O-RAN next Generation Research Group (nGRG)

Contributed Research Report

Evolution of O-RAN Near-RT RIC toward 6G

RR-2025-04

Contributors:

CMCC

Qualcomm

CICT

Release date: 2025.10

Authors

Xiaofei Xu (Editor-in-Chief), CMCC

James Li, CMCC

Su Gu, CMCC

Geetha Rajendran, Qualcomm

Satashu Goel, Qualcomm

Douglas Knisely, Qualcomm

Zhangyu Luo, CICT

Yuanfang Huang, CICT

Reviewers

Rahul Soundrarajan, Tejas Networks

Michele Polese, Northeastern University

Paul Stephens, Nokia

Jan Plachy, DTAG

Qingtian Wang, China Telcom

Mike Garyantes, Qualcomm

Nurit Sprecher, Nokia

Disclaimer

Copyright

Executive summary

An intelligent and open radio access network (RAN) is crucial for the next-level user experience and global innovations in the 6G era. It is expected that the O-RAN Near-Real-Time RAN Intelligent Controller (Near-RT RIC) will contribute to realizing that vision. This report explores the architectural and functional enhancements expected for the Near-RT RIC to support emerging 6G use cases and technologies.

The key drivers for the evolution of Near-RT RIC include:

- AI/ML integration: Enhanced support for distributed AI/ML model training, federated learning, and data collection coordination to optimize RAN performance and resource efficiency. The Near-RT RIC is positioned as an edge AI/ML hub for local 6G base stations and UEs, enabling cost-effective model sharing and compute resource pooling.
- Sensing / Integrated Sensing and Communication (ISAC): Integration of sensing capabilities with communication functions to enable environment-aware optimizations, such as beam steering and interference management, by enabling collaborative sensing and low-latency edge processing.
- Service-Based Architecture (SBA): Simplification of interfaces (e.g., E2, A1, Y1) to reduce redundancy, improve interoperability, and enable direct interactions between xApps, base stations, and external domains.
- Integration with Non-Terrestrial Networks (NTN): Coordination between terrestrial and satellite-based RAN for dynamic spectrum sharing, mobility management, and energy saving in heterogeneous networks.
- Communication and Computing Integrated Network (CCIN): Support for joint optimization of communication and computing resources, including load balancing, service-aware handovers, and edge AI/ML service provisioning.

The report also discusses several perspectives that may influence the evolution of Near-RT RIC towards 6G:

- The split architecture of RAN, and the potential impact on Near-RT RIC design;
- The need for efficient AI/ML data transfer;
- The relationship between Near-RT RIC and Self-Organizing Network (SON);
- The synergy of Near-RT RIC and real-time RAN intelligence;
- The limitation in current O-RAN interfaces between SMO/Non-RT RIC and Near-RT RIC;
- The need for interworking among multiple Near-RT RICs.

A potential architecture design for the 6G Near-RT RIC is proposed in this report. The new design inherits the modularity of xApps in the 5G Near-RT RIC, and significantly improves the functionalities of the Near-RT RIC platform, including the following:

- Expanded service exposure framework;
- Unified data management across diverse data types and sources;
- AI/ML support for RAN and UE;
- Enhanced RAN service exposure;
- Support for local positioning and sensing;
- Support for interactions with neighbor Near-RT RICs

Several implementation options for the 6G Near-RT RIC are also presented, with considerations on the architecture evolution in both 3GPP RAN and O-RAN.

In conclusion, the next-generation Near-RT RIC can serve as a cornerstone for 6G RAN, featuring seamless integration of advanced technologies, simplified and unified architecture design, and new services/capabilities. This research report can provide valuable information for O-RAN evolution towards 6G, especially in making Near-RT RIC available on 6G Day 1.

Table of Contents

List of abbreviations

3GPP	3rd Generation Partnership Project
AI/ML	Artificial Intelligence / Machine Learning
AMC	Adaptive Modulation and Coding
ASN.1	Abstract Syntax Notation One
API	Application Programming Interface
CCIN	Communication and Computing Integrated Network
CP	Control Plane
CBD	Central Business District
CSI	Channel State Information
CU	Central Unit
DU	Distributed Unit
DME	Data Management and Exposure
E2SM	E2 Service Model
E2SM-CCC	E2 Service Model – Cell Configuration and Control
E2SM-LLC	E2 Service Model – Low Layer Control
FL	Federated Learning
FR2	Frequency Range 2
FR3	Frequency Range 3
FTP	File Transfer Protocol
GEO	Geostationary Earth Orbit
GPB	Google Protocol Buffers
GPU	Graphics Processing Unit
HARQ	Hybrid Automatic Repeat Request
HEO	Highly Elliptical Orbiting
HTTP	HyperText Transfer Protocol
ISAC	Integrated Sensing and Communication
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
ITU	International Telecommunication Union
LEO	Low Earth Orbit
LLM	Large Language Model
MAC	Medium Access Control
MEC	Multi-Access Edge Computing
MEO	Medium Earth Orbit
MIMO	Multiple-Input Multiple-Output
MNO	Mobile Network Operator
Near-RT RIC	Near-real-time RAN Intelligent Controller
NF	Network Function
NFV	Network Function Virtualization
NTN	Non-Terrestrial Networks
NPU	Neural Processing Unit
OAM	Operation and Maintenance
OTT	Over-The-Top
PDU	Packet Data Unit
QUIC	Quick UDP Internet Connections
RACH	Random Access Channel
RAN	Radio Access Network
RF	Radio Frequency
RRM	Radio Resource Management
SBA	Service-based Architecture
SBI	Service-based Interface
SCTP	Stream Control Transmission Protocol

SDO	Standards Developing Organization
SME	Service Management and Exposure
SMO	Service Management and Orchestration
SON	Self-Organizing Network
TA	Timing Advance
TCP	Transmission Control Protocol
TN	Terrestrial Network
TRP	Transmission-reception point
TTI	Transmission Time Interval
UE	User Equipment
UP	User Plane
UWB	Ultra-wideband
V2X	Vehicle To Everything
VSAT	Very Small Aperture Terminal
WLAN	Wireless Local Area Network

List of figures

1 Background

Apart from openness, the vision of O-RAN ALLIANCE highlights embedded intelligence in RAN architecture to optimize network operation [1]. The Near-RT RIC is one of the key enablers for RAN intelligence defined by O-RAN, which enables near-real-time control and optimization of the services and resources of RAN via fine-grained data collection and actions over the E2 interface with control loops in the order of 10 ms-1s [2]. It is deemed as an extension to the 3GPP 4G/5G RAN.

The added value of Near-RT RIC mainly comes from the following aspects:

- Support for AI/ML-enabled solutions for a wide range of RRM related use cases [3] across multiple gNBs.
- Support for third-party xApps, service-based architecture and standardized Near-RT RIC APIs [4], which provides flexible programmability in RAN.
- Support for collaboration between RAN and services. Near-RT RIC can perform RAN analytics, and expose the analytics information via the Y1 interface [5] for RAN-aware service optimization.

On the integration of AI/ML and RAN, 3GPP RAN has also made significant progress, e.g., as documented in [6][7][8][9]. It has been identified for a few use cases that model training may be performed in Operation and Maintenance (OAM) or gNB-CU, and model inference may be performed in gNB-CU. Further AI/ML enabled enhancements for the air interface have entered normative phase in Release 19. The related work will continue toward 6G and impact on the 3GPP RAN evolution.

ITU-R M.2160 [10] states the global vision on the overall objectives for 6G, identifying its key usage scenarios and capabilities, including ubiquitous connectivity, ISAC, and AI/ML.

From the O-RAN perspective, it is expected that Near-RT RIC will continue as an indispensable part for the value of O-RAN in the 6G era, and hence this study is carried out.

2 Objective and scope of this research report

The objective of the research report is to identify and analyze aspects that may impact the evolution of Near-RT RIC toward 6G. The identification and analysis would provide an outlook for the future of Near-RT RIC, and could be beneficial to the relevant standardization activities in O-RAN WGs and potentially 3GPP RAN in the next few years.

