input data, the complexity, pro-activeness as well as non- and near-real time requirements suggest the application of machine learning techniques of input data analytics as well as use case decision generation.

The objective of this use case is to allow the operator to flexibly configure Massive MIMO system parameters by means of policies, configuration or machine learning techniques, according to objectives defined by the operator.

3.6.3 Proposed Solution

1) The performance of BF is strongly dependent on the choice of the beam pattern
   Traditional GoB solutions do not take contextual/per-site information into account and operate with a narrow set of manually predefined beam patterns. Contextual/per-site information is, among others, comprised of cell geometry and other cell parameters, antenna array parameters, real-time or long-term (e.g., seasonal) traffic patterns, real-time or long-term user mobility patterns etc. One goal of this use case is to demonstrate that by taking contextual/per-site information into account, GoB/mMIMO BF can enhance the network performance by allowing the operator to perform extensive BF customization and optimization.

2) A key factor in multiuser (MU) mMIMO is the management of Beam Failures (BFAs)
   In order to prevent connectivity and user experience degradation, two AI/ML based Beam Mobility Optimization solutions are proposed.

In this use case three optimization loops for mMIMO BF are proposed:

1) An outer, Non-RT loop Massive MIMO GoB Beam Forming Optimization
2) An inner, Near-RT loop Massive MIMO Beam-based Mobility Robustness Optimization
3) An inner, Near-RT loop Massive MIMO Beam Selection Optimization

The two optimization loops can be implemented and deployed either individually or jointly in a nested fashion.

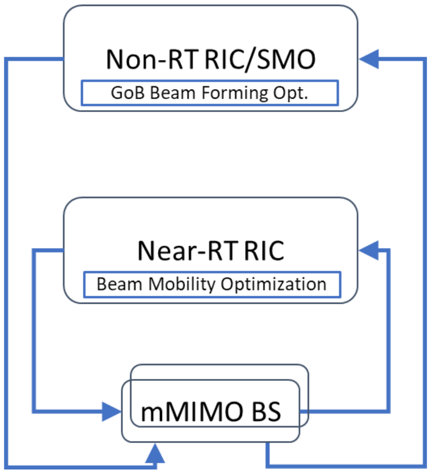


Figure 3.6.3-1: Non-RT and Near-RT Optimization loops

3.6.3.1 Non-RT Massive MIMO GoB Beam Forming Optimization

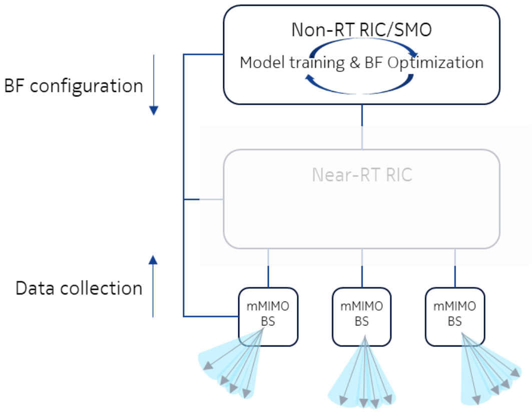


Figure 3.6.3-2: Non-RT Beam Forming (BF) Optimization

Non-RT RIC hosts an application with long-term analytics function (=ML training, Non-RT RIC), whose task is to collect, process and analyze antenna array parameters, cell performance KPIs, UE mobility/spatial density data, traffic density data, interference data and BF gain/beam RSRP and MDT measurement data.

The input data for BF optimization training and inference can be comprised of [14] antenna timing advance and Angle-of-Arrival measurements (for positioning estimation unless derived from another entity), aggregated & preprocessed beam-based reference signal measurements (average) traffic density measurements (across the respective mMIMO spatial grid) with associated positioning information, CSI measurements or, power headroom reports (PHRs), neighboring cells' beams/interference information and beam RSRPs/gains as well as cell performance measurements such as handover and beam failure statistics.

The output of the BF optimization inference can be optimized BF configuration, number of beams, beam elevation, beam horizontal & vertical widths and power allocation of beams.

The long-term analytics function and the BF optimization function can be thought of as ML training and ML inference, respectively.

Operator objectives may include desired coverage, defined in terms of the cell geometry (SSB beams), cell capacity requirements (CSI-RS beams), per UE cell edge throughput and traffic density weighted average RSRP and BF gain requirements. (E.g., the operator might wish to implement the strategy to cover low-density areas with wide beams and high-density areas with narrow beams.)

3.6.3.2 Near-RT Massive MIMO BeamBased Mobility Robustness Optimization (bMRO)

Near-RT RIC hosts an xApp with BF optimization function (=ML inference, Near-RT RIC), whose task is to monitor UE/traffic density measurements, UE positioning information, beam based mobility and beam failure KPIs and select/deploy optimized mMIMO BF parameters to E2 nodes.

The near-RT RIC may host an xApp to optimize inter-cell beam mobility such as a beam-based Mobility Robustness Optimization. In this case the Near-RT RIC might for instance configure beam individual offsets for inter-cell mobility decisions.

The Near-RT bMRO function can run individually, without the Non-RT BF Optimization. However, it can be deployed in a nested fashion within a Non-RT BF Optimization loop. In that case, upon change of the GoB configuration, the Near-RT bMRO function is reset or reconfigured. There are several options for coordinating the outer and the inner loops such as:

1) The outer loop's output comes from a finite set of configurations, and to each configuration the inner loop employs a trained model.
2) The inner loop employs a reinforcement or adaptive learning technique which is reset upon change of the GoB configuration, or, depending on implementation, adapted to it.

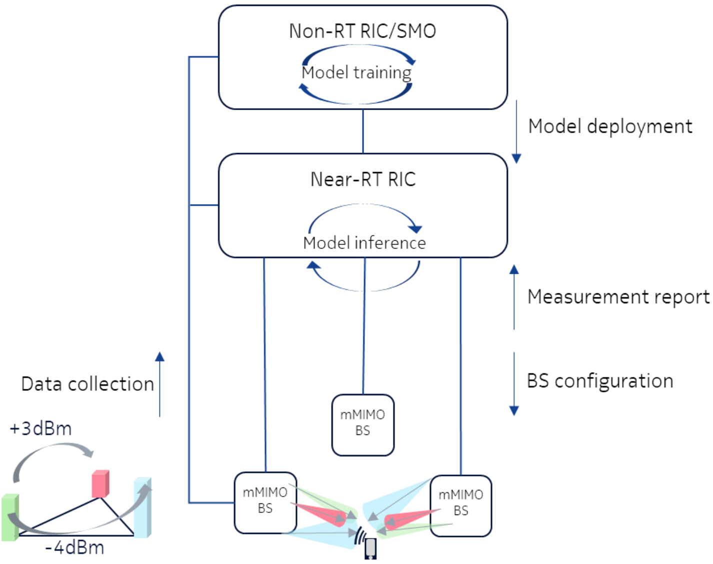


Figure 3.6.3-3: Near-RT Beam-based Mobility Robustness Optimization

Non-RT RIC hosts an application with long-term analytics function (=ML training, Non-RT RIC), whose task is to collect and analyze underlying GoB configuration, if GoB configuration exists, beam mobility and failure statistics, L1/L2 RSRP values, potential source-target beam pairs.

Near-RT RIC hosts an xApp with bMRO Optimization function (=ML inference, near-RT RIC), whose task is to monitor potential source-target beam pairs and optimize beam mobility for scheduling by managing user-beam pairing.

The input data for Beam Mobility Robustness Optimization training and inference can be comprised of [14] per-user measurements (e.g. RSRP, SINR, etc.), handover failure statistics (e.g. overall HO failures, too early or to late or HO to wrong cell), such as number of BFAs, times of BFAs, BFA rate etc., per-user potential source-target beam pairs and neighbouring cells' beams/interference information.

The output of the bMRO optimization function can be [9] adjusted offsets for candidate source-target beam pairs for beam mobility.

The long-term analytics function and the bMRO function can be thought of as ML training and ML inference, respectively.

Operator objectives may include minimization of BFA rate for a group of users (e.g., high-mobility users).

3.6.3.3 Near-RT Massive MIMO Beam Selection Optimization (BSO)

Near-RT RIC hosts an xApp with BF optimization function (=ML inference, Near-RT RIC), whose task is to monitor UE/traffic density measurements, UE positioning information, beam based mobility and beam failure KPIs and select/deploy optimized mMIMO BF parameters to E2 nodes.

The Near-RT RIC may host an xApp to optimize intra-cell beam mobility. In this case the Near-RT RIC might for instance configure parameters for intra-cell beam switching decisions.

The Near-RT BSO function can run individually, without the Non-RT BF Optimization. However, it can be deployed in a nested fashion within a Non-RT BF Optimization loop. In that case, upon change of the GoB configuration, the Near-RT BSO function is reset or reconfigured. There are several options for coordinating the outer and the inner loops such as:

1) The outer loop's output comes from a finite set of configurations, and to each configuration the inner loop employs a trained model.
2) The inner loop employs a reinforcement or adaptive learning technique which is reset upon change of the GoB configuration, or, depending on implementation, adapted to it.

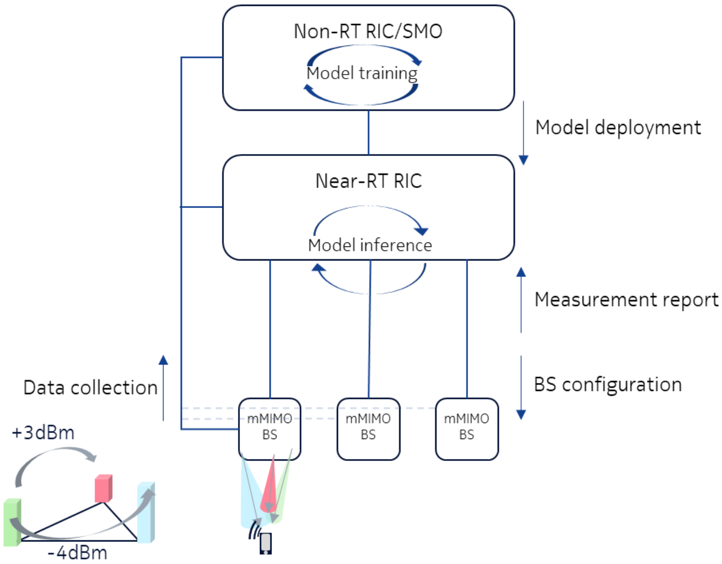


Figure 3.6.3-4: Near-RT Beam Selection Optimized (BSO) function

Non-RT RIC hosts an application with long-term analytics function (=ML training, Non-RT RIC), whose task is to collect and analyze underlying GoB configuration, if GoB configuration exists, beam mobility and failure statistics, L1/L2 RSRP values, potential source-target beam pairs.

Near-RT RIC hosts an xApp with BSO function (=ML inference, near-RT RIC), whose task is to monitor potential source-target beam pairs, and to optimize beam mobility for scheduling by managing user-beam pairing.

The input data for BSO training and inference can be comprised of per-user measurements (e.g. RSRP, SINR etc.), beam failure statistics, per-user potential source-target beam pairs and neighbouring cells' beams/interference information.

The output of the BSO optimization function can be adjusted offsets for candidate source-target beam pairs for beam mobility.

The long-term analytics function and the BSO function can be thought of as ML training and ML inference, respectively.

Operator objectives may include minimization of BFA rate for a group of users (e.g., high-mobility users).

3.6.4 Benefits of O-RAN Architecture

The advantages the O-RAN architecture provides to this use case include the chance to apply and combine both, non- and near real-time analytics, machine-learning, and decision making for various sub-tasks of this use case for a cell-centric and /or user (group) centric point of view. The Non-RT RIC oversees by definition the long-term traffic-, coverage-, and interference situation etc. in a whole cluster of cells. O-RAN interfaces such as O1, A1, and E2 will support necessary data, policy, and configuration exchanges between the architectural elements. By taking contextual information and past failure statistics into account, mMIMO BF can be much more customized and thus achieve purposeful operator desires/requirements. Optimized mMIMO BF depends on the collection, processing, and the analysis of both non-real-time and real-time cell-/UE-level data, which is facilitated by the O-RAN architecture.

3.7 Use case 7: RAN Sharing

3.7.1 Background Information

One of the main challenges for network operators is to deploy a massive number of services, while providing different Quality of Service (QoS) requirements and keeping reasonable the level of investment in the network. Network sharing is envisioned as an efficient and sustainable way to accelerate the deployment of 5G, while taking advantage of a common pool of physical infrastructures and resources, made available for two or more partner operators.

Besides, regulatory requirements often force operators to provide coverage in not business attractive areas, causing profitability issues. To this end, RAN sharing is seen as a promising solution that should reduce network costs, increase network capacity and coverage, while enhancing customer satisfaction. Accordingly, the open and multivendor nature of the O-RAN architecture can accelerate the introduction and development of RAN sharing solutions, by enhancing the deployment of virtual network functions (VNF) on commodity shared hardware, while taking into account diverse QoS requirements.

