O-RAN.WG1.CCIN-R004-v01.00

*Technical Report*

**O-RAN Work Group 1 (Use Cases and Overall Architecture) Communication and Computing Integrated Networks**

Copyright © 2024 by the O-RAN ALLIANCE e.V.

The copying or incorporation into any other work of part or all of the material available in this document in any form without the prior written permission of O-RAN ALLIANCE e.V. is prohibited, save that you may print or download extracts of the material of this document for your personal use, or copy the material of this document for the purpose of sending to individual third parties for their information provided that you acknowledge O-RAN ALLIANCE as the source of the material and that you inform the third party that these conditions apply to them and that they must comply with them.

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

Contents

[Foreword 4](#_TOC_250088)

[Modal verbs terminology 4](#_TOC_250087)

1. [Scope 5](#_TOC_250086)
2. [References 5](#_TOC_250085)
   1. [Normative references 5](#_TOC_250084)
   2. [Informative references 5](#_TOC_250083)
3. [Definition of terms, symbols and abbreviations 7](#_TOC_250082)
   1. [Terms 7](#_TOC_250081)
   2. [Symbols 7](#_TOC_250080)
   3. [Abbreviations 7](#_TOC_250079)
4. [General Concept 7](#_TOC_250078)
   1. [The convergence of communication and computing in current industry landscape 7](#_TOC_250077)
      1. [Network Function Virtualization (NFV) in ETSI 8](#_TOC_250076)
      2. [Multi-Access Edge Computing (MEC) in ETSI 8](#_TOC_250075)
      3. [Edge Computing in 3GPP 8](#_TOC_250074)
      4. [Coordination of computing and networking in IMT-2020 and beyond in ITU-T 9](#_TOC_250073)
      5. [Computing aware traffic steering in IETF 9](#_TOC_250072)
   2. [Communication capabilities supported by O-RAN 10](#_TOC_250071)
   3. [Computing capabilities in O-RAN 10](#_TOC_250070)
   4. [The convergence of communication and computing in O-RAN 11](#_TOC_250069)
5. [Scenarios and Use Cases 11](#_TOC_250068)
   1. [Use Case 1: Multi aspect handover optimization 11](#_TOC_250067)
      1. [Background Information 11](#_TOC_250066)
      2. [Motivation 12](#_TOC_250065)
      3. [Proposed Solution 13](#_TOC_250064)
      4. [Benefits of O-RAN Architecture 14](#_TOC_250063)
   2. [Use Case 2: Radio Edge Computing Resource Information Exposure 14](#_TOC_250062)
      1. [Background Information 14](#_TOC_250061)
      2. [Motivation 14](#_TOC_250060)
      3. [Proposed Solution 14](#_TOC_250059)
      4. [Benefits of O-RAN Architecture 16](#_TOC_250058)
   3. [Use Case 3: O-RAN and Mobile Devices collaborative AI/ML inference 16](#_TOC_250057)
      1. [Background Information 16](#_TOC_250056)
      2. [Motivation 18](#_TOC_250055)
      3. [Proposed Solution #1 19](#_TOC_250054)
      4. [Proposed Solution #2 20](#_TOC_250053)
      5. [Benefits of O-RAN Architecture 21](#_TOC_250052)
   4. [Use Case 4: XR rendering collaboration 22](#_TOC_250051)
      1. [Background Information 22](#_TOC_250050)
      2. [Motivation 23](#_TOC_250049)
      3. [Proposed Solution 24](#_TOC_250048)
      4. [Benefits of O-RAN Architecture 25](#_TOC_250047)
   5. [Use Case 5: Computing Resource Scheduling in V2X Communications 26](#_TOC_250046)
      1. [Background Information 26](#_TOC_250045)
      2. [Motivation 27](#_TOC_250044)
      3. [Proposed Solution 27](#_TOC_250043)
      4. [Benefits of O-RAN Architecture 28](#_TOC_250042)
   6. [Use case 6: Exposure of performance information to assist computing resource scheduling 28](#_TOC_250041)
      1. [Background Information 28](#_TOC_250040)
      2. [Motivation 28](#_TOC_250039)
      3. [Proposed Solution 29](#_TOC_250038)
         1. [Performance information reporting triggers redeployment of a computing task 29](#_TOC_250037)
         2. [Performance information reporting triggers to scale a computing task 33](#_TOC_250036)
      4. [Benefits of O-RAN Architecture 36](#_TOC_250035)
6. [Key Issues 36](#_TOC_250034)
   1. [Key Issue #1: Joint Communication and Computing Optimization 36](#_TOC_250033)
   2. [Key Issue #2: O-RAN computing service provision 36](#_TOC_250032)
   3. [Key Issue #3: O-RAN integrated communication and computing service provision 37](#_TOC_250031)
   4. [Key Issue #4: Edge application orchestration in O-RAN 37](#_TOC_250030)
7. [Potential Solutions 37](#_TOC_250029)
   1. [Solution #1: O1 configurations for multi-aspect handover optimization 37](#_TOC_250028)
      1. [Key Issue Mapping 37](#_TOC_250027)
      2. [Description 37](#_TOC_250026)
      3. [Procedures 38](#_TOC_250025)
      4. [Impacts on entities, services and interfaces 39](#_TOC_250024)
   2. [Solution #2: Computing information exposure via Y1 interface 40](#_TOC_250023)
      1. [Key Issue Mapping 40](#_TOC_250022)
      2. [Description 40](#_TOC_250021)
      3. [Procedures 40](#_TOC_250020)
      4. [Impacts on entities, services and interfaces 41](#_TOC_250019)
   3. [Solution #3: Computing Service Provisioning 42](#_TOC_250018)
      1. [Key Issue Mapping 42](#_TOC_250017)
      2. [Description 42](#_TOC_250016)
      3. [Procedures 42](#_TOC_250015)
         1. [Computing service registration and discovery 42](#_TOC_250014)
         2. [Computing service consumption 44](#_TOC_250013)
      4. [Impacts on entities, services and interfaces 45](#_TOC_250012)
   4. [Solution #4: Integrated communication & computing service provisioning via SMO 45](#_TOC_250011)
      1. [Key Issue Mapping 45](#_TOC_250010)
      2. [Description 46](#_TOC_250009)
      3. [Procedures 46](#_TOC_250008)
      4. [Impacts on entities, services and interfaces 47](#_TOC_250007)
8. [Analysis on the Potential Architecture Impacts 48](#_TOC_250006)
9. [Deployment Scenarios 49](#_TOC_250005)
   1. [Scenario A: Local O-Cloud for private networks 49](#_TOC_250004)
   2. [Scenario B: Device Edge Cloud and local O-Cloud for private networks 49](#_TOC_250003)
   3. [Scenario C: O-Cloud coordination with global orchestrator 50](#_TOC_250002)
10. [Potential Requirements 51](#_TOC_250001)
11. [Conclusion and Recommendations 51](#_TOC_250000)

Annex: Change history/Change request (history) 52

# Foreword

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

The content of the present document is subject to continuing work within O-RAN and may change following formal O- RAN approval. Should the O-RAN Alliance modify the contents of the present document, it will be re-released by O- RAN with an identifying change of version date and an increase in version number as follows:

version xx.yy.zz where:

xx: the first digit-group is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc. (the initial approved document will have xx=01). Always 2 digits with leading zero if needed.

yy: the second digit-group is incremented when editorial only changes have been incorporated in the document.

Always 2 digits with leading zero if needed.

zz: the third digit-group included only in working versions of the document indicating incremental changes during the editing process. External versions never include the third digit-group. Always 2 digits with leading zero if needed.

# Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the O-RAN Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in O-RAN deliverables except when used in direct citation.

# Scope

The present document provides a technical report for integrated communication and computing in O-RAN. The use cases are studied and potential requirements are identified.

# References

## Normative references

Not applicable.

## Informative references

References are either specific (identified by date of publication and/or edition number or version number) or

non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, O-RAN cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[i.2] Chih-Lin I, Guanding Yu, Shuangfeng Han, Geoffrey Ye Li, "Green and Software-defined Wireless Networks: From Theory to Practice", Cambridge University Press, 2019.

[i.3] ETSI White Paper No. 54，Evolving NFV towards the next decade, May, 2023.

[i.4] ETSI GS MEC 003: "Multi-access Edge Computing (MEC); Framework and Reference Architecture".

[i.5] ITU-T Y.IMT2020-CNC-req, Recommendation, Requirements of coordination of computing and networking for IMT-2020 and beyond.

[i.6] ITU-T Y.IMT2020-QoS-CNC-req, Recommendation, QoS assurance-related requirements and framework for coordination of computing and networking supported by IMT-2020 and beyond.

[i.7] O-RAN TS: "O-RAN Architecture Description".

[i.8] 3GPP TS 23.558: "Architecture for enabling Edge Applications".

[i.9] 3GPP TS 23.501: "System architecture for the 5G System (5GS); Stage 2".

[i.10] 3GPP TS 28.538: "Management and orchestration; Edge Computing Management (ECM)". [i.11] Draft-ietf-cats-usecases-requirements-00, Computing-Aware Traffic Steering (CATS) Problem

Statement, Use Cases, and Requirements, July 2023. [i.12] 3GPP TR 26.928: "Extended Reality (XR) in 5G"

[i.13] 3GPP TR 38.885: "Study on NR Vehicle-to-Everything (V2X)" [i.14] O-RAN TS: "Use Cases Detailed Specification".

[i.15] O-RAN TS: "Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN".

[i.16] ITU-R, “Framework and overall objectives of the future development of imt for 2030 and beyond,” Recommendation ITU, 2023.

[i.17] X. Zhou et al., “A surface defect detection framework for glass bottle bottom using visual attention model and wavelet transform,” IEEE Trans. Ind. Informat., vol. 16, no. 4, pp. 2189–2201, Apr.

2020.

[i.18] J. Luo, Z. Yang, S. Li, and Y. Wu, “FPCB surface defect detection: A decoupled two-stage object detection framework,” IEEE Trans. Instrum. Meas., vol. 70, pp. 1–11, 2021.

[i.19] 3GPP TR 22.874: “Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS ”,” Dec. 2021.

[i.20] N. Li, X. Li, Y. Yan, Q. Sun, Y. Han and K. Cheng, "Joint Communication and Computing Resource Optimization for Collaborative AI Inference in Mobile Networks," 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), Hong Kong, Hong Kong, 2023, pp. 1-5, doi: 10.1109/VTC2023-Fall60731.2023.10333702.

[i.21] Gapeyenko M, Petrov V, Paris S, et al. Standardization of Extended Reality (XR) over 5G and 5G- Advanced 3GPP New Radio[J]. IEEE Network, 2023, 37(4): 22-28.

[i.22] Huang Y, Li N, Sun Q, et al. Communication and Computing Integrated RAN: A New Paradigm Shift for Mobile Network[J]. IEEE Network, 2024.

[i.23] 3GPP TR 38.838: "Study on XR (Extended Reality) Evaluations for NR (Release 17)". [i.24] 3GPP TR 22.261: "Service requirements for the 5G system; Stage 1".

[i.25] 3GPP TR 26.928: "Extended Reality (XR) in 5G".

[i.26] 3GPP TR 23.748: "Study on enhancement of support for Edge Computing in 5G Core network (5GC)".

[i.27] 3GPP TR 23.758: "Study on application architecture for enabling Edge Applications". [i.28] 3GPP TS 23.558: "Architecture for enabling Edge Applications".

[i.29] https:/[/www.synopsys.com/automotive/autonomous-driving-levels.html.](http://www.synopsys.com/automotive/autonomous-driving-levels.html)

[i.30] 3GPP TS 22.186: “Enhancement of 3GPP support for V2X scenarios”, Mar. 2022. [i.31] O-RAN TS: "Non-RT RIC Architecture".

[i.32] O-RAN TR: "SMO Intents-driven Management". [i.33] O-RAN TR: "Decoupled SMO Architecture".

[i.34] O-RAN TS: "Operations and Maintenance Architecture".

[i.35] ETSI GS NFV-IFA 011. Management and Orchestration; VNF Descriptor and Packaging Specification.

[i.36] ETSI GS NFV-IFA 008. Management and Orchestration; Ve-Vnfm reference point -Interface and Information Model Specification.

[i.37] O-RAN TS: "Y1 interface: Type Definitions".

[i.38] O-RAN TS: "Y1 interface: General Aspects and Principles".

[i.39] O-RAN TS: "O-Cloud Notification API Specification for Event Consumers".

# Definition of terms, symbols and abbreviations

## Terms

For the purposes of the present document, the following terms apply:

## Symbols

For the purposes of the present document, the following symbols apply:

## Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [i.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 [i.1].

CCIN Communication and Computing Integrated Networks MEC Multi-Access Edge Computing

NFV Network Function Virtualization

Near-RT RIC Near-Real-Time RAN Intelligent Controller O-Cloud O-RAN Cloud

O-CU-CP O-RAN Central Unit – Control Plane. O-CU-UP O-RAN Central Unit – User Plane

O-DU O-RAN Distributed Unit

XR Extended Reality

# General Concept

## The convergence of communication and computing in current industry landscape

The past decades witnessed the convergence of communication and computing trends in mobile network. Great efforts were made and good momentum were achieved. C-RAN featured by centralization, cooperative, clean and cloud were first proposed in 2009, which depicts the vision of software based virtualized RAN [i.1]. The concept of Network Function Virtualization (NFV) was initiated from 2012. It introduced the information technology (IT) to enable the network transformation from the traditional physical hardware network into a new software based virtualized network [i.2]. This further steers up the pace for the convergence of communication and computing in RAN. Multiple-access edge computing (MEC), which enables compute-intensive applications for mobile devices and access to cloud computing service at the edge of communication networks, had attracted great attention [i.3]. The virtualized/cloudification RAN infrastructure can server as perfect edge computing platform. Thanks to the well- developed IT tools/frameworks, such as the containers, kubernets, the general-purpose hardware infrastructure could be easily and dynamically shared by the RAN functionality and the edge applications, providing the converged capabilities of mobile communication and computing. Edge computing was also incorporated into 5G system in 3GPP since Release

15. However, the communication and computing services offered by the existing design, such as NFV and MEC, were designed separately and developed independently. To fulfill the gap, the coordination of computing and networking is proposed in the context of IMT-2020 and beyond, and provides network enhancements based on the functional

architecture of IMT-2020, to achieve the computing and network joint optimization based on the awareness, control and management over computing resources [i.4] [i.5]. Besides the above, IETF initiated a new working group on computing aware traffic steering to study how the network edge can steer traffic between clients of a service and sites offering the services, considering the integrated network and compute resource status [i.10]. This chapter provides an overview of the related efforts in various standard developing organizations.

### Network Function Virtualization (NFV) in ETSI

Network function virtualization is the one of the key pillars that accelerating the convergence of communication and computing in mobile networks. By adopting the cutting-edge IT technologies, the network functionalities are realized based on the general-purpose servers, which can be flexibly deployed on distributed cloud infrastructure. This revolutionized the way mobile communication networks are developed, deployed and operated. Series of NFV specifications are released by ETSI ISG NFV, from the concept, architecture, a set of interoperable implementation solutions, operational matters. Additional support for containerized VNFs and more cloud native deployment and automation are also specified in recent years. Regarding the further evolvement, the requirements from telecom service providers, including: 1) unified network management; 2) user of latest cloud-native IT, automation and Artificial Intelligence (AI) open source software; 3) multi-vendor interoperability and migration among different nodes, are being considered as the key drivers [i.3].

It is worth mentioning that the NFV work is more focusing on the network virtualization to offer the RAN services in IT manner with automation. It lays a good reference framework for RAN virtualization. But the dedicated requirements for RAN virtualization are not fully considered. This is further studied and specified in O-RAN. How computing services would be extended and how to provide the integrated communication and computing services based on the telecom cloud infrastructure for the both computing and communication intensive emerging services, such as metaverse, XR and AI, are not investigated.

### Multi-Access Edge Computing (MEC) in ETSI

ETSI Industry Specifications Group (ISG) MEC defines the specifications relevant to the edge computing, where the specification is access agnostic and is not confined to mobile communications. ETSI ISG MEC specifies an open and standardized IT service environment that enables application to be deployed at the network edge close to edge user [i.4]. A common edge application enablement framework for multiple network technologies is developed with MEC service APIs. More specifically, network related service APIs are specified to provide added value to the edge application enabling the network aware service optimization. For example, radio network information (3GPP access), WLAN information, UE identity service APIs, V2X information service, etc., are specified. The MEC service APIs mostly focus on the how the MEC applications can be enabled and the orchestration and management of the MEC platform and the application. How the network would interact with the MEC functionalities and how the information required for network related service APIs can be obtained from the network is not fully investigated. To fill the gap, Y1 interfaces is introduced in O-RAN so that MEC can obtain the RAN analytics information via the Y1 interface [i.7]. However, the joint consideration of the edge service continuity and the communication service continuity, the unified orchestration and management of the mobile network and edge computing required for the operators are not yet investigated.

In summary, MEC as another key pillar of the convergence of communication and computing, are now separately designed, deployed and operated on top of the mobile network. Thus, the dynamic coordination of the mobile network and application processing is not supported to ensure the end to end service QoS.