With the uncertainty of 6G RAN architecture in mind, this research report would consider the recent 3GPP RAN WI/SI progress, 6G use cases and technologies, AI/ML related advances, and technical trends in the telco industry (e.g., CCIN, service-based RAN) and their impact on evolution of Near-RT RIC.

The scope is the architectural and functional aspects related to Near-RT RIC, including its external interfaces and internal functions (i.e., Near-RT RIC platform and xApps [4]), for its evolution toward 6G.

3 High-level targets for Near-RT RIC's evolution

The following high-level targets should be considered in the evolution of Near-RT RIC:

- **Compatibility with 3GPP RAN architecture**: The 6G Near-RT RIC should maintain compatibility with the 3GPP 6G RAN architecture. This is vital to avoid fragmentation in the telecom industry.
- **Interworking with 4G/5G RAN**: The 6G Near-RT RIC is expected to support the current 4G/5G E2 Nodes based on gNB/eNB, with no or minimal upgrades for them.
- **Interoperability**: An open multi-vendor ecosystem requires full interoperability between Near-RT RIC and E2 Nodes, and between Near-RT RIC platform and xApps.
- **Energy efficiency**: Near-RT RIC and the associated AI/ML accelerators contribute to the overall energy consumption in RAN. It is desired that the deployment of Near-RT RIC will help make the RAN more energy-efficient.
- **Enhanced distributed intelligence**: AI/ML capabilities become more pervasive in UE and base stations, and hence new opportunities emerge for collaborations, e.g., between Near-RT RIC and base stations (and associated real-time RAN intelligence, e.g., proposed dApps [11]), or between Near-RT RIC and UEs.
- **Cross-domain collaboration**: The collaboration between Non-RT RIC and Near-RT RIC highlights the collaboration between RAN and its management system, and is underway in O-RAN. The collaboration between Near-RT RIC and core network functions could bring further benefits to network optimization.
- **Support for new spectrum and heterogeneous network**: The expanded spectrum and ever-increasing complexity in the network topology (e.g., NTN, sidelink, etc.) pose challenges for RAN, and Near-RT RIC is expected to help addressing them.
- **Boosted RAN performance**: 6G Near-RT RIC is expected to optimize 6G RAN KPIs.
- **Support for new services beyond communication**: In addition to the value-added services for communication, 6G Near-RT RIC should also enable novel services, e.g. AI/ML computing and data services for verticals.

4 Aspects that may impact Near-RT RIC evolution

4.1 AI/ML

4.1.1 Overview

AI is a branch of computer science to create systems for tasks which traditionally require human intelligence [12], and ML focuses more on a data-driven approach to develop algorithms (a.k.a models) that allow computers to learn from and infer with data. The most well-known AI/ML techniques include supervised learning, unsupervised learning, reinforcement learning, deep learning and federated learning, each of which has distinctive characteristics [13]. In the past decade, AI/ML techniques have been widely applied in medical care, finance, autonomous driving, and many other fields.

AI/ML is recognized as the key enabler for 6G, to address complicated problems in telco networks, which are typically high-dimensional and/or non-linear. Such problems are seen from RAN management to Layer 3/2/1 and the radio frequency (RF) hardware. In RAN management, cell planning, network and slice provisioning, anomaly detection, and service experience analysis traditionally rely on specialists understanding the thousands of parameters and their influence. In Layer 3, radio resource utilization may benefit from AI/ML-enabled load balancing, beam pattern optimization, and mobility optimization, etc., as these

optimization problems are in general intractable and classical algorithms do not provide sufficient solutions. In Layer 2, Layer 1 and below, channel state information (CSI) compression, beam prediction, localization, RF amplifier non-linearity compensation, and adaptive modulation and coding (AMC) are among the use cases where AI/ML may come into play.

Several challenges remain for the integration of AI/ML with telco networks in 6G. MNOs need to find the right balance between the cost and energy consumption of advanced AI/ML hardware (e.g., GPU), and the return on investment. The conventional network architecture and topology may not be optimal for the efficient utilization of AI/ML computing resources. The explainability of complex AI/ML models raises concerns on their robustness in telco-grade services, compared with the classical algorithms. The quality and amount of training data determine the effectiveness of AI/ML models, while the availability of data may be limited by user privacy, storage, access to cross-domain data, etc. The concerns on the disclosure of proprietary AI/ML model designs sometimes hinders the collaborative model training among base stations, or between base station and UE [7].

In the recent years, generative AI has been proved a success for various content creation tasks, and investigated for application in the telco network. One use case is related to semantic communications, where the key information from images or videos may be extracted and sent over the air, and restored at the receiver. Another direction is more dedicated to RAN, for example, CSI data augmentation could help generate abundant synthetic CSI data for AI/ML model training, from a small set of real data from commercial network.

4.1.2 Opportunities for 6G Near-RT RIC on AI/ML

A major motivation of Near-RT RIC in 5G was to enable AI/ML for RAN resource optimization. Apart from the relevant functionalities supported by today's Near-RT RIC, further opportunities for Near-RT RIC may rely on efficient utilization of AI compute resource, and extensive support for AI/ML data and models. Some examples are as follows:

AI/ML model provider for RAN and UE

Base stations may perform AI/ML model training as specified by 3GPP TS 38.300 [8]. The tasks are usually compute-intensive but mostly intermittent, i.e., when a model's performance deteriorates below a threshold. It could lead to low resource utilization if powerful AI/ML hardware is mounted at every base station to support model training. A better approach would be to have an AI/ML-oriented edge node, which enables statistical multiplexing of the model training tasks from nearby base stations. Naturally, the Near-RT RIC could be the local RAN node to host a cluster of AI/ML hardware and serve the base stations for model training/refinement as well as associated data collection and cleaning. In this way, the 6G Near-RT RIC would play a role to relieve the unnecessary burdens from RAN, and reduce overall investment from operators. In addition, the Near-RT RIC may also serve the UEs for their model training needs.

RAN data collection coordination and storage

The AI/ML enabled RAN optimization solutions in 3GPP 5G RAN require a base station to collect data from adjacent base stations. As the base stations are densified due to reduced cell radius and additional capacity layers, a single base station may have to handle similar data collection requests from a large number of neighbors. Together with that burden is the lack of support for historic data collection, which may limit the performance of AI/ML

solutions. A Near-RT RIC may be designed to play a role similar to the Data Collection and Coordination Function (DCCF) and Analytics Data Repository Function (ADRF) in the core network [14], but dedicated to data collection and storage ain RAN, and sufficiently flexible to support specialized requirements from various xApps.

Cross-domain data collection

Near-RT RIC may help in data collection from other domains, e.g., core network, OAM and application services. Enabling RAN's awareness of the UE's service performance may help improve user's quality of experience (QoE), a concept already reflected in the standardization of RAN visible QoE measurements from UE [8]. In addition, the dynamic information from the network-side application servers could also be useful for RAN to optimize its scheduling. That would require the RAN to have access to such dynamic information. A possible mechanism for that is to extend the network exposure framework in the core network [14]. The enhanced mechanism could support RAN to collect useful information (e.g., predicted traffic load) from the core network as well as the service applications, to enhance efficiency in RAN operation.

Federated learning (FL)

Federated learning enables collaboration among multiple FL clients to train a global model by exchanging model parameters. The advantage is the training data are processed locally at the FL clients. When a Near-RT RIC is connected with multiple base stations, it may act as the FL server. As another example, Near-RT RIC may act as a FL client, when the RAN OAM or a core network function launches a wide-area model training task.

Edge repository of AI/ML models

The AI/ML models for future RAN can be well generalized across different scenarios (locations) and vendors, but more likely to require additional fine-tuning. A fine-tuned model could apply for multiple devices in the same or similar scenarios (e.g., a stripe of base stations serving a high-speed railway). Obviously, such fine-tuning may be performed once, and shared locally with all the similar devices. Near-RT RIC may be used as the model repository for that purpose. Besides, the advanced generative AI models (e.g., Large Language Models (LLM)) are typically large in size, and hence bandwidth-consuming for the backhaul if they are stored in the central cloud.

