O-RAN Use Cases and Deployment Scenarios

Towards Open and Smart RAN

White Paper, February 2020

Acknowledgements

Special thanks must be given to the numerous people contributing to the White Paper, including

Co-editors: Chai Li (CMCC), Arda Akman (Netsia)

Major contributors: Arda Akman (Netsia), Chai Li (CMCC), Lyndon Ong (Ciena), Lucian Suciu (Orange), Burcu Yargicoglu Sahin (Netsia), Tong Li (Lenovo), Paul Stjernholm (Ericsson), Jens Voigt (Amdocs), Andrea Buldorini (TIM), Qi Sun (CMCC), Ming Li (Inspur), Haruhisa Hirayama (KDDI), Gabor Hannak (Nokia), Salvatore Costanzo (Orange), Mathias Sintorn (Ericsson), Jin Yang (Verizon)

Major reviewers: Christian Gallard (Orange), Dhruv Gupta (AT&T), Paul Smith (AT&T), Jinri Huang (CMCC), Cagatay Buyukkoc (AT&T)

Contents

Executive Summary

O-RAN Alliance is committed to evolving radio access networks, making them more open and smarter than currently deployed networks. The first whitepaper “O-RAN: Towards an Open and Smart RAN” introduced the key O-RAN concepts: Openness to build a more cost effective and agile RAN through open interfaces, open hardware and open source; and Intelligence to meet the requirements for increasingly complex, denser and richer networks through deep learning techniques and embedded intelligence in every layer of the RAN architecture.

This whitepaper introduces the initial set of O-RAN use cases and cloud native deployment support options.

An O-RAN use case is a use case that both leverages the O-RAN architecture and demonstrates its unique benefits. Key benefits of O-RAN architecture include the ability to utilize machine learning systems and artificial intelligence back end modules to empower network intelligence through open and standardized interfaces in a multi-vendor network. The use cases presented in this whitepaper demonstrate these capabilities either by utilizing learning technologies to generate and deploy machine learning models and policies to control the real time behavior of RAN or by focusing on optimizing RAN via configurations along with high level policies and triggers, and some others incorporate both of these approaches to meet the service requirements of 5G and beyond. It should be noted that not all of the use cases presented in this whitepaper are currently supported by O-RAN specifications and these use cases will be supported with upcoming versions of these specs (see Annex for more details).

Along with the use cases, O-RAN cloudification and orchestration platform O-Cloud is introduced which facilitates flexible deployment options and service provisioning models of O-RAN virtualized network elements in telco clouds. The O-Cloud is the cloud computing platform comprising a collection of physical infrastructure nodes that can host the relevant O-RAN functions, the supporting software components and the appropriate management and orchestration functions.

Introduction

O-RAN is committed to evolving radio access networks with its core principles being intelligence and openness. It aims to drive the mobile industry towards an ecosystem of innovative, multi-vendor, interoperable, and autonomous RAN, with reduced cost, improved performance and greater agility.

The key principles of the O-RAN Alliance include:
1. Leading the industry towards open, interoperable interfaces, RAN virtualization, and big data and AI enabled RAN intelligence;
2. Maximizing the use of common-off-the-shelf hardware and merchant silicon and minimizing proprietary hardware;
3. Specifying APIs and interfaces, driving standards to adopt them as appropriate, and exploring open source where appropriate.

By opening up the RAN from a closed vendor environment to a standardized, multi-vendor, AI powered hierarchical controller structure along with the ability for 3rd parties to access what used to be closed RAN data, O-RAN enables not only RAN vendors but also Operators and 3rd parties to deploy innovative services as RAN applications that can leverage emerging AI/ML based technologies, hence making O-RAN the first application store of Radio Access Networks.

In O-RAN architecture, as shown in Figure 1, Service Management and Orchestration (SMO) framework contains the Non-RT RIC function which has the goal of supporting intelligent RAN optimization in non-real-time (i.e., greater than one second) by providing policy-based guidance using data analytics and AI/ML training/inference. Non-RT RIC can leverage SMO services such as data collection and provisioning services of O-RAN nodes.

Near-RT RIC, O-CU-CP, O-CU-UP, O-DU, and O-RU are the network functions for the radio access side. Near-RT RIC enables near real-time control and optimization of O-RAN (O-CU and O-DU) nodes and resources over the E2 interface with near real-time control loops (i.e., 10ms to 1s). The Near-RT RIC uses monitor, suspend/stop, override and/or control primitives to control the behaviors of O-RAN nodes. The Near-RT RIC hosts xApps that use E2 interface to collect near real-time RAN information to provide value added services using these primitives, guided by the policies and the enrichment data provided by the A1 interface from the Non-RT RIC.

In order to introduce the rich capabilities of O-RAN architecture, this paper outlines possible use cases and cloud native deployment scenarios. While some of these use cases are utilizing learning technologies to generate and deploy models using long term data along with policies to control the real time behavior of RAN, some other use cases are focused on optimizing RAN via configurations and policies, and some others are incorporating both of these approaches to support closed loop optimizations and Service Level Agreement assurance.

