O-RAN.WG6.CAD-v02.00

Technical Report

Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN

Revision History

Date	Revision	Company	Description
2019.01.18	V0000.00	AT&T, Orange, Lenovo, …	Template with initial scenarios.
2019.01.29	V00.00.01	Editor (AT&T)	Updates to terminology, miscellaneous other updates
2019.02.07	V00.00.02	Editor (AT&T)	More definitions in 2.1, New Sec 4 on Overall Architecture, expansion/ updates of sec 5 Profiles, added Sec 6 OAM placeholder.
2019.03.18	V00.00.03	Editor (AT&T)	Many additions in content and section structure.
2019.04.01	V00.00.04	Editor (AT&T)	Some restructuring and combining of early sections, and more discussion on scope and context. Addition of implementation consideration section, including performance. Added optional Fronthaul GW. Provided framework discussion in each scenario's subsection. Other updates.
2019.04.10	V00.00.05	Aricent, Red Hat, KDDI, Ciena	Updates to include comments before April 11 review. Comments from RaviKanth (Aricent), Pasi (Red Hat), Shinobu (KDDI), and Lyndon (Ciena).
2019.04.15	V00.00.06	Editor (AT&T)	Updates to include some updates from comments from April 11 review.
2019.04.24	V00.00.07	Editor (AT&T)	Updates of diagrams to address comments, additional figures on scope, and other changes to address April 11 review comments.
2019.05.01	V00.00.08	KDDI	Updates to diagrams for Scenarios A and B. Modifications per KDDI regarding C.2.
2019.05.12	V00.00.09	KDDI, Red Hat, Editor (AT&T)	Updates based on meeting discussions, subsection additions based on proposals.
2019.05.15	V00.00.10	Editor (AT&T)	Clean-up in preparation of creating a baseline document – marking of many comments as done, adding editor notes where needed, and other clarifications.
2019.05.20	V00.00.11	Editor (AT&T)	Continued clean-up in preparation of a baseline.
2019.05.29	V00.00.12	Editor (AT&T)	Continued clean-up in preparation of a baseline.
2019.06.04	V00.00.13	Wind River, China Mobile	Major additions to the Cloud requirements in section 5.4 and Appendix B by Wind River, plus updates to the Fronthaul section from China Mobile. Various additional minor updates.
2019.06.13	V00.01.00	Editor (AT&T)	This is the same as V00.00.13, but with renumbering to indicate this is the initial baseline for comment, V00.01.00
2019.06.14	V00.01.01	Wind River, AT&T	This includes updates from CRs discussed and agreed to on the June 13 call: • Wind River contributions on adding a figure for NUMA illustration and a major enhancement of Sec 9.1 on cache • AT&T contribution to add material on centralization of O-DU/O-CU resources, to Sections 5.1 and 6.2 • Update of figures to address Open Fronthaul comments (discussed June 6)
2019.07.05	V00.01.02	Editor (AT&T), based on meeting	Updates to address several CRs: • Multiple editorial items:

		discussion	○ Draft text to address 5G/4G scope in Sec 1.2 – further discussion via separate CR ○ Statement in 5.2 about performance to focus on delay ○ Statement in 5.7 about transport ○ 5.8; update of Figure 13 to indicate cloud locations. Added MEC text that to address MEC comment during call. ○ Delay and loss table updates in 6, and statement in 5.2 • Former 9.1 and 9.3 sections of Appendix B (on cache and storage details) will be transferred to Tong's document (Reference Design). • Update the O-DU pooling analysis in Section 5.1.3.
2019.07.18	V00.01.03	AT&T, Red Hat, TIM, Intel, Ericsson	Updates to address multiple CRs, through July 18: • Address NSA aspects in scope • Addition of 5.3 (Acceleration) • Removal of Scale up/down appendix, and note for future study • Update of delay figure in 5.2. • Update of Figure 4 • Replacement of Zbox concept with O-Cloud, and all related updates.
2019.08.02	V00.01.04	AT&T, Wind River, Red Hat	Updates to address multiple CRs, discussed on Aug 1: • Update Section 5.6, merge in sec 7, explain some fundamental operations concepts. • Update the sync section to point to work in other WGs, and say that text will wait until CAD version 2. • Update the delay section (5.2.1) • Remove notes that refer to items that will not receive contributions in version 1. Remove comments that are no longer relevant. • Remove Appendix A
2019.08.09	V00.01.05	Red Hat, TIM, DT, Editor (AT&T)	Updates to address multiple CRs and DT review comments, discussed on Aug 8. • Update 5.2.1 to address non-optimal fronthaul, and to correct some equations • Update 5.6 to add a figure showing the O1* interface • Addressed a range of comments by DT, some editorial, some more involved.
2019.08.16	V00.01.06	Ericsson, Wind River, AT&T	Updates to address multiple CRs and DT review comments, discussed on Aug 15. • Updates to address Ericsson's comments • Update to address DT's request to define vO-DU tile • Update of the Cloud Considerations section (5.4), mostly for restructuring to remove duplication, but to also add material for VMs or Containers where necessary to provide balanced coverage. • Additional updates: Many resolved and obsolete Word comments have been removed in anticipation of finalization. • References to documents that are not finalized have been removed.
2019.08.23	V00.01.07	AT&T	Updates to reflect: • Updates of the O-DU pooling section based on Aug 20 discussion • Management section updates are to address comments made on Aug 15 discussion, particularly regarding the use of the term domain manager and its role in an ME, and the location of O1 terminations • Edits to remove references to O-RAN WGs, and make