4.2 Sensing / ISAC

4.2.1 Overview

Sensing techniques may be used to detect range, angle and velocity of target objects, based on the reflection of electromagnetic waves by them. Due to the advances in semiconductor technology, low-cost sensing is possible and its application has extended from military (i.e., radar) to civil scenarios like automatic driving. The radio frequencies used for today's civil sensing are largely 7.163-8.812 GHz (UWB), 24 GHz, 60 GHz and 67-79 GHz, where directional antenna and narrow beams can be exploited. Meanwhile, beamforming techniques have become an indispensable part of 3GPP standards with its exploration in FR2 and FR3, leading to the convergence of sensing and communication.

ITU has identified ISAC as one of the 6G usage scenarios [10], and use cases including intruder detection, gesture recognition, and heart/respiration rate monitoring have been studied in 3GPP [15]. Six sensing modes are defined for 3GPP TRPs and UEs, namely, TRP-TRP bistatic, TRP monostatic, TRP-UE bistatic, UE-TRP bistatic, UE-UE bistatic, and

UE monostatic. It should be noted that, while the integration of sensing and communication may lead to performance trade-offs between the two purposes, sensing can also be helpful to communication, for instance, optimized radio configurations for beam patterns and transmit power, multiuser scheduling, digital twin, etc.

ISAC poses new challenges as well as opportunities for 6G RAN. Apart from the co-design challenges in the hardware, the physical layer and the networking aspects also require innovations. In the physical layer, a single waveform that may serve both purposes and support flexible allocation between is desirable. In particular, when monostatic sensing (i.e., a single node is used for both transmitting and receiving the sensing signal) is conducted, the transceiver design has to address the strong self-interference. The challenges in terms of the network sensing include interference management, switching of sensing beams/modes/frequencies, joint signal reception/combination, and synchronization.

A number of synergies between ISAC and other techniques may be envisioned. The synergy with positioning, which has been well standardized in 5G RAN [16], will enable the environment awareness about unconnected objects as well as connected devices. The non-3GPP sensing techniques (e.g., 802.11bf WLAN sensing) may collaborate with 3GPP sensing to improve sensing accuracy. What is more, the advent of reconfigurable intelligent surfaces (RIS) may help extend the coverage of sensing.

4.2.2 Opportunities for 6G Near-RT RIC on sensing/ISAC

Among the multitude of potential enhancements on different parts of RAN, the opportunities for Near-RT RIC from sensing/ISAC could arise from the network level. Besides, as sensing data could have various forms and sources (e.g., raw signals from the base stations, or intermediate estimates from UE via user plane, and so on), new designs in Near-RT RIC could be necessary to support this technology, with considerations on data rate, security, privacy, etc. Some examples are as follows:

Multi-static sensing

Multi-static sensing aims to improve sensing accuracy by the fusion of sensing data from multiple sensing nodes. Some analysis [17] reveals that multi-static sensing could outperform mono-static and bistatic sensing due to the non-uniform radar cross-section (RCS) in different angles. To harvest the gain from such collaborative sensing schemes, joint processing of the received signals is essential. The joint processing could be executed with raw signal samples, intermediate estimates, or sensing results from the multiple receiving nodes, with a descending order in sensing accuracy. For some extreme scenarios (e.g., centimeter-level sensitivity), joint processing of the raw signal samples would be desired, but the huge data volume would be an issue if the samples are aggregated to the core network for processing. In such cases, a preferable approach could be a RAN node capable of collecting and processing the data samples from the coordinating base stations. Near-RT RIC could play as such a RAN node for centralized processing of sensing data. Meanwhile, a joint controller to select the participant base stations and to allocate proper radio resources for sensing is required, and the Near-RT RIC may also serve this purpose.

AI/ML-assisted sensing

AI/ML techniques may be helpful in improving the performance of sensing as well as communication. However, the AI/ML operations, especially model training, is compute-intensive and storage-consuming for training data. It may be not economical to design every base station with sufficient compute and storage resources for AI/ML-assisted sensing, especially when an affordable network is desirable for the less-developed areas. A more

cost-effective approach could be a local Near-RT RIC that serves multiple base stations with sensing data storage and processing.

Low-latency sensing

Quite a few vertical and entertainment use cases for sensing [18] require a sensing latency of 10~1000 ms. An even smaller latency may be desired, when the sensing results are used by RAN for scheduling (e.g., beam steering). The lower end of such latency requirement could be difficult to achieve, if the sensing result is calculated and exposed from the sensing function placed in the central cloud. For such use cases, a local node at the network edge would help with the prompt delivery of sensing results. This node could be a localized core network function, or possibly a RAN function, given that some sensing results may be not concerned with UE privacy. Near-RT RIC could be the local node to offer sensing results. Moreover, when non-3GPP sensing is available, the collaboration between two sensing systems could also be bridged by Near-RT RIC.

Interference management

As the radio frequency used by ISAC could be in FR2 and FR3, beamforming techniques will be extensively used for high directivity. Accordingly, the intercell interference may be severe, as the sensing receiver of the echo signal also receives signals from the neighbor cells. Such interference in ISAC deployments could be even worse, as the conventional antenna down-tilt design has to take into account many UAV related use cases [18]. To that end, the interference management for ISAC requires more attention in the system design. The existence of a centralized RAN node, like Near-RT RIC, may be more efficient than a fully distributed RAN, in terms of coordinating the time and frequency resources scheduled by different cells for interference avoidance.

4.3 SBA

4.3.1 Overview

Service-based architecture is an architecture paradigm originated from the IT domain, which attempts to improve the flexibility of a complex system by decoupling it into a set of services. A few guidelines may be applied for SBA:

- A service can be consumed by authorized service consumers via a well-defined interface;
- The services should be loosely coupled with minimized dependencies between;
- The services should be stateless, separating the service invocation and session context;
- The services should be self-contained for independent deployment and upgrade;
- The services should be able to be discovered.

3GPP has integrated this paradigm into the control plane of the 5G core network (5GC) since its Release 15. A network function (NF) hosts NF services and exposes them via its service-based interface (SBI). The Network Repository Function (NRF) offers the functionalities of service framework (i.e., service registration and discovery).

The SBA for 5GC is successful despite some remaining issues. The architecture enables on-demand deployment, dynamic adaptation to service availability, flexible horizonal scaling and independent upgrade of the NFs. On the other hand, a practical rule might be absent for properly decoupling and combining the services (into NFs). Nowadays, more than 60 NF

types have been defined with considerable signaling overhead, and when a new feature is introduced, many NFs are often subject to upgrade. Moreover, the signaling protocol could be imperfect. The HTTP/2 protocol still suffers from the head-of-line blocking issue from TCP, and the JSON encoding is less efficient on the wire in spite of its readability.

With its potential benefits, SBA has become a topic for the 6G RAN architecture. Possible options for service-based RAN include partial service-based and full service-based. The partial service-based option could start from the control plane interface between RAN and core network (i.e., N2 interface in 5G), with part of or all of its functionalities. The other option, full service-based, may be accompanied with the refactorization of some RAN and core network functions (e.g., simplified connection management). Note that the feasibility of service-based RAN should take into account the nature of many RAN functions:

- Most processing in RAN is a chained workflow. One example is the channel coding – rate matching – modulation – resource mapping flow in the physical layer, and as another example, a lower protocol layer typically treats PDUs from the upper layer as payload and adds some header. In contrast, SBI is more efficient when multiple service consumers exist for a single service producer.
- The low layers of RAN are real-time (i.e. per TTI). Converting the data and signaling to/from network interfaces could incur additional latency.
- The UE context in the low layers is kept proximate to the UE for mobility performance. Separation of such state information from the processing logic could be inefficient.

4.3.2 Opportunities for 6G Near-RT RIC on SBA

Near-RT RIC in 5G leverages quite different approaches for its interfaces. The A1 and E2 interfaces are defined with the reference point approach, between Near-RT RIC and a specific entity (i.e., Non-RT RIC and E2 Node, respectively), while a set of services are defined on A1. Y1 interface is a pure SBI with non-specific service consumers. The Near-RT RIC internal architecture (for the interactions of xApp and Near-RT RIC platform) is defined with the SBA approach, where the Enablement APIs serve for service registration and discovery purposes.