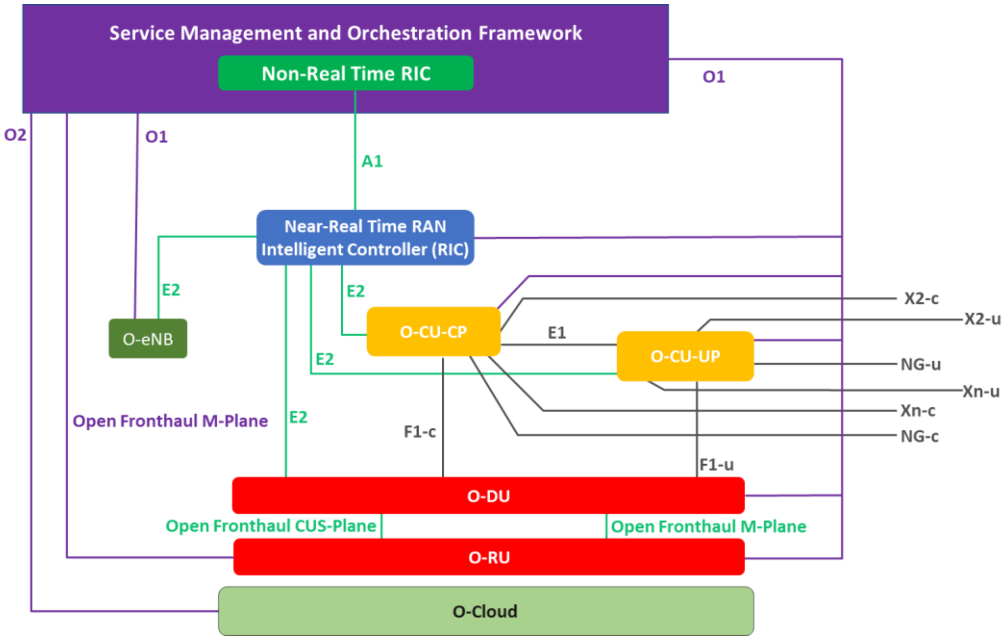


Figure 1: O-RAN Overall Logical Architecture

“Low Cost Radio Access Network White-box Hardware” and “RAN Sharing” are two use cases demonstrating CAPEX reduction. Exemplary use cases like handover management for V2X, QoE optimization are showcasing the generation of AI/ML models at Non-RT RIC using long term data gathered from both RAN and external sources, the deployment of these models and policies to Near-RT RIC to optimize and fine tune RAN in real time and the policy feedbacks to continuously update the AI/ML models. Use cases like Traffic Steering and Massive MIMO Optimization are using long term data and specific triggers to request configuration changes along with policies to optimize RAN. Finally use cases like RAN Slice SLA Assurance are using both approaches. It should be noted that not all of the use cases presented in this whitepaper are currently supported by O-RAN specifications and these use cases will be supported with upcoming versions these specs (see Annex for more details).

Building upon the ETSI NFV reference architecture, O-RAN utilizes commercial off-the-shelf (COTS) hardware and virtualization software that enables abstraction, in the form of virtual machines (VM) or containers to provide multiple hierarchical cloud deployment options for Operators. These options include a number of possible configurations for the placement of Near-RT RIC, O-CU and O-DU at regional and edge clouds. O-RAN aims to provide the necessary cloudification framework to help Operators automate the deployment and provisioning of O-RAN based radio access networks.

The O-Cloud is the cloud computing platform comprising a collection of physical infrastructure nodes that can host the relevant O-RAN functions, the supporting software components and the appropriate management and orchestration functions which is outlined in the following sections of this whitepaper.

Benefits of O-RAN Architecture

First, O-RAN is capable of reducing network CAPEX and OPEX

O-RAN is capable of reducing CAPEX through a prosperous multi-vendor ecosystem with scale economics:

1. O-RAN open interfaces eliminate single factory, and facilitate multi-vendor cooperative deployments, thus enabling a more competitive and vibrant supplier system.
2. Open source software and hardware reference designs enable faster innovation through a larger ecosystem.
3. Native cloud feature of O-RAN, enables scale out designs for capacity, reliability and availability, rather than expensive scale up designs.

O-RAN reduces OPEX by RAN automation:

O-RAN introduces embedded intelligence in every layer of the RAN architecture, and leverages new learning based technologies to highly automate operational network functions, reduce operational activities, hence reducing OPEX.

Second, O-RAN is capable of improving network efficiency and performance

O-RAN improves network efficiency and performance with RAN automation:

With O-RAN, status of network performance and network resources are continually monitored and more real-time close-loop control could take effect with little human intervention. Even for the most complex network, O-RAN will have the inherent ability to offer efficient, optimized radio resource management through closed-loop control, to enhance network performance and user experience. Interactions between Non-RT RIC and Near-RT RIC can be used to optimize and fine-tune control algorithms such as the one related to load balancing, mobility management, multi-connection control, QoS management and network energy saving.

Third, O-RAN is capable of importing new capability with great agility

O-RAN makes it easy to import new network capability via easy software upgrade, with its native cloud infrastructures.

