AWS Whitepaper

# Open Radio Access Network Architecture on AWS

# Open Radio Access Network Architecture on AWS: AWS Whitepaper

# Table of Contents

**..... iv**
**Abstract and introduction ..... i**
Abstract ..... 1
Are you Well-Architected? ..... 1
Introduction ..... 1
AWS and O-RAN ..... 2
**O-RAN evolution ..... 4**
O-RAN architecture ..... 5
O-RAN management and infrastructure components ..... 8
Infrastructure decoupling and deployment flexibility in O-RAN ..... 10
**O-RAN architecture on AWS ..... 13**
O-RAN components on AWS ..... 13
RIC ..... 13
Near-RT RIC ..... 14
Non-RT RIC ..... 17
SMO ..... 18
O-CU ..... 20
O-DU and O-RU ..... 22
**O-RAN use cases ..... 24**
MEC-RAN integration for low-latency use case ..... 24
RIC-CU/DU operation to optimize radio resources (traffic steering and QoE optimization) ..... 25
Network slicing, and service level specifications (SLS) fulfillment ..... 26
DevOps, CI/CD, and network management (a single pane of glass) ..... 27
**Conclusion ..... 29**
**Glossary ..... 30**
**Contributors ..... 33**
**Document revisions ..... 34**
**Notices ..... 35**
**AWS Glossary ..... 36**

This whitepaper is for historical reference only. Some content might be outdated and some links might not be available.

# Open Radio Access Network Architecture on AWS

Publication date: **December 02, 2022** (*Document revisions*)

## Abstract

The Open Radio Access Network, or O-RAN, is an approach to transform the radio access network to a disaggregated, open, virtualized, and fully inter-operable mobile network. The O-RAN uses cloud technologies as its foundation to achieve the architecture goals and provide low-cost and fully automated 5G networks. AWS is an ideal cloud platform for the O-RAN network innovation, with more than 200 featured services and globally adopted infrastructure for telecom customers. This whitepaper explores the concept of the O-RAN, offers a reference architecture for the O-RAN on AWS, and presents best practices on Amazon Web Services (AWS) for key features of the O-RAN.

## Are you Well-Architected?

The AWS Well-Architected Framework helps you understand the pros and cons of the decisions you make when building systems in the cloud. The six pillars of the Framework allow you to learn architectural best practices for designing and operating reliable, secure, efficient, cost-effective, and sustainable systems. Using the AWS Well-Architected Tool, available at no charge in the AWS Management Console, you can review your workloads against these best practices by answering a set of questions for each pillar.

For more expert guidance and best practices for your cloud architecture—reference architecture deployments, diagrams, and whitepapers—refer to the AWS Architecture Center.

## Introduction

Communication service providers (CSPs) are embarking on a digital transformation journey, supporting a new wave of services enabled by 5G technologies. These consist of millimeter wave spectrum (mmWave) for better throughput and a reduced latency in the radio technology side, rearchitecting of the core network based on service-based architecture and Control-User Plane Separation (CUPS) to maximize the utilization, agility, and efficiency of 5G network for network slicing use cases, and, as a final frontier, the opening of the Radio Access Network (RAN) to enable an intelligent and fully interoperable RAN.

The last part has been a key objective of the O-RAN alliance, with the intent to use the modern innovations of cloud-native software technologies such as microservices, containerized, service-based, and stateless architecture. Among those key components, the O-RAN initiative is expected to bring a significant change to the telecom industry, because it causes a decomposed and software-driven RAN architecture.

For the RAN Network to be open, the following must occur:

- Decomposition of the Central Unit (CU) and Distributed Unit (DU).
- The use of enhanced Common Public Radio Interface (eCPRI).
- The disaggregation of RAN software such as Radio Connection Management and Mobility Management from the CU and DU.
- The use of artificial intelligence (AI) to support resource discovery and self-optimization.
- The use of cloud concepts to increase innovation and reduce time to market.

As described in the whitepaper *5G Network Evolution with AWS*, AWS can provide an ideal platform for building O-RAN components, along with other 5G innovation. Because RAN has to be placed at the edge (CU) and far-edge (DU) sites, it requires various options and form-factors of edge deployment, which can be met with AWS Outposts and Amazon EKS Anywhere.

For the microservice-based RAN implementation, various options for the container service such as Amazon Elastic Container Service (Amazon ECS), Elastic Kubernetes Service (Amazon EKS), Amazon EKS Distro, and Amazon EKS Anywhere can be a hosting platform for DU and CU software, using the advantage of container orchestration. More importantly, as the O-RAN architecture intends to, using the artificial intelligence/machine learning (AI/ML) and data lakes services of AWS in the architecture contributes to the telecom industry and CSPs' ability to advance to the next level of innovation.

## AWS and O-RAN

To realize the O-RAN vision, the industry expects not only the creation of an open ecosystem for the RAN, but also ways to build digital transformation frameworks on top of the RAN. AWS is an ideal cloud platform to achieve these goals, because AWS provides a breadth and depth of digital components, from Amazon Elastic Compute Cloud (Amazon EC2) (flexible compute capacity) to Amazon SageMaker AI (the ML platform for industry-wide use cases). Many CSPs think of potential pain points such as:

- Data Lifecycle Management (LCM) is additional overhead and requires huge investment and operation teams, if CSPs build the RAN Intelligent Controller (RIC) by themselves.
- Scalability, elasticity, and reliability of RAN are the keys to sustain the business and improve the experience of end-customers.
- Data analytics and ML is a differentiator for RIC.

In this sense, this whitepaper describes the reference architecture of O-RAN implementation on AWS and its benefits, in relation to relevant services of AWS. This paper provides an O-RAN reference architecture, an overview of O-RAN components and their characteristics, use cases, and best practices for architecting O-RAN on AWS. Best practices include high-availability, scalability, security, performance, and operational excellence. Use the information in this paper to develop O-RAN solutions on AWS, providing a cost-efficient and agile path to CSPs so they can achieve an end-to-end malleable network, enabling a multitude of 5G services.

# O-RAN evolution

The O-RAN is an idea for fully interoperable radio access networks with interface standards meant to transform the RAN industry toward open, intelligent, virtualized, and fully interoperable RAN. O-RAN specifications are driven by the O-RAN alliance, which was founded in 2018 and has become a world-wide operator and vendor community, with 237 mobile network operators and equipment providers as of 2020. The O-RAN alliance is actively working on the standardization of open interfaces, the development of open software including the 3rd Generation Partnership Project (3GPP) protocol stacks, and the testing and integration to support O-RAN member companies to build O-RAN-based 5G networks.

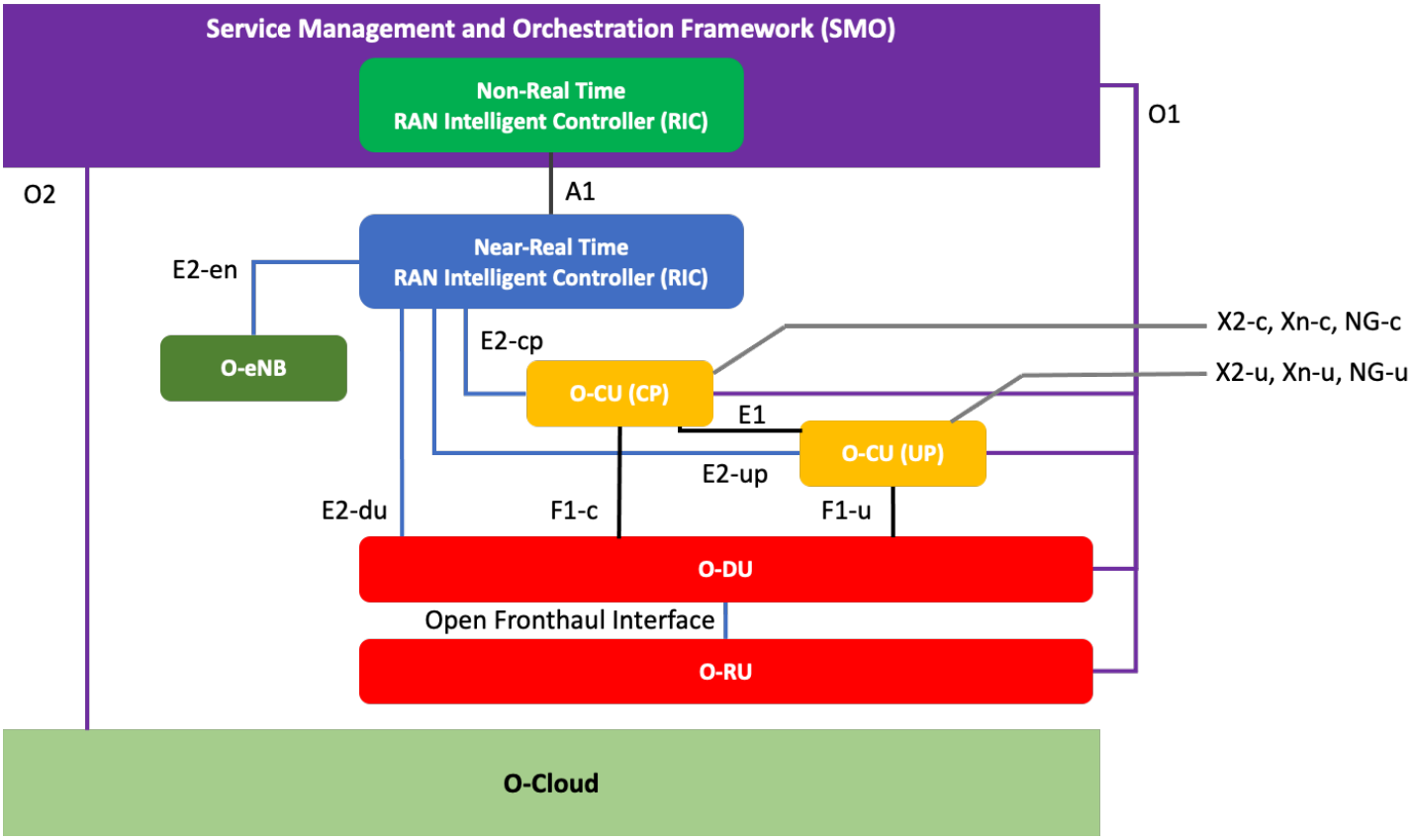
Compared to the 3GPP standard interfaces and architecture, the O-RAN alliance focuses on a disaggregated and fully interoperable RAN architecture. Regarding RAN interface standardization, the 3GPP mainly develops the interface between the mobile and the network node, which is the eNodeB in long-term evolution (LTE) or the gNodeB in 3GPP New Radio technology (NR) and the inter-network node interface. The network node, the eNodeB or gNodeB, has several layers of the 3GPP protocol stacks, but has been working as a monolithic network entity that provides all the radio access services.