The further adoption of SBA for Near-RT RIC may simplify the architecture and improve efficiency. Some examples are as follows:

Reduced redundancy across E2, A1 and Y1 interfaces

There may be similar services produced by a Near-RT RIC for different service consumers. A use case investigated by the work item “AI/ML for O-RAN” may lead to an AI/ML model training service defined on A1 interface, for which Near-RT RIC is the service producer. Another use case would study the possibility of Near-RT RIC to train AI/ML models for the E2 Node, which might introduce enhancement for E2 interface to expose such a capability. The situation might occur also for the RAN Analytics Information services, which is already defined on the Y1 interface but it is also being discussed to backport similar services to E2 interface to facilitate E2 Nodes to access the analytics. With the SBA, it could be favorable to consolidate those similar services into a single SBI offered by the Near-RT RIC, which would reduce the potential overlap.

Direct communication between xApp and E2 Node

Currently, SBA is only applied for internal Near-RT RIC architecture, invisible from outside. With that, the Near-RT RIC platform isolates the xApps from the E2 Node. Such a design has its merits but might cause inefficiency in some cases. One example is the xApp queries

E2 Node information. Under the current architecture, an xApp has to request the Near-RT RIC platform with the query, and the Near-RT RIC platform essentially forwards the query to the E2 Node. The query response comes back to the xApp in two hops as well. Efficiency may be improved by allowing the xApp to query the E2 Nodes directly. In another example, supposing the E2 Node would like to access a service produced by a specific xApp, it is currently unable to discover such services. The issue could be solved by allowing the E2 Node to leverage the service exposure framework of the Near-RT RIC platform. From this perspective, it is probably worth a future study to bridge the xApps and E2 Nodes with SBA.

Single protocol stack

The interfaces of Near-RT RIC have different protocol stacks due to different considerations. The E2 interface follows mostly the 3GPP RAN network interfaces based on Stream Control Transmission Protocol (SCTP) / Abstract Syntax Notation One (ASN.1), with an exception that uses JSON as the encoding format (i.e., E2SM-CCC). The A1 and Y1 interfaces both use HTTP/JSON, the de-facto solution for cloud and used by 3GPP OAM and core network. The Near-RT RIC APIs used both SCTP / Google Protocol Buffers (GPB) and gRPC [19]/ GPB stacks, which is a compromise between performance and developer-friendliness. It is expected that the SBIs for Near-RT RIC in 6G could have a single protocol stack that strikes a best balance among various requirements, and hence minimize the implementation efforts and improve operation efficiency. So far, HTTP/3-based solutions are quite promising.

Simplified E2 interface design

The complete application protocol of E2 interface is quite complicated with its nested structure. Tier-1 (outermost) is the generic procedures and messages with a set of container information elements (IEs). Tier-2 (middle) is the so-called E2 service model (E2SM), encapsulated in the tier-1 container IEs, each of which is associated with a specific data collection / control functionality exposed by E2 Node. In some E2SMs, a tier-3 (innermost) structure is used to convey the actual 3GPP parameters or measurements. An outer tier is agnostic to the content of the inner tier. With SBI, it is possible to simplify the design by manifesting the E2SMs as individual services produced by the E2 Node, since the E2SMs have been divided to minimize interdependency. That could improve the readability of the standards as well as reducing complexity in implementations.

4.4 NTN

4.4.1 Overview

NTN refers to the networks that leverage satellites or unmanned aircraft acting as a relay (transparent payload) or base station (regenerative payload) [20]. It can help provide ubiquitous coverage as well as resiliency for traditional terrestrial networks.

Satellite-based NTN has attracted wide interests from the telco industry, and is the topic in this report. In terms of the altitude of the satellite orbits, NTN may use LEO (Low Earth Orbit), MEO (Medium Earth Orbit), GEO (Geostationary Earth Orbit) and HEO (Highly Elliptical Orbiting) satellites, with different propagation delay/loss and ground coverage. In terms of the payload type, NTN may have transparent or regenerative payloads.

While NTN often benefit from line-of-sight (LOS) transmission, challenges come from a few aspects. Besides severe Doppler shift, the high velocity of LEO and MEO leads to rapid variation in propagation delay/loss, which necessitates advanced beam/power control, mobility management and feeder-link management. The high altitude introduces additional

propagation delay, which requires adaptation on the legacy Random Access Channel (RACH), Timing Advance (TA) and Hybrid Automatic Repeat Request (HARQ) designs. The tropospheric and ionospheric effects on the frequencies also need to be considered in the system design.

A generic architecture for NTN consists of the terrestrial segment (terminated by NTN gateway), the satellite segment, and the user segment (which may be Very Small Aperture Terminal (VSAT) or UE). The segments are connected by the feeder link and the service link. The satellites can be interconnected via inter-satellite links (ISL). From a RAN perspective, different architecture options have been discussed, roughly divided into two groups, i.e., “gNB on earth”, and “gNB (part or all) on satellite”:

- “gNB on earth”: the feeder and service links both convey the Uu interface, or both the NG interface;
- “gNB (part or all) on satellite”: one option is the whole gNB on satellite, where the feeder link conveys the NG interface and the service link conveys Uu interface. Another option may be the DU on satellite and the CU on earth, where the feeder link conveys the F1 interface and the service link conveys Uu interface.

3GPP started normative work on NTN support since its Release 17, with an emphasis on transparent payload [20]. From Release 19, NTN support for regenerative payload will be available as well as various enhancements (e.g., downlink coverage, RedCap UE, etc.). Relevant work has also started in O-RAN, with transparent payload as a first step.

4.4.2 Opportunities for 6G Near-RT RIC on NTN

The current design of Near-RT RIC is dedicated to terrestrial networks. While the Near-RT RIC for 6G might be a satellite payload, a realistic consideration may be to start from Near-RT RIC deployed in terrestrial network (TN). In this direction, some research work has emerged (e.g., [21]), and some examples are given as follows:

Coordination between NTN and TN: The integration of NTN and TN-based RAN will entail innovative solutions for the existing use cases and features, e.g., traffic steering and multi-connectivity. In particular, mobility optimization is essential to ensure service continuity for UEs, and dynamic spectrum sharing between TN and NTN could help improve spectrum utilization.

Coordination between radio network layer and transport network layer: The existing RAN network interfaces (i.e., NG, E1 and F1) assume reliable transport (e.g., optical fiber) for terrestrial networks, but the assumption may not hold for the service link of NTN. For instance, Near-RT RIC may be enhanced to assist the dynamic routing configurations for the LEO/MEO NTN gateways, in order to minimize the possible packet loss/delay in service link switching, based on the known satellite movement (ephemeris) and the availability of NTN gateways.

Energy saving: Energy saving is a critical design aspect for NTN. The satellite platforms are typically power- and energy-constrained with limited batteries and intermittent supply from photovoltaic unit, and so the on-board NFs have to adapt to that. In addition to the features investigated for terrestrial networks (e.g., cell switch-on/off and RF channel reconfiguration), dynamic reconfiguration of the NFs may be exploited. As a possible scenario, the satellite network operator may indicate the battery’s state of energy and/or power headroom to mobile network operators (MNOs). MNOs may choose to trigger low-complexity signal processing algorithms, or even the migration of some on-board functions (e.g., CU) to the ground if the MNOs have proper infrastructure.

4.5 CCIN

4.5.1 Overview

As a significant trend for IMT-2030 [10], ubiquitous computing involves the computing capability at the far edge of networks. ETSI MEC [22] has pioneered in this direction, offering support for local services with latency down to ~10 ms. New concepts are also emerging, including CCIN [23], and also more recently, AI-RAN initiated by Nvidia [24]. The common idea among such concepts is a deeper integration between communication and computing in RAN, which promises even lower latency for local services and improved utilization of RAN infrastructure.