O-RAN Key Use Cases for Phase I

White-box Hardware Design

Low Cost Radio Access Network White-box Hardware

The promotion of white-box hardware is a potential way to reduce the cost of 5G deployment. O-RAN aims to specify and release a complete reference design of a high performance, spectral and energy efficient white-box base station. By releasing and application of the reference design, the ecosystem can get the scale effect of hardware selection and reduce the cost of the whole industry chain. Releasing the reference designs can also reduce the R&D difficulties and costs, which can attract more small and medium enterprises into the telecom industry and bring industrial prosperity and innovation.
The white-box hardware design will focus on:

1. O-DU: Downlink and uplink baseband processing, supply system synchronization clock, signaling processing, OM function, interface with O-CU and the fronthaul gateway.
2. O-RU: Downlink baseband signal convert to RF signal, uplink RF signal convert to baseband signal, and interface with the fronthaul gateway.
3. Fronthaul gateway (if needed): Downlink broadcasting, uplink combining, power supply for O-RU, cascade with other fronthaul gateway, synchronization clock.

The white-box hardware design requires support of O-RAN fronthaul open interface, O-DU open-source software, and facilitates network functions deployed as virtualized applications in the cloud:

1. It will be based on O-RAN specified open fronthaul interface. Split option 7-2x is currently supported, split option 6 and split option 8 are being considered as future study items;
2. It tries to carry out open O-DU software protocol function and finally realize open-source and openness of software and hardware;
3. It will help to realize virtualization application, and will be in collaboration with cloudification and orchestration.

AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)

Traffic Steering

Traffic steering, an evolution of mobile load balancing, is a widely used network solution to achieve optimal traffic distribution based on desired objectives. With increasing network capacity and traffic, mobile networks have become more and more complex. How to effectively steer traffic to improve network efficiency and enhance user experience is a great challenge. Traditional network traffic control optimization requires a lot of manual intervention, which is inefficient and passive, and the feedback response after processing is also slow. Moreover, the RRM (Radio Resource Management) features of existing cellular networks are cell-centric. Even in different areas of a cell, there are changes in the radio environment, such as the coverage of adjacent cells, signal strength, interference status, etc. However, base stations based on traditional control strategies treat all UEs in a similar way and are usually focused on average cell-centric performance, rather than UE-centric.

O-RAN architecture improves the flexibility and agility of the network, and provides the possibility of RAN automation. In terms of traffic steering, first of all, RAN intelligence reduces manual intervention, saving OPEX while avoiding human errors. It has faster response to traffic problems and more efficient in processing these problems. Secondly, RAN intelligence customizes UE-centric strategies and provides proactive optimization by predicting the network condition and UE performance, thus improving user experience. Finally, the RAN intelligent module trained by machine learning can control the adaptation of diverse scenarios and objectives, which better supports the application of vertical industry and provides more powerful guarantee for the business development of operators.

In order to achieve intelligent and proactive traffic steering control, Non-RT RIC and Near-RT RIC control traffic steering strategies through AI/ML learning and provide AI models/policies as well as RAN control/guidance for related A1 and E2 interfaces. The data used for machine learning is collected via the O1 interface from the O-CUs and O-DUs.

QoE Optimization

The highly demanding 5G native applications like Cloud VR are both bandwidth consuming and latency sensitive. However, such traffic-intensive and highly interactive applications could not be efficiently supported with current semi-static QoS framework. This is because dynamic traffic volume could be generated by user interactions, not to mention the fast fluctuating radio transmission capabilities. Therefore, the service QoE could not be always guaranteed, except with overprovisioning QoS considering the peak traffic demand.

Vertical application specific QoE prediction and QoE based proactive close-loop network optimization in real-time would help deal with the issue. Before QoE is degraded, the radio resources could be allocated to the user and services where the radio resources are most urgently required. In this way, QoE is optimized while the radio resources are utilized in a more efficient way.

With the software defined RAN intelligent controller and the open interfaces of O-RAN, AI models could be easily deployed and upgraded to optimize QoE of emerging vertical services. Multi-dimensional data can be acquired and processed via ML algorithms to support traffic recognition, QoE prediction, and finally guiding close-loop QoS enforcement decisions. ML models can be trained offline in Non-RT RIC, while model inference will be executed real-time in Near-RT RIC. A1 and E2 interfaces help to deliver the policy/intents/AI models and RAN control to enforce the QoS for the QoE optimization. O1 interface collects data for training in Non-RT RIC. The O-CU software module helps support data provisioning to Near-RT RIC and Non-RT RIC, and executes QoS enforcement decisions from Near-RT RIC.

QoS Based Resource Optimization

For the network to appropriately support high bandwidth, low latency 5G services and applications, rigorous planning and configuration is needed to provide the needed coverage and capacity. However, because of the varying nature of traffic demand and radio environment throughout the RAN there might be situations where the default configuration and planning are not enough to provide the resources needed for services with highly demanding requirements.