			updates of the revision history. • Addition of standard O-RAN Annex ZZZ
2019.08.26	V00.01.08	Editor (AT&T)	• Clean up of references and cross references to them • Removed Word comments • Removed cardinality questions in Scenarios A (removed 6.1.1) and Scenario B
2019.08.26	V00.01.09	Editor (AT&T)	Final minor comments during Aug 27 WG6 call, in preparation for vote.
2019.10.01	V01.00.00	Editor (AT&T)	Update of Annex ZZZ, page footers, and addition of title page disclaimer.
2020.01.17	V01.00.01	Editor (AT&T)	Merged the following CRs, but with • ATT-2019-11-19 CADS-C CR ATT-CAD-010 acceleration 01.00.00 • WRS 2019-12-04 CAD-C 01.00.00 rev 1
2020.02.09	V01.00.02	Editor (AT&T)	Simplified 5.6. • Removed 5.6.1, 5.6.2 – replaced it with pointers to O1, and O2 specification. • Incorporated NVD comments on 5.3 and 5.4 addressing inline acceleration as an option
2020.03.03	V01.00.03	Editor (AT&T)	• Updated 4, 4.1 to reflect the latest O-RAN architecture • Incorporated comments on 5.6 to include O1, O2 references. • Updated 4.3 with O-Cloud description and definitions of key components of O-Cloud • Updated 5.3, Figure 15 to reflect O-Cloud reference figure in 4.3
2020.03.09	V01.00.04	Orange	• Various minor editorial modifications, take them as suggestions for better readability…
2020.03.10	V01.05	Editor (AT&T)	• Incorporated Ericsson comments provided on v01.00.02 • Updated 1.1 to include O-RAN Architecture description • Added definitions for O-RAN Physical NF, O-RAN Cloudified NF
2020.03.14	V01.06	AT&T, Orange	Minor editorial modifications, make this version ready for WG6 internal review and voting
2020.03.20	V02.00	AT&T, Orange	Minor editorial, make this version ready for TSC review and voting

Table of Contents

Table of Figures

Table of Tables

1 Scope

This Technical Report has been produced by the O-RAN Alliance.

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

Version x.y.z

where:

x the first digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc. (the initial approved document will have x=01).

y the second digit is incremented when editorial only changes have been incorporated in the document.

z the third digit included only in working versions of the document indicating incremental changes during the editing process.

1.1 Context; Relationship to Other O-RAN Work

This document introduces and examines different scenarios and use cases for O-RAN deployments of Network Functionality into Cloud Platforms, O-RAN Cloudified NFs and O-RAN Physical NFs. Deployment scenarios are associated with meeting customer and service requirements, while considering technological constraints and the need to create cost-effective solutions. It will also reference management considerations covered in more depth elsewhere.

The following O-RAN documents will be referenced (see Section 5.6):

- OAM architecture specification [8]
- OAM interface specification (O1) [9]
- O-RAN Architecture Description [10]

The details of implementing each identified scenario will be covered in separate Scenario documents, shown in green in Figure 1.

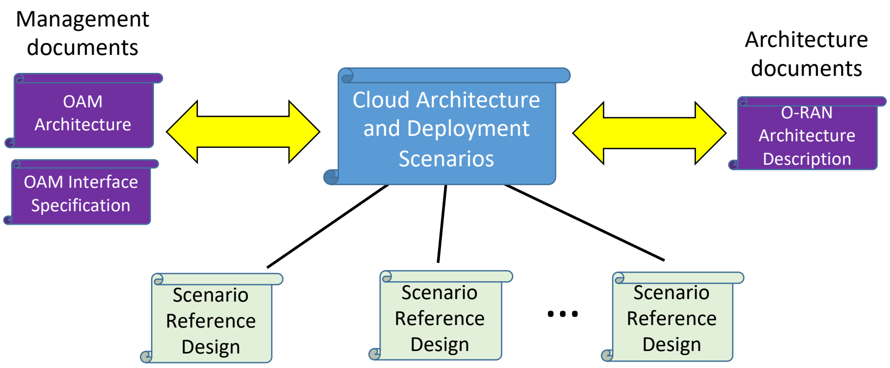


Figure 1: Relationship of this Document to Scenario Documents and O-RAN Management Documents

This document also draws on some other work from other O-RAN working groups, as well as sources from other industry bodies.

1.2 Objectives

The O-RAN Alliance seeks to improve RAN flexibility and deployment velocity, while at the same time reducing the capital and operating costs through the adoption of cloud architectures. The structure of the Orchestration and Cloudification work is shown graphically below. This document focuses on the Cloudification deployment aspects as indicated.

Editor's note: O-RU cloudification and O-RU AAL are future study items.

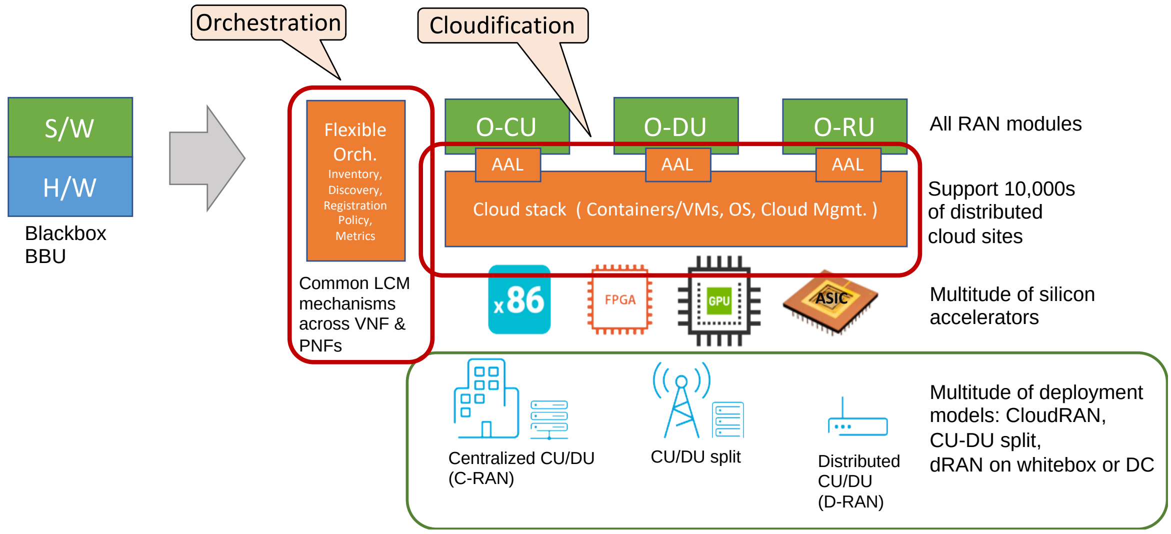


Figure 2: Major Components Related to the Orchestration and Cloudification Effort

A key principle is the decoupling of RAN hardware and software for all components including near-RT RIC, O-CU (O-CU-CP and O-CU-UP), O-DU, and O-RU, and the deployment of software components on commodity server architectures supplemented with programmable accelerators where necessary.