### Edge Computing in 3GPP

3GPP edge computing work defines the 5G access specific interactions and interfaces as a complement part of the ETSI MEC work. The overall framework generally consists of five layers, i.e., application layer, edge enabler layer, edge hosting environment, edge management layer and 3GPP transport layer [i.8]. .3GPP develops specifications for the edge enabler layer, 3GPP transport layer, edge management layer, and some part of application layer (relevant to enabling of edge capabilities to the applications).

In the transport layer, 5G supports to select/reselect user plane functions to steer/route traffic locally and connect to the local area data network in selected regions. It also supports Edge Application Server Discovery Function (EADSF), which enables the discovery of the IP address(es) of the EASs for the edge application [i.9].

Regarding the management layer, different business roles involved in edge computing and the mobile network, and their relationship are also discussed along with the edge computing management specification [i.10]. Generally, PLMN operator is responsible for the 3GPP transport layer and deployment of the 5G network functions, edge computing service provider (ECSP) is responsible for the deployment of edge data network that contain edge enabler client (EEC) in the UE and edge enabler server (EES), and application service provider (ASP) is responsible for the creation of edge application servers (EASs) and application clients (ACs). The management system framework for edge computing defined in 3GPP consists of the PLMN management system and the ECSP management system. The two systems can interact with ETSI Network Function Virtualization Management and Orchestration (MANO) to perform lifecycle management functions.

The edge enabler layer provides edge capability related Application Programming Interfaces (APIs) for the application developers [i.8]. With the well-defined services and APIs, the application developers can locate and connect to the most suitable edge application server, and to explore the service optimization leveraging the potential network information provided by underlying 3GPP network.

Similar to the MEC work in ETSI, the joint optimization of network and computing resources are not yet considered to guarantee the end-to-end service quality that depends on both the computing and communication process.

### Coordination of computing and networking in IMT-2020 and beyond in ITU-T

The requirements, framework and QoS assurance related requirement and framework for the coordination of computing and networking in IMT-2020 and beyond are studied in ITU-T [i.5] [i.6]. The coordination of computing and networking is defined as the coordination of utilization, control and management of computing, storage and networking resources. It provides the capability to jointly schedule the computing, storage and networking resources, and is expected to enhance the QoS support for IMT-2020 networks and beyond. The coordination of computing and networking is envisioned for the resource layer, control layer, service layer, and management and orchestration layer.

The general requirements identified in the draft recommendations [i.5] includes: awareness of networking, computing and storage resources; joint scheduling of resources; unified management of resources; AI/ML integration; unified standard interfaces for resources; support resource transaction services so that the resource users can order computing services; energy saving and security and privacy related requirements.

The high level QoS assurance mechanisms investigated for the coordination of computing and networking in IMT-2020 and beyond in [i.6] includes:

* Perception of computing and networking status
* Perception of service requirements
* Mapping among service, computing and network requirements
* Joint scheduling of computing and networking
* Unified management and orchestration of computing and networking

### Computing aware traffic steering in IETF

IETF initiated computing aware traffic steering working group in Mar. 2023. The main motivation is to study how to integrate both network and compute conditions in the optimization function that informs the steering applied by the network edge nodes with the desire to operate that function on nodes within the service provider's network, logically separated from service operation. The main motivation is the performance experienced by clients will depend on both

network metrics such as bandwidth and latency, and compute metrics such as processing, storage capabilities, and capacity. For some services, especially, the XR and autonomous vehicles assisted by the network, it is desired to optimize network edge node traffic steering based on the awareness of both the communication and computing related metrics.

The use cases and requirements for the computing aware traffic steering are studied in [i.11]. Four typical use cases, namely, computing-aware AR/VR, computing-aware intelligent transportation, computing-aware digital twin and computing-aware SD-WAN are discussed considering a distributed computing scenario. The requirements to support computing aware traffic steering are also identified:

* Support dynamic and effective selection among multiple service instances
* Support agreement on metric representation of both the compute capability related metrics and network metrics
* Support flexible use of network and computing metrics for service instance selection
* Support session and service continuity
* Preserve communication confidentiality.

## Communication capabilities supported by O-RAN

O-RAN is committed to evolving radio access networks with its core principles being intelligence and openness. The overall O-RAN architecture is specified in [i.7], where the key elements including the service management and orchestration framework, O-RAN network functions and O-Cloud are described. In general, the O-RAN architecture supports the 4G and 5G RAN communication services. With the hierarchical RAN intelligent controllers, namely Non- RIC and Near-RT RIC introduced in O-RAN, various AI/ML empowered RAN optimization use cases are explored in [i.14], offering improved RAN performance, user experiences, network resource utilization efficiency and energy efficiency.

## Computing capabilities in O-RAN

Diversified 5G services call for new level of agility in RAN deployment and scaling. RAN cloudification emerges as a promising solution and offers new innovation paradigm. O-RAN architecture adopts the RAN cloudification design. O- Cloud is defined as a cloud computing platform comprising a collection of physical infrastructure nodes that meet O- RAN requirements to host the relevant O-RAN NFs (i.e., Near-RT RIC, O-CU-CP, O-CU-UP, and O-DU) [i.7].

O-Cloud platform hardware includes compute, network and storage components and may also include hardware accelerators for improving the computing efficiency by offloading the most computation-intensive or timing-intensive tasks in the radio protocol stack. O-Cloud platform software consists of cloud platform runtime, cloud platform management and virtual machine (VM)/container management and orchestration. The cloud platform runtime includes the hypervisor/container runtime, accelerator driver, network driver and precision time protocol to host the RAN functions. The cloud platform management is responsible for the life cycle management, high availability, fault management and configuration of the cloud infrastructure. The VM/container management and orchestration is typically implemented with OpenStack for VM deployments and Kubernetes (K8s) for container deployment. The cloud architecture and deployments scenarios for O-RAN virtualized RAN are described in [i.15].

The cloudification in O-RAN greatly improves the RAN deployment flexibility, and transforms O-RAN into a native and cost-effective edge computing platform. Thus, edge application services can be deployed on the existing O-Cloud infrastructure along with the O-RAN functions. This lays the foundation of “beyond communication” service provision, offering the opportunities for RAN to provide the integrated communication and computing services.

## The convergence of communication and computing in O- RAN

The CCIN is a network which could provide not only traditional RAN services such as connection and mobility control but also applications services by leveraging the communication and computing resources in an integrated manner. The goal of CCIN is to provide the beyond-connection services and bring new values to operators. A CCIN network has the following features.

1. Sharable resources. Resources include not only the communication resources such as the radio bearers, the TTL, the frequency etc. but also those of computing resources including various hardware from CPU, GPU to switches, storage and so on. Thanks to the cloud technology, such resources could be scheduled and allocated for different usage, i.e. communication and computing on demand. For example, in the daytime, the computing power of the CPU should be used to process the radio stack for 5G coverage. While at night when the traffic volume drops, the CPU is then freed from the 5G computing and changed for AI model training.
2. Holistic orchestration of communication and computing resources. The resources of CCIN should be scheduled and orchestrated in a holistic way. The CCIN itself should be capable of scheduling and allocating the communication and computing resources for different workloads. In addition, the CCIN should also be able to expose such resources and status to external in collaboration with the other network functions, e.g. core networks. A global orchestrator should be aware of the resources, orchestrate them together with the functions and services in a holistic way to optimize the system performance, efficiency and/or user experience.
3. Integrated service provisioning. The CCIN could provide services externally and the services could be of different types. It could be in the very basic format of computing capability in the unit of Floating-point Operation Per Second (FLOP). It could be further provided in a more abstract yet user-sensible way. For example, the services could be the capability of training certain type of AI models with certain number of parameters. Eventually, such service provisioning would bring the beyond-connection value to operators.

NOTE: If and how O-RAN addresses this topic of integrated service provisioning is for further study.

In essence, CCIN features the integration of communication and computing. There could be several levels of integration. On the first level, the computing functions simply co-locate with the communication functions on the same hardware platform, i.e. the COTS servers. There is little interaction in between except that the user-plane local break- out by means of User-plane Function (UPF) could be implemented for local processing. On the second level, interaction between the communication and application services is enabled to provide RAN and application mutually aware optimization. On this level, the interaction is mainly via an established interface or Application Program Interface (API). For the third and the last level of integration in CCIN, the communication and computing relate to each other in a symbiotic way, i.e, they are jointly scheduled in a real-time way, managed and orchestrated in a holistic manner. With the orchestrator, the computing resources could be scaled in and out in a much more dynamic way with smaller granularity as the workload changes with the applications than the second level.

# Scenarios and Use Cases

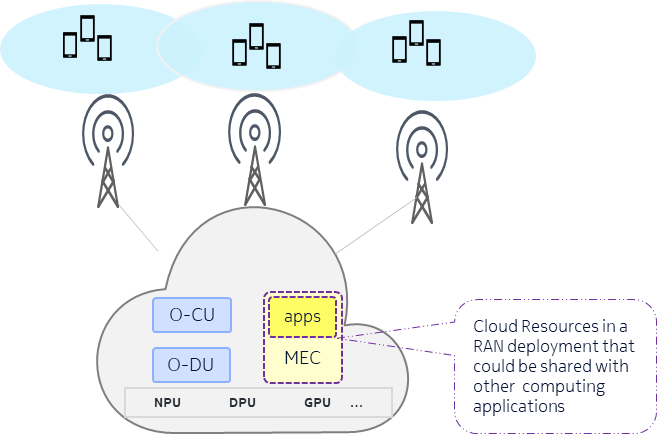
## Use Case 1: Multi aspect handover optimization

### Background Information

Communication and Computing Integrated Networks enables the collaboration of communication network and computing. It is an approach of softening the boundaries between the cloud computing and networking, such that their resources can be utilized in the most efficient way. This allows the applications to leverage on such convergence of compute and network resources. Cloud based RAN solutions can play an important role in the Communication and

Computing Integrated Networks with its potential to share some of its vacant computing resources as the edge computing platform. New age applications based on XR (see 3GPP TR 26.928 [i.12], clause 4.1.1), V2X (see 3GPP TR 38.885 [i.13]) can benefit from Communication and Computing Integrated Networks but they require application mobility in addition to posing stringent requirements on both communication and computing. New network architecture for Communication and Computing Integrated Networks needs to address these application requirements.

The types of computing resources required for different applications can be diverse based on the nature of the applications. However, the computing resources available in different entities might be of different computing resource type than what is required for specific applications. It is also possible that capabilities of a computing resource type required for specific applications is different even if an entity supports required computing resource type. Computing resource types referred here could be but not limited to the support for Graphical Processing Unit, Neural Processing Unit, Deep Processing Unit, etc. An example illustration of the cloud based joint RAN + Computing entity used in Communication and Computing Integrated Networks (CCIN) is given below in the Figure 5.1.1-1.



##### Figure 5.1.1-1: Example illustration of cloud based joint RAN + Computing entity.

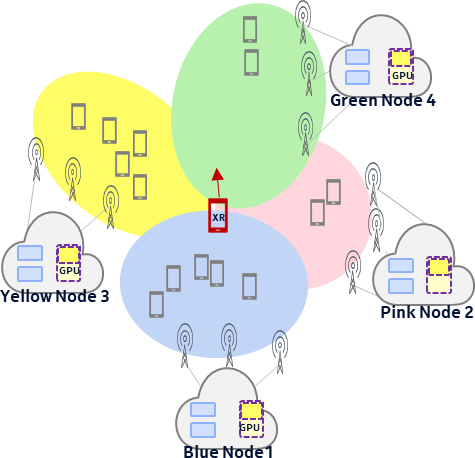
### Motivation

In a Communication and Computing Integrated Networks, the idea is to use the vacant computing resources from the cloud-based RAN (e.g., if the network load is low), to perform some edge computing related tasks so that cloud resources are used efficiently. In particular, leveraging the proximity of the cloud-based RAN entities to the end users/devices is a key factor to address low latency requirements of the tasks/applications. But in order to take advantage of this opportunity the following needs to be considered:

* If the vacant resources match the task requirements in terms of the type of the computing resource and related capabilities.
* If the amount of vacant resources of required type/capability suffices the task requirements
* If the nodes are within the proximity required to support the continuity of the service and service latency requirements of the applications.

In the use case of UE handover especially in the scenario where the UE needs to run applications with high computing and low latency requirements such as XR based applications the above stated challenges are more prominent. Thus, in such use case it is desirable that system considers the aspects of computing capability and availability of cloud-based RAN entities while supporting the mobility of UE and applications in it.

In particular, it is important to create the awareness in a cloud-based RAN entity of the information regarding the computing capability and availability of other cloud-based RAN entities in the proximity and use these multiple aspects as part of the UE handover decisions. Proximity here means that coverage area of the cells served by source cloud-based RAN entity has some overlap with the coverage area of the cells served by target cloud-based RAN entity.



##### Figure 5.1.2-1: Example illustration of multi aspect handover

### Proposed Solution

In order to address aforementioned use case and related challenges it is proposed extend the Communication and Computing Integrated Networks framework. Such extensions comprise introduction of proximal computing table which stores the information on the compute capabilities of cloud-based RAN entities and its usage in UE handover decisions. This helps the handover of the UE to a target cloud-based RAN entity in the proximity with computing resource types of relevant capabilities and availability required for the applications in that UE. The proximal computing table can be of following types:

1. Proximal Computing Capability Table which is a neighbor relation table indicative of computing resource types and/or capability of compute resource types of the target cloud-based RAN entities in the proximity of the source cloud- based Radio entity. Examples for resource types referred here are GPU, NPU, DPU, AI accelerator, etc. Examples for the capability of resource types referred here are supported types of AI algorithms, supported number of Ray Tracing Cores, etc.
2. Proximal Computing Availability Table which is a neighbor relation table indicative of current availability of computing resource types of the target cloud-based RAN entities in the proximity of the source cloud-based RAN entity. This may comprise a predicted availability of the computing resource types for predetermined time periods.

Proposed solution consists of rApp, SMO/Non-RT RIC Framework, Near-RT RIC, E2 nodes and O-Cloud. rApp retrieves CM data, PM data, location data of different cloud-based RAN entities and capability of computing resource types, load/availability of computing resource types in different cloud-based RAN entities. rApp derives Proximal Computing Capability Table for each cloud-based RAN entity based on the proximity analysis of different cloud-based RAN entities in the neighborhood and the capability analysis of computing resource types in different cloud-based RAN entities within the proximity. Proximity analysis uses different data including but not limited to PM data and location data. Similarly, rApp derives Proximal Computing Availability table for a cloud-based RAN entity based on the load/availability analysis of the computing resource types in different target entities in the corresponding Proximal Computing Capability table. rApp may rely on AI/ML model-based load/availability pattern analysis of the computing resource types, this includes prediction of available vacant resources at specific time window in a specific day e.g., during which time window it is expected to have the most vacant resources, with which confidence that can be determined and exposed etc.

rApp creates different O1 based configurations including policies for multi-aspect handover optimization based on the Proximal Computing Capability Table and Availability Table at one or more near-RT RIC instances. Near-RT RIC will then use E2 control or E2 policy to ensure the UE handover to most optimum target cloud-based Radio entity.

### Benefits of O-RAN Architecture

O-RAN architecture allows rApp based use cases to build intelligence based on both traditional RAN network data available over O1 interface as well as cloud-based computing capabilitiy data available over O2 interface. In addition, E2 control or E2 policies facilitates the finer control of UE level handover decisions in E2 nodes. In summary, O-RAN architecture provides necessary ingredients to facilitate multi aspect handover optimization.

## Use Case 2: Radio Edge Computing Resource Information Exposure

### Background Information

AI/ML processing for network intelligence and automation evolution, as well as new 5G edge applications supporting e.g., immersive Virtual Reality (VR), autonomous driving, and intelligent internet of things (IoT), etc. put an unprecedented demand on the edge computing resources. The O-Cloud architecture of O-RAN and the spatial and temporal tidal effects of RAN communication services offer great opportunities for sharing underutilized computing resources in the RAN infrastructure. This helps to improve the resource utilization efficiency and reduce operators’ investment on the edge computing infrastructure.

### Motivation

The main objective is to provide radio edge computing resource information exposure service based on O-RAN architecture and its open interface. In O-RAN, SMO is a great candidate to provide radio edge computing resource information exposure service to the service consumers for perceiving unused computing resources in the O-Cloud infrastructure. The service consumers might be other orchestrators, e.g., central cloud orchestrator, edge cloud orchestrator, radio edge cloud orchestrator, device edge cloud orchestrator etc. The orchestrators could request computing resource information for multi-cloud collaboration, such as task migration between multiple clouds or edge intelligence. The service consumers could also be other 3rd applications, e.g., vertical customers, VR consumers etc., which could request computing resource information for application deployment using shared computing resources in a secured way.

### Proposed Solution

Entities/resources involved in the use case:

1. SMO:
   1. Support collect and analysis computing related resources from O-Cloud.
   2. Support provide radio edge computing resource information service, send the radio computing information for managed O-Cloud related computing resources to the service consumer, e.g., Orchestrator/Application Server.
2. O-Cloud:
   1. Support computing related inventory and updated resources report with required granularity to SMO over O2 interface.
3. Radio Edge Computing Resource Information Exposure Service Consumer:
   1. Request/subscribe radio edge computing resource information from SMO.
   2. The service consumer might be Orchestrator/Application Server.

