RIMEDO LABS

Quality of Service-based Resource Allocator xApp (QRA-xApp)

TECHNICAL SPECIFICATION

Overview

The 5G networks are expected to be capable of serving highly diversified traffic. There are many use cases related to the specific Quality of Service (QoS) requirements, e.g., buffered video streaming requires large bandwidth, and allows relatively relaxed latency, while voice connection demands small bandwidth, but low and predictive latency is of high importance for this type of network traffic. As 5G networks follow the concept of Network Function Virtualization (NFV), and Software Defined Networking (SDN) it is possible to create a virtual network that would serve users requesting similar QoS demands, i.e., utilize Network Slicing concept, where one slice can be dedicated to the services demanding large bandwidth, and another for the ones requesting low latency.

The QoS requirements related to the particular network slice are defined at the stage of slice creation using the so-called Service Level Agreements (SLA). Moreover, within each slice different user demands on QoS can be distinguished in more detail due to the introduction of QoS flows. While each 5G cell has defined radio resources, it is an important task to effectively allocate them to network slices, so as to meet the SLAs, and individual QoS flow's requirements, e.g., minimum required user throughput, packet loss rate, maximum delay. An efficient approach to this aspect is to split the radio resources dynamically according to the actual traffic demands. This document describes a solution dedicated for this problem developed by Rimedo Labs in the form of xApp operating within the O-RAN architecture, at the Near-Real-Time RAN Intelligent Controller (Near-RT RIC).

QoS-based Resource Allocator xApp

For the purpose of dynamic allocation of radio resources Rimedo Labs proposes a QoS-based Resource Allocator xApp (QRA-xApp). The xApp dynamically controls the quota of Physical Resource Blocks (PRBs) that should be allocated to the different network slices, to meet their SLA requirements, while adjusting them based on the temporary traffic demand. See example operation in Figure 1.

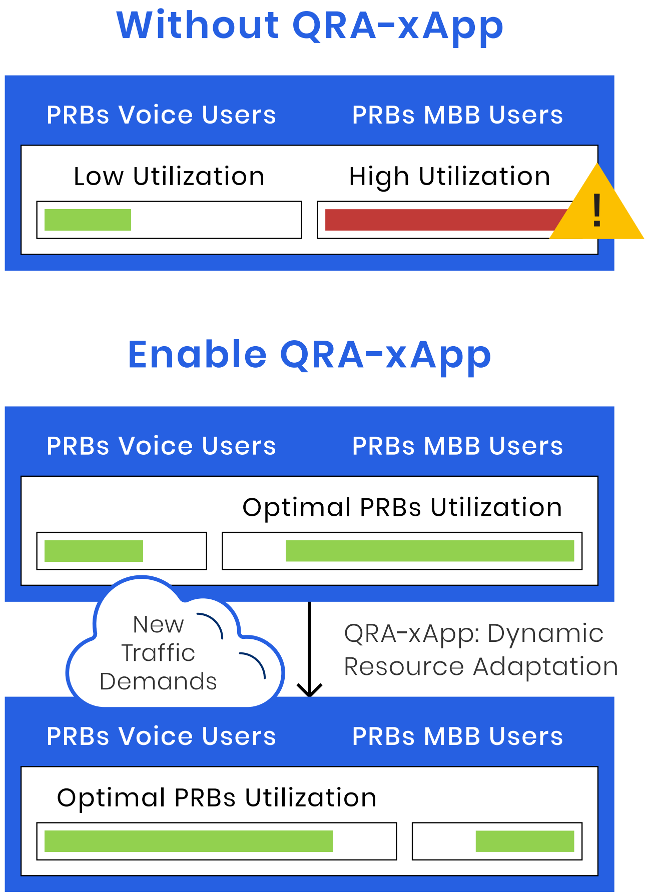


Figure 1. Example of the QRA-xApp operation

In this example, there are two network slices: one dedicated for the Voice users (Slice 1) and another dedicated for the Mobile Broadband (MBB) users (Slice 2). Initially, PRBs available at the gNB are split equally between the two slices.