Different from the conventional RAN designs dedicated to communication, the sharing of RAN infrastructure between communication and computing poses new challenges. Performance assurance for both communication and computing services may be difficult, given the limitation in the available resources and the fluctuation in the service loads. UE mobility may require prompt context transfer for the computing services across the RAN sites. The user plane traffic would need local breakout from RAN to data network, raising concerns on charging and user privacy. Increased complexities could incur in the collaboration of the such computing capabilities with those in MEC, central cloud and UE.

4.5.2 Opportunities for 6G Near-RT RIC on CCIN

In the current O-RAN architecture, SMO is responsible for the orchestration of RAN infrastructure, and Near-RT RIC is solely a NF for communications. Despite that, Near-RT RIC could possibly play a role in the deep convergence of communication and computing, with its potential for near-real-time optimization.

Load balancing across RAN sites

The available resource at a RAN site is much less than that in a central/edge cloud. Careful resource allocation would be a must for RAN in order to provide computing services beyond communication services. A simple and reasonable rule may be to guarantee the basic communication services, and then offer best-effort computing services. When a RAN site needs more resources for communication at some point of time, and hence cannot provide sufficient resources for computing, it may be possible to migrate its computing workload to a nearby under-utilized RAN site or edge cloud. The Near-RT RIC may be enhanced to monitor the resource utilization of multiple RAN sites, and to trigger the proper migration and routing configurations.

Information exposure for joint communication and computing optimization

XR rendering collaboration is analyzed as an interesting use case in the CCIN TR [23], which enables lower end-to-end delay by balancing the communication load and the computing load. An enabler of the solution is to have Near-RT RIC for analytics information exposure. The Near-RT RIC may collect near-real-time load information and provide recommendations for the application server, so that the application server may optimize its operation mode.

Enhanced multi-aspect handover:

Another use case studied by the CCIN TR [23] is the handover optimization that considers the available computing resource for the UE's service(s). The solution therein involves Near-RT RIC for handling the handovers based on policies from Non-RT RIC. An enhanced

solution may be to leverage the Near-RT RIC to collect the location information as well as the computing resource load information, and predict UE trajectory in a more real-time manner, so that the handover performance may be further improved. Besides, Near-RT RIC may be used to trigger proactive context transfer for the UE services across the RAN sites, which could help with better service continuity.

AI/ML service and generic computing service provider:

Near-RT RIC may be utilized to provide AI/ML service (e.g., AI/ML model training and inference) for over-the-top (OTT) applications, especially those from the verticals. The vacant AI/ML compute, the AI/ML tools and environments, and data access could make the Near-RT RIC a favorable node to deploy AI/ML-based OTT applications in a cost-efficient manner. Apart from the AI/ML services, Near-RT RIC is also equipped with generic compute. Such generic compute may also be used for local deployment of OTT applications, so as to maximize the resource utilization in the Near-RT RIC.

5 Other considerations

5.1 CU/DU split

Before 5G took its shape, 3GPP RAN studied the split architecture of gNB in 3GPP TR 38.801 [25]. Eight functional split options plus a flexible functional split options were evaluated, of which Option 2 was favored as the higher layer split (HLS) and shaped the 5G CU/DU. The lower layer split (LLS) was not concluded in 3GPP, but further developed in O-RAN.

Despite the standardization efforts on the F1 interface, the main-stream 5G deployments come with non-split gNBs. The reasons behind could be multifold, e.g.:

- Additional investment in the centralized site (difficult to reuse 4G sites for CU, due to the high demand for room and power supply), and increased operational complexity;
- Limited pooling gain from small-scale CUs (where the tide effect of UE traffic is not significant within the coverage), but the cost of NFV-based solutions remains higher than purpose-built hardware;
- Low demand from verticals for customization and programmability.

As technologies and service requirements progress toward 6G, there is an opportunity to revisit the split design. The applicability of NFV was one of the considerations for the HLS, but nowadays the DU may be highly virtualized/cloudified with the aid of accelerators. Sensing has emerged as a prospective 6G service, and one possible 5G-A solution under validation is to upgrade some gNBs with extra sensing processing capability, each serving for several nearby gNBs. The local services (e.g., V2X) may boost the deployment of edge service platforms like MEC, which could help reducing the investment by sharing the sites and infrastructure with CU. The increased network densification and heterogeneity could intensify the interference issues and may call for extension of centralized RRM to a more real-time realm, as reflected in the cell-free MIMO concept. All the above may have implications on the need for different split architecture in 6G, which may be similar to the 5G HLS (i.e., centralized RRM) or not (e.g., centralized RRM and centralized scheduling, or centralized compute for specific tasks). Meanwhile, we might have to recognize that the non-split base stations may remain as a cost-efficient deployment option for the basic communication services in most parts of the world. To that end, the 6G era may see both split and non-split base stations in commercial deployments, catering to the diversified business requirements.

SDOs also have an opportunity to adjust their roles in the process. The purpose of standardization is interoperability, and it could be a burden to standardize every possible functional split. It remains imperative for 3GPP RAN to play as the overarching leadership role, which specifies the overall function in RAN (e.g., eNB or gNB, which can offer all RAN services on its own), and ensure the essential interoperability (i.e., with the peer entities, the core network, and the UE). However, when it comes down to the finer-grained split options, a wait-and-see strategy could be considered unless the market demand is validated, and such standardization work may well be conducted by SDOs other than 3GPP (e.g., the O-RAN ALLIANCE) for efficiency.

As in the 5G era, the 3GPP RAN architecture in 6G will be foundational for the future of Near-RT RIC. Depending on the potential functional splits within and beyond 3GPP RAN, it is possible that the notion of Near-RT RIC in 6G may be different from today, and could be realized in more than one option. For example, when the 6G base stations are fully distributed, a Near-RT RIC can serve multiple base stations; or when a centralized RAN node is defined in 3GPP, O-RAN can consider extending the centralized RAN node to host the xApps. Whatever its shape will be, it is expected that the Near-RT RIC for 6G will maintain compatibility with, and add value to, the 3GPP RAN architecture.

5.2 Beyond CP and UP separation

As a communication-oriented network, 5G RAN introduced the separation of control plane and user plane for its CU part, with the understanding of the complication in separating the planes for low layers. In O-RAN, the Near-RT RIC and its associated interfaces (E2, Y1, Near-RT RIC APIs, etc.) are specified solely on the control plane.

As mentioned in clause 4, 6G RAN will be evolved with native beyond-communication capabilities (AI/ML, sensing, etc.). Accompanied with those capabilities are the additional requirements on bulk data collection/storage (e.g., model training data) as well as processing. In particular, AI/ML processing would be highly dependent on dedicated accelerators (e.g., GPU, NPU).

Similar to the one-to-many association between 5G CU-CP and CU-UPs, further flexibility and extensibility could be considered to separate the data and AI/ML related functions into the control part and the execution part. The execution part mostly accounts for the compute and storage resources. It would have implications on both architecture and interface designs. In terms of architecture, a new type of RAN NF might be defined for the execution part. For interface design, the signaling and the bulk data could have distinct protocol stacks (e.g., Kafka, FTP, etc.). Such designs, if realized in 3GPP, will surely have impact on the evolution of Near-RT RIC.

5.3 Relationship with SON

SON has been introduced into 3GPP since around 2010. It has evolved with mainly three functional groups, i.e., self-configuration, self-optimization and self-healing. In 3GPP, SON is mostly specified at concept level, and various implementations may exist. From an architectural perspective, SON may be realized with three styles, i.e., centralized SON (C-SON), distributed SON (D-SON) and hybrid SON (H-SON) [26]. The realistic deployments of SON have been limited, especially for multi-vendor RAN, partly because the management interfaces are typically vendor-specific or not fully open.

O-RAN contributes to the application of SON with its open interfaces associated with Non-RT RIC and Near-RT RIC. Apart from its role in RRM enhancements, Near-RT RIC is also

related to D-SON, which operates at the NF layer. The Near-RT RIC also collaborates with Non-RT RIC to support the H-SON concept. In addition, the added value of Near-RT RIC for the mainstream 5G RAN deployments also come from its flexibility (i.e., xApps). It is expected that Near-RT RIC in 6G will continue as an open and customizable technical realization of SON, and go beyond that with new added values (e.g., local AI/ML resource pool).