Among the different RAN-sharing models that have been experimented so far, a special focus is put here on the evaluation of the compatibility of the “Geographical Split” RAN sharing model with the O-RAN architecture. In such a model, a coverage area is split between two or more operators; each operator manages the RAN in a specific area, while offering access to its RAN resources to its operator partners. Two main configurations have been used worldwide [20] Multi Operator Core Network (MOCN) and Multi Operator RAN (MORAN). In MOCN, both the RAN infrastructure and carriers are pooled. Even though this model enables further cost saving, especially in rural areas due to a lower number of carriers, it requires the presence of a regulatory entity that takes care of the allocation of different parts of the shared spectrum between operators. Conversely, in MORAN, each operator utilizes a separate carrier, while getting more freedom and independency on the control of the radio resources. MORAN is the most widely used sharing configuration as it can provide appropriate independency to each sharing partner operator, while maximizing the benefits of sharing in terms of CAPEX and OPEX.

Besides, the O-RAN architecture can provide new opportunities for implementing this RAN sharing model in a more efficient way, thanks to its multi-vendor interfaces and abstraction control provided by the RIC (RAN Intelligent Controller). Accordingly, O-RAN can accelerate the deployment of 3GPP-based solutions by providing more flexibility in the management of the shared resources.

3.7.2 Motivation

Currently, in 3GPP there are ongoing discussions for identifying which RAN functions should be included in the RAN sharing procedures in 5G [4][14]. Specifically, it has been analysed the feasibility to share the whole or a part of the DU or CU functional blocks while assuming the sharing of a common physical layer (PHY). In [17], a first agreement has been achieved stating that multiple logical CU-CP/ DU can control the same PHY radio resources but coordination between logical CU-CP needs to be ensured by an appropriate implementation (not standardized).

In this context, O-RAN can be seen as the ideally enabler of the 3GPP RAN sharing model by providing the required coordination between the shared network nodes. To this end, the RIC can enable the coordination of multiple CU-CP/ DU via the E2 interface, while opening the road for diverse RAN sharing scenarios. Moreover, O-RAN can accelerate the deployment of compliant 5G RAN sharing solutions by taking advantage of the multi-vendor nature of the F1, X2 interfaces.

3.7.3 Proposed Solution

Figure 3.7.3-1 describes the logic architecture of the proposed MORAN use-case. It is assumed that two operators, denoted as Operator A and Operator B, share the same RAN infrastructure, while keeping the core network independent.

Specifically, Operator A owns the site A and shares the PHY Layer (LOW) with Operator B (Shared O-RU in Figure 3.7.3-1). Indeed, multiple PLMN IDs are broadcasted [20], while each operator operates in a different carrier.

Moreover, site A hosts VNF instances of Operator B in a shared O-DU and O-CU site. Specifically, the computing resources of the site A are shared among multiple VNFs, belonging to the operator A and B respectively. Each VNF represents a logic implementation of the O-DU and O-CU functionalities and should be controlled by each partner operator in an independent manner.

While Operator A can directly control its VNFs, Operator B needs to control its VNFs in a remote manner. The challenge here is to enable Operator B to control resources in an infrastructure that is owned by another operator. Accordingly, it is assumed that Operator B can monitor and control the remoted resources via the RIC node of site B. Note that in the proposed architecture, the RIC are not shared and kept independent at the site A and B respectively.

Such a scenario presents a set of implementation challenges:

- A common interface is needed to control and coordinate the usage of shared resources.
- An orchestrator has to be able to communicate effectively with the shared nodes regardless of the manufacturer or vendor of the used hardware devices.

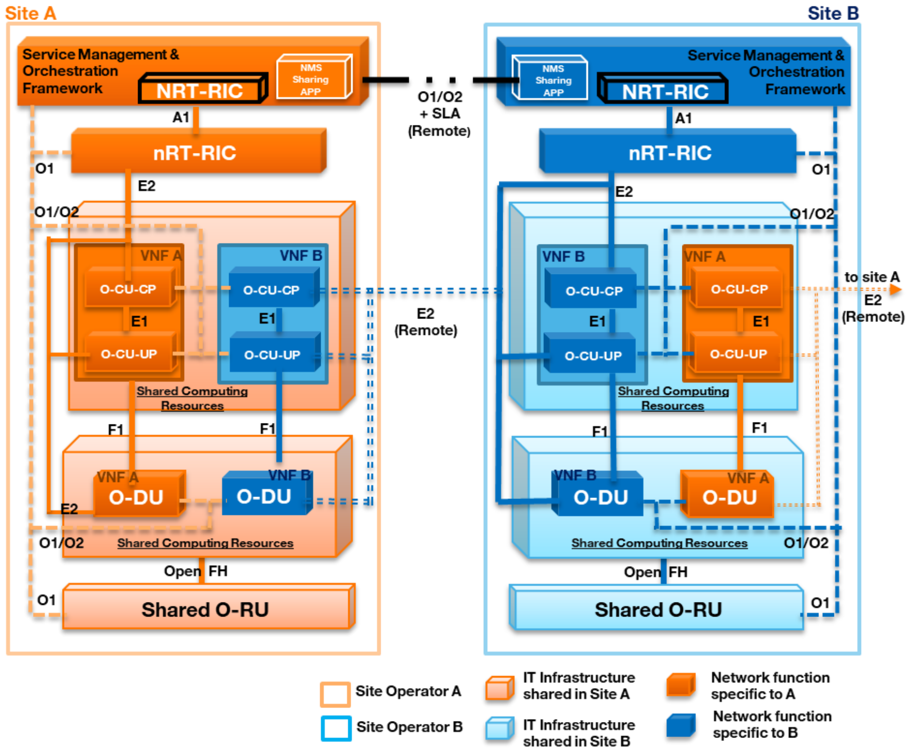


Figure 3.7.3-1: MORAN Use-Case in O-RAN

To this end, this use-case proposes to enable Operator B to remote control the O-CU and O-DU VNFs via a “remote E2 interface”, giving it the freedom to implement and configure its RIC policies in an autonomous manner.

Besides, it is assumed that Operator A instantiates all the network nodes in site A, including the VNFs of Operator B, via the O2 interface, while the management and orchestration is provided by Service & Orchestration Framework. To better orchestrate the shared resources, the Service & Orchestration Framework of Site A shall interact with the one of Operator B. This interaction may be enabled as follows:

- by designing a new interface between the Service & Orchestration Framework of the two sites (RAN-Sharing Orchestration Interface in Fig. 3.7.3-1)
- by enabling the Operator B to directly orchestrate its VNFs deployed in site A via “remote O1 & O2” interfaces.

Regarding the control plane, it is assumed that Operator A can control only the shared physical resources, while Operator B can handle only the virtual resources that belong to it.

The “remote E2” interface shall provide secure access to Op. A site, while from the Operator B RIC perspective, the O-DU and O-CU VNFs hosted at site A shall be controlled in a transparent manner as in non-shared scenario.

Finally, this use-case proposes an extension of the E2 interface in order to support the remote control of shared resources, while taking account security aspects related to the risk of: i) offering to an external actor the possibility to have access to the resources of the hosting RAN infrastructure, and ii) enabling an external actor to orchestrate the resources of the partner operator.

3.7.4 Benefits of O-RAN Architecture

The proposed MORAN sharing architecture in. Figure 3.7.3-1 lets operators to configure shared network resources independently from configuration and operating strategies of the other sharing operators. Accordingly, this architecture enables the RIC of the Operator B to monitor the radio state of the customers served by the partner operator's site, facilitating the optimization of the radio allocation process and the remote configuration of QoS parameters. Moreover, this approach can favour the deployment of slicing scenarios facilitating the differentiation of services running on the shared RAN.

3.8 Use case 8: QoS Based Resource Optimization

3.8.1 Background Information

QoS based resource optimization can used when the network has been configured to provide some kind of preferential QoS for certain users. One such scenario can be related to when the network has been configured to support e2e slices. In this case, the network has functionality that ensures resource isolation between slices as well as functionality to monitor that slice Service Level Specifications (SLS) are fulfilled.

In RAN, it is the scheduler that ensures that Physical Resource Block (PRB) resources are isolated between slices in the best possible way and also that the PRB resources are used in an optimal way to best fulfill the SLS for different slices. The desired default RAN behavior for slices is configured over O1. For example, the ratio of physical resources (PRBs) reserved for a slice is configured at slice creation (instantiation) over O1. Also, QoS can be configured to guide the RAN scheduler how to (in real-time) allocate PRB resources to different users to best fulfill the SLS of a specific slice. In the NR NRM this is described by the resource partition attribute.

Instantiation of a RAN sub-slice will be prepared by rigorous planning to understand to what extent deployed RAN resources will be able to support RAN sub-slice SLS. Part of this procedure is to configure RAN functionality according to above. With this, a default behavior of RAN is obtained that will be able to fulfill slice SLSs for most situations. However, even through rigorous planning, there will be times and places where the RAN resources are not enough to fulfill SLS given the default configuration. To understand how often (and where) this happens, the performance of a RAN slice will continuously be monitored by SMO. When SMO detects a situation when RAN SLS cannot be fulfilled, Non-RT RIC can use A1 policies to improve the situation. To understand how to utilize A1 policies and how to resolve the situation, the non RT-RIC will use additional information available in SMO.

3.8.2 Motivation

To motivate the use case an example with an emergency service as a slice tenant is used. For this example, it is understood (at slice instantiation) that 50% of the PRBs in an area should be enough to support the emergency traffic under normal circumstances. Therefore, the ratio of PRBs for the emergency users is configured to 50% as default behavior for the pre-defined group of users belonging to the emergency slice. Also, QoS is also configured in CN and RAN so that video cameras of emergency users get a minimum bitrate of 500 kbps.

Now, suppose a large fire is ongoing and emergency users are on duty. Some of the personnel capture the fire on video on site. The video streams are available to the Emergency Control Command. Because of the high traffic demand in the area from several emergency users (belonging to the same slice), the resources available for the Emergency slice is not enough to support all the traffic. In this situation, the operator has several possibilities to mitigate the situation. Depending on SLAs towards the Emergency slice compared to SLAs for other slices, the operator could reconfigure the amount of PRB reserved to Emergency slice at the expense of other slices. However, there is always a risk that Emergency video quality is not good enough irrespective if all resources are used for Emergency users. It might be that no video shows sufficient resolution due to resource limitations around the emergency site.

In this situation, the Emergency Control Command decides, based on the video content, to focus on a selected video stream to improve the resolution. The Emergency Control System gives the information about which users to up- and down-prioritized to the e2e slice assurance function (through e.g. an Edge API) of the mobile network to increase bandwidth for selected video stream(s). Given this additional information, the Non-RT RIC can influence how RAN resources are allocated to different users through a QoS target statement in an A1 policy. By good usage of the A1 policy, the Emergency Control Command can ensure that dynamically defined group of UEs provides the video resolution that is needed.

The use case can be summarized as per below:

- A fire draws a lot of emergency personnel to an area.
- Because of this RAN resources becomes congested which affects the video quality for all video feeds in the area
- The Emergency Control Command have 5 active video feeds and selects one video feed which is of specific interest
- The Emergency Control Command requests higher resolution of a selected feed, while demoting the other
- With this information, the Non-RT RIC will evaluate how to ensure higher bandwidth for the feed selected by Emergency Control Command (and lower for other feeds)
- The Non-RT RIC updates the policy for the associated UEs in the associated Near-RT RIC over the A1 interface
- Near-RT RIC enforce the modified QoS target for the associated UEs over the E2 interface to fulfill the request
- The Emergency Control Command experiences a higher resolution of the selected video feed

3.8.3 Proposed Solution

The main functions of the O-RAN components are utilized to support an improved QoS based resource optimization. QoS based resource optimization use case deploys O-CU, O-DU, the Non-RT RIC and the Near RT RIC function modules. To achieve intelligent resource optimization, Non-RT RIC should provide policies to Near-RT RIC which are used to drive QoS based resource optimization at the RAN level. Non-RT RIC shall monitor QoS related metrics from network and SMO functions. O-CU and O-DU components should provide UE performance metrics with the configured granularity to SMO via O1. In addition to performance metrics retrieved from network elements, external information sources might also be utilized to solve the problem of allocation limited RAN resources. For example, external server could provide Non-RAN data about priorities of the UEs to SMO. Finally, the E2 Nodes should execute QoS enforcement decisions received from Near-RT RIC which are expected to influence RRM behaviour.

3.8.4 Benefits of O-RAN Architecture

The main features of O-RAN architecture are pointed by the proposed solution which aims to offer more advanced QoS based resource optimization.

3.9 Use case 9: RAN Slice SLA Assurance

3.9.1 Background Information

The 3GPP standards architected a sliceable 5G infrastructure which allows creation and management of customized networks to meet specific service requirements that may be demanded by future applications, services and business verticals. Such a flexible architecture needs different requirements to be specified in terms of functionality, performance and group of users which may greatly vary from one service to the other. The 5G standardization efforts have gone into defining specific slices and their Service Level Agreements (SLAs) based on application/service type [4]. Since network slicing is conceived to be an end-to-end feature that includes the core network, the transport network and the radio access network (RAN), these requirements should be met at any slice subnet during the life-time of a network slice [5], especially in RAN side. Exemplary slice performance requirements are defined in terms of throughput, energy efficiency, latency and reliability at a high level in SDOs such as 3GPP [2] and GSMA [21]. These requirements are defined as a reference for SLA/contractual agreements for each slice, which individually need proper handling in NG-RAN.

Although network slicing support is started to be defined with 3GPP Release 15, slice assurance mechanisms in RAN needs to be further addressed to achieve deployable network slicing in an open RAN environment. It is necessary to assure the SLAs by dynamically controlling slice configurations based on slice specific performance information. Existing RAN performance measurements [7] and information model definitions [6] are not enough to support RAN slice SLA assurance use cases. This use case is intended to clarify necessary mechanisms and parameters for RAN slice SLA assurance.