The whole procedure is presented in Table 5.2.3-1, illustrated by Figure. 5.2.3-1.

##### Table 5.2.3-1: Radio Edge Computing Resource Information Exposure

|  |  |
| --- | --- |
| Use Case Step | Description of the Step |
| Goal | Radio Edge Computing Resource Information Exposure |
| Actors and Roles | SMO, O-Cloud, orchestrator/application server |
| Assumptions | * All relevant functions and components are instantiated. * O2 interface connectivity is established with SMO. |
| Pre-conditions | - SMO and O-Cloud are instantiated with O2 interface connectivity being established between them.  Editor’s Note: Security related procedure is FFS. |
| Begins when | Orchestrators or application servers want to request/subscribe Radio Edge Computing  Resource information. |
| Step 1 (M) | Orchestrators or application servers, e.g., RAN service computing center, send Radio Edge Computing Resource information request to SMO for initial or updated computing resource  information. |
| Step 2 (M) | SMO requests or subscribes the measurement data from O-Cloud, and receives computing  related resources information from O-Cloud through response or periodic notifications. |
| Step 3 (O) | Alternatively, SMO could use AI/ML or other methods to perform computing related resource  data statistics and analysis. |
| Step 4 (M) | SMO send Radio Edge Computing Resource information response, including the RAN side  computing related information managed by SMO. |
| Ends when | Orchestrators or application servers get the response. |
| Post Conditions | N/A |

@startuml Skin rose

skinparam ParticipantPadding 5

skinparam BoxPadding 10

skinparam defaultFontSize 12

Box “Service Consumer” #lightcyan

Participant “Orchestrator/Application Server” as app End box

Box “Service Management and Orchestration” #gold

Participant “SMO” as smo End box

Box “O-Cloud Platform” #lightseagreen

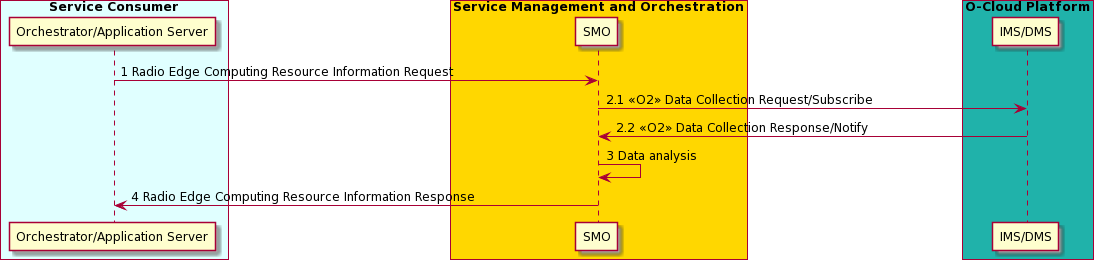
Participant ims as “IMS/DMS”

End box

app -> smo : 1 Radio Edge Computing Resource Information Request smo -> ims: 2.1 <<O2>> Data Collection Request/Subscribe

ims -> smo: 2.2 <<O2>> Data Collection Response/Notify smo -> smo: 3 Data analysis

smo -> app: 4 Radio Edge Computing Resource Information Response @enduml



##### Figure 5.2.3-1: Radio Edge Computing Resource Information Exposure

Required data:

Computing related resource data are expected to be retrieved by SMO from O-Cloud for radio computing related data statistics, analysis and information exposure, including total, allocated, reserved, and available:

1. CPU number for O-Cloud, cluster, or node
2. vCPU number for O-Cloud, cluster, or node Exposed data:
   1. Radio Edge Computing Resource information exposed by SMO includes but is not limited to,
   2. Total capacity of computing or number of computing resources
   3. Allocated capacity of computing or number of computing resources
   4. Reserved capacity of computing or number of computing resources
   5. Available capacity of computing or number of computing resources
   6. Other computing related information
      1. Computing resource identifier, e.g., Cloud identifier, Node identifier
      2. Computing resource location,
      3. vCPU number, etc.

### Benefits of O-RAN Architecture

The proposed solution supports an effective way to expose the radio edge computing resource information to the service consumer, such as other orchestrators or applications. It requires features provided by the O-RAN architecture to collect the radio computing related data, alternatively supports statistics, analysis, and expose by SMO.

## Use Case 3: O-RAN and Mobile Devices collaborative AI/ML inference

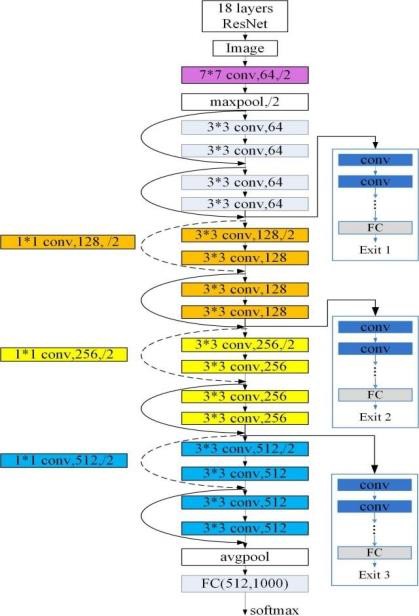
### Background Information

Integrated artificial intelligence (AI) and communication has been identified as crucial elements for IMT-2030 by ITU- R [i.16]. AI/ML-based applications are increasingly demanding in terms of computation, memory, and power consumption. However, mobile devices typically have limited energy, computing capability, and memory resources, making it challenging to run complete AI/ML inference. There is a growing need to offload AI/ML inference from

mobile devices to the central cloud or edge computing platform to support the mobile AI/ML inference application through collaborative computing.

For example, [i.17] and [i.18] give a manufacturing scenario that utilizes AI/ML technologies for analyzing image data in a product defect detection application. The product's image is initially captured by a sensor such as a camera and then it can be transmitted to the edge server equipped with an AI/ML model for identifying and classifying the potential defects in the product by analyzing the image data. Subsequently, the inference result is communicated to the console's actuator (e.g. a classifier), which directs the product to different processes based on the result. The AI/ML model can also be split between the mobile devices and the edge computing platform considering the radio transmission status and the mobile devices’ computing capabilities. In several use cases of split AI/ML operation between AI/ML endpoints studied in 3GPP TR22.874 [i.19], the AI/ML endpoints at the end user side can be mobile devices, robotics, cameras, etc., and the AI/ML endpoints on the network side can be AI/ML edge servers, cloud servers, application servers, etc.

When the performance of the AI/ML inference task decreases (e.g., the inference end-to-end latency increases and/or the inference accuracy decreases), service experience downgrades, or mobile device's computing capability cannot support current AI/ML inference computation, the AI/ML model split point between the AI/ML endpoints and the AI/ML exit point can be dynamically optimized by adjusting the computing workload of the AI/ML endpoints and the communication volume between AI/ML endpoints to ensure service quality. Taking ResNet18 as an example, Figure 5.3.1-1 shows the framework of the AI/ML model [i.20] and Table 5.3.1-1 shows data size of the intermediate outcomes and computation load for the split point inference of the ResNet18 model. The data size of AI/ML model intermediate outcomes need to be transmitted from mobile device to network side, computation load offloaded as well as the accuracy of the various AI/ML exit point are be summarized in the following table.



##### Figure 5.3.1-1: The framework of the multi-branch ResNet-18 model

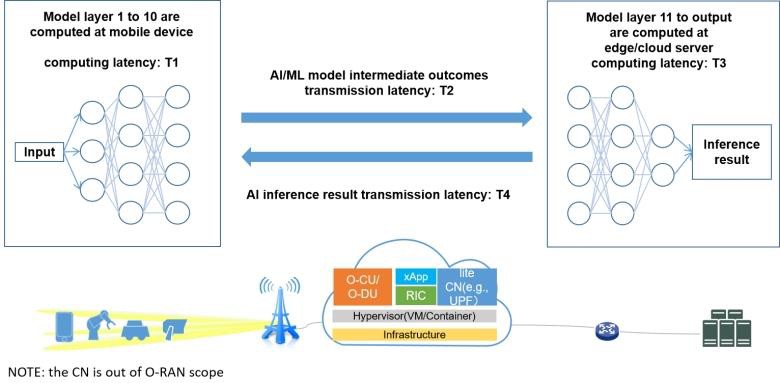
##### Table 5.3.1-1 Data size and computation load of ResNet18

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Layer ID | Type | Data size (bit) | Computation  load (FLOPs) | Accuracy of  exit point |
| 1 | conv1 | 25690112 | 118013952 |  |
|  | maxpool | 6422528 | 7225344 |  |
| 2 | conv2-1 | 6422528 | 115605504 |  |
| 3 | conv2-2 | 6422528 | 115605504 |  |
| 4 | conv2-3 | 6422528 | 115605504 |  |
| 5 (Exit1) | conv2-4 | 6422528 | 115605504 | 63.91% |
| 6 | conv3-1 | 3211264 | 57802752 |  |
| 7 | conv3-2 | 3211264 | 115605504 |  |
| 8 | conv3-3 | 3211264 | 115605504 |  |
| 9 (Exit2) | conv3-4 | 3211264 | 115605504 | 65.36% |
| 10 | conv4-1 | 1605632 | 57802752 |  |
| 11 | conv4-2 | 1605632 | 115605504 |  |
| 12 | conv4-3 | 1605632 | 115605504 |  |
| 13 (Exit3) | conv4-4 | 1605632 | 115605504 | 67.64% |
| 14 | conv5-1 | 802816 | 57802752 |  |
| 15 | conv5-2 | 802816 | 115605504 |  |
| 16 | conv5-3 | 802816 | 115605504 |  |
| 17 | conv5-4 | 802816 | 115605504 |  |
|  | avg pool | 16384 | 1229312 |  |
| 18 (Exit4) | fc | 32000 | 1023000 | 68.96% |

### Motivation

The performance of the application AI/ML inference can be evaluated by two primary metrics: end-to-end application AI/ML inference latency and inference accuracy. The end-to-end latency in collaborative AI inference involving mobile devices and cloud/edge server generally comprises four key components: the inference computation latency on the mobile device, the transmission latency for AI/ML model intermediate outcomes from the mobile devices to the network, the inference computing latency within the edge/cloud server, and the transmission latency for the inference outcomes from the network to the mobile device. Figure 5.3.2-1 illustrates the end-to-end inference latency for collaborative AI inference between mobile devices and edge/cloud server, with the total inference latency represented as the sum of T1, T2, T3, and T4. The end-to-end latency of collaborative AI/ML inference is subject to multiple determinants. Latency T1 is influenced by both the complexity of the AI/ML model deployed on the mobile device and the device's computing capability. It is worth noting that the AI/ML model inference accuracy usually increases with the AI/ML model complexity. Latency T2 depends on the data size of the AI/ML model intermediate outputs, the radio interface and backhaul connection conditions. Latency T3 is contingent upon the complexity of the AI/ML model deployed on the edge/cloud server and the computing capability of the server. Meanwhile, latency T4 hinges on the data size of the inference result, the radio interface and backhaul connection conditions. The above analysis of the impact on latency T1~T4 indicates that the end-to-end latency of collaborative AI inference highly depends on both the communication and computational capabilities of the mobile device and the network. To assure the end-to-end collaborative AI inference performance, the joint communication and computing resources optimization is desired.

The objective of this use case is to investigate how O-RAN architecture could be leveraged to support the collaborative AI/ML inference between the mobile device and the network.



##### Figure 5.3.2-1: The AI/ML collaborative inference between mobile device and edge/cloud server

### Proposed Solution #1

O-RAN could provide both the edge computing capabilities and the communication assurance for the collaborative AI inference scenarios. This also helps to facilitate the joint communication and computing resource optimization to ensure the end-to-end performance.

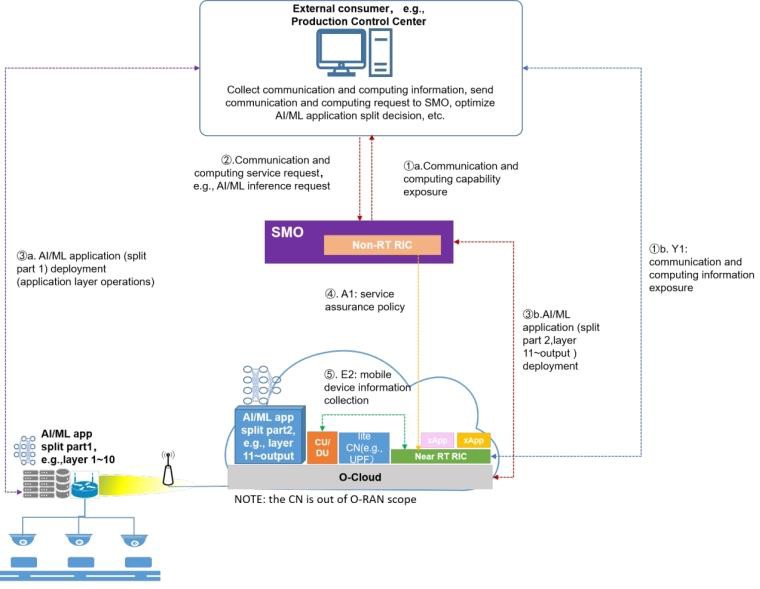
Firstly, in O-RAN system, O-Cloud infrastructure can provide high-quality computing resources for mobile device and network AI/ML collaborative inference. Idle O-Cloud resources can be leveraged to the deploy the AI/ML application. Compared with the collaborative AI inference between mobile device and central cloud server, AI/ML collaborative inference between mobile devices and O-Cloud is beneficial for reducing the transmission latency of AI/ML model intermediate outcomes and inference result. This also helps to potentially reduce the cost required for additional edge/cloud server deployment. To efficiently utilize the O-Cloud computing capabilities for the edge AI/ML applications, SMO could serve as an intermediary interface that translates external communication and computing requests into O-RAN internal operations. In detail, SMO could expose the computing services or integrated communication and computing services offered by O-RAN to the external entities or consumers so that external customers are able to convey their communication requirements and computing requirements. For example, in the AI/ML inference request, requirements such as computing resource requirements for AI/ML applications, end-to-end inference latency and accuracy requirements, service duration, and the number of served mobile devices, etc., could be provided to the SMO. With the service request input, SMO can orchestrate communication and computing resources to fulfill the service requirements. To enable the deployment of the edge applications (e.g, AI/ML application) on the O- Cloud from the external request, a new edge application orchestrator, i.e., RAN Edge Application Orchestrator (REAO), responsible for the edge application life cycle management, can be introduced in SMO or the existing mechanism could be reused. It can further consume the O-Cloud Deployment Management Service (DMS) to execute the deployment operation on O-Cloud.

Secondly, Non-RT RIC and Near-RT RIC can further enable the intelligent service assurance and network performance improvement. These capabilities can be applied to collaborative AI inference, facilitating the joint communication and computing resources optimization. For example, Non-RT RIC may predict the trajectory of mobile devices, then proactively optimize communication resources at the target BS and assist to deploy AI/ML applications in advance at the target O-Cloud, etc. Near-RT RIC is capable to gather detailed information at both the UE level and the cell level. With these data and the policies received from Non-RT RIC, Near-RT RIC can help to assure the communication performance in near real-time for AI/ML inference application.

Furthermore, O-RAN can provide RAN information including communication and computing information to network side AI/ML endpoints (which could be deployed at O-Cloud, edge or cloud servers). For example, Near-RT RIC can expose mobile device throughput, network congestion condition, radio interface latency, available computing resources,

etc via Y1 interface. This information can be used by the application to optimize the AI/ML model split and exit decisions in adaptation to the network connection conditions and computing status (e.g., mobile devices’ computing capabilities, power status, etc.). The optimized decision from application may be further notified to mobile devices and SMO and enforced by them.

Figure 5.3.3-1 shows how O-RAN support the collaborative AI inference application, such as industrial quality inspection, in the production line scenario.



##### Figure 5.3.3-1: The AI/ML collaborative inference between mobile device and O-RAN

### Proposed Solution #2

This solution aims to incorporate comprehensive management of CCIN consumers' AI/ML inference applications, specifically regarding the life-cycle management of collaborative AI/ML inference tasks by SMO.