However, PRBs dedicated for the Slice 1 are underutilized, while PRBs allocated for Slice 2 are not enough to meet the SLA. This is the place for QRA-xApp to intelligently improve the PRBs allocation. After switching on, the Rimedo QRA-xApp, PRBs are allocated, so as to provide both slices with optimal number of PRBs. QRA-xApp continuously monitors QoS parameters and radio resources utilization to split the PRBs between the network slices, e.g., when user proportion change, QRA would allocate more PRBs for the network slice serving Voice users.

The interfaces related to the QRA-xApp within O-RAN architecture is depicted in Figure 2 and is described in the next sections.

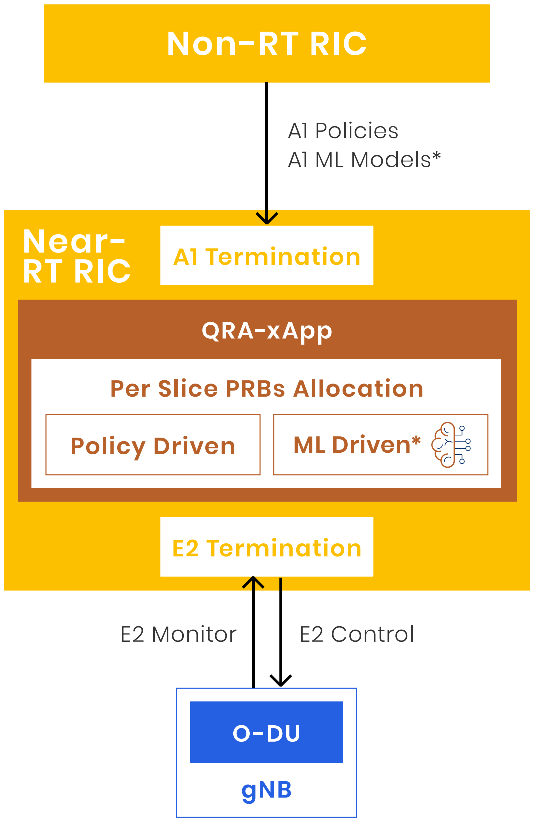


Figure 2. The deployment of QRA-xApp within the O-RAN architecture.
*ML modules are under development

Parameters Monitored Through E2 Interface – Inputs to QRA xApp from E2

The QRA-xApp is deployed at the Near-RT RIC to perform network optimization within the control loop of between 10 ms, and 1 s. To achieve this goal, the QRA-xApp must determine the current network state via the E2 interface. In compliance with [1], the following parameters are monitored by the QRA-xApp:

Mandatory Parameters – Necessary to monitor SLA-defined requirements

- **Distribution of DL/UL UE throughput in gNB** - provides the xApp with information about the distribution of user throughput in either DL or UL. These measurements can be obtained within the scope of a cell, network slice or QoS Flow [2].
- **Radio Resource Utilization** - is a group of parameters being obtained by the QRA-xApp to monitor the utilization of Physical Resource Blocks (PRBs) at a certain cell. These are [2]:
    - **DL/UL total available PRB** – provides QRA-xApp with information about the total number of PRBs available for DL/UL transmission at certain cell
    - **Mean DL/UL PRB used for data traffic** – provides QRA-xApp with information about the average number of PRBs being used in DL/UL for data traffic over a given period. These statistics can be obtained within the scope of a cell, network slice, or QoS flow.
    - **Peak DL/UL PRB used for data traffic** - provides QRA-xApp with information about the peak number of PRBs being used in DL/UL for data traffic over a given time period. This information can be provided within the scope of a cell, network slice, or QoS Flow.
- **Number of Active UEs** – is a group of parameters provided to the QRA-xApp to monitor the number of active UEs. The number of active UEs is understood as a number of active Data Radio Bearers (DRBs). These are [2]:
    - **Max number of Active UEs in the DL/UL per cell** – provides QRA-xApp with information about the maximum number of active UEs (active DRBs) in DL/UL, within the cell, network slice, or per QoS Flow.