In a situation where the deployed RAN resources and the default configuration are not appropriate to support the requirements for all users using a specific service, the QoS based resource policies can be used for the Non-RT RIC and Near-RT RIC to optimize how RAN resources are allocated between users with similar requirements.

For example, in a congestion situation, the available RAN resources in some areas may not be enough to fulfill the requirements for all users using the same service. In such a situation, there is a risk that none of the users are given enough resources to fulfill their requirements, meaning that none of the users will be satisfied. In such a situation, the QoS based resource optimization policy can be used by the Non-RT RIC to ensure that at least certain prioritized users can reach a satisfactory level of QoS.

Through a close examination of the congestion situation, analytics functions in Non-RT RIC might conclude that it is better to prioritize certain user(s) utilizing the same service in a particular situation. When so, the Non-RT RIC can use the A1 QoS policies to re-allocate RAN resources between users utilizing the same service and Near-RT RIC can realize these policies on O-CUs/O-DUs through the E2 interface.

Massive MIMO Optimization

Massive MIMO is seen as one of the key technologies for 5G. Due to the multi-antenna transmission and reception, this technology can inherently provide diversity and improve capacity by targeting high gain antenna beams towards one or multiple subscribers, thus improving the receive power levels and spatially filtering the interference from neighboring subscribers and transmission points. In addition, a spatial multiplexing operations regime can improve the network capacity by transferring multiple data streams towards/from one or different subscribers utilizing a spatial re-use of the scarce time/frequency resource blocks. Further advantages include the controlling of electromagnetic (EM) emissions or advanced network management technologies like beam-based load balancing, beam failure reductions, or adaptive cell coverage areas, especially in highly dense urban 3D environments and mobile subscribers.

The objective of the massive MIMO optimization use case is to proactively and continuously improve cell-centric network QoS and/or user (group)-centric QoE in a multi-cell and, possibly, multi-vendor massive MIMO deployment area with multiple transmission/reception points, depending on specific operator-defined objectives. This optimization is performed at different times scales from near-real-time TTI scale to non-real-time by a) adapting GoB beam configurations for initial access, control, and traffic channel beam forming (e.g., number of beams, beam boresights, vertical beam widths, horizontal beam widths, beam black lists/white lists, transmission powers), or b) adapting policies (including, but not limited to packet schedulers – PRB usage, full digital beamforming policies, or beam failure recovery). The high number of configuration parameters per array antenna in a cluster of cells, the massiveness of available measurement input data, the complexity, proactiveness as well as partly near-real-time requirements suggest the application of machine learning techniques in this process.

The advantages the O-RAN architecture provides to this use case include the possibility to apply and combine both, non- and near-real-time analytics, machine-learning, and decision making for various sub-tasks of this use case for a cell-centric and /or user (group) centric point of view. Both, the Non-RT RIC and the Near-RT RIC oversee the current traffic, coverage, and interference situation, etc., in a whole cluster of cells. Hereby; analytics, machine-learning training and decision results could be included in the near real-time sub-task as enrichment information given by the non-real-time subtasks. O-RAN interfaces such as O1, A1, and E2 will support necessary data, policy, and configuration exchanges between the architectural elements. These advantages motivate massive MIMO optimization to run on O-RAN based RAN implementations.

O-RAN Key Use Cases for Phase II

AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)

RAN Slice SLA Assurance

In the 5G era, network slicing is a prominent feature that provides end-to-end connectivity and data processing tailored to specific business requirements. These requirements include customizable network capabilities such as the support for very high data rates, traffic densities, service availability and very low latency. According to 5G standardization efforts, a 5G system should support the needs of the business through the specification of several service KPIs such as data rate, traffic capacity, user density, latency, reliability, and availability. These capabilities are specified based on a Service Level Agreement (SLA) between the mobile operator and the business customer, which has resulted in increased interest in mechanisms to ensure slice SLAs and prevent its possible violations.

O-RAN's open interfaces combined with its AI/ML based innovative architecture can enable such challenging RAN SLA assurance mechanisms. Based on RAN specific slice SLA requirements, Non-RT RIC and Near-RT RIC can fine-tune RAN behavior to assure RAN slice SLAs dynamically. Utilizing the slice specific performance metrics received from E2 Nodes, Non-RT RIC monitors long-term trends and patterns regarding RAN slice subnets' performance, and trains AI/ML models to be deployed at Near-RT RIC. Non-RT RIC also guides Near-RT RIC using A1 policies, with possible inclusion of scope identifiers (e.g. S-NSSAI) and statements such as KPI targets. Near-RT RIC enables optimized RAN actions through execution of deployed AI/ML models or other slice control/slice SLA assurance xApps in near-real-time by considering both O1 configuration (e.g. static RRM policies) and received A1 policies, as well as received slice specific E2 measurements. O-RAN slicing architecture enables such challenging mechanisms to be implemented which could help pave the way for operators to realize the opportunities of network slicing in an efficient manner as well as potentially change the way network operators do their business.