Key characteristics of cloud architectures which we will reference in this document are:

a) Decoupling of hardware from software. This aims to improve flexibility and choice for operators by decoupling selection and deployment of hardware infrastructure from software selection,

b) Standardization of hardware specifications across software implementations, to simplify physical deployment and maintenance. This aims to promote the availability of a multitude of *software* implementation choices for a given hardware configuration.

c) Sharing of hardware. This aims to promote the availability of a multitude of *hardware* implementation choices for a given software implementation.

d) Flexible instantiation and lifecycle management through orchestration automation. This aims to reduce deployment and ongoing maintenance costs by promoting simplification and automation throughout the hardware and software lifecycle through common chassis specifications and standardized orchestration interfaces.

This document will define various deployment scenarios that can be supported by the O-RAN specifications and are of either current or relatively near-term interest. Each scenario is identified by a specific grouping of functionality at different key locations (Cell Site, Edge Cloud, and Regional Cloud, which will be defined shortly), and an identification of whether functionality at a given location is provided by an O-RAN Physical NF based solution where software and hardware are tightly integrated and sharing a single identity, or by a cloud architecture that meets the above requirements.

The scope of this work clearly includes supporting all 5G technologies, i.e. E-UTRA and NR with both EPC-based Non-Standalone (NSA) and 5GC architectures. This implies that cloud/orchestration aspects of NSA (E-UTRA) are also supported. However, this version primarily addresses 5G SA deployments.

This technical report examines the constraints that drive a specific solution, and discuss the hierarchical properties of each solution, including a rough scale of the size of each cloud and a sense of the number of sub clouds expected to be served by a higher cloud. Figure 3 shows as example of how multiple cell sites feed into a smaller number of Edge Clouds, and how in turn multiple Edge Clouds feed into a Regional Cloud. For a given scenario, the Logical Functions are distributed in a certain way among each type of cloud, and the "cardinality" of the different functions will be discussed.

This has implications on the processing power needed in each type of cloud, as well as implications on the environmental requirements. This document will also discuss considerations of hardware chassis and components that are reasonable in each scenario, and the implications of managing such a cloud.

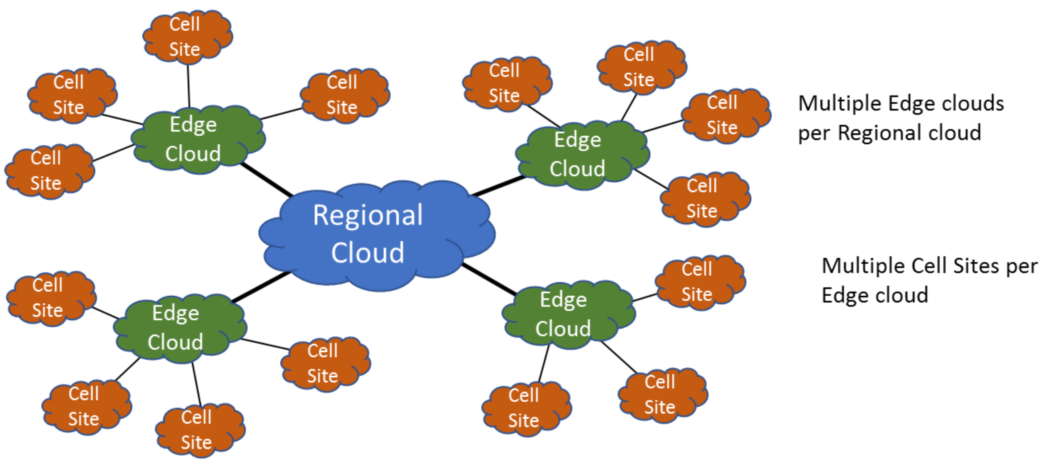


Figure 3: Different Clouds/ Sites

Additional major areas for this document are listed below:

- Mapping of logical functions to physical elements and locations, and implications of that mapping.
- High-level assessment of critical performance requirements, and how that influences architecture.
- Processor and accelerator options (e.g., x86, FPGA, GPU). In order to determine whether a Network Function is a candidate for openness, there needs to be the possibility to have multiple suppliers of software for given hardware, and multiple sources of required chip/accelerators.
    - The Hardware Abstraction Layer, aka “Acceleration Abstraction Layer” needs to be addressed in light of various hardware options that could be used.
- Cloud infrastructure makeup. This includes considerations such as:
    - Deployments are allowed to use VMs, Containers in VMs, or just Containers.
    - Multiple Operating Systems are expected to be supported; e.g., open source Ubuntu, CentOS Linux, or Yocto Linux-based distributions, or selected proprietary OSs.
- Management of a cloudified RAN introduces some new management considerations, because the mapping between Network Functionality and cloud platforms can be done in multiple ways, depending on the scenario that is chosen. Thus, management of aspects that are related to platform aspects rather than RAN functional aspects need to be designed with flexibility in mind from the start. For example, logging of physical functions, scale out actions, and survivability considerations are affected.
    - These management considerations are introduced in this document, but management documents will address the solutions.
- The transport layer will be discussed, but only to the extent that it affects the architecture and design of the network. For example, the chosen L1 technology may affect the performance of transport. As another example, the use of a Fronthaul Gateway will affect economics as well as the placement options of certain Network Functions. And of course, the existence of L2 switches in a cloud platform deployment will be required for efficient use of server resources.

Additional areas could be considered in the future.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of this report.

[1] 3GPP TS 38.470, *NG-RAN; F1 general aspects and principles (Release 15)*.

[2] 3GPP TR 21.905, *Vocabulary for 3GPP Specifications*.

[3] eCPRI Interface Specification V1.2, *Common Public Radio Interface: eCPRI Interface Specification*.