3.9.2 Motivation

In the 5G era, network slicing is a prominent feature which provides end-to-end connectivity and data processing tailored to specific business requirements. These requirements include customizable network capabilities such as the support of very high data rates, traffic densities, service availability and very low latency. According to 5G standardization efforts, the 5G system should support the needs of the business through the specification of several service needs such as data rate, traffic capacity, user density, latency, reliability, and availability. These capabilities are always provided based on a Service Level Agreement (SLA) between the mobile operator and the business customer, which brought up interest for mechanisms to ensure slice SLAs and prevent its possible violations. O-RAN's open interfaces and AI/ML based architecture will enable such challenging mechanisms to be implemented and help pave the way for operators to realize the opportunities of network slicing in an efficient manner.

3.9.3 Proposed Solution

RAN slice SLA assurance scenario involves Non-RT RIC, Near-RT RIC, E2 Nodes and SMO interaction. The scenario starts with the retrieval of RAN specific slice SLA/requirements (possibly within SMO or from NSSMF depending on Operator deployment options). Based on slice specific performance measurements from E2 Nodes, Non-RT RIC and Near-RT RIC can fine-tune RAN behavior aligned with O-RAN architectural roles to assure RAN slice SLAs dynamically. Non-RT RIC monitors long-term trends and patterns for RAN slice subnets' performance, and employs AI/ML methods to perform corrective actions through SMO (e.g. reconfiguration via O1) or via creation of A1 policies. Non-RT RIC can also construct/train relevant AI/ML models that will be deployed at Near-RT RIC. A1 policies possibly include scope identifiers (e.g. S-NSSAI) and statements such as KPI targets. On the other hand, Near-RT RIC enables optimized RAN actions through execution of deployed AI/ML models in near-real-time by considering both O1 configuration (e.g. static RRM policies) and received A1 policies, as well as received slice specific E2 measurements.

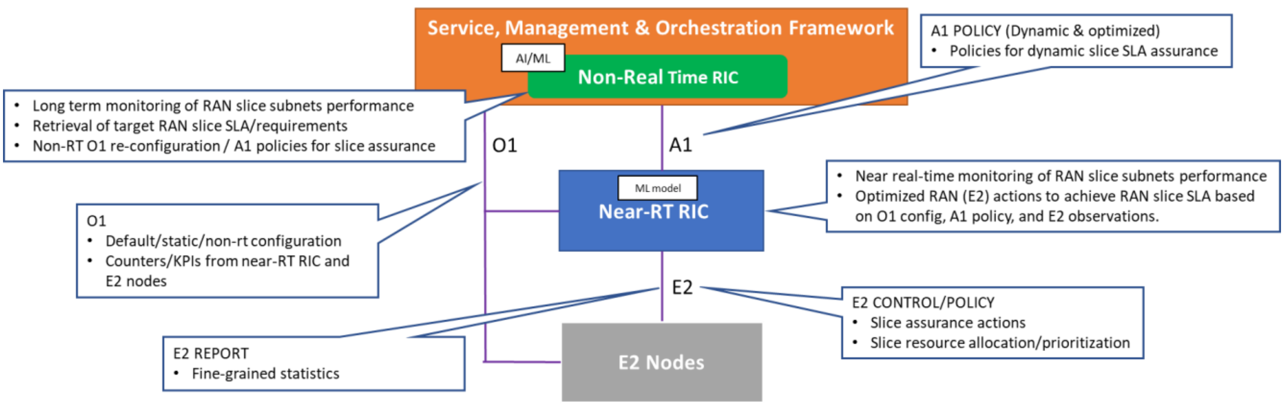


Figure 3.9.3-1: Slice SLA Assurance

3.9.4 Benefits of O-RAN Architecture

Current standardization efforts focus on defining business and system requirements, slice management functionalities and procedures, and possible RRM policies on shared resources without specifying how such a flexible envisioned system can be addressed within 5G RAN. Considering the dynamic nature of RAN, providing desired levels of service quality for each RAN slice is a challenging topic that requires further investigation and standardization efforts for a multi-vendor open RAN environment. O-RAN's open interfaces combined with its AI/ML based innovative architecture can enable such RAN SLA assurance mechanisms, which could potentially change the way network operators do their business and also enable new business models. For example, O-RAN architecture and interfaces can enable operators to manage spectrum resource allocation across slices more efficiently and dynamically in response to usage patterns, thereby allowing more efficient use of spectrum resources.

3.10 Use case 10: Multi-vendor Slices

3.10.1 Background Information

This use case enables multiple slices with functions provided by multi-vendors, such as slice #1, composed of DU(s) and CU(s), provided by vendor A and slice #2, composed of DU(s) and CU(s), provided by vendor B.

3.10.2 Motivation

When providing multiple slices, it is assumed that suitable vO-DU/scheduler and vO-CU treat each slice respectively. A vendor who provides vO-DU and vO-CU function may have a strength of a customized scheduler for a certain service. With accomplishment of multi-vendor circumstances, following benefits can be expected:

1) More flexible and time to market deployment
   Operators can maximize options to choose suitable vO-DU/scheduler and vO-CU to offer various slices. For example, some vendors may have a strength of a scheduler for eMBB service and the other may have a strength of scheduler for URLLC service. Or, vendor A can provide vO-DU/scheduler and vO-CU suitable for URLLC earlier than vendor B, therefore operators can choose vO-DU and vO-CU functions from vendor A to meet their service requirements.

   Also, when an operator wants to add a new service/slice, new functions from a new vendor can be introduced with less consideration for existing vendors if multi-vendor circumstance was realized. This may help expand vendor's business opportunities rapidly.

2) Flexible deployment when sharing RAN equipment among operators
   When operators want to share RAN equipment and resources, RAN vendors and their placement of each RAN functions may be different. If multi-vendor circumstance was introduced, then it can relax restrictions among operators to share RAN equipment and resources. This may help expanded opportunities for reaching agreements of RAN sharing among operators. With expansion of RAN sharing, operators CAPEX and OPEX can be optimized, helping with additional investment opportunities.

3) Reducing supply chain risk
   If an existing vendor providing a certain pair of vO-DU and vO-CU functions withdraws of the market due to business reasons, operators can deploy new vO-DU and vO-CU functions alternatively from other vendors under this multi-vendor circumstance. This can reduce a risk for operators' business continuity.

3.10.3 Proposed Solution

3.10.3.1 Multi-vendor Slices

Figure 3.10.3-1 depicts an architecture for multi-vendor slices use case. There are multiple slices with which have vO-DU and vO-CU functions associating respectively. As depicted in Figure 3.10.3-1, slice-1 is composed of vO-DU(s) and vO-CU(s) provided by vendor B, and slice-2 is composed of vO-DU(s) and vO-CU-UP(s) provided by vendor C. Each vO-DU/scheduler and vO-CU functions treat one slice as an example. O-RU provided by vendor A is shared between two vO-DU(s) supplied by two different vendors, vendor B and C. The case of vO-DU and vO-CU from different vendors in a slice is for further study.

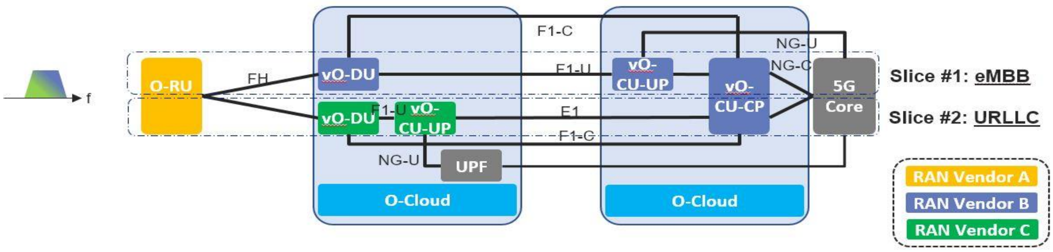


Figure 3.10.3-1: Multi-vendor Slices use case

3.10.3.2 RAN Sharing

An additional application of multi-vendor slices use case is RAN sharing where, operator A has a pair of vO-DU and vO-CU from vendor A, and operator B has a different pair of vO-DU and vO-CU from vendor B, and O-RU is shared among these two operators.

As mentioned in section 3.10.2, through RAN sharing, CAPEX and OPEX are expected to be reduced. Savings achieved by RAN sharing can then be invested again for additional expansions of RAN sharing area.

This use case considers single slice in each operator and multiple slices in each operator is for further study.

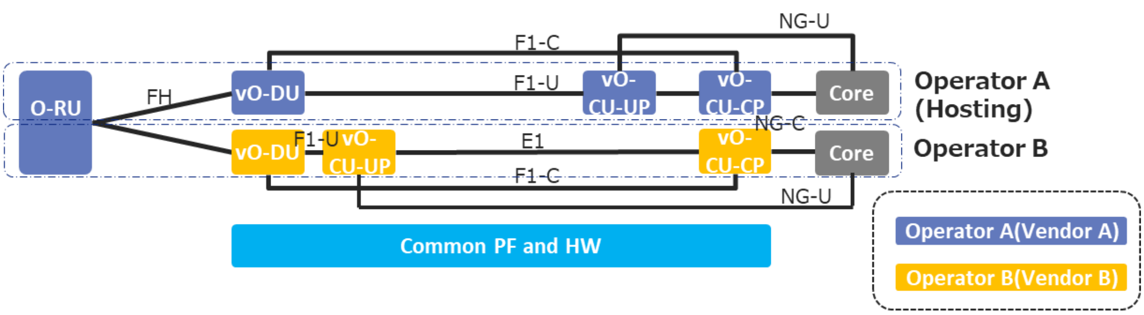


Figure 3.10.3-2: RAN Sharing use case

To realise multi-vendor slices, some coordination between vO-DU/vO-CUs will be required since radio resource shall be assigned properly and without any conflicts. Depending on different service goals and the potential impact on O-RAN architecture, a required coordination scheme needs to be determined. The possible cases are:

1) Loose coordination through O1//A1 interface (Case 1 in Figure 3.10.3-3)
2) Moderate coordination through E2/X2/F1 interface (Case 2 in Figure 3.10.3-3)
3) Tight coordination through a new interface between vO-DUs (Case 3 in Figure 3.10.3-3)

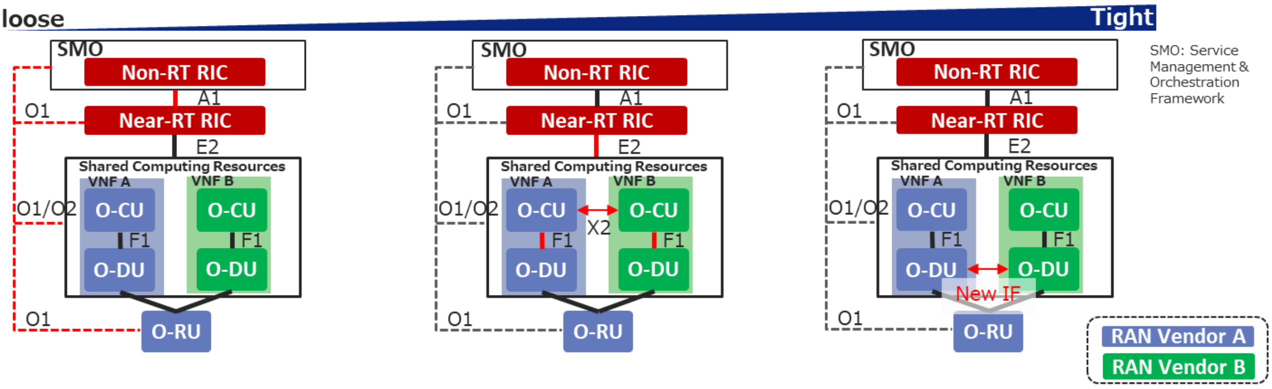


Figure 3.10.3-3: Multi-vendor Slices Coordination Scheme Options

In case 1, a resource allocation between slices or vO-DU/vO-CUs is provisioned through O1/A1 interface and each pair of vO-DU and vO-CU will allocate radio resources to each customer within radio resources allocated by Near-RT RIC and/or Non-RT RIC.

In case 2, a resource allocation can be negotiated between slices or vO-DU/vO-CUs through E2, X2 and F1 after provisioned through O1/A1 interface.

If a more adaptive radio resource allocation is needed (case 3), a more frequent negotiation would be required. This can potentially be achieved via an interface or API extension between vO-DU(s), which would be for FFS in WG1 and WG4.

3.10.4 Benefits of O-RAN Architecture

The proposed multi-vendor slices approach will not only enable operators to utilize different vendor's strengths for particular service/slice type but also provide opportunities for additional RAN Sharing use cases. The coordination needed between different vendors CU(s)/DU(s) can be realized through O-RAN architecture where Near-RT RIC and Non-RT RIC can coordinate CU(s)/DU(s) to efficiently allocate and use resources across these multi-vendor slices.

3.11 Use case 11: Dynamic Spectrum Sharing (DSS)

3.11.1 Background Information

As we transition from 4G to 5G, the spectral resources used for 5G deployment is a key consideration and this situation varies from one operator to another. Though, new C-band resources between 3-6 GHz and mmWave bands have been acquired by operators, these bands suffer from great propagation and penetration loss, limiting their coverage to users in proximity of the cell site, this is compelling particularly on the UL where the UE device is power constrained. This mandates the requirement of 5G deployment on lower bands (i.e., below 2GHz), which are widely used in existing 4G

LTE deployments. Operating on lower bands along with non-standalone mode of 5G deployment helps to cover large geography, enables seamless mobility between 4G and 5G while being sensitive to overall cost of deployment. In addition, DSS offers the advantage of dynamically sharing the available spectrum adapting to the changing work loads of 4G and 5G network.