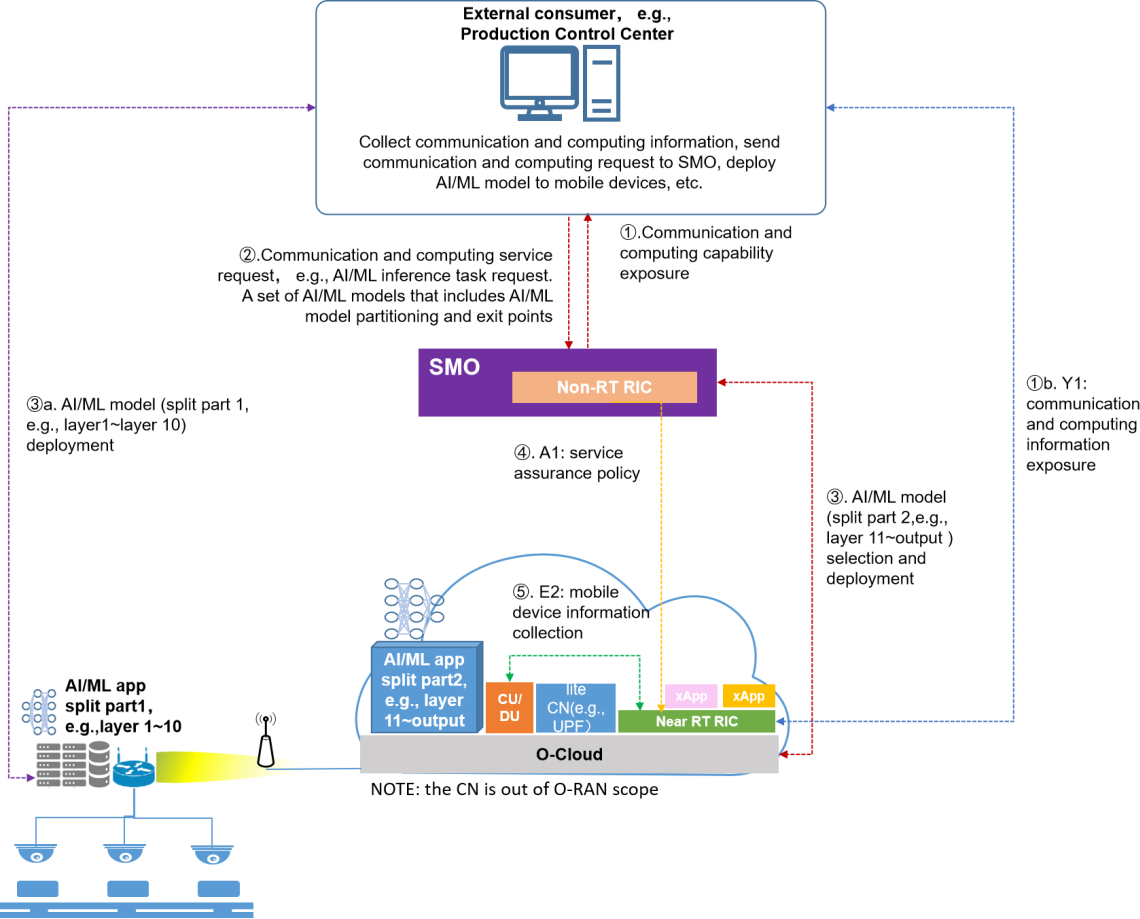
To enable O-RAN to dynamically select an AI/ML model partitioning and exit point that is suitable for the current communication and computing resource condition, and to deploy the selected AI/ML model, CCIN consumers could provide SMO with a set of certain types of AI/ML models at possible partitioning and exit points. This allows the SMO to ascertain the uplink bandwidth and computing resource requirements of the AI/ML models within the set, thereby enabling the selection of partitioning of AI/ML models based on deployment strategies generated by the Non-RT RIC.

Additionally, CCIN consumers could provide performance requirements for the AI/ML collaborative inference task, such as end-to-end latency of AI/ML inference, the area required to be covered, etc. This enables the Non-RT RIC to generate AI/ML model deployment strategies that comply with these performance metrics based on the status of communication and computing resources.

Since mobile device and O-Cloud AI/ML collaborative inference involves not only adjustments to the AI/ML models deployed on the O-Cloud side but also adjustments on the mobile device side, the AI/ML models deployed on mobile device are always intended to compute collaboratively with those on O-Cloud. If the O-Cloud deploys AI/ML inference model from the (n+1)th layer to the output layer, then the mobile device side needs to deploy the AI/ML model for

layers 1 to (n). Consequently, O-RAN needs to communicate the decision on the O-Cloud side AI/ML model selection to CCIN consumers, enabling them to inform the mobile device side about the AI/ML model that need to be deployed.

Figure 5.3.4-1 illustrates an example of O-RAN implementing AI/ML collaborative inference computing task between mobile device and O-Cloud. Herein, CCIN consumers notify the SMO of the AI/ML models (For instance, a set of AI/ML models that categorized with different AI/ML model partitioning and exit points.) and the performance requirements for the mobile device and O-Cloud AI/ML collaborative inference (such as end-to-end latency of AI/ML inference, the area desired to be covered, etc.) through the consumption of CCIN service. The Non-RT RIC, based on current communication and computational resources, generates AI/ML model deployment strategyy and notifies the CCIN consumers. CCIN consumers then complete the deployment of AI/ML model on the mobile device side, while the SMO (for example, REAO within the SMO that deploys edge applications) completes the deployment of AI/ML model on O-Cloud. By monitoring of the AI/ML collaborative computing task, the Non-RT RIC can collect the communication status of the mobile device from the Near-RT RIC and generate AI/ML model adjustment strategy based on the current communication and computing status of the mobile device and the O-Cloud. The AI/ML model adjustment strategies are also communicated to CCIN consumers to update the AI/ML model on the mobile device side.



##### Figure 5.3.4-1: AI/ML inference application selection and deployment performed by O-RAN

### Benefits of O-RAN Architecture

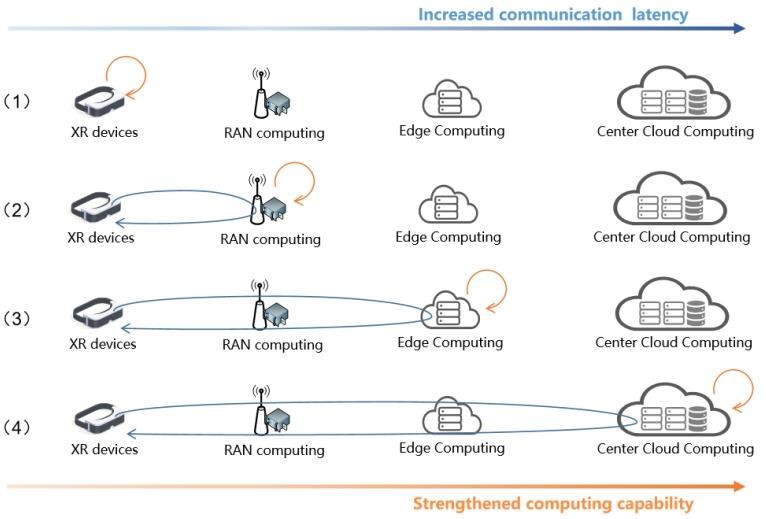
The O-RAN system presents several advantages in the collaborative AI/ML inference scenario. First, the idle computing resources of O-Cloud can be shared with AI/ML inference applications. Furthermore, Non-RT RIC and Near-RT RIC empowers joint communication and computing resources optimization to ensure the collaborative AI/ML inference end-to-end latency and accuracy performance.

## Use Case 4: XR rendering collaboration

### Background Information

As a disruptive technology that brings new interactive experience, eXtended Reality (XR) has attracted considerable attention from content providers, mobile network operators, vertical industry customers, and so on. By design, a unique feature of XR is a new device form factor replacing or complementing handheld smartphones with wearable head- mounted displays or smart glasses. New devices (i.e., glasses or head mounted displays) impose strict requirements on user equipment (UE) power consumption, performance, and heat dissipation. Consequently, XR devices cannot do all the computations (i.e., Figure 5.4.1-1(1)) but require assistance from other nodes [i.21].

XR applications often require high-quality visuals and real-time interactions, which can be demanding in terms of latency and bandwidth requirements. Figure 5.4.1-1 illustrates the collaboration between different types of computing nodes and XR devices, where blue lines represent communication latency and orange lines represent the processing in computing nodes. On the one hand, RAN computing, brings rendering and computing service closer to users, allowing for ultra-low communication latency and reliability assurance for XR applications. On the other hand, RAN computing capability is weaker than cloud computing, which may cause the larger computing delay [i.22]. In order to provide users with immersive experience and ensure real-time interaction, the delay of interaction must be small enough, which puts forward new requirements for both radio network and computing resources.



##### Figure 5.4.1-1: Collaboration between different types of computing nodes and XR devices

3GPP TR 38.838 [i.23] defines the evaluation methodology for XR studies in Release-17, including the modeling of XR applications, simulation scenarios, traffic models, KPIs, and simulation parameters. 3GPP TR 22.261 [i.24] defines the Cloud/Edge/Split Rendering as an additional high data rate and low latency service and provides corresponding requirements on the 5G system. In addition to the conversational service category, Interactive (Streaming), Passive Streaming, Download, and Split Compute/Rendering are identified as new delivery categories in 3GPP TR 26.928 [i.25], which also introduces different distributed computing architectures for XR. 3GPP TR 23.748 [i.26] defines the necessary modifications to 5G system architecture to enhance Edge Computing. Specifically, this study is investigating mechanisms to discover connectivity to available Edge Computing resources (e.g. using DNS), mobility improvements for both UE consuming Edge Computing services and for Edge Application Servers, and for network capability

exposure towards the Edge Application Server. In addition, in 3GPP TR 23.758 [i.27] and 3GPP TS 23.558 [i.28] a new set of application layer interfaces for Edge Computing are identified that may potentially be useful for integration of Edge Computing. Specifically, the interfaces will enable application-layer discovery of Edge Application Servers, capability exposure towards the Edge Application Server, and procedures for onboarding, registration, and lifecycle management of Edge Applications.

### Motivation

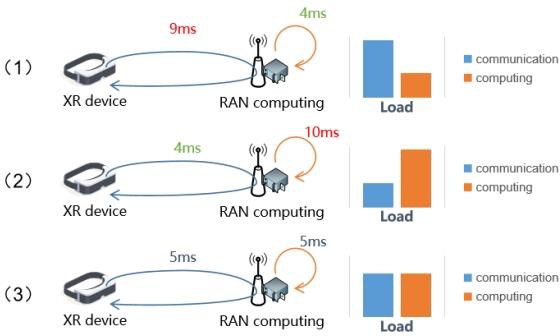
Figure 5.4.2-1 shows the XR rendering collaboration between RAN and XR devices. In this type of service, the terminal collects holographic information such as the user’s geographic location, orientation, manipulation, and five-senses through sensors. With the help of a high-speed and stable radio network, the information is transmitted to the RAN edge computing platform and the cloud through the uplink. Then, the RAN edge or cloud rendering server invokes computing resources to render the screen in real time, and the processed holographic information will be transmitted to the user’s terminal device through the downlink after being encoded and compressed. Finally, the terminal device restores the above information to the user.



##### Figure 5.4.2-1: XR rendering collaboration between RAN and XR devices

However, there are two major limitations with existing computation offloading approaches.

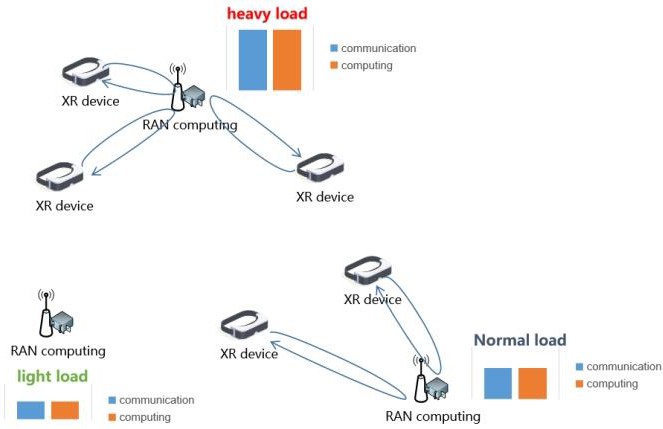
Firstly, there is a lack of alignment between communication and computing decisions. Although computing is brought closer to the RAN, the existing architecture still manages communication and computing separately, resulting in sub- optimal performance. 3GPP TR 22.261 [i.22] defines motion-to-photon latency in the range of 7ms to 15ms while maintaining the required resolution of up to 8k giving user data rate of up to 1Gbit/s. Assuming the service requirements for motion-to-photon latency is 10ms, three RAN computing nodes with different bandwidth load and computing load can be chosen as the server, as shown in Figure 5.4.2-2. If choosing the computing node based on the best computing delay it will be the node 1, the E2E delay is 13ms. If choosing the computing node based on the best communication latency it will be the node 2, the E2E delay is 14ms. If choosing the computing node based on both of the status it will be the node 3, the E2E delay is 10ms and meets the service requirements. To ensure E2E delay without resource wastage of both radio network and computing resources, it is essential to align the decisions of communication and computing.



##### Figure 5.4.2-2: RAN computing nodes with different bandwidth load and computing load

Secondly, there is a lack of united orchestration among multiple users in adapting to resource-limited edge environments. An adaptive design focusing on a single-user offloading pipeline can be duplicated for multiple users, but this can lead to poor resource efficiency. Figure 5.4.2-3 shows the situation that XR devices are associated with the closest BS. Without united orchestration among multiple users, some RAN nodes may become overloaded while others are underutilized, resulting in inefficient resource allocation. To address this issue, it is necessary to coordinate the decisions of multiple users holistically.

NOTE: Without specification, the "resource" mentioned above refers to the available communication and computing capabilities within RAN.



##### Figure 5.4.2-3: Lack of united orchestration among multiple users in adapting to resource-limited edge environments

Therefore, a joint design approach is required for computing node selection, computing resource allocation, and radio resource optimization. Leveraging modern AI technologies, computational tasks or resources can be automatically scheduled and orchestrated along with the optimization of communication resources. This integrated approach enables more efficient utilization of computing resources, thereby reducing wastage in redundant hardware capital expenditures (CAPEX) and operating expenses (OPEX).

### Proposed Solution

As shown in Figure 5.4.3-1, XR rendering collaboration over SMO consists of the following steps:

1. Application awareness: External consumer collects the XR service requirements and application status from the XR devices and XR edge applications.
2. Capability exposure: communication and computing capabilities within RAN are exposed by SMO (1a) or via Y1 (1b) interface to external consumer. For example, levels of E2E delay can be provided for XR rendering tasks.
3. Analysis and decision for service request: external consumer analyzes the XR service requirements and the capabilities within RAN, and then determines the service request for capabilities in O-RAN.
4. Service request: external consumer requests communication and computing service in O-RAN.
5. Optimization and decision analysis: potentially utilize AI/ML models to analyze the service request and determine the resource demand and policy for XR rendering assurance in O-RAN.
6. Execution:

5a) Deploy or scale the cloud resources of E2 nodes (i.e., O-CU, O-DU) via O2 interface.

5b) Reconfigure the E2 nodes via the O1 interface to guarantee radio transmission for XR rendering (e.g., cell-level spectrum allocation).

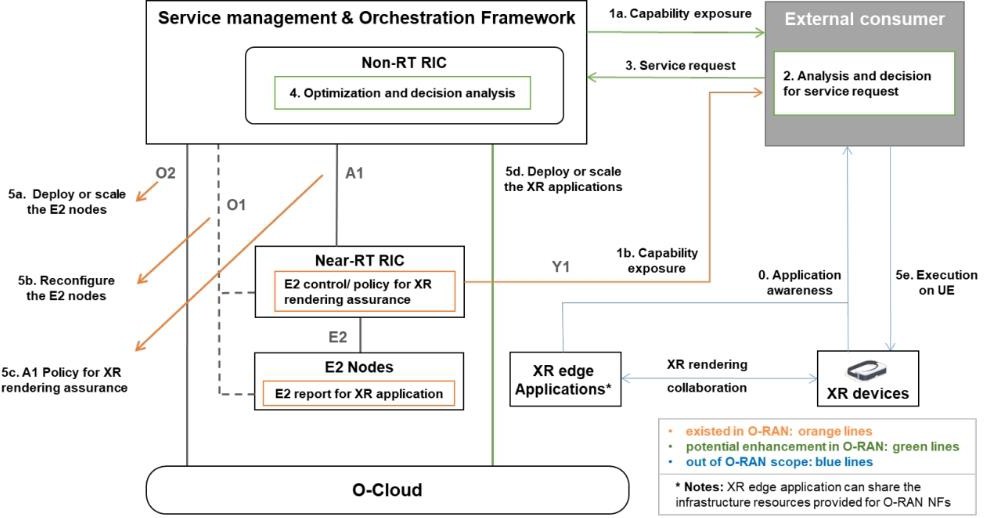
5c) Create A1 policy for XR rendering assurance.

5d) Deploy or scale the cloud resources of XR edge applications on the O-Cloud platform.

5e) (Out of O-RAN scope) Deploy XR applications on UE or notify UE the default XR edge application and/or the default communication node.

NOTE1: Step 5a. to 5d. can be freely combined or skipped to guarantee the XR rendering assurance, depending on the specific requirements of the XR service.

NOTE2: External consumers could be global orchestrator or the edge enabler server in 3GPP [i.28].



##### Figure 5.4.3-1: XR rendering collaboration over SMO

### Benefits of O-RAN Architecture

O-Cloud in O-RAN provides the necessary computing, storage, and communication resources to host and execute the virtualized functions and applications required for O-RAN operations. O-Cloud can extend its capabilities to the network edge, reducing latency and improves the performance of XR applications by processing data closer to the source.

O-RAN’s open interfaces and AI/ML based architecture will enable the challenging mechanisms of XR rendering collaboration to be implemented. For example, Near-RT RIC collects rich data on radio conditions and computing capabilities in RAN, as well as information on XR applications. By training AI models with these data, the performance of XR rendering collaboration tasks can be significantly enhanced with jointly optimizing computing node selection, computing resource allocation and radio resource optimization.