[4] eCPRI Transport Network V1.2, Requirements Specification, *Common Public Radio Interface: Requirements for the eCPRI Transport Network* .

[5] IEEE Std 802.1CM-2018, *Time-Sensitive Networking for Fronthaul*.

[6] ITU-T Technical Report, *GSTR-TN5G - Transport network support of IMT-2020/5G*.

[7] O-RAN WG4, *Control, User and Synchronization Plane Specification*, Technical Specification. See https://www.o-ran.org/specifications.

[8] O-RAN WG1, *Operations and Maintenance Architecture – v02.00*, Technical Specification. See https://www.o-ran.org/specifications.

[9] O-RAN WG1, *Operations and Maintenance Interface Specification – v1.0*, Technical Specification. See https://www.o-ran.org/specifications.

[10] O-RAN WG1, *O-RAN Architecture Description - v01.00*, Technical Specification. See https://www.o-ran.org/specifications.

[11] 3GPP TS 28.622, *Telecommunication management; Generic Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)*.

3 Definitions and Abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3GPP TR 21.905 [2] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [2].

Cell Site	This refers to the location of Radio Units (RUs); e.g., placed on same structure as the Radio Unit or at the base. The Cell Site in general will support multiple sectors and hence multiple O-RUs.
Edge Cloud	This is a location that supports virtualized RAN functions for multiple Cell Sites, and provides centralization of functions for those sites and associated economies of scale. An Edge Cloud might serve a large physical area or a relatively small one close to its cell sites, depending on the Operator's use case. However, the sites served by the Edge Cloud must be near enough to the O-RUs to meet the network latency requirements of the O-DU functions.
F1 Interface	The open interface between O-CU and O-DU in this document is the same as that defined by the CU and DU split in 3GPP TS 38.473. It consists of an F1-u part and an F1-c part.
Managed Element	The definition of a Managed Element (ME) is given in 3GPP TS 28.622 [11] section 4.3.3. The ME supports communication over management interface(s) to the manager for purposes of control and monitoring.
Managed Function	The definition of a Managed Function (MF) is given in 3GPP TS 28.622 [11] section 4.3.4. An MF instance is managed using the management interface(s) exposed by its containing ME instance.
Network Function	The near-RT RIC, O-CU-CP, O-CU-UP, O-DU, and O-RU *logical* functions that can be provided either by virtualized or non-virtualized methods.
Regional Cloud	This is a location that supports virtualized RAN functions for many Cell Sites in multiple Edge Clouds, and provides high centralization of functionality. The sites served by the Regional Cloud must be near enough to the O-DUs to meet the network latency requirements of the O-CU and near-RT RIC.
O-Cloud	This refers to a collection of O-Cloud Resource Pools at one or more location and the software to manage Nodes and Deployments hosted on them. An O-Cloud will include functionality to support both Deployment-plane and Management services. The O-Cloud

	provides a single logical reference point for all O-Cloud Resource Pools within the O-Cloud boundary.
O-RAN Physical NF	A RAN NF software deployed on tightly integrated hardware sharing a single Managed Element identity.
O-RAN Cloudified NF	A RAN NF software deployed on an O-Cloud with its own Managed Element identity, i.e., separate from the identity of the O-Cloud.

3.2 Abbreviations

For the purposes of this document, the abbreviations given in 3GPP TR 21.905 [2] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [2].

3GPP	Third Generation Partnership Project
5G	Fifth-Generation Mobile Communications
AAL	Acceleration Abstraction Layer
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuit
BBU	BaseBand Unit
BS	Base Station
CI	Cloud Infrastructure
CoMP	Co-Ordinated Multi-Point transmission/reception
CNF	Cloud-Native Network Function
CNI	Container Networking Interface
CPU	Central Processing Unit
CR	Cell Radius
CU	Centralized Unit as defined by 3GPP
DFT	Discrete Fourier Transform
DL	Downlink
DPDK	Data Plan Development Kit
DU	Distributed Unit as defined by 3GPP
eMBB	enhanced Mobile BroadBand
EPC	Evolved Packet Core
E-UTRA	Evolved UMTS Terrestrial Radio Access
FCAPS	Fault Configuration Accounting Performance Security
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FH	Fronthaul
FH GW	Fronthaul Gateway
FPGA	Field Programmable Gate Array
GNSS	Global Navigation Satellite System
GPP	General Purpose Processor
GPS	Global Positioning System
GPU	Graphics Processing Unit
HARQ	Hybrid Automatic Repeat reQuest
HW	Hardware
IEEE	Institute of Electrical and Electronics Engineers
IM	Information Modelling, or Information Model
IRQ	Interrupt ReQuest
ISA	Instruction Set Architecture
ISD	Inter-Site Distance
ITU	International Telecommunications Union
KPI	Key Performance Indicator
LCM	Life Cycle Management
LDPC	Low-Density Parity-Check
LLS	Lower Layer Split
LTE	Long Term Evolution
LVM	Logic Volume Manager
MEC	Mobile Edge Computing
mMTC	massive Machine Type Communications

MNO	Mobile Network Operator
NF	Network Function
NFD	Node Feature Discovery
NFVI	Network Function Virtualization Infrastructure
NIC	Network Interface Card
NMS	Network Management System
NR	New Radio
NSA	Non-Standalone
NTP	Network Time Protocol
NUMA	Non-Uniform Memory Access
NVMe	Non-Volatile Memory Express
O-Cloud	O-RAN Cloud Platform
OCP	Open Compute Project
O-CU	O-RAN Central Unit
O-CU-CP	O-CU Control Plane
O-CU-UP	O-CU User Plane
O-DU	O-RAN Distributed Unit (uses Lower-level Split)
O-RU	O-RAN Radio Unit
OTII	Open Telecom IT Infrastructure
OWD	One-Way Delay
PCI	Peripheral Component Interconnect
PNF	Physical Network Function
PoE	Power over Ethernet
PoP	Point of Presence
PRTC	Primary Reference Time Clock
PTP	Precision Time Protocol
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RIC	RAN Intelligent Controller
RT	Real Time
RTT	Round Trip Time
RU	Radio Unit
SA	Standalone
SFC	Service Function Chaining
SMO	Service Management and Orchestration
SMP	Symmetric MultiProcessing
SoC	System on Chip
SR-IOV	Single Root Input/ Output Virtualization
SW	Software
TCO	Total Cost of Ownership
TNE	Transport Network Element
TR	Technical Report
TRP	Transmission Reception Point
TS	Technical Specification
TSC (T-TSC)	Telecom Slave Clock
Tx	Transmitter
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communications
vCPU	virtual CPU
VIM	Virtualized Infrastructure Manager
VM	Virtual Machine
VNF	Virtualized Network Function
vO-CU	Virtualized O-RAN Central Unit
vO-CU-CP	Virtualized O-CU Control Plane
vO-CU-UP	Virtualized O-CU User Plane
vO-DU	Virtualized O-RAN Distributed Unit