The node has components such as the radio unit (RU) and the DU, but they are vendor-specific and connected over proprietary interfaces so that wireless network operators must purchase a whole entity from a single vendor. The O-RAN, however, pursues a goal to have a fully operational and interoperable architecture for RAN, with hardware and software from different vendors. The O-RAN provides an architecture as a foundation of the virtualized and disaggregated RAN on open hardware and cloud. The O-RAN specifications define the interoperable interfaces which fully support the O-RAN open architecture, and complement the 3GPP standards.

This chapter provides the overview of the O-RAN architecture, which consists of the 3GPP-defined RAN components such as the CU, the DU, and the RU, and the O-RAN specific components, such as the RIC and the O-Cloud. The functionality of each component and its interfaces are explained in depth. The last part of this chapter describes layer decoupling, which is the main concept of the O-RAN, and presents flexible deployment options to enable wireless network operators to evolve their networks to meet the demands of end customers.

# O-RAN architecture

The O-RAN architecture comprises nine network components and 19 inter-component interfaces. The architecture is based on the 3GPP RAN specifications, and has additional components and interfaces. The following figure shows the entire O-RAN architecture.




*O-RAN architecture*

The O-CU, the O-DU, and the O-RU extend the corresponding entities defined as the CU, the DU, and the RU in the 3GPP specifications. Providing the 3GPP features and interfaces, they are fully compliant with the 3GPP standards. In addition, they have O-RAN features, including E2 and O1 interfaces, to enhance RAN management and automation and to enable AI/ML-powered RAN optimization. Furthermore, the O-RAN provides a fully open and interoperable fronthaul interface between the O-DU and O-RU. While the legacy fronthaul interface standard, Common Public Radio Interface ( CPRI), needs vendor-specific information to operate, this open fronthaul interface provides full interoperability with the O-DU and O-RU from different RAN vendors to help network operators formulate true multi-vendor strategies.

Besides the O-CU, O-DU, and O-RU, the architecture has three more entities:

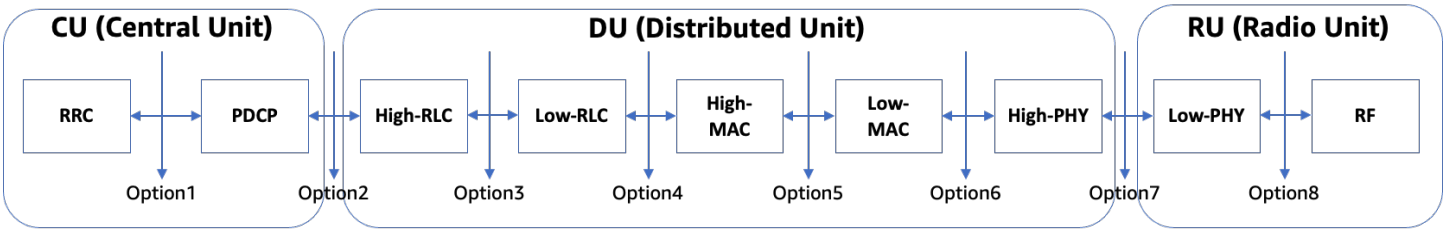
- **The O-Cloud**, or the infrastructure, including server hardware and networks, which is the fundamental layer where all O-RAN functions run.
- **The Service Management and Orchestration (SMO) framework**, which is a consolidation of a wide variety of management services that provide many network management-like functionalities. In an operator's network, the SMO may provide management services that go well beyond RAN management, and can include features such as core network management, transport management, and end-to-end network slice management. The Open Network Automation Platform (ONAP) project is one example of the SMO.
- **The RIC** (non-real-time and near-real-time RICs) is newly introduced by the O-RAN to control radio resources and improve RAN performance by mobility management, admission control, and interference management. Basically, the RIC uses SMO services such as data collection and provisioning of the O-RAN components. The RIC uses data analytics and AI/ML training and inference to determine RAN optimization actions. The RIC further controls and steers the O-CU and the O-DU via quality of service (QoS) policies. With fine-grained data collection and action reports from the O-CU and O-DU, the RIC sends QoS policies to them to allocate radio resources in a more accurate way, and to provide better end user experience.

  The RIC consists of two components: the non-real-time RIC (non-RT RIC) and near-real-time RIC (near-RT RIC). The non-RT RIC supports intelligent RAN optimization by providing policy-based guidance, ML model management, and enrichment information to the near-RT RIC, so the RAN can optimize radio resources. The non-RT RIC creates an inference model from ML training and the near-RT RIC performs RAN optimization actions based on the model. The near-RT RIC controls the O-CU and the O-DU, which allocate resources based on the policies from the near-RT RIC.

The O-RAN also defines interfaces among the O-RAN components.

- First, the E1 and F1, fully compliant with the corresponding 3GPP interface specifications, connect the O-CU Control Plane (CP) and O-CU User Plane (UP), and link the O-CU and the O-DU.
- Second, to support inter-RAN handover and core network communication, the O-RAN defines the X2, Xn, S1, and NG interfaces, which comply with the 3GPP specifications.
- Third, the open fronthaul interface between the O-DU and the O-RU is defined with the control, user, management, and synchronization planes, and comes along with the O-RAN interoperability test specifications.

The O-RAN open fronthaul specification provides wireless network operators the huge advantage of multiple RAN vendors. Regarding the function partitioning of the O-DU and the O-RU, several options were investigated to meet different implementation requirements. One option is to break the RAN physical layer into an upper and lower layer to minimize the required fronthaul bandwidth and transmission latency to the O-RU. The O-RAN chooses the split option 7-2x to support 5G features and frequency bands. Option 7-2x is a part of Option seven, where the physical layer functions up to RE mapping are implemented in the O-DU; the O-RU implements digital beamforming and all the later functions in the downlink. This split option requires fairly reasonable bandwidth and latency, compared to other split options.




*Fronthaul split options*

The O1, A1, and E2 interfaces are newly introduced in the O-RAN architecture. The O1 interface connects the SMO and the O-RAN components for component management. As all the components expose an O1 interface to the SMO, the SMO directly manages each O-RAN component, called a *managed element*, in the O-RAN OAM architecture. Working together with the O-Cloud, the SMO provides the entire management services through the O1 interface, which implements instance provisioning, fault supervision, performance assurance, tracing, software version management, and communication surveillance.

The O1 interface further supports non-virtualized elements or physical network elements. To support legacy RUs, the O-RAN OAM architecture allows two management interfaces, so that the RUs can be managed by the O-DU and by the SMO. In this case, the O-DU configures the O-RU and optimizes the operating parameters for a reliable and low-delay connection. The SMO manages the O-RU through the O1 interface, and provides generic component management services including provisioning, fault supervising, and software management.

The A1 and E2 interfaces are used for the RIC functionality. The RIC, as described in the previous section, is a new network component to achieve radio resource and mobility optimization by using AI/ML technologies. For RAN radio resource control, there are three control loops with different latency bands in the O-RAN architecture. The first loop is a closed loop with the O-DU and the UE for immediate radio resource requests and allocations. The second loop is formed with the O-DU and the near-RT RIC. The near-RT RIC provides interactive allocation policies as a function of the

current resource allocation status, the user experience and service quality, and the management policy from the non-RT RIC.

The Non-RT RIC is the controller of the last loop, and supplies resource policies and AI/ML trained models to the near-RT RIC by exploiting the gathered statistics and the existing configurations. Although the timing of these control loops is use-case dependent, it is expected that the typical run time in the non-RT RIC control loop is one second or more, the near-RT RIC control loop has a use-case run time of ten milliseconds (ms) to one second, and the last O-DU scheduler loop operates below ten ms. These control loops and their run times over the A1 and E2 interfaces should be considered when the O-RAN components are deployed.

## O-RAN management and infrastructure components

The SMO is a consolidation of a wide variety of management services for the wireless network operator. The SMO may provide management services for not only RAN but also core and transport networks to enable the establishment of end-to-end network slices and the automated management of the entire network. For the O-RAN components, the SMO performs the fault, configuration, accounting, performance, and security management. The SMO includes the non-RT RIC services for RAN optimization, and manages the infrastructure through the O2 interface with orchestration and workflow management services.

The ONAP is one of the SMO solutions. Hosted by the Linux Foundation, the ONAP has been implemented in collaboration of the O-RAN and ONAP communities. The ONAP release has aligned with the O-RAN specifications and provides managed services for the O-RAN. The ONAP further contains the non-RT RIC to support the RIC functionalities. The O-RAN Software Community, supported and funded by Linux Foundation and O-RAN, is also developing non-RT RIC features, as well as near-RT RIC and use cases.

The last new component of the O-RAN architecture is the O-Cloud. The O-Cloud manages physical resources like servers, networks, and storages, and hosts the relevant O-RAN functions. The O-Cloud exposes the O2 interface to the SMO, which provides secured communication and enables the SMO to manage the infrastructure and the life cycle of O-RAN network functions. The O2 interface is independent of specific infrastructure implementations in order to work with multiple clouds and bring a multi-vendor environment to wireless network operators.

The O-Cloud should provide physical or logical infrastructure resources and perform workload management for O-RAN network functions. The service of the O-Cloud includes resource discovery and administration, network function provisioning, network function Fault, Configuration,

Accounting, Performance, and Security (FCAPS), and software life cycle management. These services are further divided into two classes:

- The infrastructure management.
- The network function deployment.

The Infrastructure Management Services (IMS) communicates with the entity of the SMO, called the Federated O-Cloud Orchestration and Management (FOCOM). The sub-interface between them is defined as the O2-M interface. The second service class, the network function deployment, works with the Network Function Orchestrator (NFO) of the SMO via the O2-D sub-interface.