The purpose of this use case is to perform distributed rendering of co-orchestration of communication resources and computing resources in O-RAN to improve the experience of XR users.

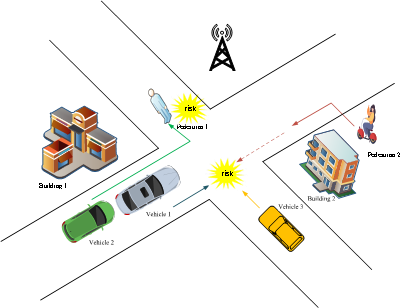
## Use Case 5: Computing Resource Scheduling in V2X Communications

### Background Information

It is expected that there will be approximately 8 million autonomous or semi-autonomous vehicles on the road by 2025 [i.29]. As an enabling technology, Vehicle-to-Everything (V2X) communication is pivotal in paving the way for the Intelligent Traffic System and autonomous vehicles by facilitating the exchange of crucial information among vehicles, infrastructure, and pedestrians. This exchange encompasses critical data, including real-time updates on traffic conditions, impending road hazards, and precise speed limits, to improve decision-making and prevent accidents.

Within the realm of V2X communications, the intersection is one of the most challenging scenarios. The various ways of transportation (e.g. vehicles, pedestrians) coupled with densely constructed buildings contributes to the complexity of traffic conditions, resulting in a large amount of data gathered from diverse sources for comprehensive sensing and collision avoidance. As shown in Figure 5.5.1-1, Vehicle 1 travels straight through the intersection, while Vehicle 2 maintains a safe distance behind Vehicle 1, intending to turn left. During the drive, there are several risks.

* Vehicle 1 risks colliding with Vehicle 3.
* Vehicle 1 or Vehicle 3 risks colliding with the pedestrian 2 (if he suddenly rushes from behind Building 2 and turns left). However, either Vehicle 1 and Vehicle 3 cannot sense the pedestrian timely due to the blocking by Building 2.
* Vehicle 2 risks colliding with other vehicles and pedestrian (e,g., Vehicle 3 and pedestrian 1). Due to the blocking by Vehicle 1 and Building 1, Vehicle 2 is unable to accurately sense traffic conditions ahead and to the left in real-time and has to obtain information by exchanging information with other vehicles or infrastructures, which may result in high latency and high risk of collision.



##### Figure 5.5.1-1: An illustration of the potential collisions at intersections

In collision warning scenarios, the latency requirement for data collection and computing are in few milliseconds (e.g., the maximum end-to-end latency for cooperative collision avoidance in the advanced driving scenario is 10ms.) [i.30], so that the results and warns can be fed back to vehicles and pedestrians in timely manner. To fulfill the strict latency demand, it is essential to process these tasks in one or several optimal computing nodes, e.g., the central cloud, edge cloud, radio edge cloud, device edge cloud etc. Since the end-to-end latency is limited by the computing latency and network latency, the optimal offloading choice cannot be found by either optimize the network or computing resource. Hence, joint optimization of network and computing resource is needed. For example, the road status sensing data from Vehicle 1, Vehicle 3 and Pedestrian 2 must be communicated and processed and the result fed back within milliseconds to avoid the collision between them.

To reduce network latency, it is desirable to process the data locally in vehicles. However, the computing resources of vehicles may be limited, which may not satisfy the task requirements in terms of computing resources and capabilities and could also lead to high computing latency. Therefore, it may be desired to offload all or part of the data processing to the best edge server (e.g., the server completes data processing with the lowest latency, lowest energy consumption.) through joint optimization of network and computing resources. However, since the task requirements on computing resources (e.g., the required type of computing resources) are diverse and the availability of computing resources is dynamic, it is of great important to enable joint communication and computing resources optimization. What is worse, there are other data to be processed, such as navigation requests from other vehicles, requests for entertainment-related services from passengers, etc., which also consume limited computing and communication resources and increase the difficulties.

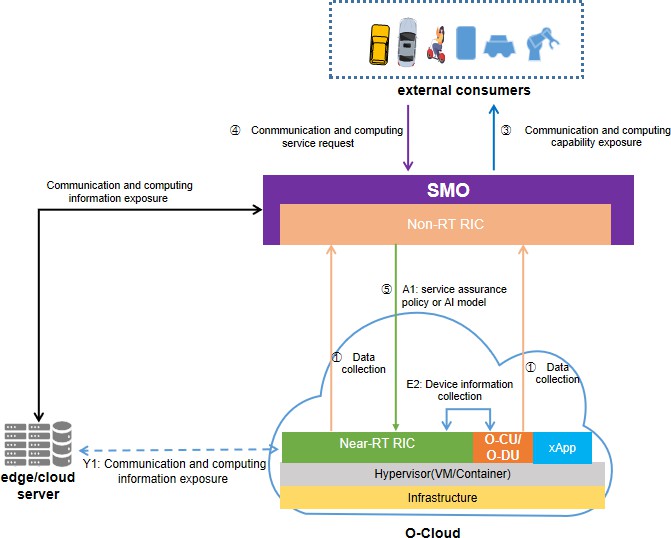
### Motivation

The main objective is to enable joint communication and computing resources optimization for V2X communications. In O-RAN, the global information on communication and computing resources is available over O1 and O2 interface, thus facilitating the joint communication and computing resources optimization. With the above global information, resource scheduling policies for both communication and computing can be developed in Non-RT RIC (e.g., policies in terms of CPU frequency, bandwidth and power, etc.), and communication optimization policies can be enforced by Near-RT RIC based on the near real-time UE level information (e.g., real-time link quality) . In addition, the vacant computing resources in O-Cloud can provide edge computing services in case of low network load, which further facilitates the low-latency computing in V2X communications.

### Proposed Solution

O-RAN can provide edge computing capabilities and communication assurance for computing in V2X. The vacant O- Cloud resources and collaboration between Non-RT RIC and Near-RT RIC facilitate on-demand resource allocation adjustments and joint optimization of communication and computing resource scheduling to ensure end-to-end performance.

Fig. 5.5.3-1 illustrates how O-RAN can facilitate computing in V2X communications, where the O-Cloud is used as the computing entity in O-RAN.



##### Figure 5.5.3-1: An illustration of O-RAN assisted computing in V2X communications

First, SMO retrieves data from Near-RT RIC, E2 nodes, O-Cloud, and/or edge/cloud servers to get the communication and computing resources information, which mainly includes but is not limited to information about available computing resources (e.g., resource type, CPU/GPU frequency and storage) and communication resources (e.g.,

network load, available bandwidth, link quality). The mentioned information of O-Cloud can be retrieved over the O2 interface, while the communication resource information can be obtained via the O1 interface from O-CU and O-DU. Also, the related information of external edge/cloud servers can be retrieved through the information exposure between external edge/cloud servers and SMO.

Next, SMO exposes the computing services or integrated communication computing services provided by O-RAN to external entities, e.g., vehicles, pedestrians and other third-party consumers, so that external customers can send their service requests. The service request information can include but is not limited to data size, required computing resource type, latency deadline of the contained tasks.

NOTE: The security related consideration is for further study.

Then, Non-RT RIC trains the AI model to generate policies based on task requirements and available computing and communication resources. The policies mainly include but are not limited to task offloading strategies and/or the joint resource allocation/adjustment strategies of O-CUs/DUs (e.g. offload tasks to one of the O-DUs and allocate CPU frequency and bandwidth to these tasks based on non-real-time information), where the communication resources allocation/adjustment strategies can be further enforced and sent from Near-RT RIC to O-CUs/DUs (e.g., RAN optimization polices based on near real-time link quality).

### Benefits of O-RAN Architecture

The O-RAN system presents several advantages in joint communication and computing resources optimization for V2X scenarios with a large amount of data to be processed in low latency. The vacant O-Cloud-based computing resources can be shared with edge applications. In addition, the Non-RT RIC can perform joint communication and computing resources optimization, and Near-RT RIC facilitates the RAN behaviour optimization based on the information obtained through E2 interface. In summary, O-RAN architecture provides necessary ingredients to facilitate computing for V2X scenarios.

## Use case 6: Exposure of performance information to assist computing resource scheduling

### Background Information

Computing task performance information are used to monitor the loading conditions of virtual compute resources of computing tasks that include computing task memory usage, CPU occupancy rate, data throughput etc...

ETSI NFV defines the use case of performance managements and performance metrics. Monitoring parameters associated with the compute related measurements and of interest for a VNF can be declared in the VNFD (see ETSI GS NFV-IFA 011 [i.35]). The measurements can be used as the triggers to the scale VNF operation (see clause 7.2.4 in ETSI GS NFV-IFA 008 [i.36].

The Kubernetes (K8s) API server provides supervision data about the Kubernetes [objects,](https://kubernetes.io/docs/concepts/#kubernetes-objects) such as pods. processes and components use this information to track whether pods are being launched and maintained as expected. These cluster state metrics can also provide a high-level view of cluster and its state.

### Motivation

In CCIN, computing resource orchestrator should continuously supervising the performance of running computing tasks. Reporting of running computing tasks performance information will assist the computing resource orchestrator to make decision on deploying a computing task to a new cloud node or triggering to scale a computing task.

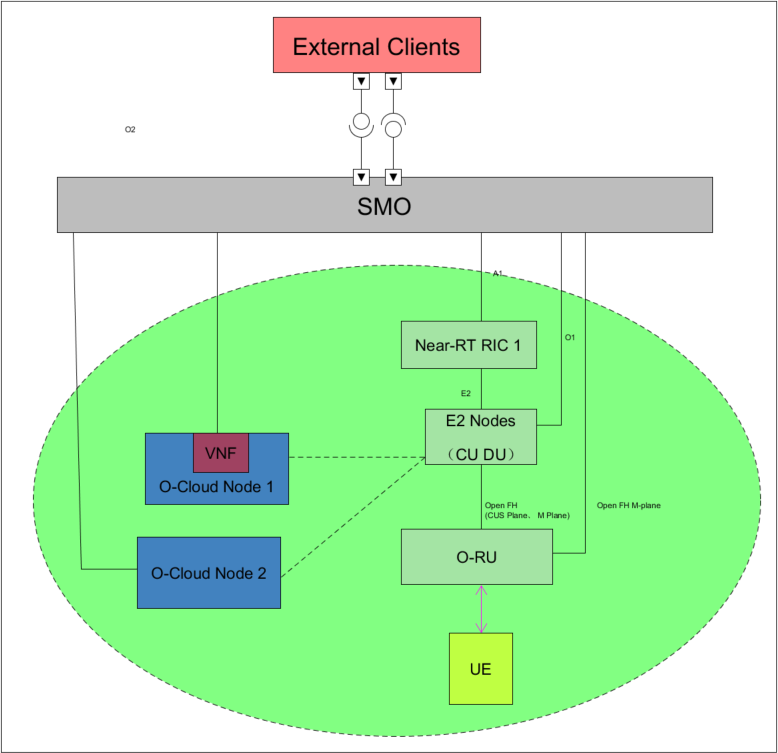
And this can also help the computing resource orchestrator to measure computing task’s actual resource usage, which helps for evaluating the adequacy of a candidate cloud node to execute the computing task during mobility.

### Proposed Solution

#### Performance information reporting triggers redeployment of a computing task

When a VNF/App is successfully deployed on an O-Cloud node, SMO collects the performance information of this VNF/App and computing resource information from Cloud Nodes distributed in various geographical locations on the O-Cloud through the O1 or O2 interface. If current O-Cloud node’s performance does not meet the requirements of the VNF/App, computing resource orchestrator may trigger SMO to redeploy the VNF/App via the O2 interface.

As shown in Figure 5.6.3.1-1, Node1 and Node2 provide computing resource at the edge cloud for UEs, and the coverage scope of cell shows as green in the figure. Through performance information reporting, when the orchestrators find that Node1 does not meet the computing resource requirements, the network may configure Node2 to serve the UE.



##### Figure 5.6.3.1-1: Computing task orchestration based on performance information

Entities/resources involved in the use case:

* + - * 1. SMO:

Support for collecting and analysing VNF/App performance information from O-Cloud.

Support for redeployment or scale-in/scale-out VNF/App via O2.

* + - * 1. O-Cloud:

Support for computing related inventory and updated VNF/App performance information with required granularity to SMO over O2 interface.

Support for deploying of a VNF/App or scaling a VNF/App

* + - * 1. Orchestrator/Application Server:

Request/subscribe VNF/App performance/computing information from SMO.

Response for deploy/scale VNF/App on O-Cloud.

The whole procedure is presented in Table 5.6.3.1-1, illustrated by Figure. 5.6.3.1-2.

##### Table 5.6.3.1-1: Performance information reporting triggers redeployment of a computing task

|  |  |
| --- | --- |
| Use Case Step | Description of the Step |
| Goal | Performance information reporting assist redeployment of a computing task |
| Actors and Roles | SMO, O-Cloud, orchestrators from other clouds/application server |
| Assumptions | All relevant functions and components are instantiated.  O2 interface connectivity is established with SMO. |
| Pre-conditions | SMO and O-Cloud are instantiated with O2 interface connectivity being established between  them. |
| Begins when | Orchestrators external to the O-Cloud want to subscribe VNFs/Apps performance information  of O-Cloud. |
| Step 1 (M) | Orchestrators external to the O-Cloud send VNFs/Apps performance information subscription  to SMO. |
| Step 2 (M) | SMO subscribe VNFs/Apps performance information from O-Cloud Nodes. |
| Step 3 (M) | SMO response for VNFs/Apps performance information subscription to orchestrators external  to the O-Cloud. |
| Step 4 (L) | SMO collect VNF/App performance information via O2/O1 interface and send the notification  to indicate VNF/App performance information to orchestrators external to the O-Cloud. |
| Step 5 (M) | Performance can not fulfill the requirement, orchestrators external to the O-Cloud requests SMO to redeploy the VNF/App or App to cloud node2. SMO deploy VNF/App on node2 and  VNF/App route switch from node1 to node2. |
| Step 6 (L) | SMO terminate VNF/App on node1. |
| Step 7 (M) | SMO response of redeployment of VNF/App success to orchestrators form other clouds or  application servers. |
| Ends when | Orchestrators form other clouds or application servers get the response. |
| Post Conditions | N/A |

@startuml

Skin rose

skinparam ParticipantPadding 5

skinparam BoxPadding 10

skinparam defaultFontSize 12

Box “Orchestrator” #lightcyan Participant "Orchestrator” as client

End box

Box SMO #gold

Participant “SMO” as smo

End box

Box O\_Cloud Node #lightseagreen Participant ”IMS/DMS” as ims #orange Participant ”Node 1” as ocloud1

Participant ”Node 2...Node N” as ocloud2

End box autonumber 1

|||

activate ocloud1

== VNF/Apps performance Subscription ==

client ->smo : Subscription of VNF/App performance information\n<color #red> report periodic or event trigger

autonumber 2.1

smo ->ims : Subscription of VNF/App performance information\n <color #red> report periodic or event trigger

smo <-ims : Response of VNF/App performance information autonumber 3

client <-smo : Response of VNF/App performance information

|||

loop performance data accumulation autonumber 4.1

ims <-> ocloud1: <<O2>> or <<O1>> VNF/App performance Information <color #red> ims <-> ocloud2: <<O2>> or <<O1>> VNF/App performance Information <color #red> ims -> smo: <<O2>> or <<O1>> VNF/App performance Information<color #red> client <- smo: Notifications to indicatee VNF/App performance<color #red>

end

|||

== Performance cant fulfill the resource requirement， deploy the VNF/App or App to other Cloud Node ==

|||

autonumber 5.1

client -> smo: Request to change the cloud node <color #red>to o-Cloud node 2 loop for each NF need to be deployed on o-Cloud node 2

alt ETSI

note over smo, ocloud2:ETSI o-cloud Instantiate VNF/Apps refer to WG6 O2DMS ETSI IF 2.4.2.1

else K8s（Kubernetes）

note over smo, ocloud2: K8s（Kubernetes）o-cloud Instantiate NF deployment refer to WG6 O2DMS Kubernetes IF 3.1.2

end

ocloud2<->ocloud1:Swtich service route to ocloud2 deactivate ocloud1

activate ocloud2 end

|||

|||

== Swtich service route to ocloud2 and terminate VNF/App instance on ocloud2 == autonumber 6.1