Context Based Dynamic Handover Management for V2X

V2X (vehicle to vehicle/infrastructure/anything) communications is a current key topic for the modern car manufacturers, network technology companies, legislation and regulation bodies. It promises numerous benefits such as increased road safety, reducing emissions, and saving time. These are facilitated by orchestration of the traffic and assistance of individual user decisions based on real-time information on the road and traffic conditions, driver intentions etc. Part of the V2X architecture is the V2X UE (SIM + device attached to vehicle) which communicates with the V2X Application Server (V2X AS). The exchanged information comprises Cooperative Awareness Messages (CAMs) (from UE to V2X AS) [ETSI EN 302 637-2], radio cell IDs, connection IDs, and basic radio measurements (RSRP, RSPQ, etc.)

As vehicles traverse along a highway, due to their high speed and the heterogeneous natural environment, they might be handed over frequently or in suboptimal ways, which may cause handover (HO) anomalies: e.g., short stay, ping-pong, and remote cell. For the V2X applications and in general the V2X ecosystem to function properly, continuous and reliable connectivity (low delay, good coverage etc.) of the vehicular UE to the network (and so the V2X AS) is necessary. Suboptimal HO sequences and anomalies might substantially impair the connectivity, the performance of the V2X applications, and eventually the general reliability of the whole V2X ecosystem. Thus, their mitigation is of prime importance.

The ORAN architecture allows for i.) the collection and maintenance of, and the access to historical traffic/navigation and radio/HO data (in Near-RT RIC and via the O1 interface), ii.) the deployment, continuous retraining, and evaluation of AI/ML based applications that detect, predict HO anomalies on a

UE level (in the Near-RT RIC and via the A1 interface), iii.) the real-time monitoring of traffic and radio conditions on a UE level (in the Near-RT RIC and via the E2 and A1 interfaces), and iv.) the deployment and maintenance of real-time applications that predict, prevent, and mitigate anomalous situations on a UE level (in the Near-RT RIC).

Flight Path Based Dynamic UAV Resource Allocation

The application of UAV has played a great role in civil applications including agricultural plant protection, power inspection, police enforcement, geological exploration, environmental monitoring, and the applicable fields are rapidly expanding. And 5G can provide high-speed networked data transmission channels for low-altitude UAV, replacing the traditional point-to-point communication link between traditional UAV and ground control station. However, since the site along the flight is mainly for ground cover construction, the altitude of the UAV is not within the main lobe of the ground station antenna, the air signal is disorderly, and there is no main coverage cell, which causes the terminal downlink interference to be large, and may not be solved in some areas. Hence, there is a potential disconnection problem, which may trigger a UAV that has been in the area for a long time to land or return.

In O-RAN architecture, multi-dimensional data can be acquired, for example, Non-RT RIC can retrieve necessary Aerial Vehicles related measurement metrics from the network based on UE's measurement report and SMO, and flight path information of Aerial vehicle, climate information, flight forbidden/limitation area information and Space Load information etc. from the application, e.g. UTM (Unmanned Traffic Management) for constructing and training relevant AI/ML models that will be deployed in RAN. Near-RT RIC can support the deployment and execution of the AI/ML models from Non-RT RIC. Based on this, the Near-RT RIC can perform the radio resource allocation for on-demand coverage for UAV considering the radio channel condition, flight path information and other application information.

Radio Resource Allocation for UAV Applications

In the UAV control vehicle scenario, the O-RAN architecture meets the needs of wireless resource adjustment. This scenario refers to a Rotor UAV flying at low altitude and low speed, and carrying cameras, sensors and other devices mounted. The Operation terminals work in the 5.8GHz to remote control the UAV for border/forest inspection, high voltage/base station inspection, field mapping, pollution sampling, and HD live broadcast. At the same time, the UAV mobile control stations and the anti-UAV weapons jointly provide the service of fighting against illegal UAVs to ensure low-altitude safety in special areas. The UAV Operation terminals, the anti-UAV weapons, and the UAV mobile control stations are connected with the UAV Control Vehicle using the 5G network.

UAV Control Vehicle deploys the Non-RT RIC, Near-RT RIC, O-CU, O-DU, and the Application Servers. The user plane data transmitted over the network includes 4K high-definition video data, which has obvious uplink and downlink service asymmetry, and the uplink has high requirements on network bandwidth. The Near-RT RIC function module provides radio resource management functions of the O-CU and O-DU side.

In the UAV Control Vehicle Application scenario, there is a small amount of control data interaction requirement between the terminal and the network interaction, as well as the large bandwidth requirements for uploading HD video. The service asymmetry raises new requirements for resource allocation of the gNB. At the same time, the existing network operation and maintenance management platform (e.g., OSS systems) can only optimize the parameters of a specific group of UEs, but not for individual users. In the O-RAN architecture, the radio resource requirements for different terminals are sent to the O-CU and O-DU for execution by means of the Near-RT RIC function module.

Virtual RAN Network

RAN Sharing

RAN sharing is envisioned as an efficient and sustainable way to reduce the network deployment costs, while increasing network capacity and coverage. Accordingly, the open and multivendor nature of the O-RAN architecture can accelerate the introduction and development of RAN sharing solutions, by enhancing the deployment and configuration of virtual RAN network functions on commodity shared hardware.