The IMS is responsible for physical resource allocation based on the request from the SMO and resource tracking and management. The IMS builds physical and logical inventories and shares them with the SMO through the O2-M interface. The SMO receives the inventory information from the IMS, updates its inventory accordingly, and makes a request to allocate a resource based on the inventory updates. The IMS also provisions infrastructure resources and flexibly matches the resource demands of the O-RAN network functions.

This elastic and flexible resource provisioning brings benefits, including agility, scalability, and cost saving. The Deployment Management Services perform deployment lifecycle management. The DMS deploys O-RAN network functions on the assigned resources of the infrastructure. The DMS scales up and down according to the demand of the network function. The DMS also monitors the status of network functions, and performs a proper recovery scheme if a network function is out of service.

To monitor the network functions, the O-RAN defines three types of telemetry data:

- Managed element telemetry.
- Deployment telemetry.
- Infrastructure telemetry.

*Managed element telemetry* is related to the network function status and collected via the O1 interface. *Deployment telemetry* monitors the deployed resource status, including CPU, network, and memory usage. It also monitors the ongoing deployment. *Infrastructure telemetry* monitors the health of the infrastructure components. For example, the telemetry data includes the capacity and resource utilization of the infrastructure. Deployment and infrastructure telemetry data is collected via the O2 interface.

# Infrastructure decoupling and deployment flexibility in O-RAN

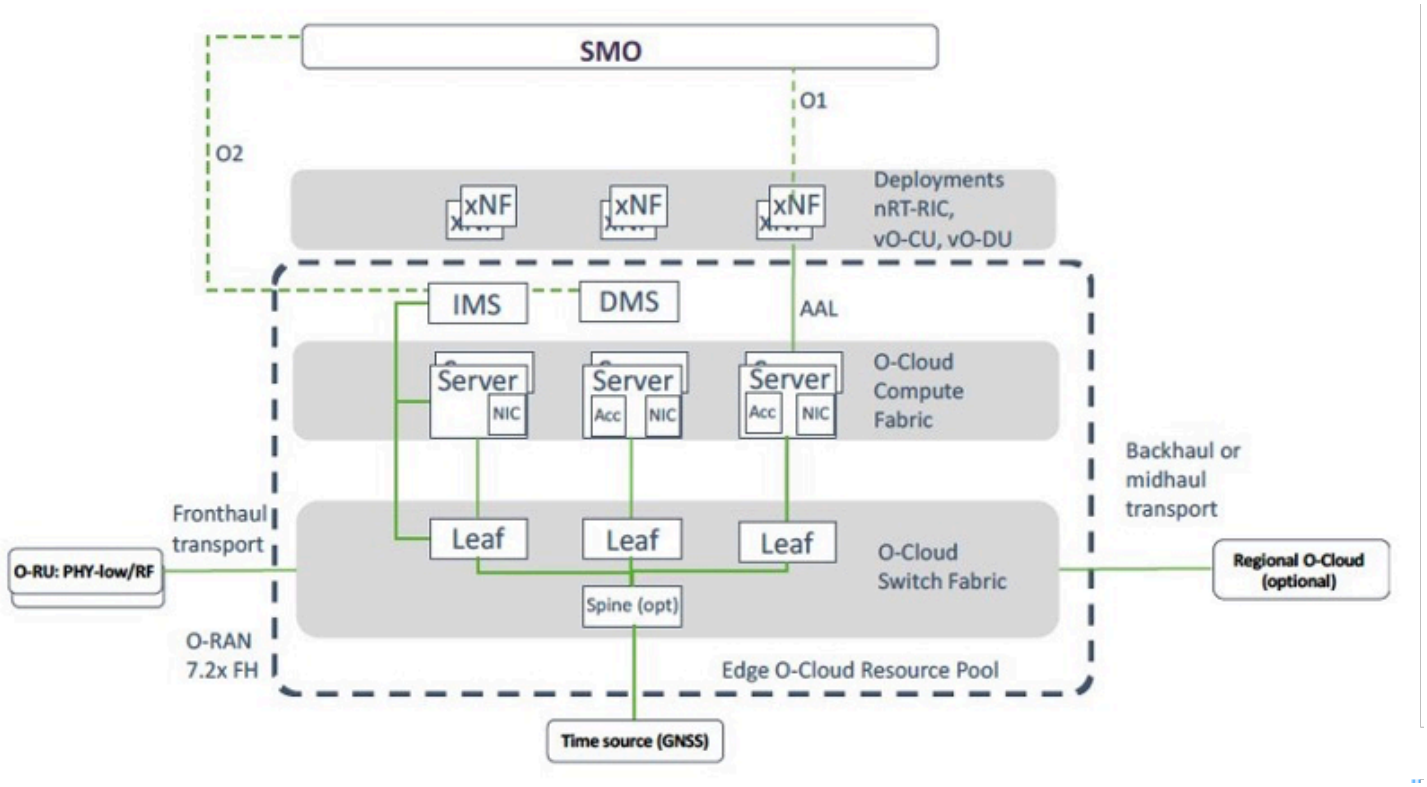
O-RAN pursues the complete openness in the RAN architecture. The interfaces among the O-RAN components are fully open. For example, the open fronthaul interface between the O-DU and the O-RU has complete functionality and interoperability. All other O-RAN interfaces develop openness with full functional features. From the perspective of infrastructure, O-RAN decouples the infrastructure and O-RAN network functions in pursuit of true network function virtualization as O-RAN defines three layers:

- The hardware layer.
- The middle layer.
- The top layer.

The *hardware layer* includes server blade and sled, network switches, storages, and so on. The *middle layer* or *cloud stack* consists of operating systems, cloud management software, and container or VM management software. The middle layer also includes the *accelerator abstraction layer* to enable the *top layer* to exploit accelerators for task offloading and high performance. The top layer comprises the O-RAN network function applications for the O-CU and O-DU.

The infrastructure, or O-Cloud, has the first two layers: the hardware layer and the middle layer. The O-Cloud is a set of hardware and software to provide cloud computing capabilities to host and run O-RAN network functions. The O-Cloud includes compute, networking, and storage components, and also provides specific hardware and software functions such as RAN physical layer accelerators and encryption/decryption accelerators. The O-Cloud may have sub-processors such as SoCs, GPUs, and Field-Programmable Gate Arrays (FPGAs) to enhance complex computations required in the RAN. The O-Cloud manages the hardware and software that belongs to the hardware and middle layers, and exposes the O2 interface and open APIs to the SMO and the O-RAN network functions. The interfaces between the O-Cloud and the O-RAN network functions are open APIs driven by open-source communities such as Kubernetes and Linux.

The hardware layer of the O-Cloud may have the O-Cloud Compute fabric with a collection of physical servers, or the O-Cloud Nodes, and the O-Cloud switch fabric with leaf and spine switches. The O-Cloud Nodes have hardware components such as CPUs, memory, storage, network interfaces, and accelerators. The switch fabric provides connectivity to other O-RAN components like the O-CU and the O-RU. The switch fabric also can be connected to the regional O-Cloud and the time source for GPS synchronization. The diagram from the O-RAN Cloud Architecture specification (O-RAN.WG6.CAD-v02.01) depicts this concept:




*O-RAN/O-Cloud function decoupling*

The software components of the O-Cloud include the Infrastructure Management Services (IMS) or Deployment Management Services (DMS), which work with the SMO for resource provisioning and function deployment. The O-Cloud also has lifecycle management of O-RAN network functions. Amazon Elastic Kubernetes Service (Amazon EKS) is a good choice to manage the O-RAN network functions. Amazon EKS is a managed Kubernetes control plane service that makes it easy to deploy, manage, and scale containerized applications. Amazon EKS runs native upstream Kubernetes, and is certified to be Kubernetes conformant. Applications running on Amazon EKS are fully compatible with applications running on any standard Kubernetes environment, whether running in on-premises data centers or public or private clouds. In collaboration of Amazon EKS, continuous integration and continuous delivery (CI/CD) pipelines can build automated lifecycle management. This whitepaper discusses lifecycle management in the next chapter.

With infrastructure management and resources, the O-Cloud provides multiple locations of service from cell sites to edge and regional clouds, and flexible deployment options to operators. In 5G networks, ultra-low latency is the key enabler for new services such as autonomous car driving, robot control, virtual reality, and augmented reality. Depending on the latency requirements of the service, the operator deploys O-RAN network functions to the place with the required latency.

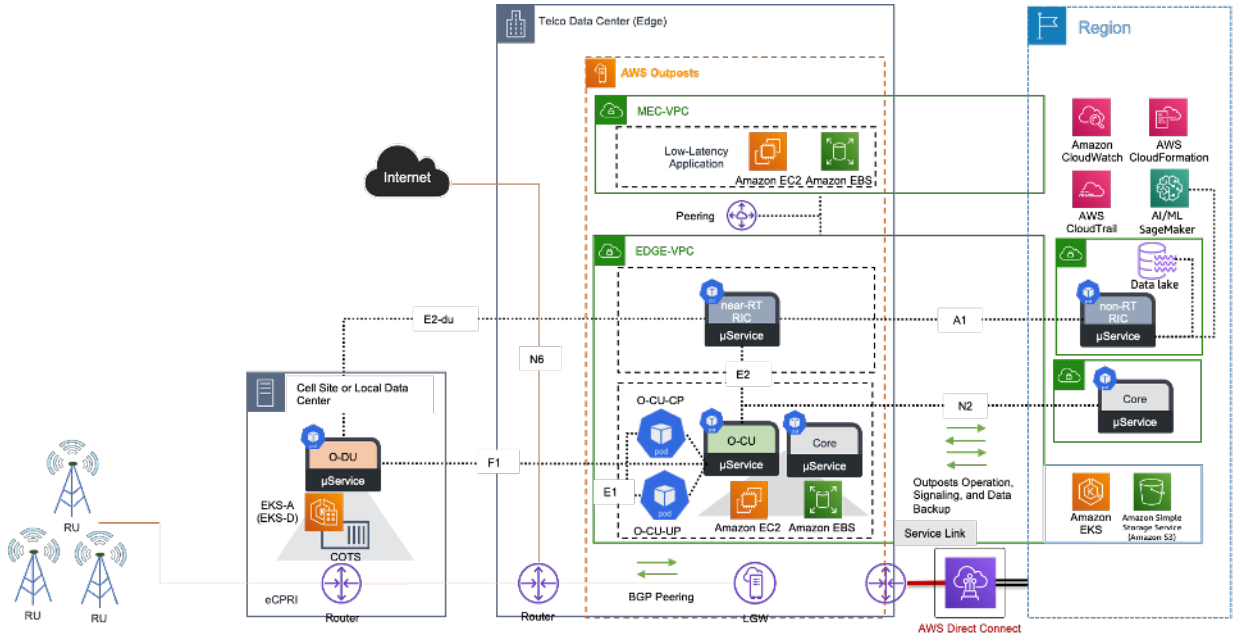
The O-DU can be either in cell sites or edge clouds, if the edge cloud meets the requirement of transmission latency or distance to the O-RU.