3.11.2 Motivation

DSS is compelling considering the need for operators to dynamically share already deployed spectral resources between LTE and NR devices without degrading the QoE of the current 4G subscribers while offering the same level of coverage to support NR devices, with the consideration that they will be rolled out in an incremental manner. The objective of this use case is to propose DSS in the context of the O-RAN architecture, specifically to realize it as an application in the RIC framework. This would particularly benefit vRAN implementations when the 4G/5G CU/DU are from different vendors and one could leverage RAN data over ORAN’s framework for traffic prediction, resource management and control functions. Towards this, we identify the intelligent control functions which can be realized as a DSS application to augment the L2/L1 control functions defined as part of LTE-NR coexistence in Rel-15/16.

3.11.3 Proposed Solution

The architectural context for the proposed solution is depicted in Figure 3.11.3-1, where DSS enables 4G and 5G UEs to operate over the same spectrum identified as X (typically low band), while 5G itself could operate on new bands Y (typically high band) not used by current 4G deployment. In a typical setting, Y would offer higher capacity, low latency and smaller coverage, while X would be used to offer reasonable capacity along with larger coverage. 3GPP specifications offers DSS support over X2/Xn interface to enable dynamic sharing of spectrum resource, in addition to L2/L1 adaptation for 5G-NR to co-exist with LTE subscriber.

When DSS is enabled in the SA mode, 5G UE would be capable of operating on lower LTE bands (below 2GHz), C and mmWave bands and requires connectivity only to the gNBs. The sharing of the LTE bands between LTE and 5G data channels are achieved by both 4G scheduler and 5G scheduler, assisted by the coordination function, complimented by the RIC, between them.

For 5G NSA mode, the 5G UE is required to have dual connectivity capability and be able to connect to eNBs on LTE bands for control plane requirements and user plane connectivity towards the LTE and/or 5G depending on deployment requirements. In the scenario where gNB only operates on 5G C or mmWave bands, the sharing of the LTE frequency band between 4G and 5G UEs can be solely fulfilled by eNB MAC scheduler, as the two devices are indistinguishable from L2/L1 perspective. While, if the gNB is required to operate on lower LTE bands as well, then spectral sharing needs to be coordinated between the LTE and 5G schedulers.

The use case proposes to conduct DSS related policy, configuration, resource management and control functions using the Non-RT and Near-RT functions over open interfaces proposed by ORAN.

An abstracted view of how DSS application can be realized using the Non-RT and Near-RT RIC components is shown in Figure 3.11.3-2. The DSS over RIC can be realized as multiple applications considering its multiple optimization and operational objectives. One possible logical breakdown is as a resource management application (DSS-App) managing the shared spectrum resource adapting to dynamic 4G and 5G specific workload requirements in various local contexts, and another application (RAT-App) to configure, control and monitor DSS rules in the CU/DU corresponding to the LTE and 5G cells. The DSS-App engineers to translate the global DSS objectives such as workload requirements for a region and time-of-day to spectrum sharing policies such as max/min bandwidth threshold at a local level (e.g. central office). The RAT-App then translates the DSS-App’s resource policies to RAT specific configuration and control actions communicating with the respective CU/DU instances.

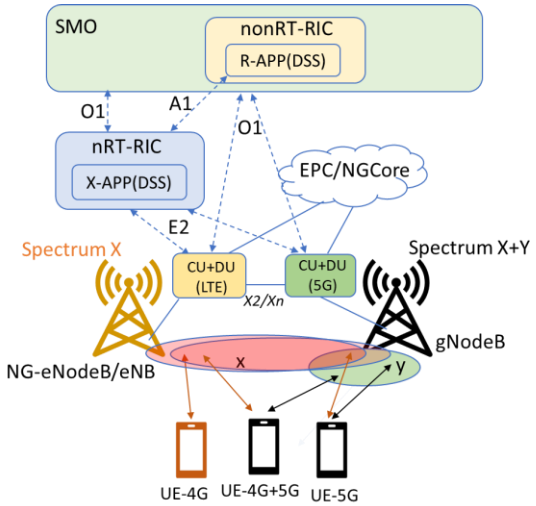


Figure 3.11.3-1: RIC Based DSS Architecture

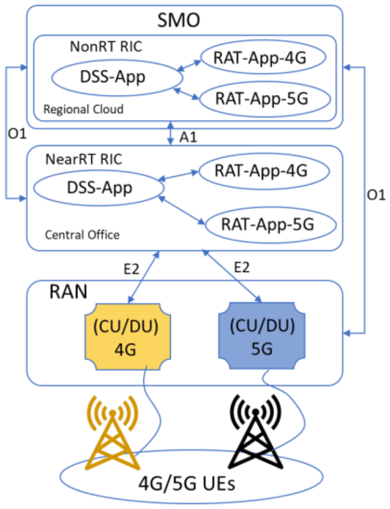


Figure 3.11.3-2: RIC Based DSS Realization

The main goal of the Non-RT DSS-App is to provide long-term policy or intent as a scheduling guidance to 4G and 5G scheduler considering business, user, spatial and temporal workload factors and the main functionality of Non-RT RAT-App is to translate the global DSS policies from Non-RT DSS-App to RAT specific policies to the RAT-App in the Near-RT RIC over A1.

The main functionalities of the Near-RT DSS-App include policy translation between Non-RT DSS-App to RAT specific configuration to the Near-RT RAT-App. Furthermore, it is actively involved in closed loop decision using the KPIs from the RAN adapting to the needs of the 4G and 5G cells. The main functionality of Near-RT RAT-App is to perform RAT specific configuration, control and data subscription over E2 interface with RAN (CU/DU components).

3.11.4 Benefits of O-RAN Architecture

Using RIC based implementation of DSS benefits from using the non-real-time and near-real-time data from the RAN to influence the 4G/5G MAC scheduler's near and long-term resource management policies and scheduler configuration. This is in addition to the value driven from control over open interfaces and using third party DSS applications. This is in contrast to synchronizing the 4G and 5G schedulers using closed resource management functions. Some scenarios where we foresee the advantage of a RIC based DSS include

1) DSS over RIC allows policy driven resource management between the 4G and 5G schedulers, aided by predictive intelligence. Furthermore, DSS requires synchronizing the MAC schedulers to avoid scheduling interference. However, the granularity of synchronization depends on the nature of workload. While workloads that are bursty in nature may require close to per-TTI synchronization, predictable workload could be handled without trading-off significant efficiency with coarser TTI synchronization. The latter would be workloads DSS over RIC would be suitable for, which would also be in line with Near-RT RIC's operational requirement of 10ms-1s control loop latency.
2) DSS over RIC can also improve resource management efficiency over multiple 4G/5G spectrum sharing cells. In this scenario, the RIC could use the global data spanning distributed cell sites, aided with predictive models to share the spectral resources dynamically. For example, data on predictive LTE or 5G user mobility can be used to predict the bandwidth requirement in adjacent 4G and 5G cells. This bandwidth decision can be then coordinated in advance across multiple distributed 4G/5G cells.
3) DSS over RIC can also be useful when the 4G/5G cells may only overlap over the cell edge, in which case UE related information such as RSRP/RSRQ, PDCP throughput can be used by the Near-RT RIC to identify the set of

overlapping cells along with the projected workload. This information is then used to slice the shared resource among the cell edge users to avoid interference by the respective schedulers while optimizing resources for the cell center users.

3.12 Use case 12: NSSI Resource Allocation Optimization

3.12.1 Background Information

5G networks are becoming increasingly complex with the densification of millimeter wave small cells, and various new services, such as eMBB (enhanced Mobile Broadband), URLLC (Ultra Reliable Low Latency Communications), and mMTC (massive Machine Type Communications) that are characterized by high speed high data volume, low speed ultra-low latency, and infrequent transmitting low data volume from huge number of emerging smart devices, respectively. It is a challenging task for 5G networks to allocate resources dynamically and efficiently among multiple network nodes to support various services.

3.12.2 Motivation

eMBB, URLLC, and mMTC services in 5G are typically realized as NSI(s) (Network Slice instance(s)) and the resources allocated to NSSI (Network Slice Subnet Instance) to support the O-RAN nodes can be optimized according to the service requirements. As the new 5G services have different characteristics, the network traffic tends to be sporadic, where there may be different usage pattern in terms of time, location, UE distribution, and types of applications. For example, most IoT sensor applications may run during off-peak hours or weekends. Special events, such as sport games, concerts, can cause traffic demand to shoot up at certain times and locations. Therefore, NSSI resource allocation optimization function trains the AI/ML model, based on the huge volume of performance data collected over days, weeks, months from O-RAN nodes. It then uses the AI/ML model to predict the traffic demand patterns of 5G networks in different times and locations for each network slice subnet, and automatically re-allocates the network resources ahead of the network issues surfaced.

3.12.3 Proposed Solution

Figure 3.12.3-1 shows the NSSI resource allocation optimization on the Non-RT RIC, and may consist of the following steps:

1) Monitoring: monitor the radio network(s) by collecting data via the O1 interface, including performance measurements (that are measured per NSSI (TS 28.552 [7])) such as DL PRB used for data traffic, UL PRB used for data traffic, Average DL UE throughput in gNB, Average UL UE throughput in gNB, Number of PDU Sessions requested to setup, Number of PDU Sessions successfully setup, Number of PDU Sessions failed to setup, etc.
2) Analysis & Decision: analyze the data to train the AI/ML model, and then determine the actions needed to add or reduce the resources (e.g. capacity, VNF resources, slice subnet attributes (TS 28.541 [6]), etc.) for the NSSI at the given time, and location.
3) Execution: execute the following actions to reallocate the NSSI resources:
    3a. Re-configure the NSSI attributes via the O1 interface
    3b. Update the cloud resources via the O2 interface

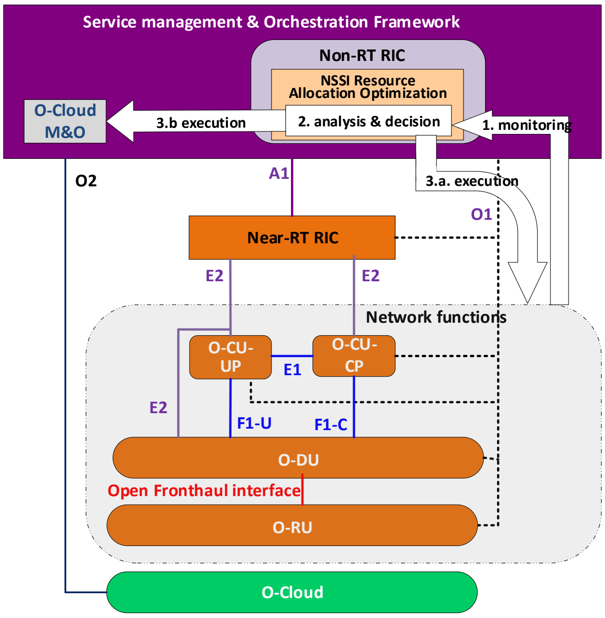


Figure 3.12.3-1: The realization of NSSI resource allocation optimization over Non-Real Time RIC

3.12.4 Benefits of O-RAN Architecture

This use case is based on the retrieval and analysis of long term performance data captured from RAN nodes and the generation of AI/ML models for re-allocation of network resources based on the prediction of traffic demand patterns of 5G networks, which is made possible by O-RAN's AI/ML based architecture.

3.13 Use case 13: Local Indoor Positioning in RAN

3.13.1 Background Information

Cellular network based positioning is an important technology for 5G vertical industries, individuals and operators, especially in local indoor scenarios. For instance, Mall can provide value-added services, such as local indoor navigation and shop recommendation by leverage real-time indoor positioning. Industrial manufacturing field shall send real-time safe warning to remind operators keep away from the dangerous area which would also need real-time indoor positioning.

NR positioning is introduced by 3GPP Rel.16. The location management function (LMF) resides in core network acts as a location server. LTE positioning protocol (LPP) is reused for the UE measurements and NR positioning Protocol "a" (NRPPa) based on LPPa is used for the gNB measurements [8]. The NR Positioning Protocol A (NRPPa) PDUs carry E-CID, OTDOA are routed between eNB/gNB and the LMF via AMF. The AMF routes the NRPPa PDUs transparently based on a Routing ID corresponding to the involved LMF over NG-C interface without knowledge of the involved NRPPa transaction. This long route messages between eNB/gNB and centralized LMF may suffer network jitters and leads to un-real-time UE location results. For some local indoor scenario, e.g., mall and industrial manufacturing, it would be better to directly deploy the positioning function in the RAN side and expose UE location to local applications.

3.13.2 Motivation

For local indoor scenarios, the positioning function inside RAN is envisioned to be a promising solution. It not only reduces the latency of positioning but also can reuse the edge cloud infrastructure in RAN. In the context of O-RAN architecture, the positioning function can be deployed as a positioning xApp in the Near-RT RIC. The positioning xApp computes the UE location and optional velocity based on the positioning measurement obtained via the E2 interface.