4 Overall Architecture

This section addresses the overall architecture in terms of the Network Functions and infrastructure (O-RAN Physical NFs, servers, and clouds) that are in scope. Figure 4 provides a high-level view of the O-RAN architecture as depicted in [10].

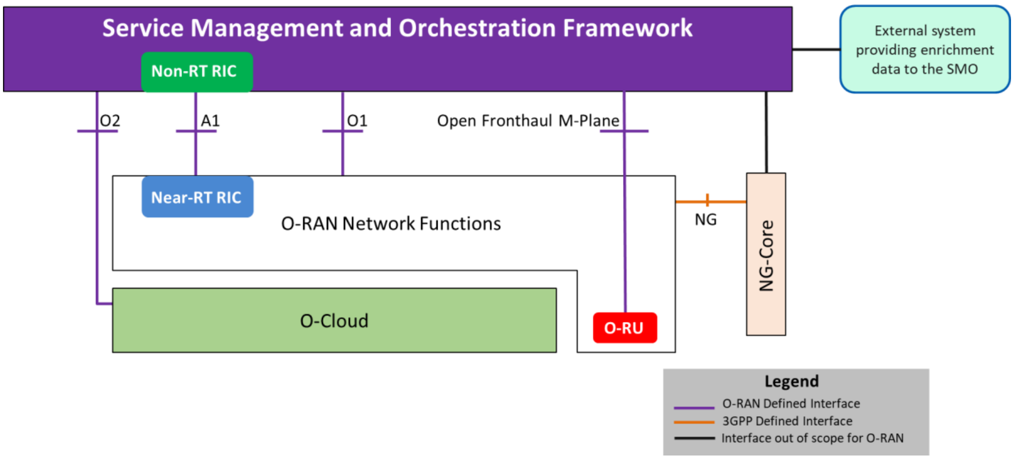


Figure 4: High Level Architecture of O-RAN

4.1 O-RAN Functions Definitions

This section reviews key O-RAN functions definitions in O-RAN.

- The O-DU/ O-RU split is defined as using Option 7-2x. See [7].
- The O-CU/ O-DU split is defined as using the CU/ DU split F1 as defined in 3GPP TS 38.470 [1].

This document assumes these two splits.

Figure 5 shows the logical architecture of O-RAN (as depicted in [10]) with O-Cloud platform at the bottom, where any given O-RAN function could be supported by O-Cloud, depending on the deployment scenario. For example, the figure here illustrates a case where the O-RU is implemented as an O-RAN Physical NF, and the other functions within the dashed line are supported by O-Cloud.

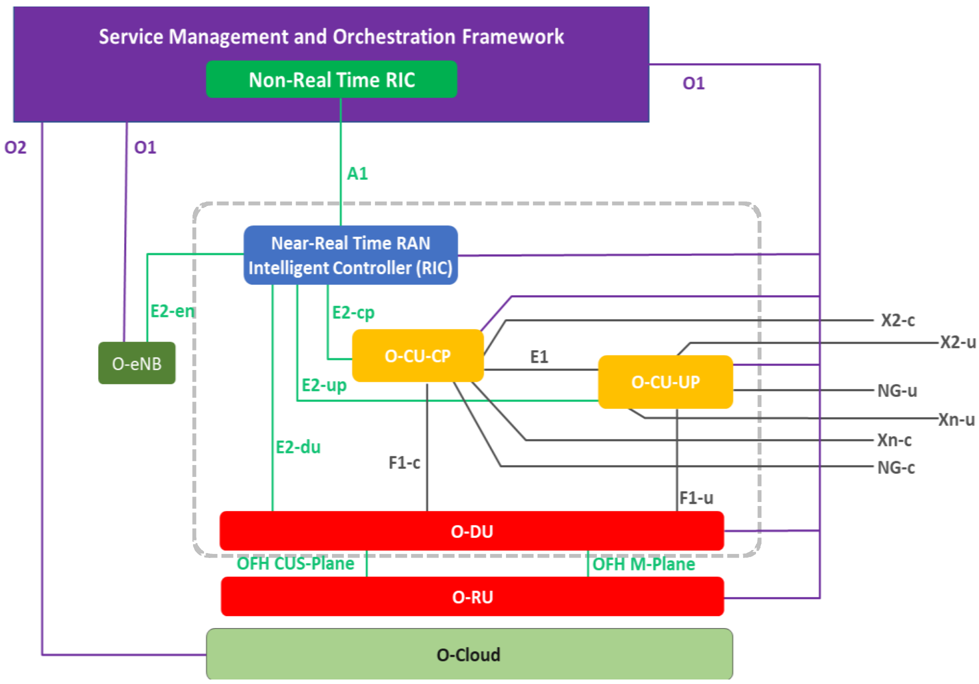


Figure 5: Logical Architecture of O-RAN

4.2 Degree of Openness

In theory, every architecture component could be open in every sense imaginable, but in practice it is likely that different components will have varying degrees of openness due to economic and other implementation considerations. Some factors are significantly affected by the deployment scenario; for example, what might be viable in an indoor deployment might not be viable in an outdoor deployment.