Centralization of the O-DU brings benefits to the wireless network operators in terms of capital expenditures (CAPEX) and operating expenses (OPEX). This centralization enables cluster placement of the O-DU to provide not only pooling of resources, but also easy maintenance and management. Edge clouds further host the O-CU and the user plane function of the core network or user plane function (UPF) to reduce the end-to-end latency to a few milliseconds. If even lower latency is required, the O-DU, O-CU, and UPF can sit in the cell site and provide low round-trip latency of less than one or two ms, as 3GPP and ITU-T defines Ultra-Reliable Low-Latency 5G for Industrial Automation.

AWS provides multiple solutions regarding O-RAN component deployment and open interfaces. The AWS Cloud is the most secure, extensive, and reliable cloud platform, offering over 200 fully-featured services from data centers globally. In addition, AWS Local Zones provide low-latency edge cloud services to allow for single-digit millisecond latency services for end users. AWS also has Amazon EC2 as a compute service, suitable for the O-DU of cell sites, so that operators can have the same experience from the AWS Cloud and a single pane of glass to manage all the infrastructure. The next chapter shows AWS functionalities and components for O-RAN operation on the AWS cloud.

# O-RAN architecture on AWS

As shown in the preceding figure, AWS can provide all required building blocks for O-RAN development and deployment. While non-RT RIC is located in the Region to use all the benefit of AWS data lakes and AI/ML gears, near-RT RIC and O-CU can be hosted at the edge site using the AWS Outposts and Amazon EKS services. At the far-edge site, O-DU can be placed on Amazon EKS Anywhere. Because this architecture is fully empowered by AWS services, service deployment and monitoring of each component can be done through AWS management and orchestration services such as AWS Software Development Kits (AWS SDKs), Amazon CloudWatch, and AWS CloudFormation, which provide true single panes of glass for the entire RAN operation. The following figure shows a reference architecture of O-RAN on AWS infrastructure hosting the O-RAN components.




*O-RAN reference architecture*

# O-RAN components on AWS

## RIC

Disintegration of RAN as defined by the O-RAN alliances provides an opportunity to use cloud concepts farther away from the radio stations. The RIC, near-RT and non-RT, provide mobility

operators with the ability to customize their RAN network, automate their network operations and optimization, and use AI/ML capabilities. The following sections expand on how AWS services enable RIC architecture.

Radio network conditions are constantly changing based on users' behaviors, environmental changes, interferences, and more. The randomness of the radio channels' characterization and the random nature of events impacting signal quality requires a RAN that reacts to these events to maximize the quality of its users' experience. Traditionally, CSPs tackle this problem through centralized self-organizing network (SON) capabilities, and network equipment provider (NEP)s features delivery and densification. The latter is often limited to individual NEPs, resulting in CSPs choosing radio equipment vendors for entire countries (or Regions) rather than mixing and matching based on the localized environmental behaviors.

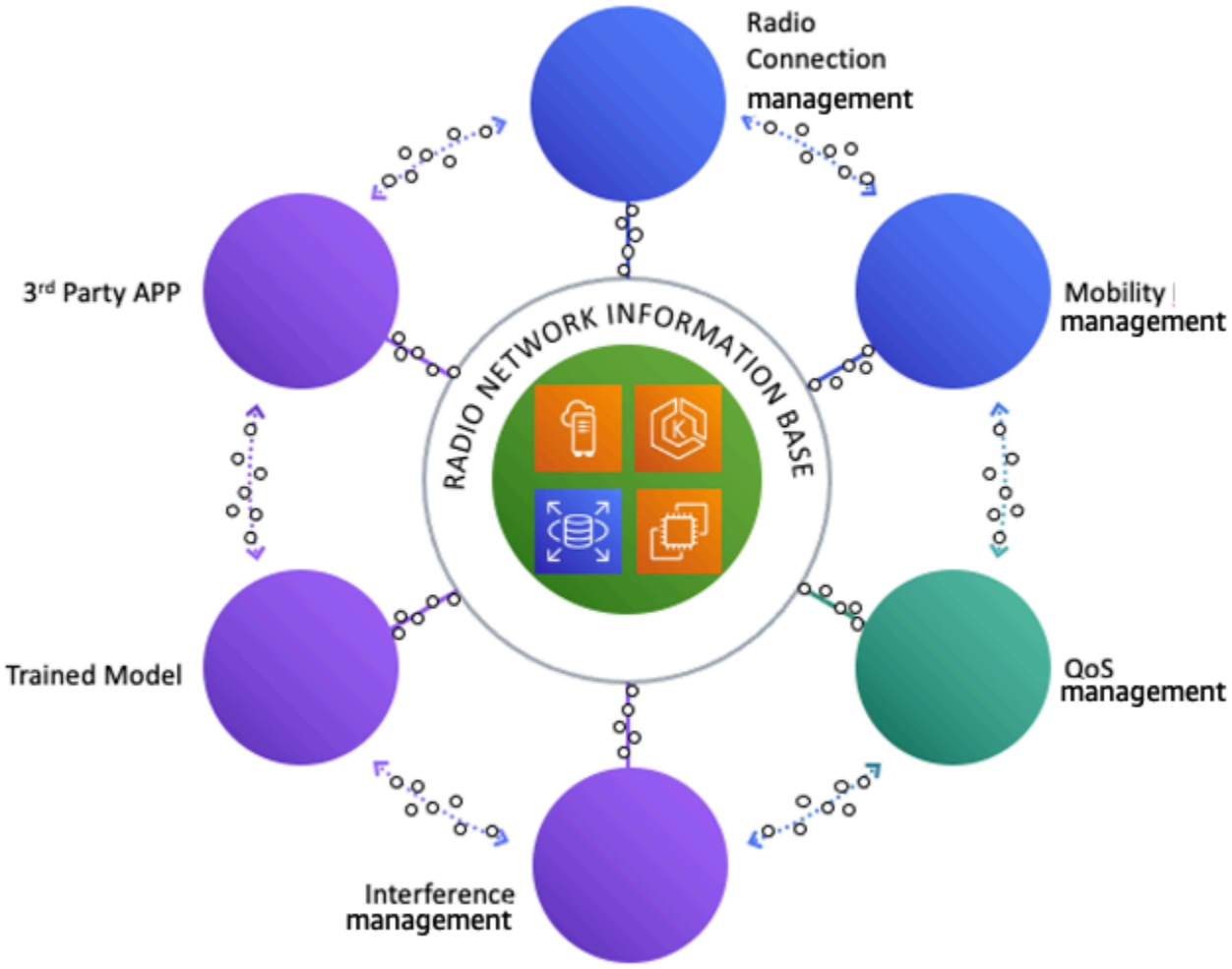
This idea of customizing approaches based on localized environments is the foundational idea behind RIC. As such, the RIC architecture needs to support agility, innovation, and elasticity, while reducing the overall cost and complexity of managing a mobility network. These characteristics are aligned with cloud benefits enabled by AWS.

## Near-RT RIC

The near-RT RIC has the following functions:

- Mobility management.
- Radio connection management.
- Quality of Service (QoS) management.
- Interference management.
- Trained models.
- Independent and extensible software plug-ins.

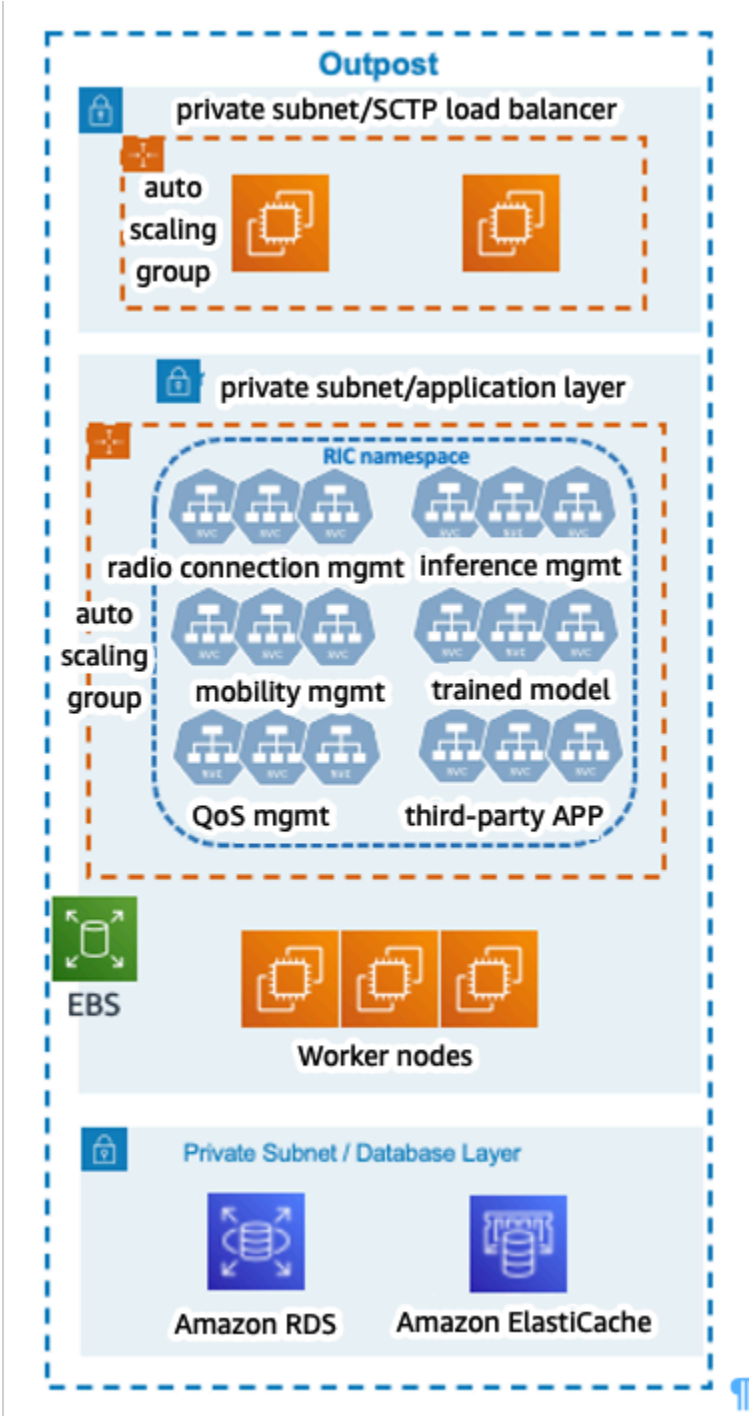
All these functions interact with one another, and are enabled by a common radio network information base. The latter provides near-RT RIC functions with an overview of the network RIC supports. This is illustrated in the following figure.