One example is that distributed indoor small base station in the operator's domain can provide UE positioning service by adding Near-RT RIC with positioning xApp. The positioning related information report from small base station to the positioning xApp inside Near-RT RIC can be measured based on the uplink SRS information. By leveraging multi-point positioning, field strength positioning and other positioning algorithms, positioning xApp inside Near-RT RIC can conduct a real-time position of UE.

3.13.3 Proposed Solution

Cellular network based positioning procedure mainly includes two main steps: 1) positioning measurements reports; and 2) position computation and optional velocity estimation based on the measurements. The local indoor positioning in RAN use case mainly involves the Near-RT RIC, E2 nodes and potentially Non-RT RIC in the O-RAN architecture. To enable the local indoor positioning in RAN, Near-RT RIC should be able to discover the location measurement capability of the E2 nodes and subscribe the positioning measurements based on the chosen positioning algorithm and positioning QoS requirements.

Non-RT RIC can also be leveraged to provide the AI/ML model if machine learning based algorithm is selected for the positioning. In this case, Non-RT RIC trains the AI/ML position model based on the historical positioning measurements (e.g., RSSI) and the labeled user location (e.g. by manual or by minimal drive test (MDT). Then, the trained positioning AI/ML model can be deployed to Near-RT RIC for real-time positioning inference.

The E2 nodes are expected to provide positioning measurements to Near-RT RIC as required. The measurements report can be periodical or event driven based on Near-RT RIC subscription. Near-RT RIC may also adjust the corresponding positioning measurement configurations in E2 nodes to assure the measure requirement. What kind of positioning measurements need to be reported from E2 Node depend on the subscription request from positioning xApp in Near-RT RIC. For instance, the positioning measurements may include E-CID，OTDOA, UTDOA, TOA, RSSI, RSRQ and RRU antenna ID if it is the indoor system. The report granularity is expected in the order of 10-100 ms.

The positioning xApp in Near-RT RIC can pass the positioning results to the SMO for further exposure. Furthermore, Near-RT RIC may also expose the positioning results to edge application nearby in a secure manner (via. API gateway with firewall).

3.13.4 Benefits of O-RAN Architecture

By enabling the local indoor positioning xApp in Near-RT RIC, it helps to reduce the latency of positioning in local indoor scenarios. Moreover, Near-RT RIC covers multiple E2 nodes, it can track UE location in case of mobility within certain areas.

3.14 Use case 14: Massive MIMO SU/MU-MIMO Grouping Optimization

3.14.1 Background Information

The Massive MIMO is one of the key technologies for 4G and 5G. Due to the multi-antenna transmission and reception, this technology can inherently provide diversity and improve capacity by targeting high gain antenna beams towards one or multiple subscribers, thus improving the receive power levels and spatially filtering the interference from neighboring subscribers and transmission points. The case that only one subscriber is served by a given physical resource block of a massive MIMO cell in a transmission interval is called single-user MIMO (SU-MIMO). While the case that multiple subscribers are served by a given physical resource block of a massive MIMO cell in a transmission interval is called multi-user MIMO (MU-MIMO).

In the commercial deployment, the massive MIMO eNB/gNB support both SU-MIMO and MU-MIMO transmission. The SU-MIMO and MU-MIMO are targeting for different scenarios, and can achieve different spectral efficiency. Some subscribers can be stationary, some subscribers can be pedestrian, and some subscribers can be moving at a high speed. The performance of MU-MIMO is very susceptible to subscriber's moving speed, while the performance of SU-MIMO is relatively robust to subscriber's moving speed. Additionally, some subscribers may have VoLTE or VoNR traffic, some subscribers may have download/FTP traffic, and some subscribers may have YouTube-like traffic. The performance of MU-MIMO considerably reduces in the low traffic volume scenario, while the performance of SU-MIMO is relatively robust to low traffic scenarios.

The spectral efficiency of SU-MIMO and MU-MIMO differ from each other. The spectral efficiency of MU-MIMO can be several times of spectral efficiency of SU-MIMO. Furthermore, SU-MIMO and MU-MIMO have different requirement for C-DRX, SRS, and other RRC configurations. These RRC configurations can be optimized based on scenario information.

3.14.2 Motivation

The massive SU/MU-MIMO grouping optimization use case aims at proactively and continuously improving cell and/or user-centric transmission efficiency and reliability in a massive MIMO deployment area by adapting appropriate transmission methods (e.g., SU-MIMO, MU-MIMO) for each user. The massiveness of available measurement input data, the complexity, pro-activeness as well as near-real time requirement suggest the application of machine learning techniques of input data analytics as well as use case decision generation.

The objective of this use case is to allow the operator and vendor to optimize SU/MU-MIMO transmission by means of policies, configuration, or machine learning techniques, according to objectives defined by the operator.

3.14.3 Proposed Solution

3.14.3.1 Solution 1

Figure 3.14.3-1 shows Non-RT RIC Inference solution. In this solution, Non-RT RIC is for SU/MU-MIMO grouping decision.

The SMO shall collect the necessary configurations, performance indicators, measurement reports data from RAN nodes triggered by Non-RT RIC if required, and Non-RT RIC shall retrieve user enrichment information (e.g. GPS Information, traffic information) from the application server. When the optimization objective fails, it triggers the AI/ML model re-training, data analytics and optimization in Non-RT RIC.

The Non-RT RIC shall enable to retrieve necessary configurations, performance indicators, measurement reports and other information (e.g. user GPS Information, traffic information) for the purpose of constructing/training relevant AI/ML models. The Non-RT RIC shall use the trained AI/ML model to decide the UE list for SU-MIMO group and MU-MIMO group by inferring the mobility, traffic model of each user. Additionally, Non-RT RIC should also decide RRC configuration for SU-MIMO group and MU-MIMO group, such as SRS and C-DRX configuration etc.

The Near-RT RIC shall be able to retrieve SU/MU-MIMO grouping, and related RRC configurations from non-RT RIC (if it configures over A1). Moreover, it can send the configurations to E2 nodes by policy.

The RAN Nodes shall send proper RRC configuration accordingly for UE in both SU-MIMO and MU-MIMO groups, do SU-MIMO scheduling for UE in SU-MIMO group, and do SU-MIMO or MU-MIMO scheduling for UE in MU-MIMO group dynamically. The RAN nodes shall collect and report the performance measurement to SMO related to SU- MIMO and MU-MIMO spectral efficiency. For example, average layer, rank, and throughput for SU-MIMO and MU-MIMO.

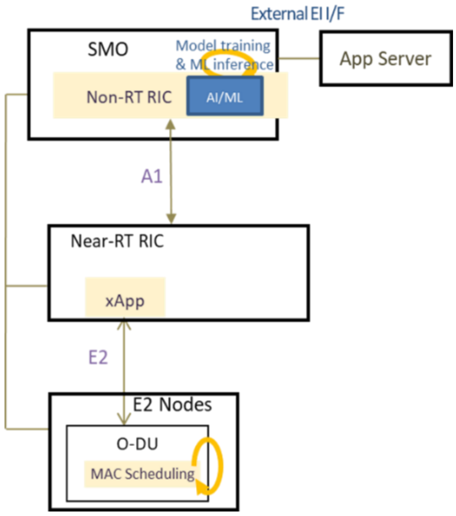


Figure 3.14.3-1: Solution 1: Non-RT RIC centered SU/MU-MIMO grouping

3.14.3.2 Solution 2

Figure 3.14.3-2 shows Near-RT RIC Inference solution. This solution is Near-RT RIC centered SU/MU-MIMO grouping decision. Another alternative is that the Near-RT RIC just provides the mobility and the traffic model prediction result over E2 interface. The E2 node makes final decision on SU/MU-MIMO grouping.

The SMO shall collect the necessary configurations, performance indicators, measurement reports data from RAN nodes triggered by Non-RT RIC if required, and Non-RT RIC shall retrieve user Enrichment Information (e.g. GPS Information, traffic information) from the application server when the optimization objective fails, and it triggers the AI/ML model re-training, data analytics and optimization in Non-RT RIC.

The Non-RT RIC shall enable the operators to retrieve necessary configurations, performance indicators, measurement reports and other information (e.g. user GPS Information, traffic information) for the purpose of constructing/training relevant AI/ML models. The Non-RT RIC shall sends Enrichment Information to Near-RT RIC by A1-EI I/F. The Near-RT RIC supports deployment and execution of AI/ML model from Non-RT RIC. The Near-RT RIC does ML inference to decide the UE list for SU-MIMO group and MU-MIMO group by inferring the mobility, the traffic model of each user. Additionally, Near-RT RIC shall decide RRC configuration for SU-MIMO group and MU-MIMO group, such as SRS and C-DRX configuration, etc.

The Near-RT RIC shall be able to send the configurations to E2 nodes by policy.

E2 Nodes shall send proper RRC configuration accordingly for UE in both SU-MIMO and MU-MIMO groups, do SU-MIMO scheduling for UE in SU-MIMO group, and do SU-MIMO or MU-MIMO scheduling for UE in MU-MIMO group dynamically. E2 nodes shall collect and report performance measurement to SMO related to SU- MIMO and MU-MIMO spectral efficiency. For example, average layer, rank, and throughput for SU-MIMO and MU-MIMO. An E2 node may also decide on the SU/MU-MIMO grouping; the E2 nodes should retrieve the mobility and the traffic model prediction result over the E2. E2 nodes should support the advanced MAC scheduling algorithms that decide to do SU-MIMO or MU-MIMO transmission for each user considering the user mobility and traffic model prediction.

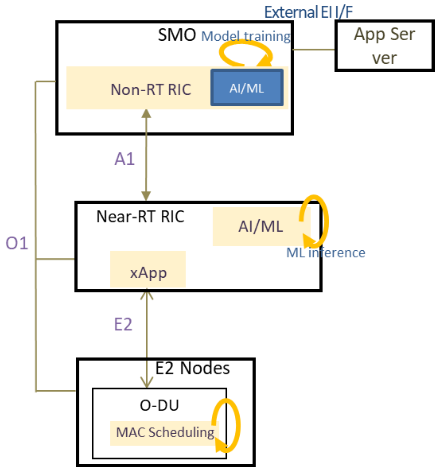


Figure 3.14.3-2: Solution 2: Near-RT RIC cantered SU/MU-MIMO grouping

3.14.4 Benefits of O-RAN Architecture

The advantages the O-RAN architecture provides to massive MIMO SU/MU-MIMO grouping optimization use case include the ability to apply and combine both, non- and near real-time analytics, machine-learning, and decision making for various sub-tasks of this use case for a cell-centric and /or user (group) centric point of view. The Non-RT RIC supports long-term AI/ML training based on enrichment information. O-RAN interfaces such as O1, A1, and E2 has support for necessary data, policy, and configuration exchanges between the architectural elements. These advantages motivate massive MIMO SU/MU-MIMO grouping optimization to run on O-RAN based RAN implementations.

3.15 Use case 15: O-RAN Signalling Storm Protection

3.15.1 Background Information

Society is increasingly dependent on network connectivity at any time and in any place and increasing diversity of device types ranging from complex devices such as smart phone to very simple and low-cost IoT devices are connecting to the network. The sheer number of connected devices, as well as the wide range of device types, makes the mobility network subject to accidental or intentional attacks that may disrupt the regular usage of the network. Given that life-critical applications are moving to wireless networks, such network disruptions are not only an inconvenience but may have impact on life and health of individuals. The O-RAN architecture offers an opportunity to address such security challenges in customizable and creative ways by utilizing the near-real time RIC xApps and non-real time RIC rApps.

3.15.2 Motivation

The main defense mechanism against attacks coming from the devices toward the network is based on configuration of the devices themselves and trust that the devices will indeed comply to restrictions defined by mobility standards. One such defense mechanism is the back-off timer that restricts the number of repeated device registrations, thus preventing devices from overloading the network with attaches. If this trust is breached there are no other options for defending the network rather than rejecting (denying service) randomly to both benign and malicious devices, a state which is equivalent to DDoS. Unfortunately, even today the network has few hundreds of device types that accidently breach this trust and allow devices to aggressively attach to the network in a rate of few thousand times per hour (the maximum allowed number by standard is less than 20 attaches per hour). An attacker that finds a way to manipulate a large set of these vulnerable devices remotely can cause an attach storm that would lead to a long outage of large parts of the network. Furthermore, this attacker can continue this attack over many hours, each time picking few thousand of devices from a large pool of millions of devices available; the network carrier will not be able to stop this attack without intelligent and fine-grained controls to act against a certain patterns of behaviour. The good news is that detecting these aggressive devices is possible as their behaviour is very different from the other devices in the network. What the network really needs is to apply dynamic restriction over these devices to prevent them from overloading the control plane of the network. This restriction should be smart enough to still allow benign devices to register to the network without interruption. Having smart security control at the RAN can stop such attack and without overloading deeper parts of the network in the core.

3.15.3 Proposed Solution

In order to protect the network from such UE originated signalling storms, an xApp can be built with two main functionalities: a DDoS detection capability and a DDoS mitigation capability. The DDoS detection capability has two parts: the near real time detection, which takes place in a RIC xApp and a non-real time detection, which takes place at the SMO and relies on enrichment data originated in external system (e.g., 5G Core or OAM system) (see Figure 1). The reason for this functionality split is that some detection logic relies on information that is available only at the core (such as IMSI, IMEI, device types, PLMN etc.), while for other detection logic, the information available at the RAN will suffice. It is assumed that the Near Real-Time RIC DDoS Detection xApp provides a faster response with coarse grained aggressive device detection while the 5G Core detection results computed by the SMO rApp provide slower response with finer grained aggressive device detection. It is noted that for some attack scenarios the xApp can perform the detection without the help of external information coming from the Non Real-Time RIC. The DDoS Mitigation xApp can implement an E2 INSERT-CONTROL control loop where the xApp can decide for each attach request if it should be accepted or rejected, or it may update an appropriate E2 POLICY when a UE is determined to be suspect. An exemplary E2 POLICY in this use case may allow suspected UEs to be blocked completely or throttled at a given rate at the E2 Node.