    - **Number of Active UEs in the DL/UL per cell** - provides QRA-xApp with information about the mean number of active UEs (active DRBs) in DL/UL, within the cell, network slice, or per QoS Flow.
- **PDU Session Management** – is a group of parameters provided to the QRA-xApp to monitor the number of PDU sessions (requested, successfully setup, and failed to setup) within the scope of gNB or network slice. These are [2]:
    - **Number of PDU Sessions requested to setup** – provides QRA-xApp with information about the number of PDU sessions that are requested to be set up within the gNB or per network slice.
    - **Number of PDU Sessions successfully setup** - provides QRA-xApp with information about the number of PDU sessions that are successfully set up within the gNB or per network slice.
    - **Number of PDU Sessions failed to setup** - provides QRA-xApp with information about the number of PDU sessions that failed to set up within the gNB or per network slice.
- **Cell Global Identity (CGI)** – the combination of Public Land Mobile Network Identity (PLMN ID) and E-UTRAN Cell Identity (ECI) or New Radio Cell Identity (NCI) [3].
- **Single – Network Slice Selection Assistance Information (S-NSSAI)** – identifies network slice within a PLMN. It is a combination of the slice/service type (SST) and a slice differentiator (SD) [4].

Optional Parameters – additional insights into the demands of QoS Flows within the network slice

- **Distribution of DL/UL Packet Drop/Loss Rate** – provides the xApp with the information about the distribution of packet drop/loss for DL, and UL respectively. This measurement is O-RAN-Specific, and can be provided within the scope of gNB, network slice, or QoS Flow [1].
- **Distribution of DL/UL delay between NG-RAN and UE** – provides the xApp with information about the distribution of the end-to-end delay between RAN and UE, either in DL or UL. This measurement is provided within the scope of gNB, network slice, or QoS Flow [2].
- **5G QoS Identifier (5QI)** – a parameter associated with a particular 5G QoS characteristic. These characteristics define requirements for a QoS Flow, e.g., packet error rate, and priority [4].

Parameters Controlled Through E2 Interface – Outputs from QRA xApp to E2

To fulfill QoS requirements of the mobile network users, the QRA-xApp adjusts the radio resources available for a certain cell between the associated network slices, with different QoS demands. In compliance with [5], the QRA-xApp can set the following scheduler parameters through the E2 interface:

- **Min PRB Policy Ratio** is the minimal percentage of PRBs that must be guaranteed for a given network slice. It includes shared, prioritized and dedicated PRBs. The sum of **Min PRB Policy Ratio** values of all slices should be less or equal to 100.
- **Max PRB Policy Ratio** is the maximum percentage of PRBs that can be assigned to the given network slice. It includes shared, prioritized, and dedicated PRBs.
- **Dedicated PRB Policy Ratio** it is a percentage of PRBs that is dedicated for a given slice. It includes only dedicated PRBs. The sum of **Dedicated PRB Policy Ratio** values of all slices should be less or equal to 100.

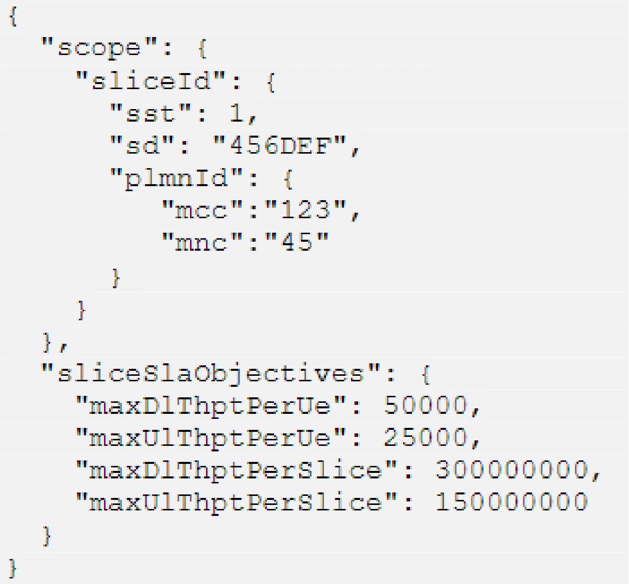
Note: QRA-xApp is also capable of dealing with splitting the radio resources between different users, group of users or QoS Flows within a single network slice. This functionality goes beyond the current scope of the O-RAN Alliance specification as defined in [5].