*Near-RT RIC components*

The near-RT RIC requires rapid low latency access to the network. This is enabled by AWS Outposts, which provides CSPs with a fully-managed service that offers the same AWS infrastructure, AWS services, APIs, and tools to their edge locations. AWS Outposts provides services such as Amazon EKS and Amazon ECS to support container-based applications, Amazon EMR clusters to support data analytics effort requiring immediate local processing, and Amazon RDS for relational databases.

Near-RT RIC ISVs can use Amazon EKS on Outposts to deliver RIC functions such as QoS managements. Amazon EKS enables non-RT RIC to scale with network conditions, upgrades RIC functions independently from one another, supports canary deployment of the new RIC version, and supports third-party RAN applications.




*Near-RT RIC AWS architecture*

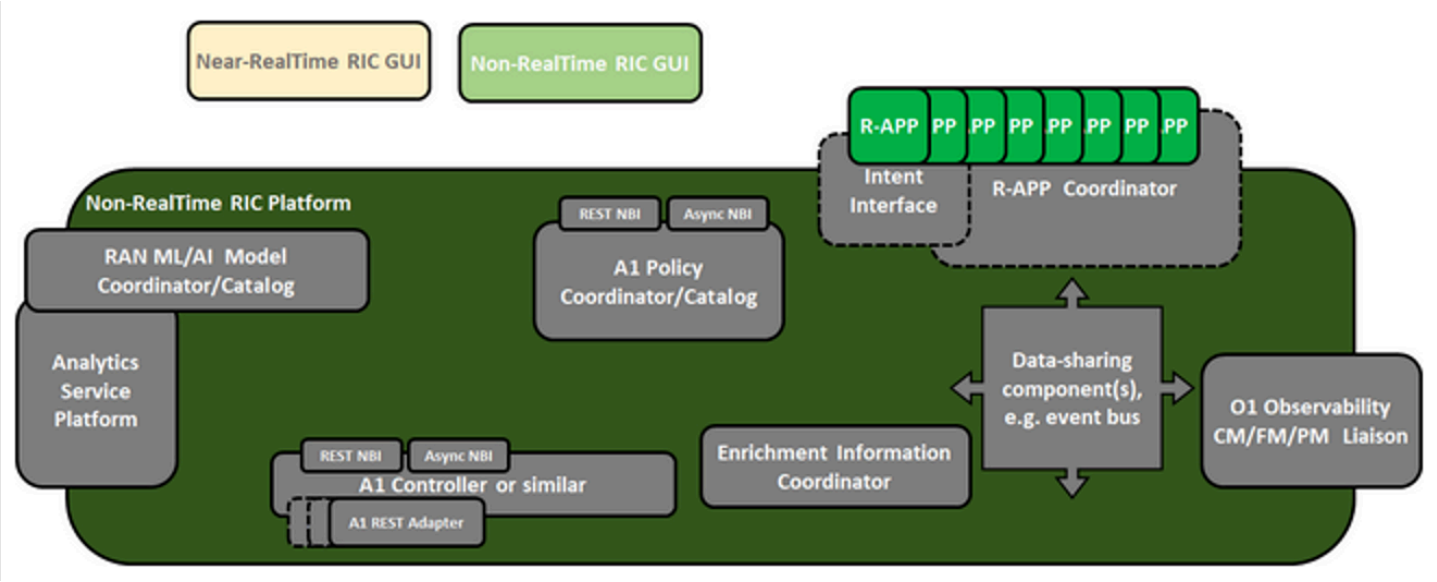
Amazon ElastiCache provides a fully managed Redis database, simplifying the hosting of RIC Shared Data Storage in support of the O-RAN SDL libraries. It provides sub-millisecond latency to support near real-time RIC applications such as mobility management, and supports the scaling necessary to support the entire set of RIC applications.

Amazon RDS provides a scalable, easy-to-set-up, and operationally efficient solution to support static and semi-static network configuration. Amazon RDS on Outposts supports mySQL and PostreSQL database engines.

## Non-RT RIC

The non-RT RIC provides the required intelligence to perform optimization of the RAN networks, uses data from across the operation support system (OSS) stack, and has access to AI/ML resources to build a model that can be applied to a given near-RT RIC, or a set of near-RT RIC. As illustrated in the following figure, the non-RT RIC has three logical roles:

- Ingestion of A1 messages.
- Hosting of non-real time applications (R-APP).
- Enrichment of A1 messages.



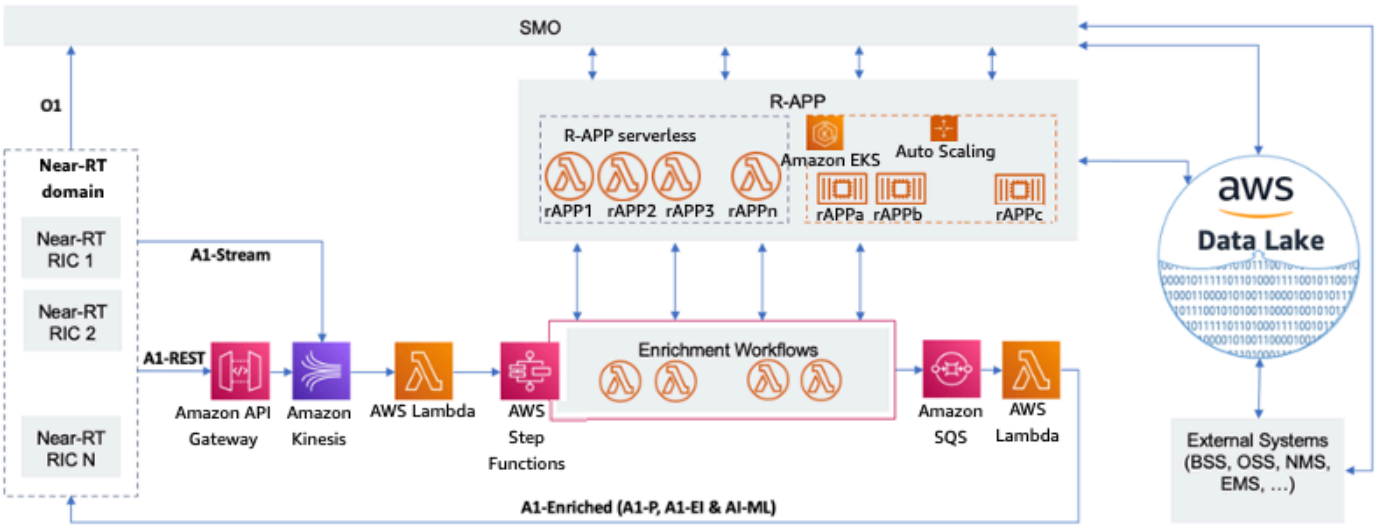

*Non-RT RIC functional view (ONAP)*

To support operators that plan for a one-to-many relationship between non-RT RIC and near-RT-RIC, you can use AWS services such as Amazon API Gateway, AWS Lambda, AWS Step Functions, and Amazon Simple Queue Service (Amazon SQS).

- Amazon API Gateway provides you with scalable services that make it easy to publish, maintain, and monitor the A1 interfaces between non-RT and near-RT RIC.

- AWS Lambda, a serverless event-driven compute service, provides you with the ability to perform A1 enrichment without having to provision or manage dedicated servers.
- AWS Step Functions provides you with a low-code visual workflow service that helps you orchestrate and automate A1 enrichment procedure.
- Amazon SQS is a fully-managed queueing service that enables you to communicate between R-APP at scale, while allowing you to set different prioritization for given A1 messages, both native and enriched.

The following figure illustrates a reference architecture for developing Non-RT RIC on AWS:




*Non-RT RIC on AWS*

The AWS Cloud enables you to host modern R-APP applications, specially designed to run on the Non-RT RIC as code by using AWS Lambda. Similarly, Amazon EKS provides you with a managed container service to run and scale your Kubernetes-based R-APP applications.