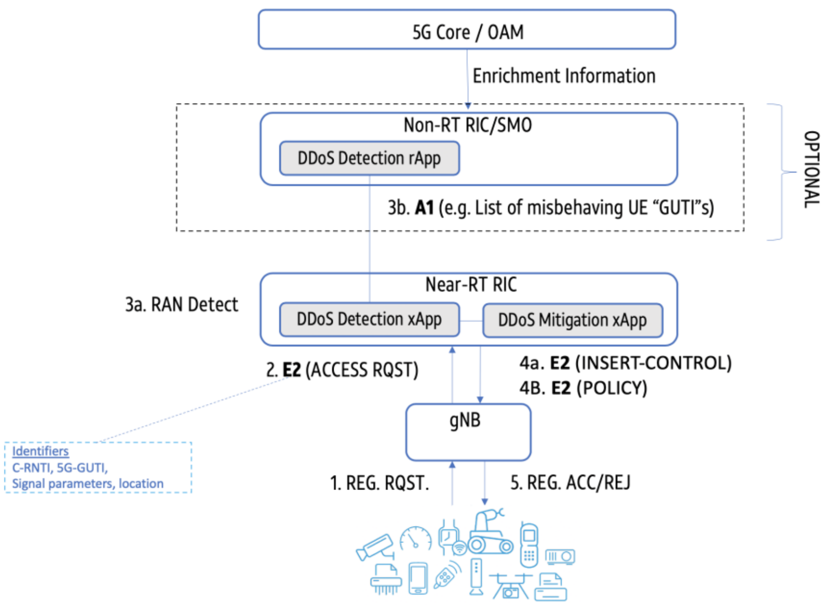


Figure 3.15.3-1: RIC DDoS Capability for 5G Stand Alone Setup

The detection algorithms should detect abnormal activity of high volume of control plane messages originating from a set of devices. The detection algorithm may change based on the network environment (metro, rural, enterprise, etc.) as well as dynamic usage trends and available devices. The output of the detection algorithm may also be of different levels of quality, depending on the attack type, in some cases the detection algorithm will produce list of specific devices while in other cases, it would only point to problematic geographic locations, specific gNBs or cell towers where the attack seems to be originated. Generally speaking, a detection algorithm should track registration behavior of individual devices and alert in the case that a threshold is exceeded for a given time threshold (e.g. one minute). The thresholds per tracking unit (i.e. cell site, gNB, AMF group etc.) should be calculated dynamically based on normal behavior of the tracking unit. The learning of the normal behavior may be achieved using ML algorithm over time.

An important concern regarding the detection report quality is the identifiers lifetime. Identifiers of network elements such as a gNB or cell tower are considered persistent over long periods of time. When considering device or subscriber related identifiers their life span depends on where they are collected. Specifically, equipment identifier (such as IMEI) or subscriber identifier (such as IMSI) are only visible at the network core and are not available at the RAN. Some RAN identifiers change upon every network attach and are allocated randomly (for example C-RNTI), while others persist longer such as the 5G TMSI. The available identifiers in the detection report depends on the nature of the attack: location, number of devices used, exploited vulnerability type and more. For example, if an aggressive device uses an RRC Connection Re-Establishment procedure, the C-RNTI remains the same over the numerous connection attempts, while in the case of RRC Connection Establishment procedure there is a fresh randomly selected C-RNTI for each attempt. There are detection techniques to identify an aggressive UE that don't rely solely on its identifiers, rather on signal parameters such as signal quality or timing advance, which may in turn, help build a unique fingerprint of the aggressive device. An important disclaimer is that not all attack scenarios can be prevented. Specifically, in the case where the modem or chipset itself are compromised and communication parameters can be faked by the attacker, the detection methods mentioned above could be avoided.

The mitigation functionality should support a set of actions, depending on the policy applied. These actions can be applied to a single UE, or a set of UEs that apply to a certain criterion. The criteria may be a location, a cell tower or a gNB. A UE set may be defined in a custom manner to include a list of UEs, where the criteria is based on considerations external to RIC, such as UE reputation, vendor, or vulnerability type. The actions applied to a single UE or set of UE should include (but not be limited to) rejecting attach requests, throttling attach requests (attach rate should not exceed a certain threshold), handover a set of devices to a different radio or applying a different slice to set.

3.15.4 Benefits of O-RAN Architecture

The O-RAN architecture allows applying protection to the network at the edge, thus minimizing the effect of these signaling storm DDoS attacks on network resources.

The combination of near real time logic at the Near Real-Time RIC for fast detection, with slower scale analysis and input from the Non Real-Time RIC and SMO provides a good mixture between quick reaction and advanced detection schemes. Lighter detection processes can be applied as xApps while more advanced heavy processing ML analytics can run externally and send input to RIC over the A1 interface.

3.16 Use case 16: Congestion Prediction & Management

3.16.1 Background Information

This use case provides a proactive approach to congestion handling in the base station by analyzing the radio resource utilization and taking timely corrective action so as to mitigate any potential congestion in the system.

3.16.2 Motivation

Large-scale commercial cellular networks have many problems - like cell congestions leading to RLF (Radio Link Failure etc.), Handover Failure, poor data rates etc. Network congestion is a crucial problem for the telecom operators as it affects the Quality of Service (QoS) of the users directly. An operator has many solutions like Offload (across carriers, Wi-Fi, etc.) or Antenna techniques (Cell Split, Higher order MIMO, etc.) to handle congestion. However, the congestion patterns in the network are not fully understood and mitigation is done post facto at the expense of the prolonged user experience degradation. The congestion mitigation is critical for operators to retain their subscribers for operator and individual user experience. With 5G, the congestion needs to be handled to as to best utilize the radio resources. Today operators do not have a well-defined mechanism to predict congestion. The main objective of this use case is to use the embedded intelligence of O-RAN to predict the congestion ahead of time, so that operators can keep the cell congestion mitigation solution in place before the congestion is predicted to happen.

3.16.3 Proposed Solution

CPM (Congestion Prediction & Management) architecture is proposed to detect and mitigate congestion pro-actively. In the CPM architecture, E2 node statistics [counters] are collected by the data collector of SMO. This is done over the O1 interface. The pre-processing of data is also done in the same place. Pre-processing includes adding VNF/cell names/numbers and ids to the data and converting counters into KPI using KPI logic. After preprocessing data is shared with Non-RT RIC deployed in the SMO using a data sharing entity. Non-RT RIC will invoke the corresponding training model/application in an AI server inside SMO (it can be placed outside SMO also). The data cleaning and training will happen and the predicted KPIs will be sent back to CPM rApp in Non-RT RIC. Machine learning models, can be used to learn and predict the future traffic for the next hour/day/month. The prediction window can be configured by operators as per the available data and its periodicity. CPM rAPP in Non-RT RIC will form the inference. Inference logic to define cell congestion can be like:

1. The average user-perceived IP throughput < P Mbps
2. DL PRB utilization > Q%
3. Average RRC User > R.

The inference will contain the cell ids, information about whether the cell is congested or not, time stamp of cell congestion and predicted KPI value (to decide the congestion intensity). As per the CPM rApp information in Non-RT RIC, there are two options to mitigate cell congestion as follows:

Option a: CPM rApp in Non-RT RIC transfers the inference to the CPM xApp in Near-RT RIC through A1 interface. Near-RT RIC can decide the mitigation solutions as per the inference. Mitigation solutions can include

1. Switching to dual connectivity mode
2. Debarring of user access
3. Load sharing.

These solutions can be controlled over E2 interface.

Option b: Non-RT RIC can also directly help to mitigate the congestion with the help of O1 interface. Some of the mitigation solutions can be

1. Splitting a cell (assuming hardware support available)
2. Add more carriers
3. Switching to higher order MIMO
4. Switching some of the users to Wi-Fi.

Finally, the E2 nodes feedback can be sent to Non-RT RIC through O1 interface. Figure 3.16.3-1 depicts the proposed CPM solution in detail.

Please note: There can be more mitigation solutions apart from the one provided in steps 6a and 6b (see Fig 3.16.3-1). Choosing the mitigation solution is up to the operator to decide as per the congestion intensity (specified in the inference).

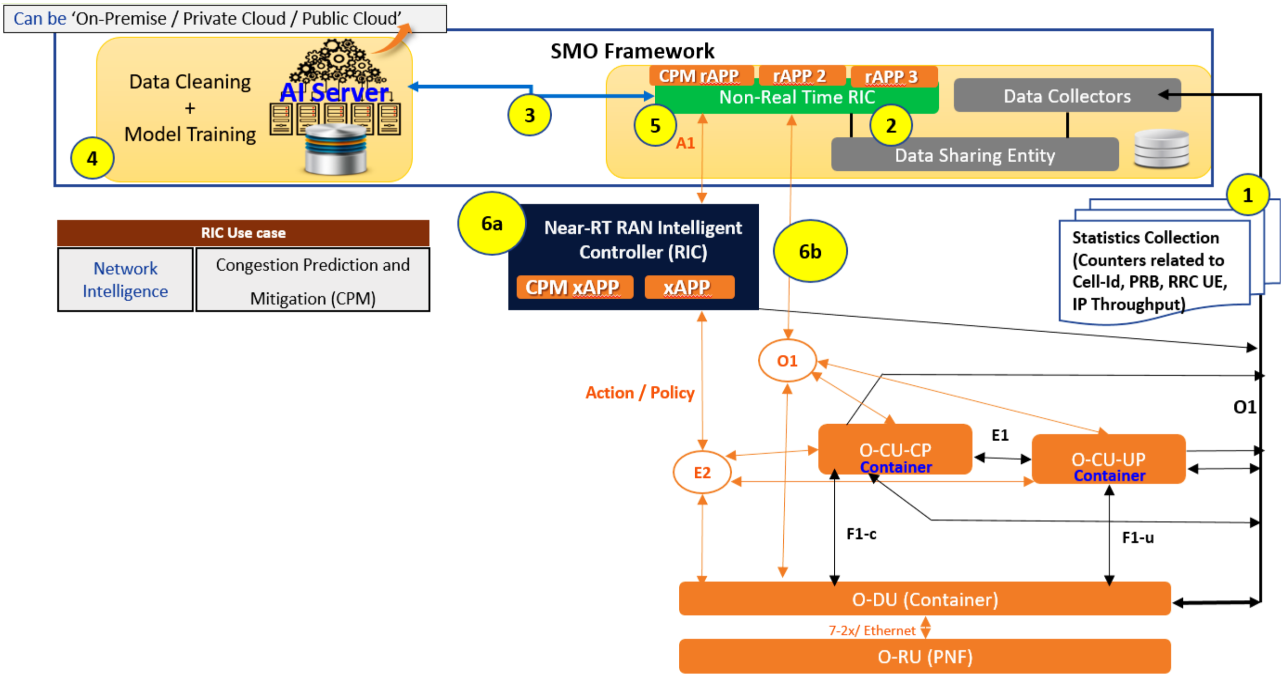


Figure 3.16.3-1: Proposed Congestion Prediction Management Solution

3.16.4 Benefits of O-RAN Architecture

All the main O-RAN components can leverage from proposed CPM use case. It predicts congestion before its occurrence so that an operator can configure the cell and prevent the degradation of QoS of a cell. AI server in SMO can be utilized for model training of KPIs. Similar models can be used to predict KPIs for other use cases as well. Non-RT RIC can be utilized to prepare inference based on predicted KPIs and congestion logic. These inference can be helpful for operators to choose on what solutions can be applied for a give congestion severity. CPM utilize Near-RT RIC for configuring a cell well before the congestion based on the inference and infrastructure dynamics.

3.17 Use case 17: Industrial IoT Optimization

3.17.1 Background Information

In 3GPP Industrial Internet of Thing (IIoT) item, new scenarios with high reliability (i.e. 10^{-5}~10^{-6}) have been considered including factory automation, transport industry. To satisfy these scenarios requirements, several key features as below have been supported for IIoT in 5G system, such as data duplication and multi-connectivity enhancement (PDCP duplication, PDU session duplication, QoS flow duplication), Time Sensitive Networking (Ethernet Header Compression (EHC), Accurate reference timing provisioning, QoS and scheduling enhancements) and different prioritized transmission multiplexing.

Based on O-RAN architecture some of these features can be optimized, e.g., PDCP duplication, EHC, and different prioritized transmission multiplexing. For PDCP duplication, up to 4 RLC entities/legs can be supported. RRC signaling is used for initial configurations. And MAC CE is used for dynamic control. According to 3GPP specifications definition, PDCP duplication can only be applicable for NR. For EHC, PDCP entities associated with DRBs can be configured by RRC to use EHC. Each PDCP entity carrying user plane data may be configured to use EHC. Every PDCP entity uses at most one EHC compressor instance and at most one EHC decompressor instance. For different prioritized transmission multiplexing, the lower priority transmission can be cancelled and preempted by the higher priority transmissions. RRC signaling is used for the related configurations.