5.4 Relationship with real-time RAN intelligence

O-RAN for 5G outlines the three control loops for RAN in terms of time scale, i.e., the non-real-time, the near-real-time and the real-time. Non-RT RIC and Near-RT RIC are specified so far for the corresponding RAN intelligence, while the real-time RAN intelligence is left out of scope due to its tight coupling with the MAC/PHY layers.

In recent years, O-RAN has started to explore the possibility of improving real-time RAN intelligence with more flexibility, represented by the research on dApps [11]. Several exciting use cases have been identified, including but not limited to, direct handling of I/Q samples (for communication as well as sensing), and real-time scheduling.

It is desirable that the potential dApps will have synergy with the xApps in the 6G O-RAN. One possibility is that the two types of software could be unified under a common framework, and SBA as discussed in clause 4.3 might be useful (though the latency could be a challenge). The dApps and the xApps could be paired to collaborate in some use cases, e.g., for sensing/ISAC, the dApps can act as a first processing stage on the raw radio signals, and share with the xApps a processed version of them for refinement/fusion. The introduction of dApp could also help with the refinement of some Near-RT RIC functionalities, e.g., the E2SM-LLC (the E2 service model for low-layer control) is currently defined for E2 interface, which could be reallocated as part of the RAN services for dApps and with more capabilities.

5.5 Limitations of current O-RAN architecture interfaces between SMO/Non-RT RIC and Near-RT RIC

SMO and Non-RT RIC are the key functionalities in O-RAN architecture for RAN management and orchestration, which have close collaboration with Near-RT RIC as indicated by the non-real-time and near-real-time control loop [2]. However, several limitations can be observed in the current architecture as described below:

- **SMO services not accessible to Near-RT RIC**: Service management and exposure services (SME) and Data management and exposure services (DME) are available in SMO but are not accessible to the Near-RT RIC. New custom procedures (often unique for each separate use case) are needed over the A1 or O1 interface (or any new interface that performs similar services) to access these common services. For example, RAN analytics exposed over Y1 interface could have been easily done via DME and SME. For example, since the Near-RT RIC cannot access DME and SME in SMO, the Y1 interface had to be developed to expose RAN analytics from Near-RT RIC. As another example, AI/ML services are expected to be common services that almost any entity in O-RAN architecture should be able to use. Instead, AI/ML services over A1 are being defined to expose AI/ML services in SMO to Near-RT RIC.
- **Limited rApp and xApp coordination**: rApp in Non-RT RIC and xApp in Near-RT RIC can only interact with each other via services in corresponding RICs that

communicate over A1 interface. To coordinate on policies, the applications need to utilize A1 policies defined in O-RAN specifications. Applications are dependent on RICs and A1 interface specification for coordination which makes it difficult to extend the coordination.

- **Multiple interfaces towards E2 Nodes**: E2 Nodes (CU and DU) are connected to SMO over the O1 interface and to the Near-RT RIC over the E2 interface. It is not straightforward to integrate the information elements conveyed on these two interfaces in a flexible manner, e.g., easily develop a new use case that requires combining this information. Some use cases also duplicate information on E2 Node over O1 interface and vice versa e.g., the Filtered KPI use case.

To address those limitations, the following enhancements can be considered:

Near-RT RIC communicates directly with SMO services

- The Near-RT RIC (or its functional equivalent in 6G) could be extended or modified to communicate with services in SMO by permitting direct access to the SMO communication bus (subject to appropriate access control and security restrictions).
- This would enable the Near-RT RIC to access SMO services such as AI/ML Workflow services, Data Management and Exposure (DME), Service Management and Exposure (SME) and would avoid re-inventing the same functions in Near-RT RIC.
- Similarly, the Near-RT RIC could easily expose fine-grained RAN policy control and analytics information (e.g., processed information received over the E2 interface) to SMO services, which would avoid duplication of similar information or functionality over the O1, E2 and Y1 interfaces.

Replace P2P E2 interface by service-based interface

- In the ideal case, a service-based interface towards E2 Node(s) would allow different consumers (RIC, SMO, etc.) to leverage the same services and to access the same information/analytics without unnecessary duplication of functionality.
- SMO services could access E2 Nodes and vice versa to consume data or to provide control functions, information, and policies.
- This can be viewed an extension from the discussion of SBA in clause 4.3.

5.6 Near-RT RIC interworking

Though interworking of multiple Near-RT RICs is not supported yet in 5G, it could be essential in 6G era.

A first use case is AI/ML data transfer. It is mentioned in clause 4.1 that Near-RT RIC in 6G could act as a local RAN data storage. However, the local RAN data from a single Near-RT RIC may still be insufficient to train an AI/ML model specialized for a specific scenario, e.g. CBDs, residential areas, high-speed railways, heavy traffic areas, etc. It may be more efficient to aggregate such data for centralized processing, which requires 6G Near-RT RICs to transfer specific RAN data to a selected Near-RT RIC.

Another associated use case is AI/ML model transfer. The 6G RAN will be pervasive with numerous AI/ML models for scheduling, channel estimation, beamforming configurations, etc., among which a few well-generalized would accommodate most of the scenarios. With that assumption, the most outstanding AI/ML models, which may be trained in a single Near-RT RIC, would be deployed in multiple Near-RT RICs. Accordingly, it would be necessary to transfer such AI/ML models between Near-RT RICs.

In addition, the xApps on different Near-RT RICs may benefit from sharing information enabled by this mechanism, e.g., [27].

6 Potential Near-RT RIC architecture design for 6G

6.1 General

This clause provides a perspective on what the 6G Near-RT RIC may look like. It also gives a few implementation examples, which are based on several hypothetical 3GPP 6G RAN architecture options.

It should be noted that the 3GPP 6G RAN architecture will be the basis for the design of next-generation Near-RT RIC. It is expected that a high-level consensus on the 3GPP 6G RAN architecture will emerge when a dedicated Release 20 study item concludes, and based on that, the next-generation Near-RT RIC will take its shape at 6G's Day 1.

6.2 Functional entities of 6G Near-RT RIC

The modularity enabled by the xApps should continue for the 6G Near-RT RIC. The xApps are software packaged and highlight flexibility in development and deployment for various use cases. The Near-RT RIC platform offers generic support for the xApps in operation.

6.3 Interface design

6.3.1 Service-based control plane

For the 6G Near-RT RIC, the service-based approach may be considered on the control plane, to unify the communications among the xApps, the Near-RT RIC platform and the base stations. Each entity exposes its set of services via its service-based interface, which can be accessed by any authorized consumer entities.

A single protocol stack should be defined for those service-based interfaces. In the past, a few solution sets were evaluated for the Near-RT RIC APIs, including SCTP/ASN.1, HTTP/JSON, and gRPC/GPB (Google Protocol Buffers), but none is perfect. SCTP/ASN.1 is simple but less developer-friendly, and the support for cloud-native environment and model security solutions is limited. HTTP/JSON is the solution for 3GPP 5G core network, but JSON trades encoding efficiency for better readability. gRPC/GPB is subject to the same TCP head-of-line blocking as HTTP/JSON, and it is a company-sourced solution that is not maintained by an international standardization organization. The emergence of QUIC in the recent years addresses some of the above concerns, and could reshape the service-based interfaces in the future. So far, implementations of HTTP3 over QUIC and gRPC over HTTP3 are available in open-source communities, and further evaluations would be helpful for the selection of the 6G Near-RT RIC's interfaces.

6.3.2 Flexible data path

A potential performance bottleneck of Near-RT RIC interfaces is from the massive RAN data collection. In the current Near-RT RIC design, the collected RAN data are multiplexed with the signaling for RAN configuration/control and that for data collection control. When network

congestion occurs due to heavy load, some time-sensitive RAN configuration/control signaling may be delayed or lost, degrading the service quality of Near-RT RIC. The issue could be more severe when the AI/ML models and sensing data are transferred on the interfaces. To that end, more flexibility in data transfer options be considered, tailored to the diversified data collection and model transfer tasks.