Increasing degrees of openness for an O-RAN Physical Network Function or O-RAN Cloudified Network Function(s) are:

A. Interfaces among Network Functions are open; e.g., E2, F1, and Open Fronthaul are used. Therefore, Network Functions in different O-RAN Physical NFs/clouds from different vendors can interconnect.

B. In addition to having open connections as described above, the chassis of servers in a cloud are open and can accept blades/sleds from multiple vendors. However, the blades/sleds have RAN software that *is not* decoupled from the hardware.

C. In addition to having open connections and an open chassis, a specific blade/sled uses software that *is* decoupled from the hardware. In this scenario, the software could be from one supplier, the blade/sled could be from another, and the chassis could be from another.

Categories A and B have O-RAN Physical NFs/clouds, while Category C is an open solution that we are calling an O-Cloud, and is subject to the cloudification discussion and requirements.

In this document, the degree of openness for each O-RAN Physical NF/cloud can vary by scenario. The question of which Network Functions should be split vs. combined, and the degree of openness in each one, is addressed in the discussion of scenarios.

4.3 Decoupling of Hardware and Software

Editor's note: O-RU AAL is a future study item.

There are three layers that we must consider when we discuss decoupling of hardware and software:

- The hardware layer, shown at the bottom in Figure 6. (In the case of a VM deployment, this maps basically to the ETSI NFVI hardware sub-layer.)
- A middle layer that includes Cloud Stack functions as well as Acceleration Abstraction Layer functions. (In the case of a VM deployment, these map to the ETSI NFVI virtualization sub-layer + VIM.)
- A top layer that supports the virtual RAN functions.

Each layer can come from a different supplier. The first aspect of decoupling has to do with ensuring that a Cloud Stack can work on multiple suppliers' hardware; i.e., it does not require vendor-specific hardware.

The second aspect of decoupling has to do with ensuring that a Cloud Platform can support RAN virtualized functions from multiple RAN software suppliers. If this is possible, then we say that the Cloud Platform (which includes the hardware that it runs on) is an O-RAN Cloud Platform, or "O-Cloud". See Figure 6 below.

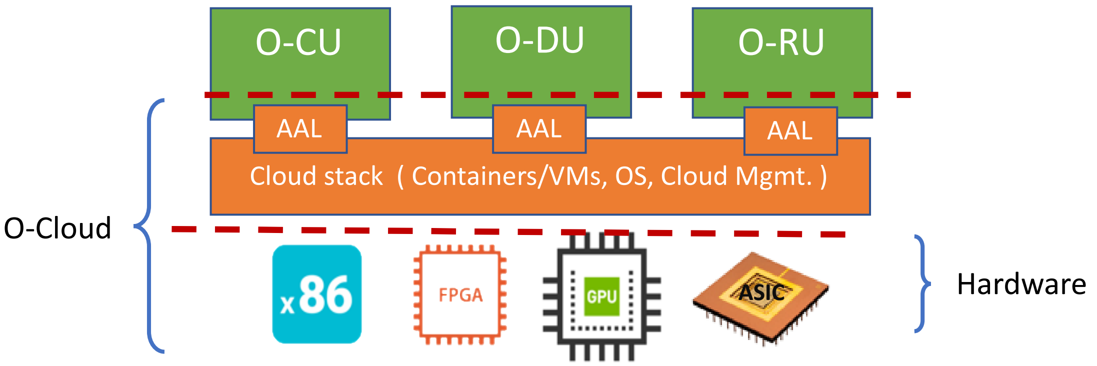


Figure 6: Decoupling, and Illustration of the O-Cloud Concept

4.3.1 The O-Cloud

The general definition of the O-Cloud Cloud Platform includes the following characteristics:

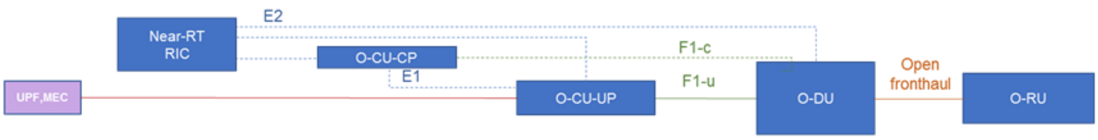
1. The Cloud Platform is a set of hardware and software components that provide cloud computing capabilities to execute RAN network functions.
2. The Cloud Platform hardware includes compute, networking and storage components, and may also include various acceleration technologies required by the RAN network functions to meet their performance objectives.
3. The Cloud Platform software exposes open and well-defined APIs that enable the management of the entire life cycle for network functions.
4. The Cloud Platform software is decoupled from the Cloud Platform hardware (i.e., it can typically be sourced from different vendors).

The management aspects of the O-Cloud platform are discussed in 5.6. The scope of this document includes listing specific requirements of the Cloud Platform to support execution of the various O-RAN Network Functions.

An example of a Cloud Platform is an OpenStack and/or a Kubernetes deployment on a set of COTS servers (including FPGA and GPU cards), interconnected by a spine/leaf networking fabric.

There is an important interplay between specific virtualized RAN functions and the hardware that is needed to meet performance requirements and to support the functionality *economically*. Therefore, a hardware/ cloud platform combination that can support, say, a vO-CU function might not be appropriate to adequately support a vO-DU function. When RAN functions are combined in different ways in each specific deployment scenario, these aspects must be considered.

Below is a high-level conceptual example of how different accelerators, along with their associated cloud capabilities, can be required for different RAN functions. Although we do not specify any particular hardware requirement or cloud capability here, we can note some general themes. For example, any RAN function that involves real-time movement of user traffic will require the cloud platform to control for delay and jitter, which may in turn require features such as real-time OSs, avoidance of frequent interrupts, CPU pinning, etc.



Cloud/ HW features	Near-RT RIC	O-CU-CP	O-CU-UP	O-DU	O-RU
Standard Cloud Infrastructure (CI) & General Purpose CPU	✓	✓			
CI + high speed UP support. Acceleration optional			✓		
CI + high speed UP, acceleration for O-DU				✓	
CI + high speed UP, acceleration for O-RU					✓

Figure 7: Relationship between RAN Functions and Demands on Cloud Infrastructure and Hardware

Please note that any cloud that has features required for a given function (e.g., for O-DU) can also support functions that do not require such features. For example, a cloud that can support O-DU can also support functions such as O-CU-CP.