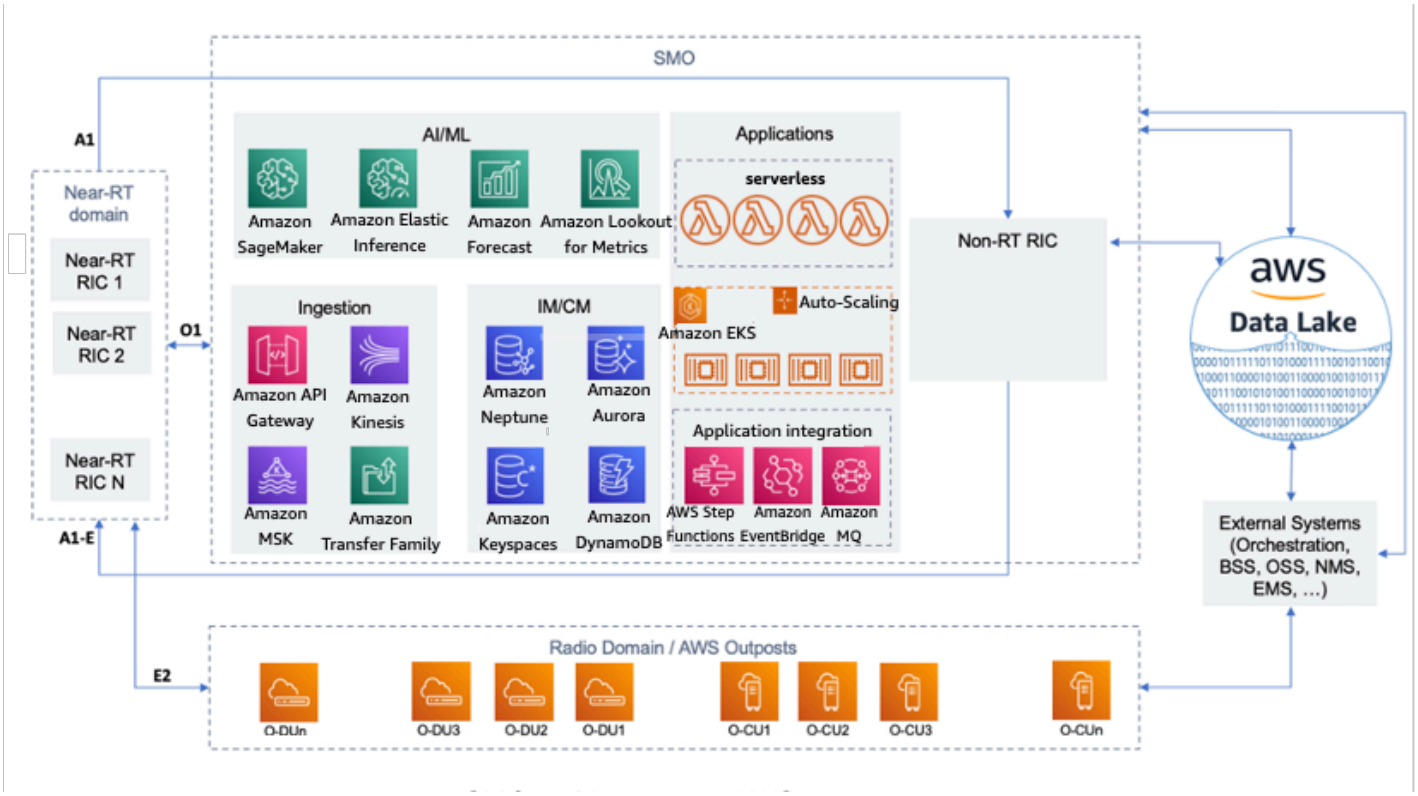
## SMO

Service Management and Orchestration (SMO) can be reduced to:

- Ingestion of OAM data.
- Applications hosting.
- Inventory and configuration data storage.

- API layer.

This section discusses how AWS services help you develop SMOs that benefit from scalability, performance, and reliability, and deploy them in minutes across your entire network. The following figure illustrates an SMO architecture on AWS.




*SMO architecture on AWS*

Ingestion of O1 messages is facilitated by Amazon API Gateway, Amazon Kinesis, Amazon MSK, and AWS Transfer Family. For example, when an O-RAN network element (NE) has a large configuration change, a large log, or a large performance data available, Amazon API Gateway facilitates the implementation of a REST request to initiate a file retrieval from the NE. The data transfer is facilitated by the AWS Transfer Family. Similarly, Amazon Kinesis and Amazon MSK provide you with scalable, reliable, managed solutions to ingest near real-time network events and near real-time configuration messages. SMO functions and applications can subscribe to a fully scalable data bus, and provide the required RAN network management and orchestration.

Use Amazon EKS to run Kubernetes-compliant SMO applications on AWS without the need to install and operate your own Kubernetes control plane. AWS Lambda provides you with the ability to benefit from the event-driven nature of SMO OAM functions by building SMO applications that

only use compute resources when needed. AWS Lambda is a serverless, event-driven compute service that lets you run code virtually without provisioning or managing servers.

Use AWS Step Functions to build orchestration logic.

Use Amazon EventBridge for event-driven integration with other Business Support Systems (BSS)/OSS/external applications, while Amazon SQS provides you with a fully-managed message queuing service that enables to decouple your SMO microservices and serverless applications.

AWS purpose-built database services for Graph DB, NoSQL, and RDBMS provide you with the ability to support inventory and configuration management use cases. They provide you with the scalability, reliability, and performance to decouple the database layer from your application to achieve a common data unification model shared with applications across the OSS stack.

Amazon API Gateway enables you to integrate SMO with external applications. For example, a REST API can easily be exposed for a service orchestrator to initiate a configuration change. Amazon API Gateway, being a fully managed service, makes it easy for you to develop, create, publish, maintain, monitor, and scale the APIs required northbound and southbound of SMO.

## O-CU

The O-CU hosts Radio Resource Control (RRC), Service Data Adaptation Protocol (SDAP), and PDCP protocols, and consists of the O-CU control plane (O-CU-CP) and the O-CU user plane (O-CU-UP). Because O-CU communicates O-DU through the F1 interface and the Core Network through N2 (for Access and Mobility Management Function (AMF)) and N3 (for UPF) in a low latency, it should be located at the edge data center. AWS Outposts provides a perfect option for building the edge data center of CSP, because it provides a fully-managed service that offers the same AWS infrastructure, AWS services, APIs, and tools to virtually any datacenter, co-location space, or on-premises facility for a truly consistent hybrid experience.

From the perspective of protocol stack, the O-Ran Central Unit control plane

(O-CU-CP) hosts the RRC and the control plane part of the PDCP protocol, while O-CU-UP hosts the user plane part of the PDCP protocol and the SDAP protocol. O-CU deals with upper layer protocol stacks, unlike O-DU and O-RU, which make it independent of the complex physical layer. This aspect enables you to virtualize and containerize O-CU easily, as you can with 5G Core Network function components.

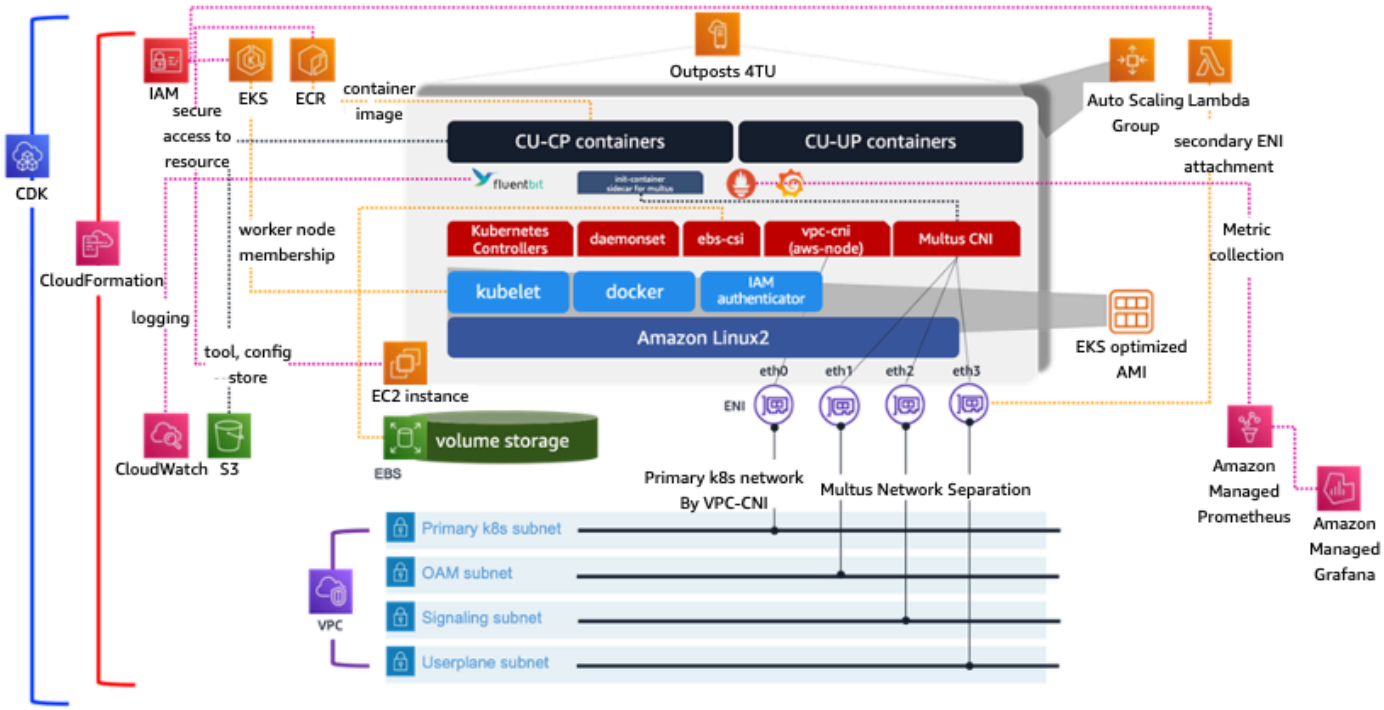
Because O-CU has to cover an aggregated set of O-DUs, the scalability, elasticity, and flexibility of using compute and storage resources would bring a huge benefit in terms of efficiency for resource

utilization and cost optimization. To maximize this benefit from the software architecture, people often choose container-based implementation. For the orchestration of containers, Kubernetes is often selected, not only for 5G Core Network providers, but also O-RAN SW providers. AWS provides Amazon EKS as a managed Kubernetes service with the control plane or primary nodes functionality. Worker nodes under the EKS cluster are created using Auto Scaling groups in cases of EKS managed node groups and self-managed node groups.

As with Core Network functions, O-CU is often required to support network segmentation such as having separate OAM, control plane, and user plane sub-networks. The best practice for having this network separation at the Kubernetes environment is using the Multus meta CNI Plugin. Amazon EKS now supports Multus CNI as an add-on package of EKS. As illustrated in the AWS official GitHub, creation of Multus-ready worker node groups can be automated with CloudFormation or CDK using a Lambda function and CloudWatch Event rule. Mostly O-CU-UP requires packet processing acceleration, generally using SR-IOV DPDK. Because of this, the best practice for selecting an instance type for a worker node group of O-CU-CP is using the latest generation of ENA-available instances, such as the C6 and M5 instance families.

For the container storage, the Amazon EBS CSI driver provides a container storage interface (CSI) interface that allows Amazon EKS clusters to manage the lifecycle of Amazon EBS volumes for persistent volumes. In the AWS environment, the O-CU container image and helm chart can be stored and managed in Amazon Elastic Container Registry (Amazon ECR). For the collection of KPI and metric for O-CU, you can use Amazon Managed Service for Prometheus and Amazon Managed Grafana.