loop for each VNF/Apps

ims <- smo: Request to terminate the VNF/App

alt ETSI

note over ims, ocloud1:ETSI o-cloud terminate VNF/Apps else K8s（Kubernetes）

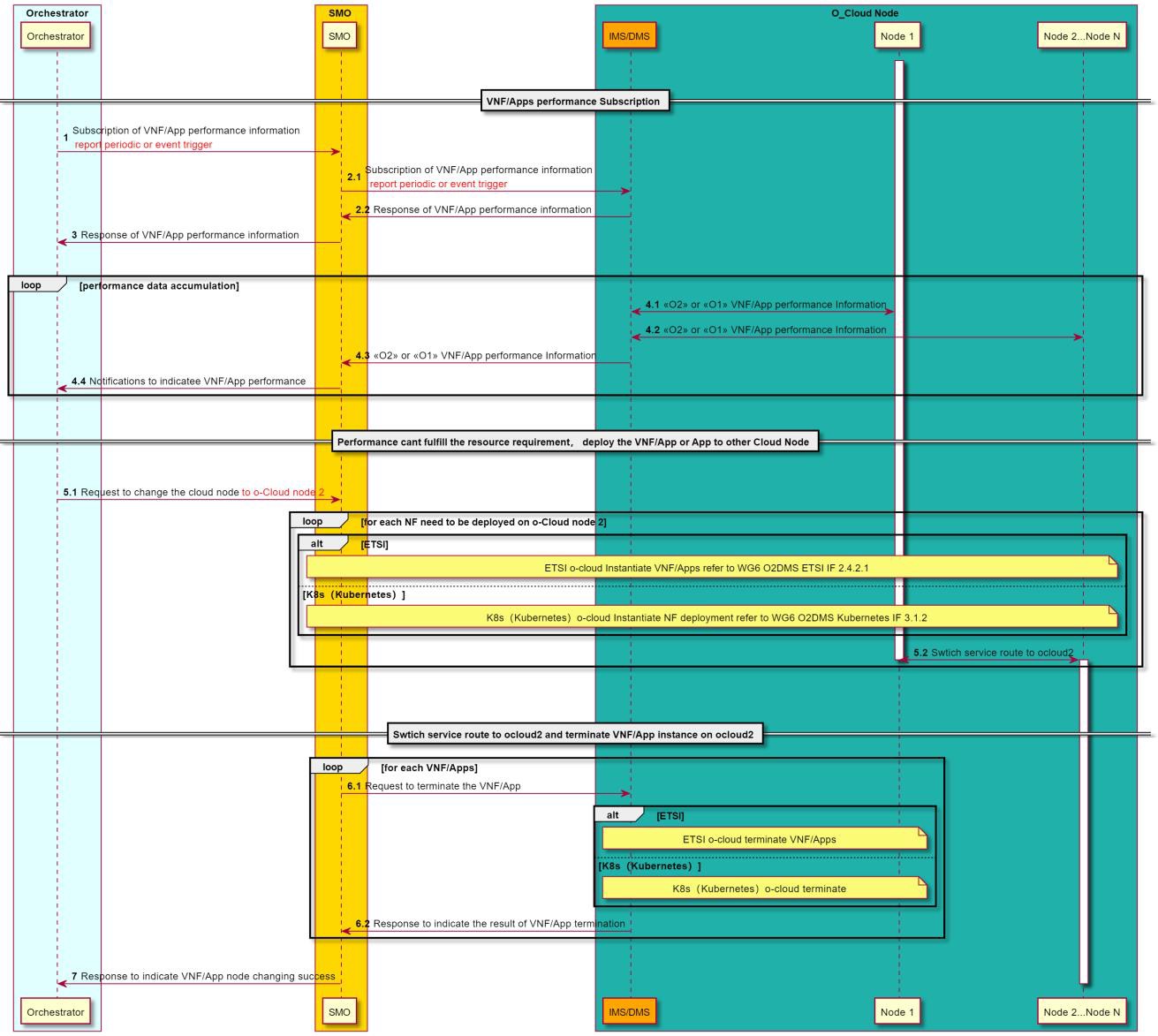
note over ims, ocloud1: K8s（Kubernetes）o-cloud terminate end

ims -> smo: Response to indicate the result of VNF/App termination end

||| autonumber 7

client <-smo :Response to indicate VNF/App node changing success deactivate ocloud2

@enduml



##### Table 5.6.3.2-2: Performance information reporting triggers redeployment of a computing task

Required data：

Some computing task performance information that proposed to be exposed is list as below:

1. K8s
   * K8S pod status:
     + Pending
     + Running
     + Succussed
     + Failed
     + unknown
   * K8S Node, Pod, containers usage:
     + Memory requests: Total memory requests (bytes) of a pod
     + Memory limits Total memory limits (bytes) of a pod
     + Allocatable memory Total allocatable memory (bytes) of the node
     + Memory utilization Total memory in use on a node or pod
     + CPU requests Total CPU requests (cores) of a pod
     + CPU limits Total CPU limits (cores) of a pod
     + Allocatable CPU total allocatable CPU (cores) of the node
     + CPU utilization total CPU in use on a node or pod
     + Disk utilization total disk space used on a node
2. ETSI
   * ETSI VNF performance:
     + Average virtual CPU usage
     + Average virtual memory usage
     + Average virtual disk usage
     + Input data size (byte)
     + Output data size (byte)
     + Number of input packets
     + Number of output packets

#### Performance information reporting triggers to scale a computing task

The orchestrators external to the O-Cloud who responsible for computing resource orchestration can also be triggered to perform Scale-In and Scale-Out by performance information reporting so as to make full use of computing resources.

The whole procedure is presented in Table 5.6.3.2-1, illustrated by Figure. 5.6.3.2-1.

##### Table 5.6.3.2-1: Performance information reporting triggers to scale a computing task

|  |  |
| --- | --- |
| Use Case Step | Description of the Step |
| Goal | Performance information reporting assist scale-in/scale-out of a computing task |
| Actors and Roles | SMO, O-Cloud, orchestrators from other clouds/application server |
| Assumptions | * All relevant functions and components are instantiated. * O2 interface connectivity is established with SMO. |
| Pre-conditions | - SMO and O-Cloud are instantiated with O2 interface connectivity being established  between them. |
| Begins when | Orchestrators form other clouds or application servers send subscribe VNFs/Apps  performance information of O-Cloud. |
| Step 1 (M) | Orchestrators external to the O-Cloud send VNFs/Apps performance information subscription  to SMO. |
| Step 2 (M) | SMO subscribe VNFs/Apps performance information from O-Cloud Nodes. |
| Step 3 (M) | SMO response for VNFs/Apps performance information subscription to orchestrators external  to the O-Cloud. |
| Step 4 (L) | SMO collect VNF/App performance information via O2/O1 interface and send the notification to indicate VNF/App performance information to orchestrators form other clouds or application  servers. |
| Step 5 (M) | Resource can not fulfill the requirement (Scale out) or less resource is needed (Scale in),  orchestrators form other clouds or application servers triggers scale action. |
| Step 6 (L) | SMO scale VNF/App via O2 interface. |
| Step 7 (M) | SMO send response of VNF/App scale-out/Scale-in complete. |
| Ends when | Orchestrators form other clouds or application servers get the response. |
| Post Conditions | N/A |

@startuml

Skin rose

skinparam ParticipantPadding 5

skinparam BoxPadding 10

skinparam defaultFontSize 12

Box “Orchestrator” #lightcyan Participant "Orchestrator” as client

End box

Box SMO #gold

Participant “SMO” as smo

End box

Box O\_Cloud Node #lightseagreen Participant ”IMS/DMS” as ims #orange Participant ”Node 1” as ocloud1

End box autonumber 1

activate ocloud1

== VNFs/Apps performance Subscription ==

client ->smo : Subscription of VNF/App performance information\n<color #red> report periodic or event trigger

autonumber 2.1

smo ->ims : Subscription of VNF/App performance information\n <color #red> report periodic or event trigger

smo <-ims : Response of VNF/App performance information autonumber 3

client <-smo : Response of VNF/App performance information

|||

loop performance data accumulation autonumber 4.1

ims <-> ocloud1: <<O2>> or <<O1>> VNF/App performance Information <color #red> ims -> smo: <<O2>> or <<O1>> VNF/App performance Information<color #red>

client <- smo: Notifications to indicatee VNF/App performance<color #red> end

|||

== Resource can not fulfill the requirement(Scale out) or less resource is needed(Scale in) ==

||| autonumber 5

client -> smo: Request to scale-out/scale-in VNF/App

loop for each NF need to be scale-out/scale-in on o-Cloud node autonumber 6.1

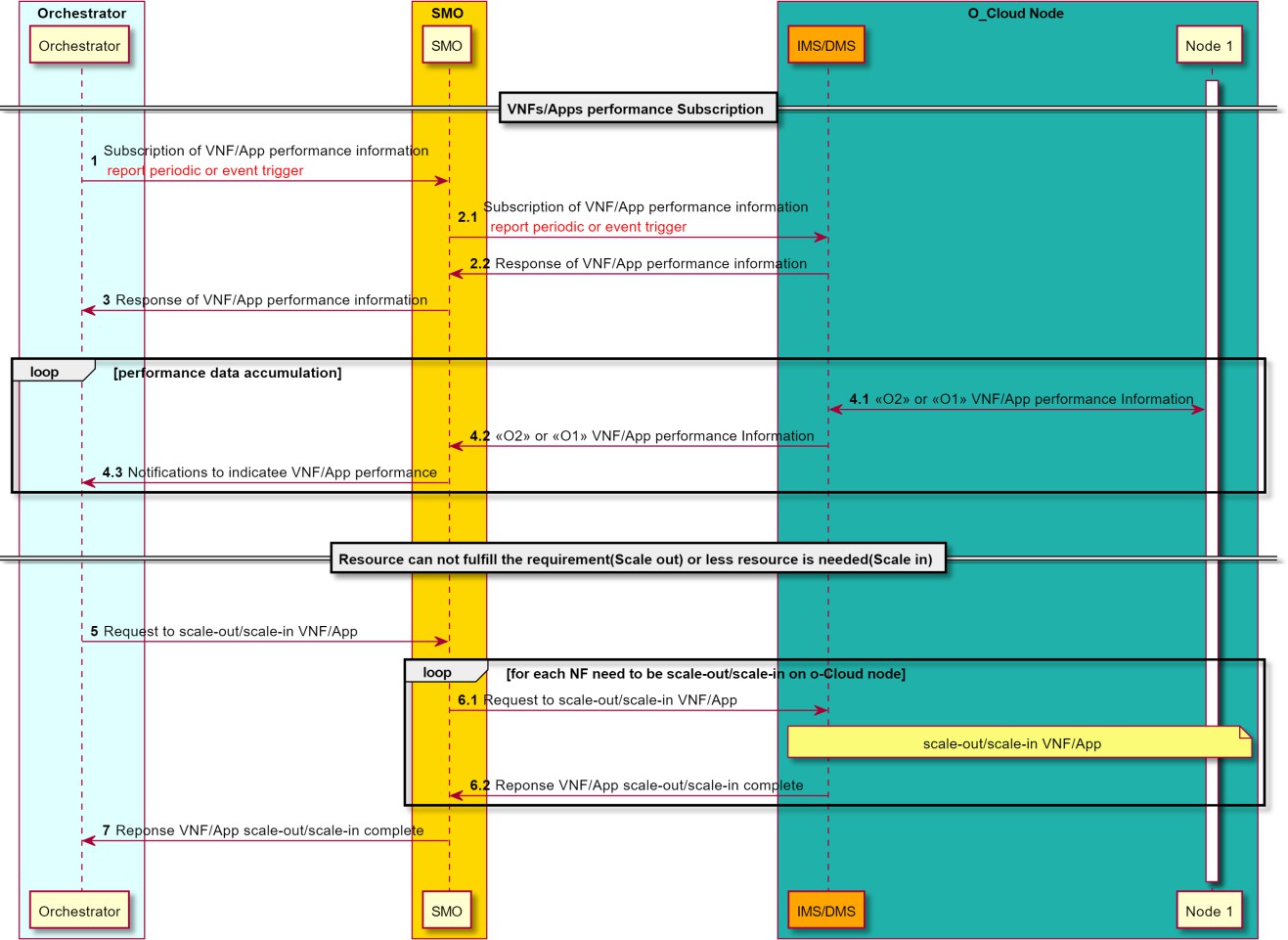
smo ->ims : Request to scale-out/scale-in VNF/App note over ims, ocloud1: scale-out/scale-in VNF/App

smo <-ims : Reponse VNF/App scale-out/scale-in complete end

autonumber 7

client <- smo:Reponse VNF/App scale-out/scale-in complete

||| @enduml



##### Figure 5.6.3.2-1: Performance information reporting triggers to scale a computing task

Required data：

Same as clause 5.6.3.1.

### Benefits of O-RAN Architecture

By providing computing task performance information to orchestrators external to the O-Cloud, orchestrators and application servers can redeploy or adjust computing tasks in O-Cloud. O-Cloud's edge computing resources can be fully and accurately scheduled. In that way, edge computing vertical businesses (e.g. XR, AI/ML, V2X, vertical industries) can be effectively supported by computing resource services.

# Key Issues

## Key Issue #1: Joint Communication and Computing Optimization

Use cases 3, 4, and 5 have a common requirements on the joint communication and computing optimization. Use case 1 investigates the computing aware handover optimization. Use case 3 discusses how to expose the communication and computing status to the external entities (e.g., multi-cloud orchestrator, applications) to enable the RAN aware application optimization.

In this key issue, the following points need to be studied:

* How to enhance the mobility management to ensure the connection & service continuity taking both the communication and computing resource status into consideration?
* What information needs to be exposed by O-RAN to external entities?
* Which entity in O-RAN will do the exposure?
* What’s the mechanism (including the security mechanism) to support the information exposure?

## Key Issue #2: O-RAN computing service provision

Use cases 3, 4 and 5 have a common requirement of exposing vacant O-RAN computing capabilities. The radio edge computing capability offered by O-RAN can be discovered by the O-RAN external consumers such as the multi-cloud orchestration or the application servers. The external consumers can monitor the O-RAN computing capability as well as other necessary information to decide whether to request edge application offloading to the O-RAN system.

In this key issue, the following mechanisms need to be studied:

* How an external consumer requests or subscribes to O-RAN system for providing the O-RAN computing capabilities?
* How O-RAN system provides/updates the O-RAN computing capabilities to the external consumer?
* How an external consumer requests or subscribes to O-RAN system for offloading the computing tasks to O- RAN system?

## Key Issue #3: O-RAN integrated communication and computing service provision

Use Cases 3, 4 and 5 in Clause 5 expect O-RAN to provide integrated communication and computing services towards the external consumers (e.g., a multi-cloud orchestrator) . Currently, SMO intent driven management was studied in [i.x], where RAN management intent and RAN performance assurance intents were investigated. However, they are all focusing on RAN communication services provision. While key issue1 targets the O-RAN computing service provision. Key issue 3 is built on key issue 2 and the RAN service provision studied in [i.32] [i.33] [i.34], trying to investigate the integrated communication and computing service provision.

In this key issue, the following mechanisms need to be studied:

* How to describe or model the integrated communication and computing service profile?
* How an external consumer requests or subscribes to the O-RAN system for integrated communication and computing services?
* How does O-RAN system jointly orchestrate the communication and computing resources to fulfill the service request?

## Key Issue #4: Edge application orchestration in O-RAN

As stated in use cases 3, 4 and 5, the vacant computing resources from the cloud-based RAN can be leveraged to offload edge computing applications or tasks so that radio cloud resources can be efficiently utilized. To achieve this vision and facilitate the computing service provision, integrated communication and computing services provision discussed in key issue 2 and 3, the edge application lifecycle management, including deployment, termination, upgrade, scale-in and out, etc., needs to be considered within O-RAN.

In this key issue, the following mechanism needs to be studied:

* How to deploy, manage and orchestrate the edge applications running on the O-Cloud?
* How the edge application orchestration functionality interacts with the SMO services (SMOS) studied in [i.33], such as the RAN analytics SMOS, service orchestration SMOS, service assurance SMOS, to facilitate the joint communication and computing orchestration?

# Potential Solutions

## Solution #1: O1 configurations for multi-aspect handover optimization

### Key Issue Mapping

This solution is for Key Issue #1: Joint Communication and Computing Optimization, specifically to enhance mobility management to ensure connection and service continuity, taking both communication and computing resource status into consideration.

### Description

The solution is based on the clause 5.1 Use Case 1: Multi aspect handover optimization. It is built based on the existing capabilities of different interfaces such as R1 interface, O1 interface, O2 interface, etc and introduces new O1 based

configurations in Near-RT RIC to support multi-aspect handover optimization in a Communication and Computing Integrated Networks.

This solution consists of rApp, SMO/Non-RT RIC Framework, near-RT RIC, E2 nodes, and O-Cloud. rApp uses the R1 interface to retrieve different input data, such as CM data, PM data, location data of different cloud-based RAN entities, supported types of computing resources, capabilities of computing resource types, and load/availability of computing resources in different cloud-based RAN entities. rApp then executes relevant algorithms to derive the proximal computing capability table and availability table for each cloud-based RAN entity. rApp uses this information to create different O1 based configurations for multi-aspect handover optimization using R1 services provided by SMO/Non-RT RIC Framework, which then creates the O1 based configurations at one or more near-RT RIC instances over O1. Near- RT RIC will then use E2 control or E2 policy to ensure the UE handover to the optimum target cloud-based Radio entity.

### Procedures

Figure 7.1.3-1 depicts the procedure summarizing the solution.