The new AI/ML data plane should support multiple types of communication. Streaming is a must for continuous collection of RAN data, a task for which the conventional HTTP-based protocols are not designed. File transfer such as FTP is also necessary for bulk data, especially when the data size is huge. Message brokers like Kafka may also be among the non-exclusive options to support a large number of xApps consuming the same RAN data. In short, multiple protocols could be defined for the data path, and selected for different type of tasks.

6.4 Key functionalities of Near-RT RIC platform

6.4.1 Service exposure framework

The Near-RT RIC platform could provide support for the service exposure framework in the control plane. It includes service registration, service discovery and event notification for the xApps. With that, authorized xApps could communicate directly with the base stations and even the functions in the other domains.

6.4.2 Unified data management

A unified framework for data collection is expected for the 6G Near-RT RIC. For historical reasons, Near-RT RIC handles the various types of data via different APIs including SDL APIs, E2 related APIs and A1 related APIs, each of which has a dedicated API design. With the 6G Near-RT RIC, we have the opportunity to refactor those designs, and come up with a generic set of APIs for data collection. It will reduce the efforts in standardization and also accelerate the development of products.

The same philosophy applies to data storage as well. A set of standard APIs may be designed to support the standardized types of data, and offer sufficient flexibility to the vast number of unstandardized types of data used by the xApps.

The AI/ML data generation/augmentation can be another relevant functionality of the future Near-RT RIC platform. By analyzing/learning the patterns in the real RAN data, the Near-RT RIC platform may produce high-quality synthetic data to improve AI/ML model training.

6.4.3 AI/ML support for RAN and UE

The 6G Near-RT RIC could become a local RAN node for AI/ML compute and model storage. In addition, more types of analytics information and recommendations could be produced by Near-RT RIC, and consumed by the other RAN functions to optimize the radio performance on their own. Toward this direction, some initial work is expected to be carried out in O-RAN's Work Group 3, exploring Near-RT RIC's AI/ML services for the 5G base stations.

The 6G UEs may also benefit from Near-RT RIC's local compute, model and data resources, though the challenges on proprietary models and private data need to be addressed. Note

that the Near-RT RIC should be transparent to the air interface, so as to minimize the impact on UEs.

6.4.4 Enhanced RAN service exposure

The 6G Near-RT RIC could be leveraged to support more interactions with other domains of the network. As of today, the Y1 interface provides support for Near-RT RIC to expose the RAN analytics information to authorized consumers (e.g., OAM, core network, trusted application servers, etc.). More data exposure services can be introduced. Furthermore, one possible enhancement is to allow RAN to consume certain information from the local application servers, enabling better service-aware optimization in RAN. Another enhancement could be to support distributed AI/ML training or inference with the core network.

6.4.5 Local positioning and sensing support

Positioning/sensing information may be further exploited in 6G for better handover and beam steering. The Near-RT RIC could be further enhanced to support the fusion of positioning/sensing results from the base stations, or optimize the coordination of positioning/sensing operations among the base stations.

6.4.6 Interactions with neighbor Near-RT RICs

RAN data and AI/ML model transfer could be supported among Near-RT RICs.

In addition, such collaboration is useful to support service continuity when a served UE moves across the serving areas of multiple Near-RT RICs. Currently, the per-UE data collection and optimization will be interrupted in that situation. The mechanisms to enable service continuity of Near-RT RIC for such UEs should be investigated in the 6G era.

6.5 Near-RT RIC implementation options

6.5.1 Implementation example 1

This example assumes a fully distributed deployment of non-split 6G base stations ("6gNB"). In this example, a Near-RT RIC platform and a few xApps are inter-connected with the 6gNBs, as shown in Figure 1.

This example also applies for the 5G gNBs.

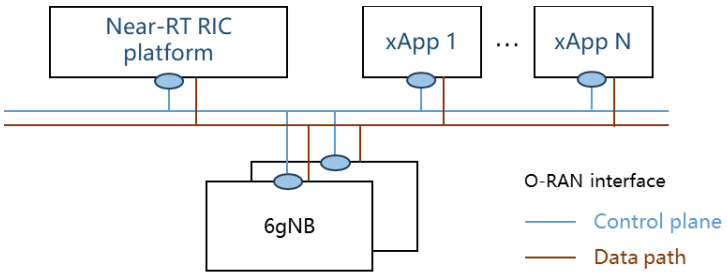

Figure 1 Near-RT RIC platform and xApps with 6gNBs

6.5.2 Implementation example 2

This example assumes a CU/DU split architecture for the 6G base station. The 6G CU is associated with multiple 6G DUs and provides a relatively large coverage area. In this example, the Near-RT RIC platform can be implemented as part of the 6G CU as shown in Figure 2. The xApps may interact with the 6G DUs depending on use cases, and with potential collaborations with the dApps.

This example also applies for the 5G gNB-CUs and gNB-DUs.

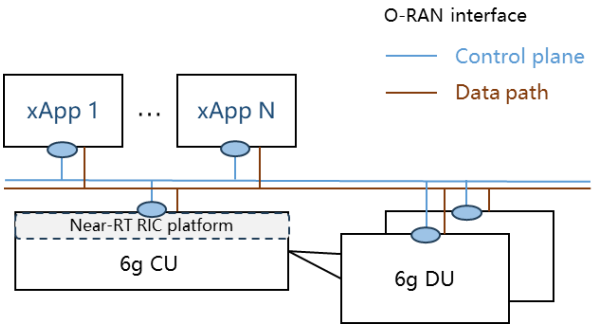

Figure 2 xApps with 6G CU and DU

6.5.3 Implementation example 3

This example assumes an AI Node connects with multiple 6gNBs. The AI Node may support the following functionalities:

- RAN data collection
- AI/ML lifecycle management
- Computing resource management
- AI bearer management
- Distributed collaboration

In this example, the Near-RT RIC platform can be implemented as part of the AI Node. O-RAN could define the interfaces between the xApps and the AI Node, as shown in Figure 3. O-RAN may further define the interfaces between an xApp and the 6gNBs based on use cases, as shown in Figure 4.

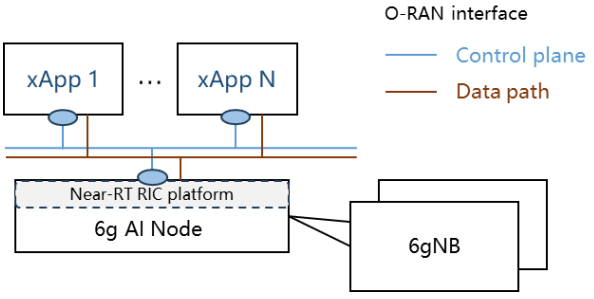

Figure 3 xApps with AI Node and 6gNB, option 1

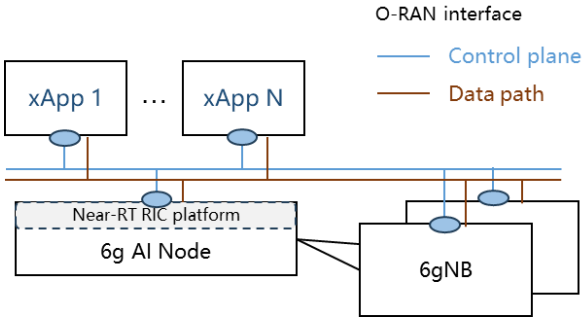

Figure 4 xApps with AI Node and 6gNB, option 2

6.5.4 Implementation example 4

This example highlights the potential unified design for Near-RT RIC with SMO and Non-RT RIC, based on the limitations and enhancements discussed in clause 5.5.

A first option is shown in Figure 5.