This full stack of AWS tools for building O-CU on AWS from the bottom layer (through AWS Outposts) to the top of O-CU application and additional monitoring and orchestration layers (through CloudFormation, CloudWatch) are shown in the following figure in a high-level view.




*Full-stack view for O-CU design in AWS*

## O-DU and O-RU

The O-DU performs the Radio Link Control (RLC), Medium Access Control (MAC), and physical layer in the 3GPP specifications. The RLC builds information packets and recovers transmission losses by retransmitting lost packets. The MAC layer controls the physical layer, and allocates radio resources to transmit and receive packets over the air. The physical layer in the O-DU is responsible for link connectivity. The performance of physical layer dominates wireless link throughput and coverage where user experience would be mostly affected. The physical layer consumes heavy computational power for complex channel estimation and interference cancelling to improve the performance. To relieve its computational complexity, hardware acceleration techniques are commonly used in the O-DU.

AWS infrastructure and services bring significant advantages and benefits to CSPs. AWS provides tightly integrated hardware infrastructure and platform software by working with multiple hardware and RAN vendors. AWS management control, programmable APIs, and tools and services such as Amazon CloudWatch improves operation efficiency and relieves undifferentiated heavy lifting of infrastructure management. AWS automation services, CI/CD with AWS CodePipeline, and automated life cycle management will speed up time to market and enhance operational resiliency.

With the cost savings by AWS, total cost of ownership (TCO) benefits will come in the form of enhanced staff productivity, less management efforts, and flexible payment like pay-as-you-go.

The O-DU can run on Amazon EKS Anywhere. AWS provides Amazon EKS Anywhere for CSPs who want to keep the existing Commercial-Off-The-Shelf (COTS) hardware, but build an end-to-end platform and management framework from cell sites to an AWS Region cloud. As a CaaS layer, EKS Anywhere runs on bare metal servers and operates clusters on the CSPs on-premises data centers with fully managed Kubernetes services. EKS Anywhere can support customer designated devices such as L1 accelerators and network interfaces supporting Precision Time Protocol (PTP), and enable low-latency access to the devices by Single Root I/O Virtualization Container Network Interface (SR-IOV CNI) and Data Plane Development Kit (DPDK). CSPs can use the EKS console to view all the Kubernetes clusters running on cell sites, in AWS Outposts, Local Zones, and Regions. This enables customers to have a single pane of glass and build an end-to-end management framework to deploy 5G CNFs from O-DU to 5G Core Network functions.

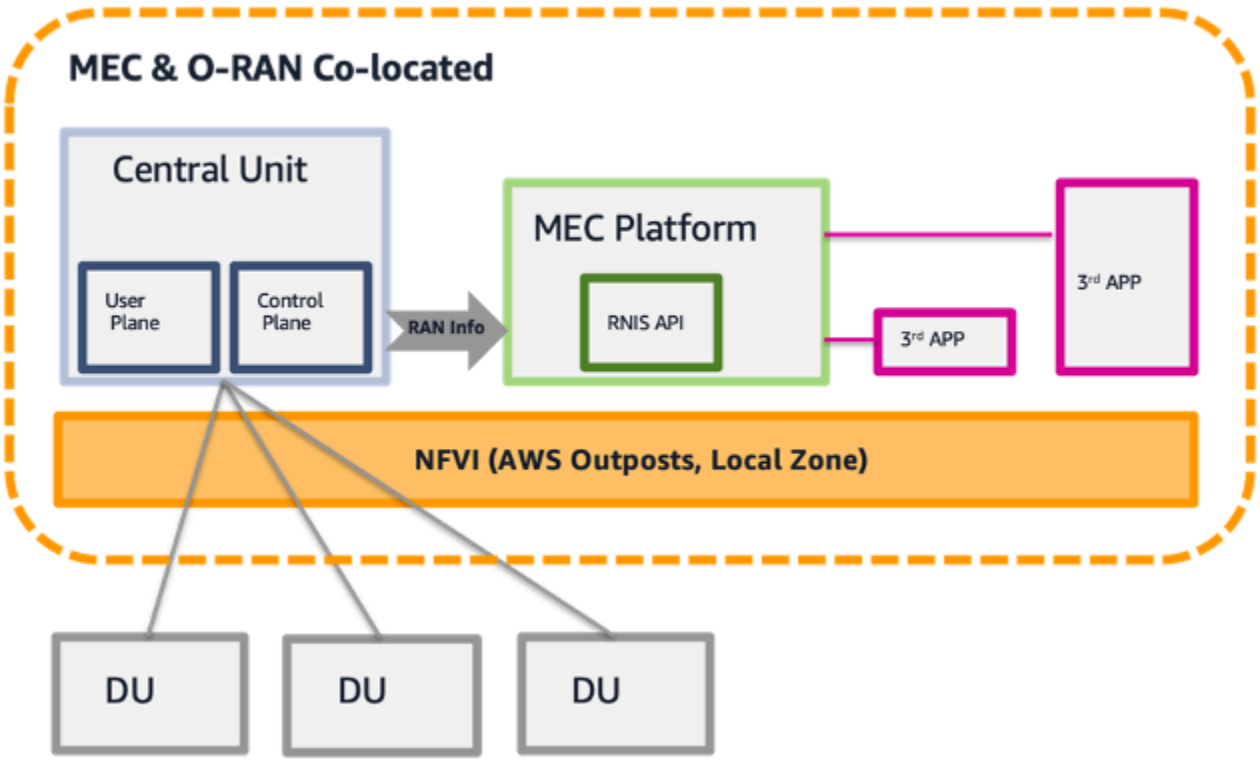
# O-RAN use cases

## MEC-RAN integration for low-latency use case

As described in the *5G Network Evolution on AWS* whitepaper, AWS compute and edge services such as EKS, EC2, Outposts, and Local Zone can provide a hosting environment for a Multi-Edge Computing (MEC) platform and application, When the CU is collocated with an UPF and MEC platform and application at the edge site, this configuration can provide a local breakout of user traffic at a distributed edge site that is closer to the mobile user, which can result in a low-latency service access such as Ultra-Reliable Low Latency (URLLC).

This configuration allows user traffic to be consumed locally at the edge site without pumping traffics to the backhaul network, which is efficiently applicable to high-bandwidth services with regard to saving the cost of backhaul network. In addition, if you collocate a CU, UPF, and MEC together, on the same network functions virtualization infrastructure (NFVI) layer such as AWS Outposts, it can help for the MEC platform and application to use a network quality status for the Radio Network Information Service (RNIS) through the API exchange within the platform.

Co-hosting of CU and UPF-like network functions (NFs) and MEC applications on the AWS brings the benefit of a single pane of glass for the orchestration for all network and service applications.




*MEC and O-RAN collocation on the AWS*

## RIC-CU/DU operation to optimize radio resources (traffic steering and QoE optimization)

An O-RAN architecture on AWS enables ISVs (and DSPs) to dynamically interact with the radio resources, allowing them to dictate how compute and network resources are allocated to steer traffic, improve QoS, and control Quality of Experience (QoE).

As discussed earlier, the separation of RU/DU/CU provides an opportunity to steer traffic from RUs to a pool of DUs, steer control traffic from a DU to a pool of CUs, and steer traffic from RAN to a pool of 5GC resources, such as from DU to a pool of UPFs.

AWS helps you achieve traffic steering, as defined in 3GPP TR 23.793. By using a Telco data lake on AWS, operators can feed telemetry, infrastructure, and application data from their access networks (5G and 4G), from UE environments (such as Wi-Fi, 5G private networks, and so on) to predict when it becomes beneficial to offload PDU sessions from the 5G radio network to an underlying 4G network, or to a Wi-Fi network.

Amazon SageMaker AI provides RF engineers, DSPs/ISVs data scientists, and DSPs/ISVs developers with the ability to build ML radio applications and models to provide the instructions to UEs, DUs, and CUs with the steering mechanisms that are advantageous for a given user. As discussed earlier, these mechanisms work with the near RT-RIC to address traffic steering use cases.

AWS enables you easily use data from a multitude of sources when combined into a data lake. This is particularly useful for QoE optimization, where solely using RAN resources doesn't provide you with the user context. By combining data across the spectrum of available data such as BSS and OSS, QoE optimization algorithms are more accurate because of that data enrichment. With the simplification that AWS Lake Formation brings to your data lake management and the ease of use of Amazon SageMaker AI, RF engineers can build QoE models that are rich and accurate.

## Network slicing, and service level specifications (SLS) fulfillment

The idea of network slicing is to create virtualized logical networks over a physical network which consists of RANs, core networks, and transport networks. This virtual network overlaying allows the end customers, including business companies, to have an isolated and tailored network connectivity for their own business purpose.