@startuml Skin rose

skinparam ParticipantPadding 5

skinparam BoxPadding 10

skinparam defaultFontSize 12 skinparam lifelineStrategy solid skinparam wrapWidth 300 autonumber

box #lightseagreen

participant “O-Cloud(s)” as OCLOUDS

endbox

box #gold

participant "rApp" as rApp

Participant "SMO / Non-RT RIC Framework" as nonRTRICFramework endbox

box "O-RAN Nodes" #lightpink

Participant "Near-RT RIC" as NearRTRIC Participant "E2-Node(s)" as RANNodes

endbox

group O1 Data Collection

rApp -> nonRTRICFramework : <<R1>> Data collection request nonRTRICFramework -> RANNodes : <<O1>> Data collection request RANNodes -> nonRTRICFramework : <<O1>> Data Collection nonRTRICFramework -> rApp : <<R1>> Data Collection

end

group O2 Data Collection

rApp -> nonRTRICFramework : <<R1>> Data collection request nonRTRICFramework -> OCLOUDS : <<O2>> Data collection request OCLOUDS -> nonRTRICFramework : <<O2>> Data Collection nonRTRICFramework -> rApp : <<R1>> Data Collection

end

group Data Analysis and Inference

rApp -> rApp : Create Proximal Computing Capability Table and Availability table for each O-Cloud based RAN nodes and prepare related O1 configurations for multi-aspect handover optimization.

end

group Policy Creation

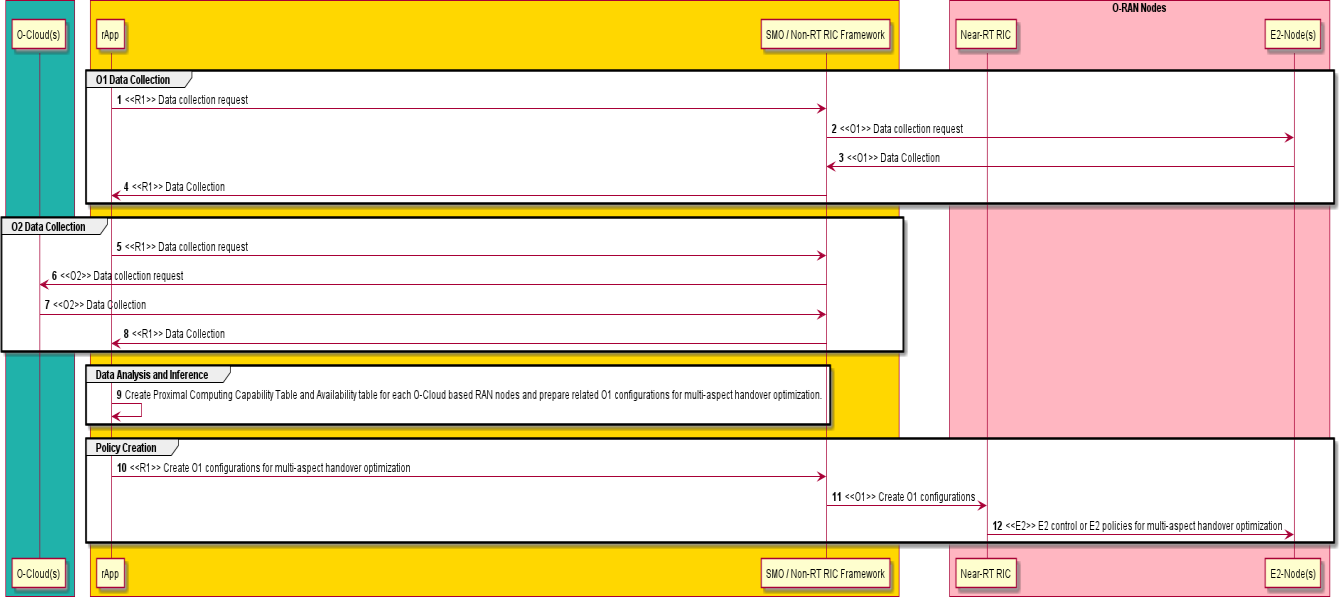
rApp -> nonRTRICFramework : <<R1>> Create O1 configurations for multi-aspect handover optimization

nonRTRICFramework -> NearRTRIC : <<O1>> Create O1 configurations

NearRTRIC -> RANNodes : <<E2>> E2 control or E2 policies for multi-aspect handover optimization

end

@enduml



##### Figure 7.1.3-1: O1 configurations for multi-aspect handover optimization

### Impacts on entities, services and interfaces

SMO/Non-RT RIC Framework:

* Collects necessary CM/PM data from different cloud-based RAN entities over O1 interface. Collects supported types of computing resources, capabilities of computing resource types, and load/availability of computing resources in different cloud-based RAN entities over O2 interface.
* Provides CM data, PM data, supported types of computing resources, capabilities of computing resource types, and load/availability of computing resources of different cloud-based RAN entities to rApp over R1 interface.
* Provides O1 configurations to Near-RT RIC via O1 interface based on the request from rApp. rApp:
* Retrieves necessary C/PM data related to different cloud-based RAN entities from SMO/Non-RT RIC Framework.
* Retrieves supported types of computing resources, capabilities of computing resource types, and load/availability of computing resources in different cloud-based RAN entities from SMO/Non-RT RIC Framework.
* Derives the Proximal Computing Capability Table and Availability Table for each cloud-based RAN entity, use this information to create related O1 configurations to support multi-aspect handover optimization in Near- RT RIC.

Near-RT RIC:

* Performs multi-aspect handover optimization based on the O1 configurations received.
* Sends E2 control or E2 policy messages to E2 nodes (O-CU) related to multi-aspect handover optimization.

E2 node:

* Supports CM data read requests from SMO over O1 interface. Collect and Report PM data to SMO over O1 interface.
* Supports UE handover related control and/or policy enforcement from near-RT RIC over E2 interface.

O-Cloud:

* Supports data requests related to supported types of computing resources, capabilities of computing resource types, and load/availability of computing resources of different cloud-based RAN entities to SMO over O2 interface.

## Solution #2: Computing information exposure via Y1 interface

### Key Issue Mapping

This solution is for KI #1 Joint Communication and Computing Optimization.

### Description

The solution aims to optimize joint communication and computing resources within O-RAN, particularly for use cases 3, 4, and 5. Building upon the existing capabilities of the Y1 interface, which is well-defined in the Y1 Type Definitions (Y1TD) and currently facilitates the exposure of communication information [i.37], this solution aims to extend the interface's functionality to include the exposure of computing information to external entities.

The implementation of this solution involves defining the data structures and formats for computing information on the Y1 interface without needing to change the protocols or the procedures [i.38]. Besides, since Y1 Termination communicates with Y1 consumers via Y1 interface and exposes RAN analytics information service(s) from Near-RT RIC, Near-RT RIC need to obtain computing information from other O-RAN entities.

The solution proposes that Near-RT RIC utilizes the O-Cloud Notification APIs to obtain computing information [i.39], which may be computing resources (containing physical and logical resources) or status (e.g., CPU load, network load, and storage consumption) information.

### Procedures

Figure 7.2.3-1 depicts the procedure summarizing the solution.

@startuml Skin rose

skinparam ParticipantPadding 5

skinparam BoxPadding 10

skinparam defaultFontSize 12

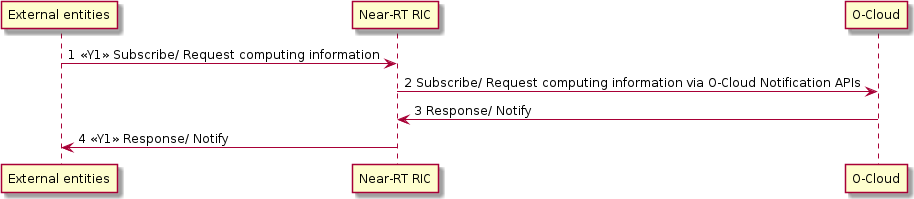
Participant “External entities” as entities Participant “Near-RT RIC” as near Participant “O-Cloud” as ocloud

entities -> near: 1 <<Y1>> Subscribe/ Request computing information

near -> ocloud: 2 Subscribe/ Request computing information via O-Cloud Notification APIs ocloud -> near: 3 Response/ Notify

near -> entities: 4 <<Y1>> Response/ Notify

@enduml



##### Figure 7.2.3-1: Computing Information Exposure via Y1 interface

1. External entities subscribe or request computing information from Near-RT RIC via Y1 interface.
2. Near-RT RIC subscribes or requests computing information via O-Cloud Notification APIs.
3. Near-RT RIC receives response or notifications from O-Cloud.
4. External entities receive response or notifications from Near-RT RIC via Y1 interface.

### Impacts on entities, services and interfaces

Near-RT RIC:

* Supports for receiving computing information subscription or request from external entities.
* Supports the management of computing information processing workflows before exposure.
* Supports subscription or request to O-Cloud via O-Cloud Notification APIs.
* Supports computing information exposure on the Y1 interface.

O-Cloud:

* Supports for receiving computing information subscription or request from Near-RT RIC. External entities:
* Supports subscription or request to Near-RT RIC via Y1 interface.

## Solution #3: Computing Service Provisioning

### Key Issue Mapping

The solution is for Key Issue #2: O-RAN computing service provision.

### Description

In clause 5, the use cases 2 to 5 describes the consumers consuming low latency communication and computing services provided by O-RAN via SMO. To facilitate the provisioning of computing services across diverse application scenarios, it is essential to develop a unified service provisioning approach for the SMO. An unified service provisioning approach of SMO will not only expedite O-RAN's ability to cater to varied business contexts and eliminate the complexity of designing different service delivery methods for different scenarios.

Document [i.31] gives an Service Based Architecture (SBA) of Non-RT RIC, and the logical function Service Management and Exposure (SME) has been defined for Non-RT RIC internal services management, and exposes service to externals, therefore, utilizing SME to provide computing service to externals consumers is a feasible solution.

This solution proposes to use SME to provide computing services externally. As a prerequisite for offering computing service, computing service firstly needs to be registered with the SME, e.g., RAN edge application orchestrator (REAO) registers computing service to SME. The SME offers service exposure to support external consumers discover computing service. External consumers obtain computing service API information through service discovery and consume computing service with the API.

### Procedures

Figures 7.3.3-1 to 7.3.3-2 illustrate the process that summarizes the solution, wherein REAO serves as the computing service producer capable of registering and deregistering computing services with SME.

#### Computing service registration and discovery

In the computing service registration procedure, REAO registers computing service to SME. The registered service could be exposed to external consumers.

SME needs to provide a service discovery mechanism to enable external consumers to discover the computing service. SME can provide the service API information to external consumers through the service discovery. The service API information may contain the URI, IP address and Port, protocol, protocol number and data format, etc. for consumer consuming the service.

The whole procedure is presented in Table 7.3.3-1, illustrated by Figure. 7.3.3-1.

##### Table 7.3.3-1: Computing service registration and service discovery

|  |  |
| --- | --- |
| Use Case Stage | Evolution / Specification |
| Goal | Computing service is registered with the SME and exposed to external consumers.  Consumers can retrieve the computing service API through service discovery. |
| Actors and Roles | SME, REAO, computing service consumers. |
| Assumptions | The REAO inside SMO produces computing service. |
| Step 1 (M) | REAO registers computing services with the SME. The computing service API information  may include the URI, IP address, port, protocol, protocol number, data format, etc. |
| Step 2 (M) | SME feeds the computing service registration result, e.g., successful or failure back to REAO |
| Step 3 (M) | Computing service consumers interacts with SME for the computing service discovery. |
| Step 4 (M) | SME feeds the computing service discovery back with service discovery result (e.g.,  successful, failure), service API, etc. |

@startuml

skinparam sequenceArrowThickness 2

skinparam ParticipantPadding 5

skinparam BoxPadding 10

collections “Computing service consumer” as Consumers #LightCyan box “Service Management and Orchestration” #Gold

Participant "SME" as SME #LightGray Collections "REAO" as REAO #LightGray

end box Autonumber 1

group "Computing service registration"

REAO -> SME: Computing service registration request SME -> REAO: Computing service registration response

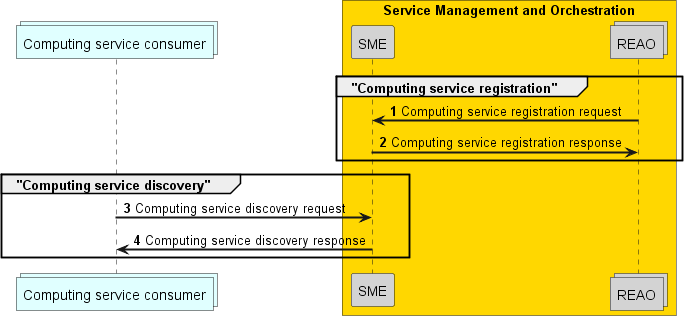
end

group "Computing service discovery"

Consumers -> SME: Computing service discovery request SME -> Consumers: Computing service discovery response

end

@enduml



##### Figure 7.3.3-1: Computing Service Registration and Discovery

#### Computing service consumption

In the computing service consumption procedure, consumers consume the computing service with the API retrieved from the service discovery procedure. The consumer provides the service requirements (e.g., required computing resource type, computing resource quantity, e2e latency, image URL, etc.) with in the computing service request, REAO allocates computing resources according to the requirements, deploys the workload and provides the consumer service endpoint information.

The whole procedure is presented in Table 7.3.3-2, illustrated by Figure. 7.3.3-2.

##### Table 7.3.3-2: Computing service consumption

|  |  |
| --- | --- |
| Use Case Stage | Evolution / Specification |
| Goal | REAO deploys workloads on the O-Cloud based on user requirements, enabling users to  access workloads. |
| Actors and Roles | REAO, O-Cloud, computing service consumers. |
| Assumptions | The REAO inside SMO produces computing service. |
| Step 1 (M) | Consumers request computing service to REAO. Computing service requirement may  includes computing resource type (e.g., CPU, GPU,etc.), computing resource quantity, e2e latency, image URL, etc. |
| Step 2 (M) | REAO feeds consumer back with a response to the received computing service request. |
| Step 3 (M) | REAO deploys the workload on O-Cloud according to the computing requirements. |
| Step 4 (M) | Once the workload deployed, REAO notifies consumer with the service availability and  service entry point. |
| Step 5 (M) | Consumer accesses the deployed workload. |

@startuml

skinparam sequenceArrowThickness 2

skinparam ParticipantPadding 5

skinparam BoxPadding 10

collections “Computing service consumer” as Consumers #LightCyan box “Service Management and Orchestration” #Gold

Participant "REAO" as REAO #LightGray end box

box "O-Cloud" #DarkCyan Collections "WorkLoad" as WLoad

end box Autonumber 1

group "Computing service request"

Consumers -> REAO: Computing service request REAO -> Consumers: Computing service response

end

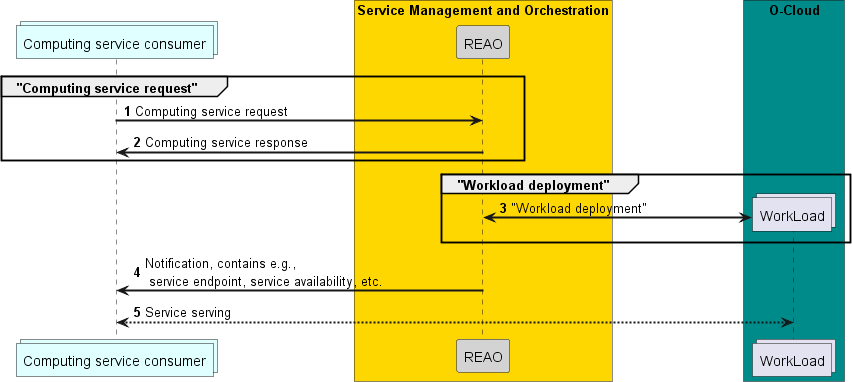
group "Workload deployment"

REAO <-> WLoad\*\*: "Workload deployment" end

REAO -> Consumers: Notification, contains e.g., \n service endpoint, service availability, etc.

Consumers <--> WLoad: Service serving

@enduml



##### Figure 7.3.3-2: Computing Service Consumption

### Impacts on entities, services and interfaces

SMO functionality:

* Introduces new functionality or reuse such as NFO to provide computing service.
* SME functions:
  + Support SMOS producer e.g., computing service producer REAO registering computing service to SME.
  + Support SMOS producer e.g., computing service producer REAO deregistering computing service from SME.
  + Support SMOS producer e.g., computing service producer REAO updating computing service to SME.
  + Support computing service exposure to external consumers.
  + Support external consumers to discover computing service.

New service provided by SMO:

* computing service.

## Solution #4: Integrated communication & computing service provisioning via SMO

### Key Issue Mapping

This solution is for KI #3 O-RAN integrated communication and computing service provision.

### Description

The solution aims to provide integrated communication and computing services towards the external consumers (e.g., a multi-cloud orchestrator) with O-RAN, particularly for use cases AI, XR, and V2X. Based on the existing research of SMO-enabled RAN communication service provisioning and its capability in retrieving the communication and computing resources information, this solution aims to extend the SMO’s function to provide integrated communication and computing services.

This solution involves illustrating the procedure of integrated communication and computing service request, as well as the orchestration of joint communication and computing resources within O-RAN.

### Procedures

The whole procedure is presented in Table 7.4.3-1, illustrated by Figure. 7.4.3-1.