For instance, let us consider a scenario where a “Home” operator (Op.A) makes available its RAN infrastructure and computing resources to host the virtual RAN functions (VNF) of a “Host” operator (Op.B), wherein each VNF represents a logic implementation of the O-DU and O-CU functionalities.

The challenge here is to enable Op.B to configure and control resources in an infrastructure that is owned by another operator. In that scenario, the adoption of the O-RAN architecture can facilitate the remote control and configuration of such VNFs. Indeed, Op.B can monitor and control the remote O-DU via the Near-RT RIC of site B, using a “remote” E2 interface, while all the remote VNF configuration procedures can be handled by a specific “sharing orchestration application”, located at the SMO of each operator. Indeed, a “remote” O1/O2 interface can be introduced to allow Op.B to communicate the desired configuration for the VNF hosted at site A.

Differently from traditional RAN architectures, O-RAN can provide a set of open interfaces (remote E2, O1/O2) that enables the monitoring of the remote users’ performance, thus facilitating the optimization of the radio allocation process and the remote configuration of QoS parameters. Accordingly, O-RAN introduces more freedom in the RAN sharing process, enabling each operator to configure the shared network resources independently from configuration and operating strategies of other sharing operators.

O-RAN Cloud Native Deployment

This section presents an overview of O-RAN cloudification, the cloud architecture, and the deployment scenarios.

Overview

5G has the potential to be a revolutionary wireless technology. But the limited radio range characteristics of mid band (3-6GHz) and high band (> 24GHz) spectrum used in 5G networks require an unprecedented degree of densification which is expected to reduce the overall utilization of baseband units and thus drive up the costs. Statistical multiplexing of workloads from multiple radios has the potential to improve utilization and the resulting economics of densification.

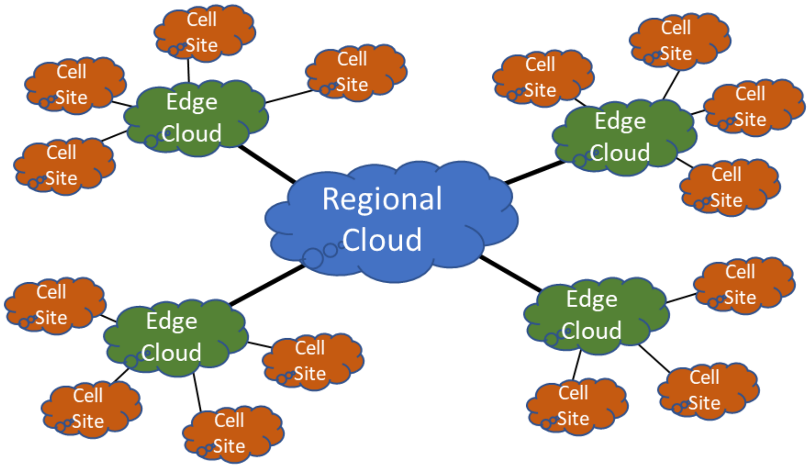
The cloudification and the flexible resource provisioning, enabled through adequate orchestration, are ideal vehicles to facilitate BBU resource pooling. Virtualization technologies have a rich history in resource multiplexing. Pooling gains enabled by server consolidation in underutilized IT infrastructures were one of the first drivers behind CPU virtualization and led to the birth of the cloud. Indeed, one of the early motivations behind Cloud RAN was also a desire to benefit from statistical multiplexing gains and reduced operational costs by centralizing RAN baseband infrastructure.

In the past, the large bandwidth demands of CPRI-based fronthaul and the high cost of DWDM infrastructures needed to support them have limited the deployability of Cloud RAN architectures. However, advances in fronthaul technology with O-RAN's adoption of an Ethernet-based transport through the 7.2x specification dramatically reduce the bandwidth needs for the fronthaul and make BBU hubbing and virtualization possible (e.g., in central offices). To further enable the case of BBU centralization and virtualization, it is also important to have more relaxed latency requirements for the fronthaul in comparison to CPRI. It is O-RAN's vision to eventually allow a single location to serve the BBU processing needs of an entire metropolitan area.

Hierarchical Cloud Deployment

O-RAN cloudification builds upon the ETSI NFV reference architecture. The NFV Infrastructure (NFVI) includes commercial off-the-shelf (COTS) hardware, such as servers and switches, and virtualization software that enables abstraction, in the form of virtual machines (VM) or containers running on the servers. The Virtual Infrastructure Manager (VIM) acts as the control plane of the NFVI for a cloud site, weaving multiple servers together and managing them as a single distributed system.