4.3.2 Key O-Cloud Concepts

Figure 8 illustrates key components of an O-Cloud and its management.

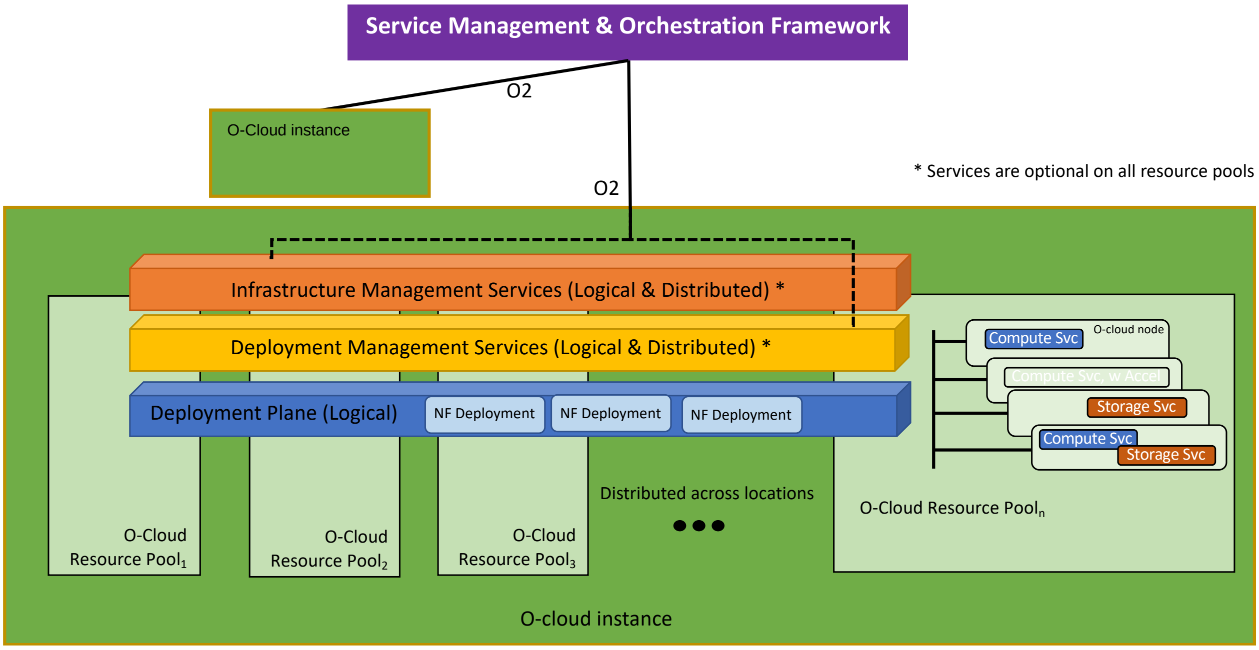


Figure 8: Key Components Involved in/with an O-Cloud

Key terms in this figure are defined below:

- An **O-Cloud** instance refers to a collection of O-Cloud Resource Pools at one or more location and the software to manage Nodes and Deployments hosted on them. An O-Cloud will include functionality to support both Deployment-plane (aka. user-plane) and Management services. The O-Cloud provides a single logical reference point for all O-Cloud Resource Pools within the O-Cloud boundary.
- An **O-Cloud Resource Pool** is a collection of O-Clouds nodes with homogeneous profiles in one location which can be used for either Management services or Deployment Plane functions. The allocation of NF deployment to a resource pool is determined by the SMO.
- An **O-Cloud Node** is a collection of CPUs, Mem, Storage, NICs, Accelerators, BIOSes, BMCs, etc., and can be thought of as a server. Each O-Cloud Node will support one or more "roles", see next.
- **O-Cloud Node Role** refers to the functionalities that a given node may support. These include Compute, Storage, Networking for the Deployment-plane (i.e., user-plane related functions such as the O-RAN NF), they may include optional acceleration functions, and they may also include the appropriate Management services.
- **O-Cloud Deployment Plane** is a logical construct representing the O-Cloud Nodes across the Resource Pools which are used to create NF Deployments.
- An **O-Cloud NF Deployment** is a deployment of a cloud native Network Function (all or partial), resources shared within a NF Function, or resource shared across network functions. The NF Deployment configures and assembles user-plane resources required for the cloud native construct used to establish the NF Deployment and manage its life cycle from creation to destruction.
- The **O2 Interface** is a collection of services and their associated interfaces that are provided by the O-Cloud platform to the SMO. The services are categorized into two logical groups: (i) **Infrastructure Management Services**, which include the subset of O2 functions that are responsible for deploying and managing cloud infrastructure. (ii) **Deployment Management Services**, which include the subset of O2 functions that are responsible for managing the lifecycle of virtualized/containerized deployments on the cloud infrastructure. The O2 services and their associated interfaces shall be specified in the upcoming O2 specification. Any definitions of SMO functional elements needed to consume these services shall be described in OAM architecture.

Figure 9 illustrates several deployment examples to show the different O-Cloud Node Roles. Note that the O-Cloud Node Roles and the O-Cloud Node names are mentioned here as examples and are neither exhaustive nor standardized.