##### Table 7.4.3-1: Integrated communication & computing service provision via SMO

|  |  |
| --- | --- |
| Use Case Stage | Evolution / Specification |
| Goal | Provide integrated communication and computing service via SMO |
| Actors and Roles | SMO, Near-RT RIC, O-Cloud，O-CU，O-DU |
| Assumptions | All relevant functions and components are instantiated. |
| Step 1-3 (M) | SMO exposes the integrated communication and computing service, and consumers send requests to SMO, where the service request includes but not limited to following information:  Service profile ID Maximum number of UEs  Coverage for communication part of communication and computing services provided by RAN  Location of UE  Service end-to-end latency requirements UE mobility level, UE trajectory  Overall uplink/downlink throughput requirements Uplink/downlink throughput requirements per UE Computing resource requirement  Storage resource requirement |
| Step 4-5 (M) | SMO retrieves communication and computing information via O1 and O2 interfaces and  performs joint communication and computing resource optimization via AI/ML inference. |
| Step 6-8 (M) | SMO sends the workload deployment policy (e.g., the assigned endpoint) and/or computing resource optimization policy to O-Cloud via the O2 interface; O-Cloud performs workload deployment and resource assignment based on the policy.  SMO sends the communication resource optimization policy to Near-RT RIC via A1 interface, and Near-RT RIC send the policy to the E2 node. Also， Near-RT RIC can perform further communication optimization. E2 node allocates the communication (e.g.,  bandwidth) for services. |
| Step 9 (M) | SMO sends notifications (e.g., service availability) to the consumer. The consumer establishes connection with the assigned endpoint and sends service load to the  endpoint. |
| Step 10 (M) | SMO evaluates the performance (e.g., network performance, QoS of service), updates  AI/ML model and /or policy accordingly. |

@startuml skin rose

skinparam defaultFontSize 12 Autonumber

Participant "Service consumer" as consumer

Participant "SMO" as SMO Participant "Near RT-RIC" as RIC Participant "O-Cloud" as cloud Participant "E2 nodes" as e2

group service exposure

consumer->SMO: service discovery request consumer<-SMO: service discovery response end

consumer->SMO: service request

group Joint communication and computing resource optimization SMO<-cloud: <O2> Computing resource information

SMO<-e2: <O1> Communication resource information ref over SMO

AI/ML inference end ref

SMO->cloud: <O2> Workload offload policy and/or computing resource optimization policy SMO->RIC:<A1> Communication resource optimization policy

RIC->e2:<E2> Communication resource optimization policy end

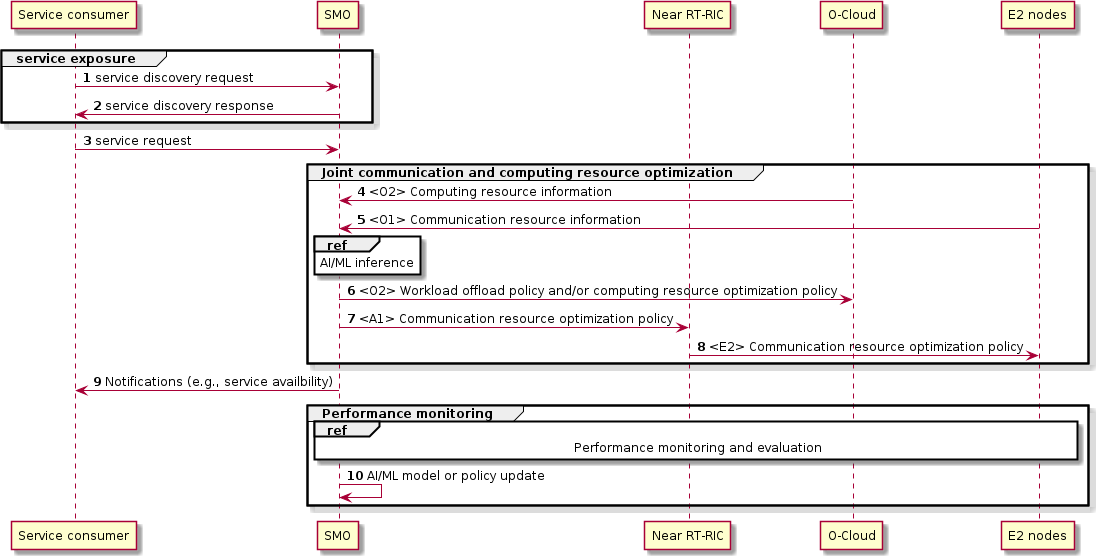
consumer<-SMO:Notifications (e.g., service availbility) group Performance monitoring

ref over SMO,RIC,cloud, e2

Performance monitoring and evaluation end ref

SMO->SMO: AI/ML model or policy update end

@enduml



##### Figure 7.4.3-1: Integrated communication and computing service via SMO

### Impacts on entities, services and interfaces

SMO:

* Supports integrated communication and computing service exposure.
* Supports receiving integrated communication and computing service requests from external entities.
* Supports joint optimization of communication and computing resources.
* Manages communication and computing information.

New service provided by SMO:

* Integrated communication and computing service.

# Analysis on the Potential Architecture Impacts

Based on the use cases study and potential solutions discussed above, the following potential impacts are identified for the O-RAN architecture.

Potential SMO Impact

* Exposure of the O-RAN computing resource information
* Exposure of the Computing services, and integrated communication and computing services
* New functionality or enhancement on existing SMO functionality to enable the deployment, management and orchestration of the applications running on the O-Cloud
* New functionality or enhancement on existing SMO functionality to facilitate the joint communication and computing orchestration

Potential Near-RT RIC Impact

* Exposure of the O-RAN communication and computing information
* New functionality or enhancement on existing functionality to facilitate the computing aware handover optimization

Potential O-Cloud Impact

* Orchestration and management of the application deployment’s life cycle
* Management of the heterogeneous hardware infrastructure for diversified applications, e.g., AI, XR rendering

Potential Impacts on rApps

* Derive proximal computing availability for O-RAN entity based on the load/availability analysis of the computing resource types

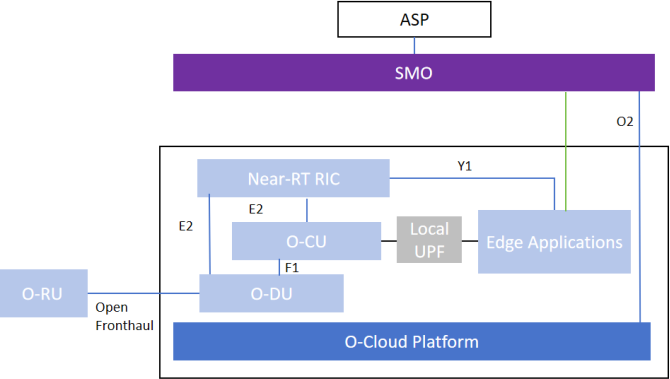
Potential O1 Impact

* O1 configurations for handover based on the proximal computing capability and availability

# Deployment Scenarios

## Scenario A: Local O-Cloud for private networks

In this scenario, the Near-RT RIC, O-CU, O-DU functions, and edge applications are all deployed on the same local O- Cloud platform as shown in Figure 9.1-1. This scenario support deployments in the industry private networks, where the edge applications, e.g., industry vision detection application can be deployed together with the 5G network to reduce the service latency and the transport bandwidth from the edge to the central cloud. In this scenario, the mobile network operator can provide the entire enterprise 5G network and edge solution. The mobile operator can choose to partner with one or more industry-specific edge application/platform provider, to deliver the end-to-end services including the 5G connection and the edge cloud capabilities. In this case, the mobile network operator can run the O-RAN NFs as well as the edge applications on its own infrastructure, i.e., O-Cloud. The O-RAN network functions and the edge applications are managed and orchestrated by the SMO. The application service provides (ASPs) may consume the services provided by SMO to orchestrate and manage the edge applications.

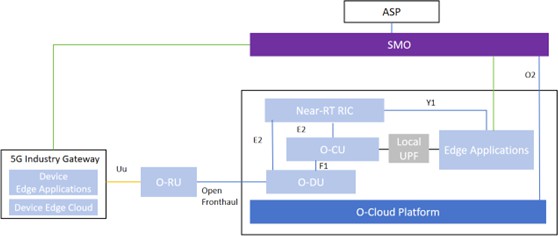


##### Figure 9.1-1: Deployment Scenario A

NOTE: UPF is out of O-RAN Scope.

## Scenario B: Device Edge Cloud and local O-Cloud for private networks

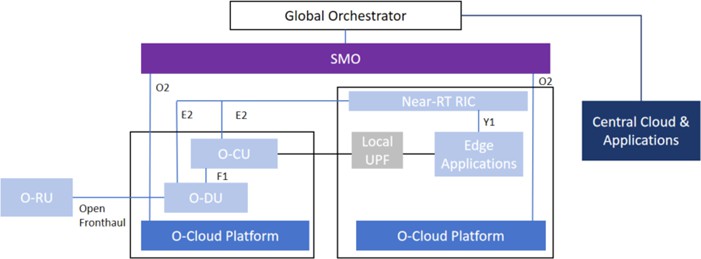
In this scenario, similar to scenario A, the Near-RT RIC, O-CU, O-DU functions, and edge applications are all deployed on the same local O-Cloud platform. Furthermore, the 5G industry gateway is introduced and serves as device edge cloud. The mobile operator can choose to partner with one or more industry-specific edge application/platform, device edge application provider, to deliver the end-to-end services including the 5G connection, the edge cloud capabilities and the device edge cloud capabilities. The SMO can be extended to also manage and orchestrate the device edge cloud and the device edge applications compared to scenario A. The deployment scenario is shown in Figure 9.2-1.



##### Figure 9.2-1: Deployment Scenario B

## Scenario C: O-Cloud coordination with global orchestrator

In this scenario, the Near-RT RIC network function is deployed on a regional cloud platform, While the O-CU and O- DU functions are deployed on an edge cloud platform that in a different location of the Near-RT RIC. Near-RT RIC can control multiple O-CUs and O-DUs and is centralized for RAN optimization based on a more global view. The deployment scenario is shown in Figure 9.3-1. SMO can manage and orchestrate the cloud infrastructure that holds both the Near-RT RIC, O-CU and O-DU. Moreover, the computing capability of the infrastructure can be exposed by SMO to the global orchestrator. So the global orchestrator can request to deploy the edge applications on the O-Cloud platform as needed.



##### Figure 9.3 -1: Deployment Scenario C

# Potential Requirements

|  |  |
| --- | --- |
|  | Description |
| REQ-SMO-01 | The SMO should support the capability to expose O-RAN computing resource information to  authorized SMO services consumers. |
| REQ-SMO-02 | The SMO should expose the RAN computing services to authorized SMO services consumers. |
| REQ-SMO-03 | The SMO should expose the RAN integrated communication and computing services to  authorized SMO services consumers. |
| REQ-SMO-04 | The SMO should provide discoverability of the RAN computing capabilities offered by the SMO  to authorized SMO services consumers. |
| REQ-SMO-05 | The SMO should provide discoverability of the RAN integrated communication and computing  capabilities offered by the SMO to authorized SMO services consumers. |
| REQ-SMO-06 | The SMO should support the capabilities to jointly orchestrate the communication and  computing resources of the O-Cloud and functionalities running over the O-Cloud . |
| REQ-SMO-07 | The SMO should support the capabilities to deploy, manage and orchestrate the applications  running on the O-Cloud. |
| REQ-RIC-01 | The Near-RT RIC should support the capability to expose the O-RAN communication and  computing analytics information. |
| REQ-RIC-02 | The Near-RT RIC should support the computing aware handover optimization. |
| REQ-OCloud-01 | The O-Cloud should support the orchestration and management of the application  deployment’s life cycle. |
| REQ-OCloud-02 | The O-Cloud should support the management of heterogeneous hardware infrastructure for  diversified applications, e.g., AI, XR rendering. |
| REQ-O1-01 | The O1 interface should support the capability of transferring the O1 configurations for  handover proximal computing capability and availability. |
| REQ-HRD-01 | The hardware reference design for O-RAN should support heterogeneous computing for O-  RAN network functions and diversified applications, e.g., edge AI, XR rendering. |

# Conclusion and Recommendations

This Technical Report presents the use cases for integrated communication and computing in O-RAN. The potential solutions are also studied for the identified use cases and key issues.

It is recommended to continue the study and specify the well identified use cases, the necessary architecture requirements and enhancements during normative work.

# Annex:

Change history/Change request (history)

|  |  |  |
| --- | --- | --- |
| Date | Revision | Description |
| 2024.07.28 | 01.00 | Version incremented for TSC. |
| 2024.07.28 | 00.00.10 | Editorial changes according to the review comments during WG1 poll. |
| 2024.07.12 | 00.00.09 | Implemented CRs   * CMCC-2024.07.04-WG1-UCTG-CR-0027-Potential Requirements for CCIN-v01.docx * CMCC-2024.07.04-WG1-UCTG-CR-0028-Add Conclusion and Recommendations for CCIN-v02.docx   Remove the optional Executive summary, introduction and the blank Annex A, Annex B |
| 2024.07.11 | 00.00.08 | Implemented CRs   * CMCC-2024.04.26-WG1-UCTG-CR-0019-Potential solution for CCIN TR key issue 1-v03.docx * CMCC-2024.04.17-ORAN-UCTG-CR-0020-addPotentialSoultionfor KeyIssue2\_v02.docx * CMCC-2024.04.01-ORAN-UCTG-CR-0021UseCase-   AddProposedSolution4MobileDeviceandRANAIInferenceCollaboration\_v02.docx   * CMCC-2024.05.30-WG1-UCTG-CR-0022-Potential solution 3 for CCIN TR-v03.docx * CMCC-2024.06.03-WG1-UCTG-CR-0023-Analysis on Potential Architecture Impacts of CCIN-v05.docx * CMCC-2024.06.03-WG1-UCTG-CR-0024-Deployment Scenarios for CCIN-v02.docx * CMCC-2024.06.03-WG1-UCTG-CR-0025-Potential Requirements for CCIN-v03.docx Editorial corrections |
| 2024.06.18 | 00.00.07 | Implemented CR  - NOK-2024-06-09-WG1-UCTG-CR-0067-CCIN TR-Solution based on multi-aspect- handover-optimization-v02.docx |
| 2024.06.04 | 00.00.06 | Implemented CRs  - CICT-2024.04.18-WG1-UCTG-CCIN-CR-0001-TR-Exposure of Performance information-v02.docx |
| 2024.05.06 | 00.00.05 | Implemented CRs   * CMCC-2024.03.21-WG1-UCTG-CR-0015-add key issue 1 for CCIN TR clause 6- v02.docx * CMCC-2024.03.21-WG1-UCTG-CR-0016-add key issue 2 for CCIN TR clause 6- v02.docx * CMCC-2024.03.21-WG1-UCTG-CR-0017-add key issue 3 for CCIN TR clause 6- v02.docx * CMCC-2024.03.21-WG1-UCTG-CR-0018-add key issue 4 for CCIN TR clause 6- v02.docx |
| 2024.03.30 | 00.00.04 | Implemented CRs   * CMCC-2023.10.11-ORAN-UCTG-CR-0008UseCase   MobileDeviceandRANAIInferenceCollaboration\_v04.docx   * CMCC-2024.01.18-WG1-UCTG-CR-0010-CCIN-add XR rendering collaboration use case for clause 5\_v04.docx * CMCC-2024.03.21-WG1-UCTG-CR-0011-CCIN- add V2X communications use case for clause 5\_v06.docx |
| 2024.03.21 | 00.00.03 | Implemented CRs   * CMCC-2024.02.13-WG1-UCTG-CR-0012-add O-RAN related work for CCIN TR   clause 4.2-v01.docx   * CMCC-2024.02.13-WG1-UCTG-CR-0013-add O-RAN related work for CCIN TR   clause 4.3-v01.docx   * CMCC-2024.03.07-ORAN-UCTG-CR-0014 CCIN phases and general challenges v3.docx |

|  |  |  |
| --- | --- | --- |
| 2024.02.13 | 00.00.02 | Implemented CRs   * CMCC-2023.10.09-WG1-UCTG-CR-0001-add scope description for CCIN TR clause 1 - v02.docx * NOK.AO-2023-10-11-WG1-CR-0001-CCIN TR-Multi aspect handover optimization- v02.docx * CMCC-2023.10.09-WG1-UCTG-CR-0002-add text for CCIN TR clause 4.1-v02.docx * CMCC-2023.10.09-WG1-UCTG-CR-0003-add NFV related work for CCIN TR clause 4.1-v02.docx * CMCC-2023.10.09-WG1-UCTG-CR-0004-add MEC related work in ETSI for CCIN   TR clause 4.1-v02.docx   * CMCC-2023.10.09-WG1-UCTG-CR-0005-add Edge Computing related work in 3GPP for CCIN TR clause 4.1-v02.docx * CMCC-2023.10.09-WG1-UCTG-CR-0006-add coordination of computing and networking related work in ITU-T for CCIN TR clause 4.1-v02.docx * CMCC-2023.10.09-WG1-UCTG-CR-0007-add computing aware traffic steering related work in IETF for CCIN TR clause 4.1-v02.docx * CMCC-2023.10.11-ORAN-CCIN-CR-0009 UseCase   RadioEdgeComputingResourceInformationExposure\_v02.docx  apply the latest TR template O-RAN.WGn.descriptor.0-v02.00-TR-template Editorial corrections |
| 2023.07.17 | 00.00.01 | Initial TR skeleton |