Policy Driven Resource Allocation – Inputs to QRA xApp from A1

The aim of the QRA-xApp is to dynamically (under near-RT control loop) split the PRBs between the network slices within the scope of a particular cell. The configuration is done through setting the proper parameters of PRB Policy Ratio in gNB through the E2 interface. By default, QRA-xApp, performs PRB control actions on the basis of policies sent by the Non-RT RIC through the A1 interface. These policies contain information about the SLA. The xApp itself maps these SLA requirements on the quota of PRBs to be assigned to each network slice within a particular cell, e.g., QRA-xApp reserves some PRBs for a given slice, and increases the amount of shared PRBs within other slices. According to the O-RAN Alliance specification, the policy type for such purpose is called the "SLA Target", and provides the xApp with at last one of the SLA parameters as defined in the following table [6]:

Parameter	Data Type	Description
maxNumberOfUes	Number	SLA target for the maximum number of UEs that can be served by the network slice concurrently
maxNumberOfPdu Sessions	Number	SLA target for the maximum number of PDU sessions to be supported by the network slice concurrently
guaDlThptPerSlice	Number	SLA target for providing guaranteed data rate (kbps) in downlink to be served by the network slice
maxDlThptPerSlice	Number	SLA target for providing maximum data rate supported by the network slice for all UEs together in downlink in kbps
maxDlThptPerUe	Number	Maximum data rate supported by the network slice per UE in downlink in kbps
guaUlThptPerSlice	Number	SLA target for providing guaranteed data rate as kbps in uplink to be served by the network slice
maxUlThptPerSlice	Number	SLA target for providing maximum data rate supported by the network slice for all UEs together in uplink in kbps
maxUlThptPerUe	Number	maximum data rate supported by the network slice per UE in uplink in kbps

The “SLA Target” policies are send from the Non-RT RIC to the QRA-xApp trough the A1 interface in the form of JSON files. The representative example of such a policy is depicted in Figure 3.



```
{
  "scope": {
    "sliceId": {
      "sst": 1,
      "sd": "456DEF",
      "plmnId": {
         "mcc":"123",
         "mnc":"45"
      }
    }
  },
  "sliceSlaObjectives": {
    "maxDlThptPerUe": 50000,
    "maxUlThptPerUe": 25000,
    "maxDlThptPerSlice": 300000000,
    "maxUlThptPerSlice": 150000000
  }
}
```

Figure 3. Example of JSON file containing SLA Target policy that defines maximum throughput levels per slice and per UE [6]

Machine Learning Driven Resource Allocation[1]

QRA-xApp can operate independently of the policies, i.e., it can use ML model to dynamically split radio resources between network slices. The ML model is continuously trained in the Non-RT RIC through interaction with 5G Network (environment), following the concept of Reinforcement Learning (RL). First, the pre-trained ML model is provided to the QRA-xApp from the Non-RT RIC trough the A1 interface. It is used to determine the number of PRBs to be assigned to each network slice (action) on the basis of input information from the E2 interface, e.g., UE throughputs, radio resource utilization, transmission delays, etc. (state). At the same time Non-RT RIC monitor the performance of actions by comparing QoS metrics obtained through O1 interface against SLAs, i.e., obtains reward. Finally, on the basis of action, state, reward sequences non-RT RIC can improve the ML model. The resultant RL Cycle is depicted in Figure 4.