O-RAN components, such as the O-DU, O-CU, Near-RT RIC, as well as the associated Mobile Edge Computing (MEC) applications and the 5G User Plane Function (UPF), are virtualized network functions (VNFs) and run in VMs or containers. For functions demanding performance and energy efficiency that



Multiple Edge clouds per Regional cloud

Multiple Cell Sites per Edge cloud

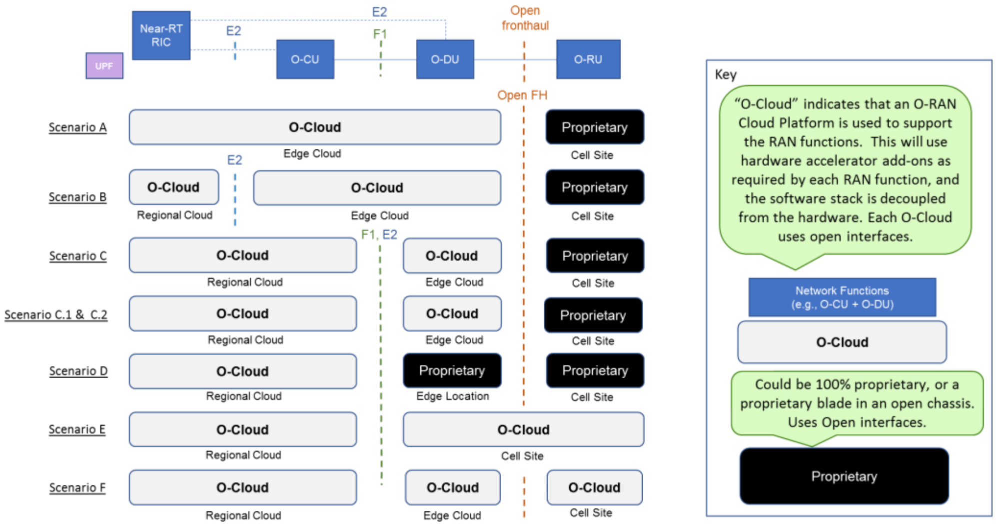
Figure 2: Hierarchical Cloud Deployment

exceed what a general-purpose CPU can achieve, purpose-built hardware accelerators can be used as additional peripherals on a given server.

Usually the deployment of the cloud architecture takes a hierarchical approach. Physical conditions at the different levels of the cloud hierarchy can be quite different, for example, a regional cloud may be hosted in a data center, whereas an edge cloud often resides in traditional telco equipment rooms with limited space, power, cooling capacity, and floor loading.

Deployment Scenarios

Figure 3 shows a list of O-RAN cloud deployment scenarios. The VNFs are shown at the top, and each scenario illustrates how the VNFs are implemented in either a proprietary network element, or on the O-RAN compliant O-Cloud. The term O-Cloud refers to an infrastructure element (or collection elements) that is based on COTS servers, uses hardware accelerator add-ons as needed, and has a software stack that is decoupled from the hardware.



"O-Cloud" indicates that an O-RAN Cloud Platform is used to support the RAN functions. This will use hardware accelerator add-ons as required by each RAN function, and the software stack is decoupled from the hardware. Each O-Cloud uses open interfaces.

Could be 100% proprietary, or a proprietary blade in an open chassis. Uses Open interfaces.

Figure 3: Cloud Deployment Scenarios

As part of the O-RAN cloudification work, Scenario B is currently the first priority of focus. In this scenario, the O-RU function is provided by a proprietary network element, located at the cell site. The O-CU and O-DU run on O-Cloud in an edge location, while the Near-RT RIC runs on a different O-Cloud in a regional location.

O-RAN Service Provisioning

This section explains how to provide O-RAN services by using the OAM architecture which includes the Service Management & Orchestration Framework (SMO) and the O-Cloud related management entities.

Figure 4 shows the O-RAN OAM logical architecture and together with the OAM interfaces or reference points being standardized, i.e., O1 and O2. Use of standardized interfaces enables multi-vendor interoperability and choice for the O-RAN operator.

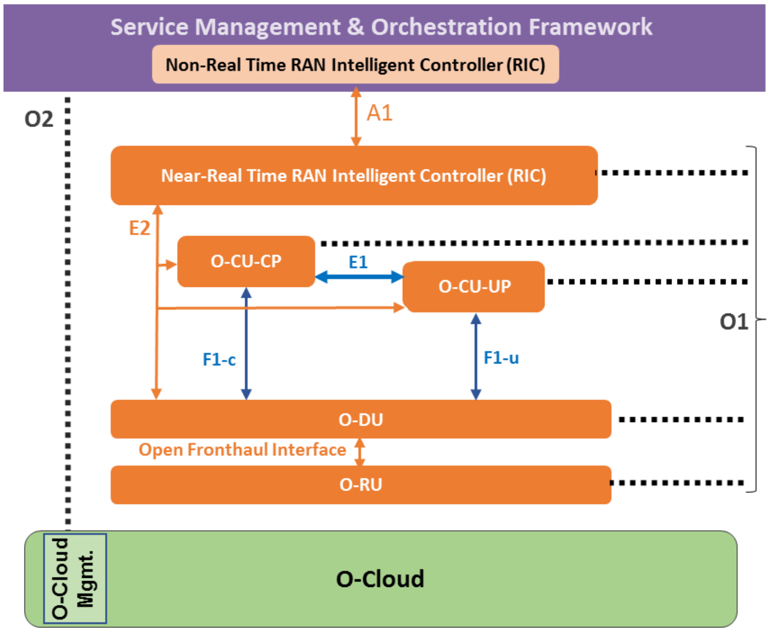


Figure 4: OAM Logical Architecture

O1 supports typical FCAPS and other management functions between the O-RAN components and the Service Management and Orchestration Framework, while O2 supports virtual resource management functions and other O-Cloud related management functions.