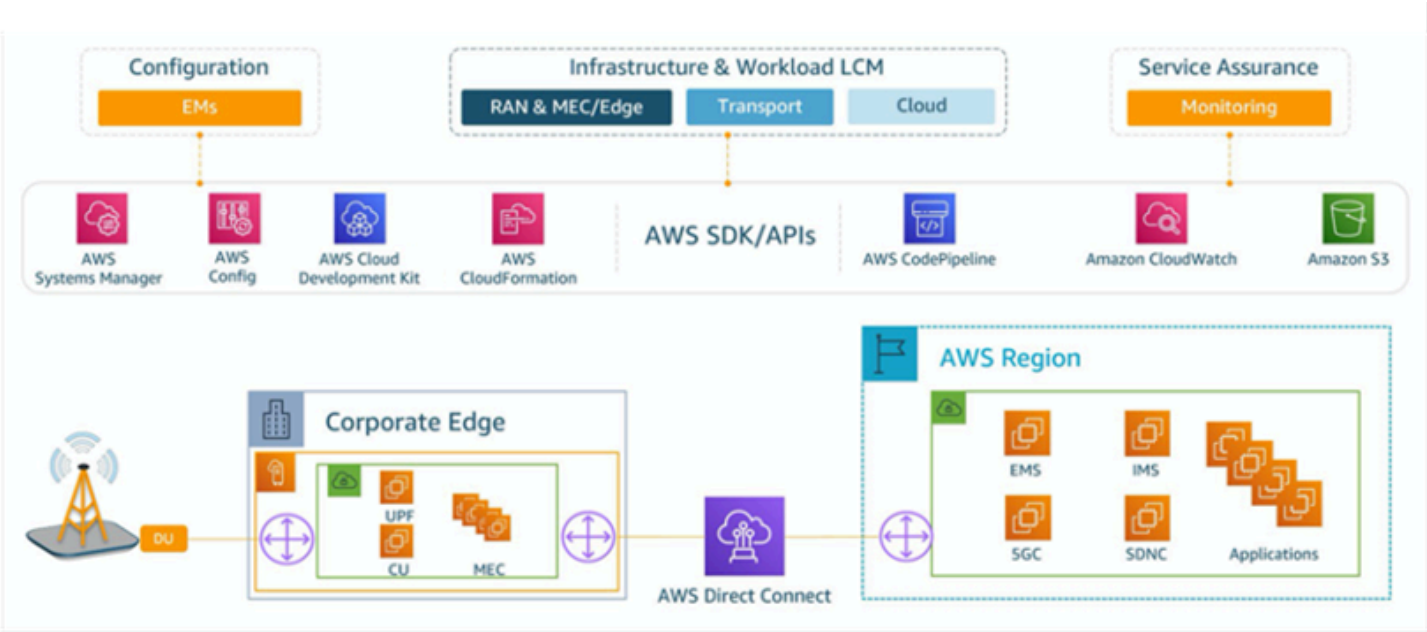
For example, a customer can establish a network enabling low latency and ultra-reliability communications for a critical Internet of Things (IoT) use case. Another example would be the networks of connected cars. The concept of connected cars has been introduced with in-car entertainment, assisted or fully automated driving, and maintenance data gathering. On the 5G network, car manufacturers could create a virtual network with the required service level specifications to accommodate millions of connected cars.

Network slicing on the 3GPP network components like RAN has been standardized in terms of interface and functionality. In addition, one important component is needed to manage and orchestrate network slicing across RAN and core network. This management and orchestration entity oversees the operator's entire network, and creates or deletes a slice based on the customers' demands. The management and orchestration are responsible for allocating network resources and managing the life cycle of network slices.

The network slicing management and orchestration can be implemented on the AWS Cloud, as shown in the following figure. By using AWS services such as Amazon EKS (for CNF management), Amazon RDS (for data management), AWS Lambda, CodePipeline (for lifecycle and resource management), and AWS CloudFormation (for infrastructure management), the network slice

manager can create a slice by allocating and configuring network resources through the AWS programmable infrastructure and well-defined APIs.

All the created slices and allocated resources can be monitored by the network slicing manager, which provides full visibility to the operator via a graphical view. In addition, AWS provides APIs to allocate resources on AWS Outposts, which are placed in the operator's own data centers or corporate edge sites so that the network slicing manager can control the on-premises resources. The following figure shows the 5G network architecture and the AWS services for network slicing.




*Network slicing manager architecture on AWS*

All management operations are performed via AWS APIs, which enables network operators to have no dependency on specific resource mapping across a wide range of network domains. The consistent infrastructure APIs further allows the development of new services such as AI-powered monitoring and service assurance, by using AWS services such as Amazon SageMaker AI.

## DevOps, CI/CD, and network management (a single pane of glass)

As described in the *5G Network Evolution on AWS* whitepaper, as 5G networks are increasing in complexity in terms of their heterogeneous network environment as well as microservice-based architecture, CNF lifecycle management should be fully automated to maximize the efficiency of operation in scale and minimize the cost. In case of a CSP, the network would be built out of

multiple instances of CNF microservices, supplied from various RAN vendors as binary Docker images. These images regularly undergo updates from individual vendors to release new features, and need to be deployed as updates into the network environment.

Multiple CD pipelines are deployed per individual vendor, that are kicked-off with the upload of updated Docker images or configuration (such as helm charts and YAML files). Each respective pipeline runs through various stages for vetting the updates by deploying it first on test environments for unit testing, later in the staging environment for system-level integration testing, and finally in production using blue/green, canary-based deployments. AWS CodePipeline is used in the source stage of the pipeline with AWS CodeCommit as a configuration repository and Amazon ECR for container registry, whereas in the test and deployment stages, AWS CodeBuild is commonly used.

From an RIC perspective, it is important to manage each characteristic of underlying network resources such as CPU, hardware accelerators (such as FPGA, GPU, ASIC), throughput, latency, jitter, and so on to optimize performance with deploying NFs on the appropriate NFVI. To achieve the closed loop automation and flexible deployment, tight integration of AWS CI/CD pipeline and DevOps tools with RIC and SMO is required.

# Conclusion

The O-RAN architecture is gaining interest from wireless network operators for implementing 5G networks. Driven by the O-RAN alliance with more than 230 operators and vendors, the O-RAN architecture defines a disaggregate RAN structure with full inter-operability, using cloud infrastructure as a common network platform.

This paper outlines O-RAN on AWS, and what the O-RAN components, including the RIC, the SMO, the O-CU, and the O-DU would look like on the AWS Cloud infrastructure. The breadth and depth of AWS services can also benefit O-RAN. AWS services and infrastructure can be not only a place to easily implement a virtualized 5G network, but also a unified management platform to control all the O-RAN components via a single pane of glass. The AWS programmable cloud can enable important 5G use cases and features, which include the edge cloud for low-latency applications, RIC network optimization, network slicing, DevOps, and continuous evolution. In addition, the AWS Cloud has a wide range of open ecosystems, with various partners and proven O-RAN components.

As specified in the O-RAN architecture, the cloud platform by AWS with fully programmable management and a wide-open ecosystem introduces the multi-vendor model, and multiple ways of efficient automation to the wireless operators who pursuit business differentiation and leadership in the 5G era. The operators on the AWS Cloud can focus their business applications and ideas to monetize the network and enhance end users' experience. The AWS O-RAN Cloud, together with the operators, accelerates time-to-market, reduces TCO, and brings the flexibility and ability to automate network operations and new service rollouts, to reinvent 5G networks and reimagine the customer experience.

# Glossary

- **3GPP** – 3rd Generation Partnership Project
- **5GC** – 5G Core network
- **AAL** – Accelerator Abstraction Layer
- **ACC** – Accelerator
- **AMF** — Access and Mobility Management Function
- **AMI** – Amazon Machine Image
- **CNF** – Containers or Cloud-Native Network Functions
- **CSP** – Communication Service Providers
- **CM** – Configuration Management
- **CNI** - Container Network Interface
- **COTS** — Commercial Off-The-Shelf
- **CP** — Control Plane
- **CSI** — Container Storage Interface
- **CU** – Central Unit
- **CUPS** — Control-User Plane Separation
- **DMS** – Deployment Management Services
- **DPDK** – Data Plane Development Kit
- **DSP** — Digital Service Provider
- **DU** – Distributed Unit
- **eCPRI** – enhanced Common Public Radio Interface
- **ENI** – Elastic Network Interface
- **eNodeB** – Evolved Node B
- **FCAPS** – Fault, Configuration, Accounting, Performance, and Security
- **FM** – Fault Management
- **FOCOM** – Federated O-Cloud Orchestration and Management
- **FPGA** — Field-Programmable Gate Array
- **GNSS** – Global Navigation Satellite System
- **gNodeB** – next Generation Node B

- **HW** — Hardware
- **IMS** – Infrastructure Management Services
- **ITU-T** – International Telecommunication Union Telecommunication standardization sector
- **LCM** — Lifecycle management
- **LTE** – 3GPP Long-Term Evolution technology
- **MAC** – Medium Access Control
- **MEC** — Multi-Edge Computing
- **NBI** – Northbound Interface
- **NE** — Network Element
- **Near-RT RIC** – Near-Real-Time RAN Intelligent Controller
- **NF** — Network Function
- **NFVI** — Network Functions Virtualization Layer
- **NFO** – Network Function Orchestrator
- **NIC** – Network Interface Controller
- **Non-RT RIC** – Non-Real-Time RAN Intelligent Controller
- **NR** – 3GPP New Radio technology
- **OAM** – Operations, Administration and Maintenance
- **O-Cloud** – A cloud computing platform that meet O-RAN requirements to host the relevant O-RAN functions
- **O-CU** – O-RAN Central Unit
- **O-CU-CP** — O-Ran Central Unit control plane
- **O-DU** – O-RAN Distributed Unit
- **O-RAN** – Open Radio Access Network
- **O-RU** – O-RAN Radio Unit
- **OSS** — Operation Support System
- **PDCP** – Packet Data Convergence Protocol
- **PHY** – Physical Layer
- **PM** – Performance Management
- **QoE** — Quality of Experience
- **QoS** — Quality of Service
- **R-APP** —Real-time Application

- **RAN** – Radio Access Network
- **RF** – Radio Frequency
- **RIC** – RAN intelligent controller
- **RLC** – Radio Link Control
- **RNIS** — Radio Network Information Service
- **RRC** – Radio Resource Control
- **RT** — Real Time
- **RU** – Radio Unit
- **SCTP** – Stream Control Transmission Protocol
- **SDAP** — Service Data Adaption Protocol
- **SDNC** – SDN Controller
- **SLS** — Service Level Specifications
- **SMO** – Service Management and Orchestration
- **SON** — Self Organizing Network
- **SR-IOV** – Single Root I/O Virtualization
- **TCO** — Total Cost of Ownership
- **UE** – User Equipment
- **UP** — User Plane
- **NEP** — Network Equipment Provider
- **UPF** – User Plane Function
- **URLLC** — Ultra-Reliable Low Latency
- **xAPP** – A software tool used by a RAN Intelligent Controller (RIC) to manage network functions in near-real time
- **xNF** – The combination of PNF and VNF; Network Function

# Contributors

Contributors to this document include:

- Hoon Chang, Ph.D., Principal SA, WWCS Telco, Amazon Web Services
- Young Jung, Ph.D., Principal Tech Leader 5G, WWCS Telco, Amazon Web Services
- Tetsuya Nakamura, Senior Manager, WWCS Telco, Amazon Web Services
- Aymen Saidi, Principal, Architecture and Product, AWS – 5G, Amazon Web Services
- Salil Sawhney, Senior Manager, AWS-5G, Amazon Web Services

# Document revisions

To be notified about updates to this whitepaper, subscribe to the RSS feed.

| Change | Description | Date |
| --- | --- | --- |
| Initial publication | Whitepaper published. | December 2, 2022 |

# Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

# AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.