[1] The Machine Learning modules are currently under the development

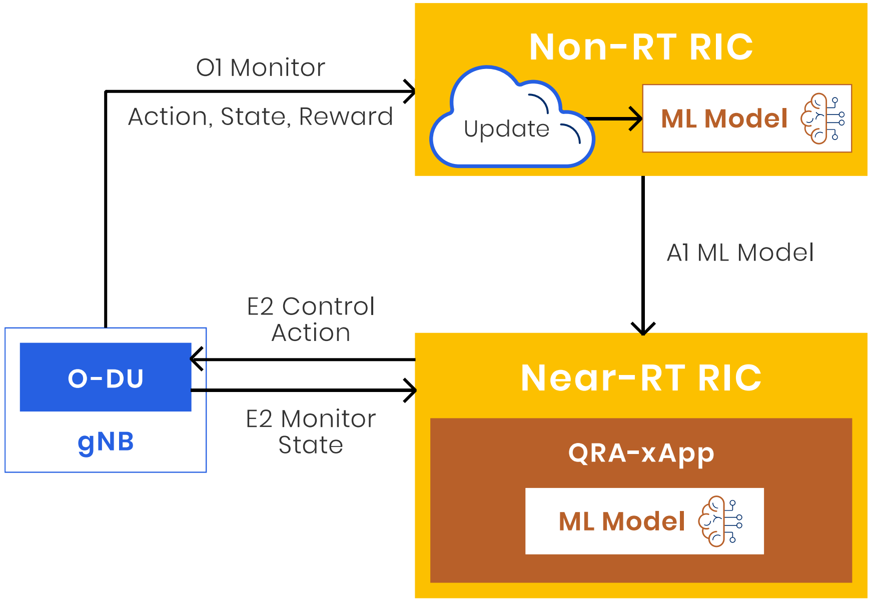


Figure 4. ML-driven resource allocation in QRA-xApp

Features and Applications

- QRA-xApp addresses a use case of RAN Slice SLA Assurance as identified by O-RAN Alliance [7].The main objective of the use case is to optimize resource allocation to fulfill Service Level Specifications (SLS).

- QRA-xApp can be used by MNOs to improve the radio resource utilization through dynamic adaptation of PRBs quota within network slices.
- The radio resource allocation can be driven either by the policies defined by the MNOs, or independently by the internal intelligence of QRA-xApp, i.e., Machine Learning
- QRA-xApp is suitable for radio resource management in the heterogenous networks, e.g., small cells can be configured to prioritize Mobile Broadband services (like video streaming), while macro cells can be configured to prioritize voice users.
- QRA-xApp can be adapted to the use-case of QoS-based resource optimization as defined by the O-RAN Alliance, by setting UE-oriented priority levels to meet the QoS Flow requirements [7].[2]
- The performance of QRA-xApp can be improved when working in the cooperation with other Rimedo xApps, and rApps e.g., Traffic Steering xApp, Frequency Band Selector rApp[2]

Bibliography

[1] O-RAN.WG3.E2SM-KPM-v02.01, "Near-Real-time RAN Intelligent Controller E2 Service Model (E2SM) KPM", O-RAN Alliance, March 2022

[2] 3GPP TS 28.552 V17.6.0, "Technical Specification Group Services and System Aspects, Management and orchestration, 5G performance measurements", 3rd Generation Partnership Project, March 2022

[3] 3GPP TS 23.003 V17.5.0, "Technical Specification Group Core Network and Terminals, Numbering, addressing and identification", 3rd Generation Partnership Project, March 2022

[4] 3GPP TS 23.501 V17.4.0, "Technical Specification Group Services and System Aspects, System architecture for the 5G System (5GS)", 3rd Generation Partnership Project, March 2022

[5] O-RAN.WG3.E2SM-RC-v01.01, "O-RAN Near-Real-time RAN Intelligent Controller E2 Service Model (E2SM), RAN Control", O-RAN Alliance, March 2022

[6] O-RAN.WG2.A1TD-v02.00, "O-RAN A1 interface: Type Definitions", O-RAN Alliance, October 2021

[7] O-RAN.WG2.Use-Case-Requirements-v05.00, "O-RAN Non-RT RIC & A1 Interface: Use Cases and Requirements", O-RAN Alliance, March 2022

[2] These features are under development

Notes:

- For cooperation models, reach out to us at: info@rimedolabs.com
- The information contained herein is the property of RIMEDO sp. z o. o. and is provided only if it is not disclosed, directly or indirectly to a third party, or used for purposes other than those for which it was prepared.
- All information discussed in the document is provided "as is" and RIMEDO makes no warranty that this information is fit for purpose. Users use this information at their own risk and responsibility.

About us

RIMEDO Labs specializes in providing high quality consulting, implementation and R&D services in the field of Open RAN, 5G and 6G. We are a spin-off from the Poznan University of Technology, Poland from the Institute of Radiocommunications.

Our services in the Open RAN area include:

- xApp and rApp development for the RAN Intelligent Controller;
- Pre-recorded an Live technical courses delivery;
- Live webinars;
- Dedicated simulations and algorithm design;
- Whitepapers and technical articles delivery

Company details

RIMEDO sp. z o. o.
ul. Polanka 3
61-131 Poznań
Poland, EU
VAT ID: PL7822883638

info@rimedolabs.com
+48 (61) 665 38 17

www.rimedolabs.com

All things wireless