For O-RAN OAM and Management, the primary use case is the initial provisioning of the O-RAN Service over a RAN consisting of both VNFs and PNFs. This use case demonstrates the orchestration of multiple O-RAN components by the Service Management and Orchestration Framework to bring up a complex service. Functions such as discovery/registration and configuration of addressing, versioning and monitoring are done through the O1 interface using standard protocols and data models. Subsequently, resource and performance updates can be sent from the O-RAN managed functions back up to the SMO and they can be re-configured as needed by the SMO using the O1 interface.

Conclusion

O-RAN Alliance aims to build an open and smarter RAN with virtualized network elements, white-box hardware and standardized interfaces powered by an innovative RAN architecture. With the support of 3rd party RAN application development and deployment capabilities, this architecture enables new and innovative services that can leverage emerging learning based technologies to dynamically fine-tune RAN behavior using near-real-time RAN data as well as optional enrichment information that can be supplied via external applications. In order to introduce the rich capabilities of O-RAN architecture, this paper outlined the initial set of use cases and cloud native deployment scenarios.

Exemplary use cases including “Context Based Dynamic Handover Management for V2X”, “QoE Optimization” and “Flight Path Based Dynamic UAV Resource Allocation” are showcasing the generation of AI/ML models at Non-RT RIC using long term data gathered from both RAN and external sources, the deployment of these models and policies to Near-RT RIC to optimize and fine tune RAN in real time and the policy feedbacks to continuously update the AI/ML models. Use cases including “Traffic Steering”, “Massive MIMO Optimization” and “Radio Resource Allocation for UAV Applications” and “QoS Based Resource Optimization” are using long term data analytics and specific triggers to request configuration changes along with policies to optimize RAN. Finally use cases like RAN Slice SLA Assurance are using both approaches to meet the SLA requirements.

“Low Cost Radio Access Network White-box Hardware” and “RAN Sharing” are two use cases demonstrating CAPEX reduction. “RAN sharing” use case illustrates how O-RAN open interfaces can enable innovative solutions to share infrastructure between operators compared to the current situation for which bi-/multi-lateral agreements have to be defined.

Building upon the ETSI NFV reference architecture, O-RAN utilizes commercial off-the-shelf (COTS) hardware and virtualization software that enables abstraction, in the form of virtual machines (VM) or containers to provide multiple hierarchical cloud deployment options for Operators. These options include a number of possible configurations for the placement of Near-RT RIC, O-CU and O-DU at regional and edge clouds. O-RAN aims to provide the necessary cloudification framework, O-Cloud, to help Operators automate the deployment and provisioning of O-RAN based radio access networks.

O-RAN alliance continues to evolve RAN through intelligence and openness in RAN and more use cases utilizing the strengths of this architecture will be captured in the next version of this whitepaper,

Annex

Table 1: O-RAN Use Case Phases and Specification Support

Use Case	Class	Type	Impacted WGs	Impacted Interfaces to support the use case with current O-RAN specs	Impacted Interfaces to support the use case with future O-RAN specs	Impacted Entities
Low Cost Radio Access Network White-box Hardware	*Phase I*	*White-box Hardware Design*	*WG7*	*Open fronthaul interface*	-	*O-DU, O-RU*
Traffic Steering	*Phase I*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*	*O1/A1*	*E2*	*Non-RT RIC, Near-RT RIC, O-CU*
QoE Optimization	*Phase I*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*	*O1/A1*	*E2*	*Non-RT RIC, Near-RT RIC, O-CU*
Massive MIMO Optimization	*Phase I*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*	*O1*	*A1/E2*	*Non-RT RIC, Near-RT RIC, O-CU*
QoS Based Resource Optimization	*Phase I*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*	*A1*	*O1/E2*	*Non-RT RIC, Near-RT RIC, O-CU*
RAN Sharing	*Phase II*	*Virtual RAN Network*	*WG1,WG2,WG3, WG6*		*O1/O2/E2*	*Non-RT RIC, Near-RT RIC, O-CU, O-DU*
RAN Slice SLA Assurance	*Phase II*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG2,WG3,WG6*		*O1/O2/A1/E2*	*Non-RT RIC, Near-RT RIC, O-*

						CU, O-DU
Context Based Dynamic Handover Management for V2X	*Phase II*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*		*O1/A1/E2*	*Non-RT RIC , Near-RT RIC , O-CU*
Flight Path Based Dynamic UAV Resource Allocation	*Phase II*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*		*O1/A1/E2*	*Non-RT RIC , Near-RT RIC , O-CU*
Radio Resource Allocation for UAV Applications	*Phase II*	*AI-enabled RAN and Open RAN Interfaces (O1/A1/E2)*	*WG1,WG2,WG3*		*O1/A1/E2*	*Non-RT RIC , Near-RT RIC , O-CU*