Network deployment of IIoT use case includes both public network and non-public network (NPN). For public network, network slicing may be used to implement the resource isolation to guarantee the performance of IIoT use case. For NPN or private network, it can have different ways to implement the network deployment. For example, one way is to deploy private network equipment completely, such as gNB, UPF and 5GC. The network equipment can only provide services for the private use. The other way is to use network slicing to implement a virtual NPN, which means only some of the network equipment are deployed for private use, such as gNB. Meanwhile, the others, such as UPF, 5GC, are shared with the public network. The network resource isolation can be implemented by network slicing. If network slicing is used, IIoT use case only focuses on optimization for the specific piece of slicing bearing IIoT traffic.

3.17.2 Motivation

The objective of the IIoT optimization use case is to guarantee reliability, and resource utilization efficiency in the above scenarios by optimizing the enhancement mechanisms of related technique features, e.g. PDCP duplication, EHC, different prioritized transmission multiplexing and etc.

PDCP duplication and EHC is applied to the corresponding DRB. Improper configuration of PDCP duplication and EHC will leads to the waste of HW/SW and transmission resources. What service or data radio bearer (DRB) needs to use PDCP duplication is not only depends on the reliability requirement, but also related to the network environment, such as network load and channel quality/interference condition. Besides, what service or DRB needs to use EHC is depends on the service characteristics, such as service type, traffic period, duration, and packet size. EHC is especially beneficial when payload sizes are relatively small compared to the overall size of the frame, which is typically the case in IIoT networks based on Ethernet. Additionally, for different prioritized transmission multiplexing, which and where users are configured to listen to the Pre-emption Indication (PI) or Cancellation Indication (CI) depends on the service characteristics and network conditions. Improper configuration of PI or CI related configurations will cause invalid monitoring and power consumptions.

Therefore, the the enhancement mechanisms for IIoT shall be semi-static or dynamic configured for some specific users and some specific services based on the service prediction and performance prediction, considering the variant network load and channel quality/interference environment. AI/ML can be used for service prediction and KPI prediction to infer and derive the strategies.

3.17.3 Proposed Solution

The proposed solutions include Non-RT RIC optimization solution and Near-RT RIC optimization solution. The Non-RT RIC optimization is a slower loop while Near-RT RIC solution has faster loop.

3.17.3.1 Non-RT RIC solution of IIoT optimization

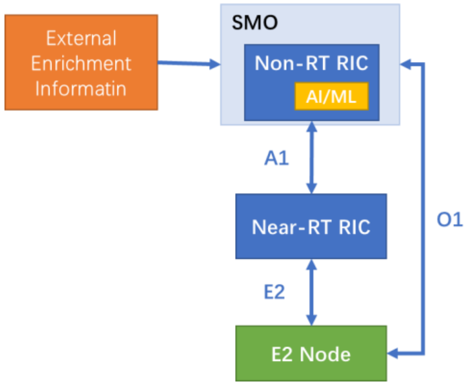


Figure 3.17.3-1: Architecture of Non-RT RIC solution of IIoT optimization use case

Non-RT RIC shall be able to retrieve network data (network load, performance metrics, QoS flow 5QI, and etc.) and external service information from SMO. The external enrichment information such as service type, traffic period, duration, packet size, KPI requirements can be provided by MEC server, APP server or industrial control platform and etc., to SMO. Non-RT RIC shall be able to retrieve configured policy from SMO, e.g., cell edge user reliability KPI target, Ethernet header compression efficiency. Additionally, Non-RT RIC shall be able to deploy and update AI/ML model from SMO. Non-RT RIC shall use AI/ML model to infer and decide related IIoT key techniques configurations, such as whether the QoS flow shall map to DRB supporting duplication and EHC, whether high priority traffic can preempt transmission resource by cancelling the low priority traffic transmission.

Near-RT RIC shall be able to retrieve IIoT related configurations from Non-RT RIC (if it configures over A1). Near-RT RIC shall be able to send the configurations to E2 nodes by policy.

E2 Node (O-CU, O-DU) shall provide UE measurement report, performance metrics, and etc. to SMO through O1 interface. E2 Node shall enforce E2 control policy or O1 configuration, if O1 configuration per UE can be supported.

3.17.3.2 Near-RT RIC solution of IIoT optimization

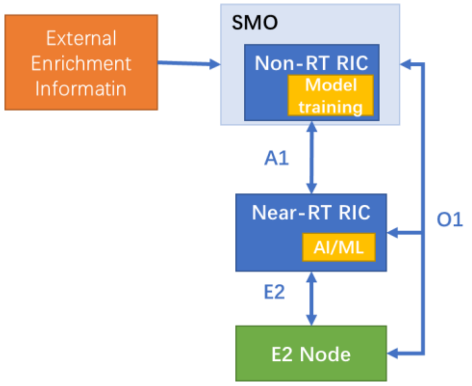


Figure 3.17.3-2: Architecture of Near-RT RIC solution of IIoT optimization use case

Non-RT RIC shall be able to retrieve network data (network load, performance metrics, QoS flow 5QI, and etc.) and external service information from SMO. The external enrichment information such as service type, traffic period, duration, packet size, KPI requirements can be provided by MEC server, APP server or industrial control platform and etc. to SMO. Non-RT RIC shall be able to train AI/ML model for Near-RT RIC. Additionally, Non-RT RIC shall be able to evaluate the collected data and A1 policy feedback, if required, and generate or update the appropriate optimization policy, such as reliability targets, and sends it to the Near-RT RIC via A1 interface.

Near-RT RIC shall be able to deploy or update AI/ML model via O1 interface. Near-RT RIC shall be able to receive an A1 policy from the Non-RT RIC, and initiate the corresponding optimization procedure. Additionally, Near-RT RIC shall be able to evaluate the performance data from the E2 Nodes and monitor whether the performance is out of KPI targets which are indicated in the A1 policy. E.g., cell edge user performance cannot meet the A1 policy. Based on service estimation, i.e. service characteristics, and channel quality estimation, and A1 policy, the Near-RT RIC shall infer and make decision of PDCP duplication on/off, RLC entity selection and high priority traffic preemption. Then, Near-RT RIC may generate new E2 policies or modify the existing ones to send them to the E2 Nodes. If required, Near-RT RIC shall send report to Non-RT RIC for evaluation and optimization.

E2 Node (O-CU, O-DU) shall provide network metrics and etc. to SMO through O1 interface. E2 Node shall provide UE measurement report, performance metrics, and cell resource utility to Near-RT RIC through E2 interface. Additionally, E2 Node shall enforce E2 control policy.

3.17.4 Benefits of O-RAN Architecture

The benefits of O-RAN architecture for IIoT use case include the ability to use external enrichment information provided by MEC server, APPs server, or industrial control platform to SMO. Based on the external service information and RAN information, AI/ML method is introduced to help generating IIoT key techniques strategies to improve performance of system and users. For SDAP sub-layer, it can provide semi-static QoS flow to DRB mapping rules. For PDCP and RLC sub-layer, semi-static EHC on/off configuration, semi-static or dynamic PDCP duplication activate/deactivate decision and RLC entities dynamic control/selection decision can be provided. For MAC and PHY, it can provide semi-static and dynamic PI and CI configurations when different prioritized transmissions are multiplexed. Besides, multi-E2 Node coordination will be beneficial based on O-RAN architecture.

3.18 Use case 18: BBU Pooling to achieve RAN Elasticity

3.18.1 Background Information

Driving cost efficiency by improving the elasticity of workloads is one of the principal goals of cloudification. This use case presents BBU pooling as an opportunity to achieve RAN elasticity. O-RAN CADS describes multiple deployment options for the O-RAN cloudified NFs, which allows for different variants of centralization of these resources in the edge and regional clouds. In particular, cloudified BBUs can be deployed on a common hardware pool, centralized at the same cloud location. This centralization opens up the opportunity for O-RUs to be flexibly mapped to BBUs, thereby enabling RAN elasticity, while potentially lowering the costs in the long term

3.18.2 Motivation

Pooling of BBU resources provides many benefits like having to maintain only a single large, controlled environment for many cell sites, better resiliency and agility to deal with unexpected failures and increases in demands, and allowing capacity to be added incrementally and assigned to cell sites as needed resulting in better overall utilization. More importantly, BBU Pooling provides significant statistical multiplexing benefits due to better resource consolidation. The total resources required from the shared pool will be less than total resources required across distributed locations, because the peak of the sum of the traffic will generally be lower than the sum of the individual cell site traffic peaks. BBU Pooling enables more aggressive capacity dimensioning based on different percentiles of resource consumption rather than for the peak consumption. This proposal addresses the scenarios where both O-CUs and O-DUs can pooled, but for simplicity the following sections focuses on the description of O-DU pooling.

3.18.3 Proposed Solution

Figure 3.18.3-1 shows a high-level depiction of approaches to O-DU Pooling. The O-DU resources for multiple cell sites are pooled at a single centralized location or the *O-DU Pool*. The Cloud Access Switch (CAS) at the O-DU Pool aggregates traffic from the multiple Front Haul Gateways and routes it to the appropriate O-DU blade based on the load balancing requirement. For the O-DU pooling use case, it is assumed that an open Front Haul exists with O-RAN defined 7-2x split. Based on the association between O-RUs and O-DUs, and the granularity at which the traffic is assigned from the O-RUs to the O-DUs, we define the following classes of O-DU pooling.

1) **Class 0 pooling**: A scenario where an O-RU is assigned to a single specific O-DU statically, and the traffic from the O-RU is not split into subsets that could be assigned to different O-DUs. This is the same as Simple Centralization as defined in CADS V2.0. With Class 0 pooling, re-assigning an O-RU to connect to a different O-DU would require significant “hands on” activity, and is performed very infrequently during specific maintenance windows.
2) **Class 1 pooling:** A scenario where *n* O-RUs are initially assigned to a single O-DU during a specific period of time, but the O-RUs can be re-assigned to different O-DU at any point of time via orchestration procedures with the help of SMO. This automated re-assignment can be triggered outside maintenance windows, based on various performance criterion or for load balancing. Like with Class 0 pooling, the traffic from the O-RU is not split into subsets that could be assigned to different O-DUs.
3) **Class 2 pooling:** A scenario where n O-RUs are assigned to m O-DUs, and subsets of traffic from one O-RU are dynamically distributed (and load-balanced) across the O-DU resources within the O-DU pool using a cluster aware scheduler.

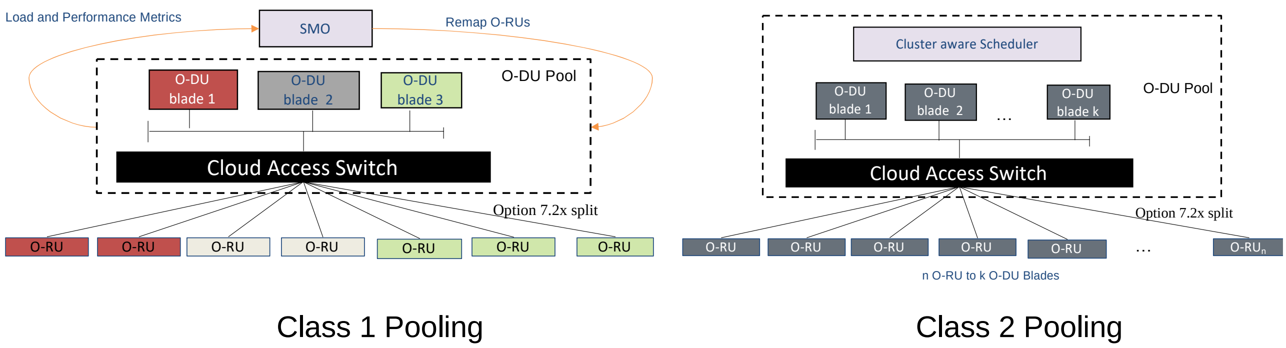


Figure 3.18.3-1:O-DU Pooling

With all classes of O-DU pooling, the total resources required from the shared pool will be less than total resources required across distributed locations, because the peak of the sum of the traffic will generally be lower than the sum of the individual cell site traffic peaks. In particular, Class 1 and Class 2 O-DU pooling enables more aggressive capacity dimensioning based on different percentiles of resource consumption rather than for the peak consumption. However, the key distinction between the classes of pooling is the granularity at which traffic can be load balanced across the O-DUs, which provides different statistical multiplexing benefits. While Class 0 pooling provides a one-time statistical multiplexing gains at the time of initial deployment due to the centralization, Class 1 pooling allows further load balancing benefits due to the automated and flexible re-mapping of O-RUs to the O-DUs periodically. However, the re-assignment of O-RUs needs to be performed during maintenance windows (or via hitless migration) to minimize service interruptions. Finally, Class 1 Pooling enables better handling of hot spots, and more efficient management of traffic loads and maintenance/upgrades.

On the other hand, Class 2 pooling dynamically assigns subsets of traffic to the O-DU resources in the shared O-DU pool. Hence, if some cell sites experience light traffic demand, while others experience high traffic demand, then subsets of traffic can be routed to O-DU resources in the shared pool such that the traffic is optimally consolidated among the O-DU resources needed to meet the performance requirements, leading to significant statistical multiplexing benefits. At one extreme, Class 2 pooling allows each O-DU in the O-DU pool to have similar load (once the optimal number of blades have been predicted for a given time period), if each UE flow or each slot processing can be mapped to the least-busy blade for the duration of that flow. In addition, there is no service interruption as with Class 1 pooling. Class 2 pooling could also be performed with more coarse-grained distribution of traffic subsets to the O-DUs. For example, traffic could be mapped to the least-utilized blade on a per UE basis or on a per-carrier basis within a TRP.