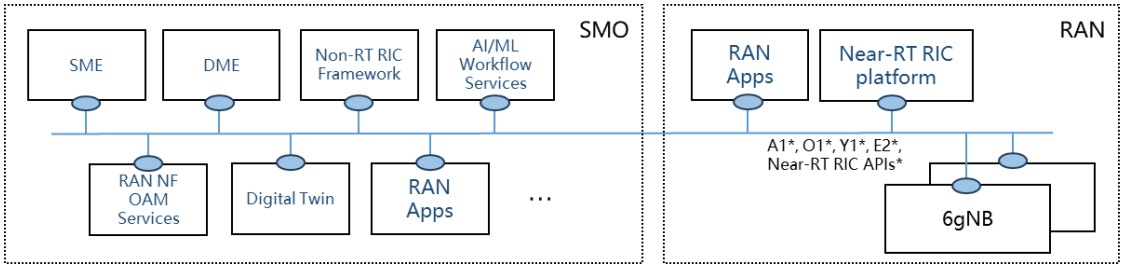

Figure 5 Unified design with SMO and Non-RT RIC, option 1

- In this architecture, the Near-RT RIC can access SMO services directly via the service bus instead of requiring support on A1 or O1 interfaces for new features.
- rApps and xApps are conceptually unified as generic RAN Apps which can be deployed in a flexible, cloud-native manner (e.g., within the evolved SMO and Non-RT/Near-RT RIC).
- A RAN App can be decomposed into multiple RAN Apps with functionality analogous to rApps and xApps and different RAN Apps can easily coordinate with each other over the service bus.
- E2 Nodes can access SMO services directly via service-based interfaces.
- Near-RT RIC and E2 nodes can interact via service-based interfaces exposed by E2 Nodes.
- E2 Nodes can be managed via service-based interface.
- The A1*, O1*, Y1*, E2*, and Near-RT RIC APIs* functionalities (subsuming the related A1, O1, Y1, E2 and Near-RT RIC APIs, and E2 functionalities) are proposed to be provided as services over service-based interface.

Note that the conversion of existing interface functions and information elements to a service-based API model could be done in a relatively mechanical and automated manner

(e.g., with wrapper functions), and would not require the development of entirely new APIs from scratch.

A second option is shown in Figure 6.

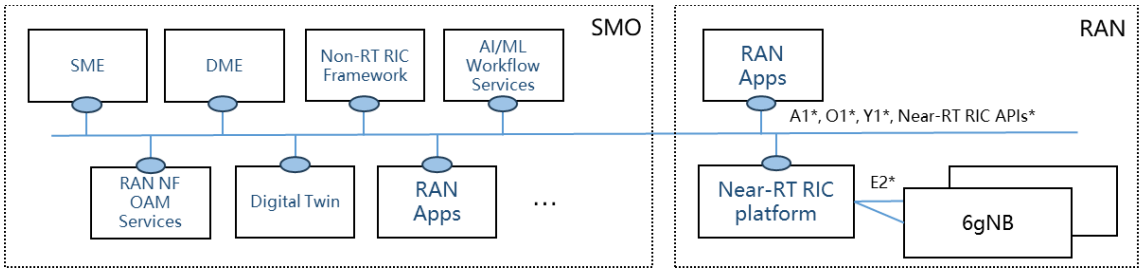

Figure 6 Unified design with SMO and Non-RT RIC, option 2

- In this architecture, the Near-RT RIC can access SMO services directly via the service bus instead of requiring support of the A1 and/or O1 interfaces.
- Similar to option 1, rApps and xApps are conceptually unified as generic RAN Apps which can be deployed in a flexible, cloud-native manner (e.g., within the evolved SMO and Non-RT/Near-RT RIC).
- Similar to option 1, a RAN App can be decomposed into multiple RAN Apps with functionality analogous to rApps and xApps and different RAN Apps can easily coordinate with each other over the service bus.
- The A1*, O1*, Y1*, Near-RT RIC APIs* functionalities (subsuming the related A1, O1, Y1, and Near-RT RIC APIs functionalities) are proposed to be provided as services over service-based interfaces as in option 1.

7 Conclusion

The evolution of Near-RT RIC towards 6G will be driven by the integration of advanced technologies and by the trends in the telco industry, including but not limited to, AI/ML, sensing/ISAC, SBA, NTN, and CCIN. It is also helpful to revisit the concepts that shaped 5G RAN and today's Near-RT RIC, and examine them in the 6G context. A potential architectural design for the next-generation Near-RT RIC is presented in this report, highlighting innovative protocol and functionalities. Along with the 6G O-RAN, the next-generation Near-RT RIC is poised to play a pivotal role in realizing the vision of intelligent 6G network.

References

[1] O-RAN Alliance. "O-RAN: Towards an Open and Smart RAN". [Online]. Available: https://mediastorage.o-ran.org/white-papers/O-RAN.White-Paper-2018-10.pdf.

[2] O-RAN Alliance. "O-RAN Work Group 1; O-RAN Architecture Description".

[3] O-RAN Alliance. "O-RAN Work Group 3; O-RAN Use Cases and Requirements".

[4] O-RAN Alliance. "O-RAN Work Group 3; Near-RT RIC Architecture".

[5] O-RAN Alliance. "O-RAN Work Group 3; Y1 interface: General Aspects and Principles".

[6] 3GPP TR 37.817: "Study on Enhancement for Data Collection for NR and EN-DC".

[7] 3GPP TR 38.843: "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface".

[8] 3GPP TS 38.300: "NR and NG-RAN Overall Description; Stage 2".

[9] 3GPP TS 38.423: "Xn application protocol (XnAP) ".

[10] ITU-R M.2160: "Framework and overall objectives of the future development of IMT for 2030 and beyond".

[11] O-RAN Alliance. "dApps for Real-Time RAN Control: Use Cases and Requirements". [Online]. Available: https://mediastorage.o-ran.org/ngrg-rr/nGRG-RR-2024-10-dApp%20use%20cases%20and%20requirements.pdf.

[12] Stanford University. “AI Demystified: Introduction to AI”. [Online]. Available: https://uit.stanford.edu/service/techtraining/ai-demystified/introduction

[13] 3GPP TS 28.105: "Artificial Intelligence / Machine Learning (AI/ML) management".

[14] 3GPP TS 23.501: "System architecture for the 5G System (5GS); Stage 2".

[15] 3GPP TS 22.137: "Service requirements for Integrated Sensing and Communication; Stage 1".

[16] 3GPP TS 38.305: "Stage 2 functional specification of User Equipment (UE) positioning in NG-RAN".

[17] Z. Han et al., "Multistatic Integrated Sensing and Communication System in Cellular Networks," 2023 IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, Malaysia, 2023, pp. 123-128, doi: 10.1109/GCWkshps58843.2023.10464728.

[18] 3GPP TR 22.837: "Feasibility Study on Integrated Sensing and Communication".

[19] Google gRPC. [Online]. Available: https://grpc.io/.

[20] 3GPP TR 38.821: "Solutions for NR to support non-terrestrial networks (NTN)".

[21] R. Campana, C. Amatetti and A. Vanelli-Coralli, "O-RAN based Non-Terrestrial Networks: Trends and Challenges," 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Gothenburg, Sweden, 2023, pp. 264-269, doi: 10.1109/EuCNC/6GSummit58263.2023.10188308.

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

[23] O-RAN Alliance. “O-RAN Work Group 1; O-RAN Communication and Computing Integrated Networks”.

[24] L. Kundu, X. Lin, R. Gadiyar, J. Lacasse and S. Chowdhury, "AI-RAN: Transforming RAN with AI-driven Computing Infrastructure". [Online]. Available: https://arxiv.org/abs/2501.09007.

[25] 3GPP TR 38.801: "Study on new radio access technology: Radio access architecture and interfaces (Release 14)".

[26] 3GPP TS 28.313: "Self-Organizing Networks (SON) for 5G networks".

[27] M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb and M. Zorzi, "Machine Learning at the Edge: A Data-Driven Architecture With Applications to 5G Cellular Networks," in *IEEE Transactions on Mobile Computing*, vol. 20, no. 12, pp. 3367-3382, 1 Dec. 2021, doi: 10.1109/TMC.2020.2999852.