3.18.4 Benefits of O-RAN Architecture

As described in section 3.18.3, both Class 1 and Class 2 Pooling are enabled by the O-RAN Open Fronthaul with LLS (option 7-2x split). Further, for Class 1 Pooling, O-RAN Architecture standardizes the SMO functions needed to enable pooled BBUs including:

- Management of O-DU & O-RU (Registration/De-registration of O-RU to O-DU, Registration of RU locations to SMO)
- Orchestration features via O1/O2 interfaces and procedures for remapping O-RUs from one O-DU to another O-DU, implementing the Orchestration control loops provided by the non-RT RIC, hitless draining of cell traffic from one O-DU to another O-DU.
- Monitoring Performance/Fault metrics to determine DU load and KPIs.

Class 2 pooling is relatively less explored in the industry, aiming to provide fine-grained load scheduling across the O-DUs in the pool. For Class 2 Pooling, O-RAN Architecture helps with:

- Affinity of control and downlink data, supported by cluster aware scheduler.
- Affinity between control and uplink data, supported by the Cloud Access Switch or in the O-RU.
- Configuration of the O-DU to support a wider range of sectors through the cluster aware scheduler.
- Monitoring Performance/Fault metrics to determine DU load and KPIs.
- Triggers for the scale-in/scale-out workflows in the SMO that manages the O-DU Pool.
- Fine-grained load balancing functions through the RIC that standardizes the role of CU-CP High.

Annex ZZZ O-RAN Adopter License Agreement

BY DOWNLOADING, USING OR OTHERWISE ACCESSING ANY O-RAN SPECIFICATION, ADOPTER AGREES TO THE TERMS OF THIS AGREEMENT.

This O-RAN Adopter License Agreement (the "Agreement") is made by and between the O-RAN Alliance and the entity that downloads, uses or otherwise accesses any O-RAN Specification, including its Affiliates (the "Adopter").

This is a license agreement for entities who wish to adopt any O-RAN Specification.

SECTION 1: DEFINITIONS

1.1 "Affiliate" means an entity that directly or indirectly controls, is controlled by, or is under common control with another entity, so long as such control exists. For the purpose of this Section, "Control" means beneficial ownership of fifty (50%) percent or more of the voting stock or equity in an entity.

1.2 "Compliant Portion" means only those specific portions of products (hardware, software or combinations thereof) that implement any O-RAN Specification.

1.3 "Adopter(s)" means all entities, who are not Members, Contributors or Academic Contributors, including their Affiliates, who wish to download, use or otherwise access O-RAN Specifications.

1.4 "Minor Update" means an update or revision to an O-RAN Specification published by O-RAN Alliance that does not add any significant new features or functionality and remains interoperable with the prior version of an O-RAN Specification. The term "O-RAN Specifications" includes Minor Updates.

1.5 "Necessary Claims" means those claims of all present and future patents and patent applications, other than design patents and design registrations, throughout the world, which (i) are owned or otherwise licensable by a Member, Contributor or Academic Contributor during the term of its Member, Contributor or Academic Contributorship; (ii) such Member, Contributor or Academic Contributor has the right to grant a license without the payment of consideration to a third party; and (iii) are necessarily infringed by implementation of a Final Specification (without considering any Contributions not included in the Final Specification). A claim is necessarily infringed only when it is not possible on technical (but not commercial) grounds, taking into account normal technical practice and the state of the art generally available at the date any Final Specification was published by the O-RAN Alliance or the date the patent claim first came into existence, whichever last occurred, to make, sell, lease, otherwise dispose of, repair, use or operate an implementation which complies with a Final Specification without infringing that claim. For the avoidance of doubt in exceptional cases where a Final Specification can only be implemented by technical solutions, all of which infringe patent claims, all such patent claims shall be considered Necessary Claims.

1.6 "Defensive Suspension" means for the purposes of any license grant pursuant to Section 3, Member, Contributor, Academic Contributor, Adopter, or any of their Affiliates, may have the discretion to include in their license a term allowing the licensor to suspend the license against a licensee who brings a patent infringement suit against the licensing Member, Contributor, Academic Contributor, Adopter, or any of their Affiliates.

SECTION 2: COPYRIGHT LICENSE

2.1 Subject to the terms and conditions of this Agreement, O-RAN Alliance hereby grants to Adopter a nonexclusive, nontransferable, irrevocable, non-sublicensable, worldwide copyright license to obtain, use and modify O-RAN Specifications, but not to further distribute such O-RAN Specification in any modified or unmodified way, solely in furtherance of implementations of an O-RAN Specification.

2.2 Adopter shall not use O-RAN Specifications except as expressly set forth in this Agreement or in a separate written agreement with O-RAN Alliance.

SECTION 3: FRAND LICENSE

3.1 Members, Contributors and Academic Contributors and their Affiliates are prepared to grant based on a separate Patent License Agreement to each Adopter under Fair, Reasonable And Non-Discriminatory (FRAND) terms and conditions with or without compensation (royalties) a nonexclusive, non-transferable, irrevocable (but subject to Defensive Suspension), non-sublicensable, worldwide license under their Necessary Claims to make, have made, use, import, offer to sell, lease, sell and otherwise distribute Compliant Portions; provided, however, that such license shall not extend: (a) to any part or function of a product in which a Compliant Portion is incorporated that is not itself part of the Compliant Portion; or (b) to any Adopter if that Adopter is not making a reciprocal grant to Members, Contributors and Academic Contributors, as set forth in Section 3.3. For the avoidance of doubt, the foregoing license includes the distribution by the Adopter's distributors and the use by the Adopter's customers of such licensed Compliant Portions.

3.2 Notwithstanding the above, if any Member, Contributor or Academic Contributor, Adopter or their Affiliates has reserved the right to charge a FRAND royalty or other fee for its license of Necessary Claims to Adopter, then Adopter is entitled to charge a FRAND royalty or other fee to such Member, Contributor or Academic Contributor, Adopter and its Affiliates for its license of Necessary Claims to its licensees.

3.3 Adopter, on behalf of itself and its Affiliates, shall be prepared to grant based on a separate Patent License Agreement to each Members, Contributors, Academic Contributors, Adopters and their Affiliates under FRAND terms and conditions with or without compensation (royalties) a nonexclusive, non-transferable, irrevocable (but subject to Defensive Suspension), non-sublicensable, worldwide license under their Necessary Claims to make, have made, use, import, offer to sell, lease, sell and otherwise distribute Compliant Portions; provided, however, that such license will not extend: (a) to any part or function of a product in which a Compliant Portion is incorporated that is not itself part of the Compliant Portion; or (b) to any Members, Contributors, Academic Contributors, Adopters and their Affiliates that is not making a reciprocal grant to Adopter, as set forth in Section 3.1. For the avoidance of doubt, the foregoing license includes the distribution by the Members', Contributors', Academic Contributors', Adopters' and their Affiliates' distributors and the use by the Members', Contributors', Academic Contributors', Adopters' and their Affiliates' customers of such licensed Compliant Portions.

SECTION 4: TERM AND TERMINATION

4.1 This Agreement shall remain in force, unless early terminated according to this Section 4.

4.2 O-RAN Alliance on behalf of its Members, Contributors and Academic Contributors may terminate this Agreement if Adopter materially breaches this Agreement and does not cure or is not capable of curing such breach within thirty (30) days after being given notice specifying the breach.

4.3 Sections 1, 3, 5 - 11 of this Agreement shall survive any termination of this Agreement. Under surviving Section 3, after termination of this Agreement, Adopter will continue to grant licenses (a) to entities who become Adopters after the date of termination; and (b) for future versions of O-RAN Specifications that are backwards compatible with the version that was current as of the date of termination.

SECTION 5: CONFIDENTIALITY

Adopter will use the same care and discretion to avoid disclosure, publication, and dissemination of O-RAN Specifications to third parties, as Adopter employs with its own confidential information, but no less than reasonable care. Any disclosure by Adopter to its Affiliates, contractors and consultants should be subject to an obligation of confidentiality at least as restrictive as those contained in this Section. The foregoing obligation shall not apply to any information which is: (1) rightfully known by Adopter without any limitation on use or disclosure prior to disclosure; (2) publicly available through no fault of Adopter; (3) rightfully received without a duty of confidentiality; (4) disclosed by O-RAN Alliance or a Member, Contributor or Academic Contributor to a third party without a duty of confidentiality on such third party; (5) independently developed by Adopter; (6) disclosed pursuant to the order of a court or other authorized governmental body, or as required by law, provided that Adopter provides reasonable prior written notice to O-RAN Alliance, and cooperates with O-RAN Alliance and/or the applicable Member, Contributor or Academic Contributor to have the opportunity to oppose any such order; or (7) disclosed by Adopter with O-RAN Alliance's prior written approval.

SECTION 6: INDEMNIFICATION

Adopter shall indemnify, defend, and hold harmless the O-RAN Alliance, its Members, Contributors or Academic Contributors, and their employees, and agents and their respective successors, heirs and assigns (the "Indemnitees"), against any liability, damage, loss, or expense (including reasonable attorneys' fees and expenses) incurred by or imposed upon any of the Indemnitees in connection with any claims, suits, investigations, actions, demands or judgments arising out of Adopter's use of the licensed O-RAN Specifications or Adopter's commercialization of products that comply with O-RAN Specifications.

SECTION 7: LIMITATIONS ON LIABILITY; NO WARRANTY

EXCEPT FOR BREACH OF CONFIDENTIALITY, ADOPTER'S BREACH OF SECTION 3, AND ADOPTER'S INDEMNIFICATION OBLIGATIONS, IN NO EVENT SHALL ANY PARTY BE LIABLE TO ANY OTHER PARTY OR THIRD PARTY FOR ANY INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE OR CONSEQUENTIAL DAMAGES RESULTING FROM ITS PERFORMANCE OR NON-PERFORMANCE UNDER THIS AGREEMENT, IN EACH CASE WHETHER UNDER CONTRACT, TORT, WARRANTY, OR OTHERWISE, AND WHETHER OR NOT SUCH PARTY HAD ADVANCE NOTICE OF THE POSSIBILITY OF SUCH DAMAGES.

O-RAN SPECIFICATIONS ARE PROVIDED "AS IS" WITH NO WARRANTIES OR CONDITIONS WHATSOEVER, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHERWISE. THE O-RAN ALLIANCE AND THE MEMBERS, CONTRIBUTORS OR ACADEMIC CONTRIBUTORS EXPRESSLY DISCLAIM ANY WARRANTY OR CONDITION OF MERCHANTABILITY, SECURITY, SATISFACTORY QUALITY, NONINFRINGEMENT, FITNESS FOR ANY

PARTICULAR PURPOSE, ERROR-FREE OPERATION, OR ANY WARRANTY OR CONDITION FOR O-RAN SPECIFICATIONS.

SECTION 8: ASSIGNMENT

Adopter may not assign the Agreement or any of its rights or obligations under this Agreement or make any grants or other sublicenses to this Agreement, except as expressly authorized hereunder, without having first received the prior, written consent of the O-RAN Alliance, which consent may be withheld in O-RAN Alliance's sole discretion. O-RAN Alliance may freely assign this Agreement.

SECTION 9: THIRD-PARTY BENEFICIARY RIGHTS

Adopter acknowledges and agrees that Members, Contributors and Academic Contributors (including future Members, Contributors and Academic Contributors) are entitled to rights as a third-party beneficiary under this Agreement, including as licensees under Section 3.

SECTION 10: BINDING ON AFFILIATES

Execution of this Agreement by Adopter in its capacity as a legal entity or association constitutes that legal entity's or association's agreement that its Affiliates are likewise bound to the obligations that are applicable to Adopter hereunder and are also entitled to the benefits of the rights of Adopter hereunder.

SECTION 11: GENERAL

This Agreement is governed by the laws of Germany without regard to its conflict or choice of law provisions.

This Agreement constitutes the entire agreement between the parties as to its express subject matter and expressly supersedes and replaces any prior or contemporaneous agreements between the parties, whether written or oral, relating to the subject matter of this Agreement.

Adopter, on behalf of itself and its Affiliates, agrees to comply at all times with all applicable laws, rules and regulations with respect to its and its Affiliates' performance under this Agreement, including without limitation, export control and antitrust laws. Without limiting the generality of the foregoing, Adopter acknowledges that this Agreement prohibits any communication that would violate the antitrust laws.

By execution hereof, no form of any partnership, joint venture or other special relationship is created between Adopter, or O-RAN Alliance or its Members, Contributors or Academic Contributors. Except as expressly set forth in this Agreement, no party is authorized to make any commitment on behalf of Adopter, or O-RAN Alliance or its Members, Contributors or Academic Contributors.

In the event that any provision of this Agreement conflicts with governing law or if any provision is held to be null, void or otherwise ineffective or invalid by a court of competent jurisdiction, (i) such provisions will be deemed stricken from the contract, and (ii) the remaining terms, provisions, covenants and restrictions of this Agreement will remain in full force and effect.

Any failure by a party or third party beneficiary to insist upon or enforce performance by another party of any of the provisions of this Agreement or to exercise any rights or remedies under this Agreement or otherwise by law shall not be construed as a waiver or relinquishment to any extent of the other parties' or third party beneficiary's right to assert or rely upon any such provision, right or remedy in that or any other instance; rather the same shall be and remain in full force and effect.

….