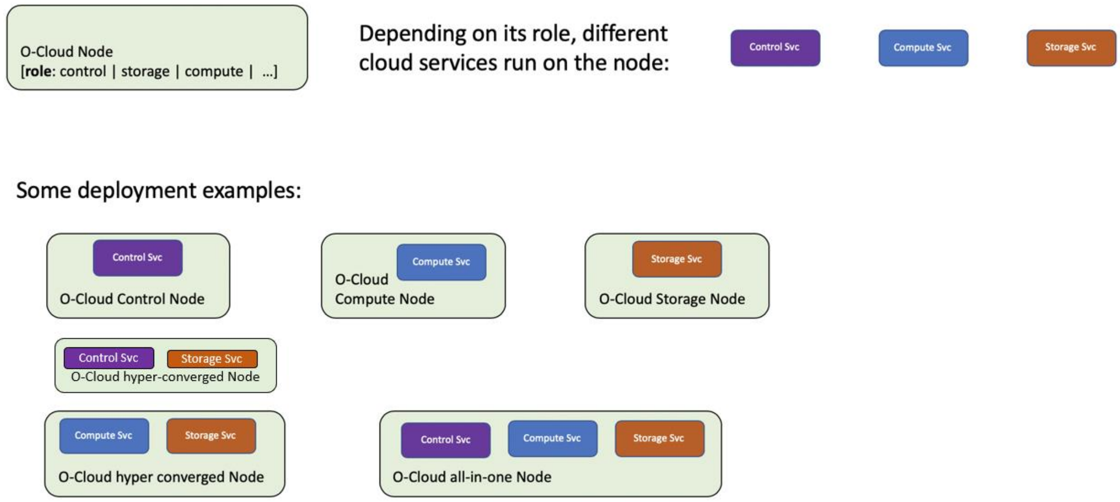


Figure 9: O-Cloud Node Roles and Deployment Examples

5 Deployment Scenarios: Common Considerations

In any implementation of logical network functionality, decisions need to be made regarding which logical functions are mapped to which Cloud Platforms, and therefore which functions are to be co-located with other logical functions. In this document we do not prescribe one specific implementation, but we do understand that in order to establish agreements and requirements, the manner in which the Network Functions are mapped to the same or different Cloud Platforms must be considered.

We refer to each specific mapping as a "deployment scenario". In this section, we examine the deployment scenarios that are receiving the most consideration. Then we will select the one or ones that should be the focus of initial scenario reference design efforts.

5.1 Mapping Logical Functionality to Physical Implementations

There are many aspects that need to be considered when deciding to implement logical functions in distinct O-Clouds. Some aspects have to do with fundamental technical constraints and economic considerations, while others have to do with the nature of the services that are being offered.

5.1.1 Technical Constraints that Affect Hardware Implementations

Below are some factors that will affect the cost of implementations, and can drive a carrier to require separation of or combining of different logical functions.

- **Environment:** Equipment may be deployed in indoor controlled environments (e.g., Central Offices), semi-controlled environments (e.g., cabinets with fans and heaters), and exposed environments (e.g., Radio Units on a tower). In general, the less controlled the environment, the more difficult and expensive the equipment will be. The required temperature range is a key design factor, and can drive higher power requirements.
- **Dimensions:** The physical dimensions can also drive deployment constraints – e.g., the need to fit into a tight cabinet, or to be placed safely on a tower or pole.
- **Transport technology:** The transport technology used for Fronthaul, Midhaul, and Backhaul is often fiber, which has an extremely low and acceptable loss rate. However, there are options other than fiber, in particular wireless/ microwave, where the potential for data loss must be considered. This will be discussed further in the next section.
- **Acceleration Hardware:** The need for acceleration hardware can be driven by the need to meet basic performance requirements, but can also be tied to some of the above considerations. For example, a hardware acceleration chip (COTS or proprietary) can result in lower power use, less generated heat, and smaller physical dimensions than if acceleration is not used. On the other hand, some types of hardware acceleration chips might not be "hardened" (i.e., they might only operate properly in a restricted environment), and could require a more controlled environment such as in a central office.

The acceleration hardware most often referred to includes:

- Field Programmable Gate Arrays (FPGAs)
- Graphical Processing Units (GPUs)
- System on Chip (SoC)

- **Standardized Hardware:** Use of standardized hardware designs and standardized form factors can have advantages such as helping to reduce operations complexity, e.g., when an operator makes periodic technology upgrades of selected components. An example would be to use an Open Compute Project (OCP) or Open Telecom IT Infrastructure (OTII) –based design.

5.1.2 Service Requirements that Affect Implementation Design

RANs can serve a wide range of services and customer requirements, and each market can drive some unique requirements. Some examples are below.

- **Indoor or outdoor deployment:** Indoor deployments (e.g., in a public venue like a sports stadium, train station, shopping mall, etc.) often enjoy a controlled environment for all elements, including the Radio Units. This can improve the economics of some indoor deployment scenarios. The distance between Network Functions tends to be much lower, and the devices that support O-RU functionality may be much easier and cheaper to install and maintain. This can affect the density of certain deployments, and the frequency that certain scenarios are deployed.
- **Bands supported, and Macro cell vs. Small cell:** The choice of bands (e.g., Sub-6 GHz vs. mmWave) might be driven by whether the target customers are mobile vs. fixed, and whether a clear line of sight to the customer is available or is needed. The bands to be supported will of course affect O-RU design. In addition, because mmWave carriers can support much higher channel width (e.g., 400 MHz vs. 20 MHz), mmWave deployments can require a great deal more O-DU and O-CU processing power. And of course the operations costs of deploying Macro cells vs. Small cells differ in other ways.
- **Performance requirements of the Application / Network Slice:** Ultimately, user applications drive performance requirements, and RANs are expected to support a very wide range of applications. For example, the delay requirements to support a Connected Car application using Ultra Reliable Low Latency Communications (URLLC) will be more demanding than the delay requirements for other types of applications. In our discussion of 5G, we can start by considering requirements separately for URLLC, enhanced Mobile Broadband (eMBB), and massive Machine Type Communications (mMTC).

The consideration of performance requirements is a primary one, and is the subject of Section 5.2.

5.1.3 Rationalization of Centralizing O-DU Functionality

Almost all Scenarios to be discussed in this document involve a degree of centralization of O-DU. In this section it is assumed that O-DU resources for a set of O-RUs are centralized at the same location.

Editor's Note: While most Scenarios also centralize O-CU-CP, O-CU-UP, and near-RT RIC in one form or another, the benefits of centralizing them are not discussed in this section.

Managing O-DU in equipment at individual cell sites (via on-site BBUs today) has multiple challenges, including:

- If changes are needed at a site (e.g., adding radio carriers), then adding equipment is a coarse-grained activity – i.e., one cannot generally just add "another 1/5 of a box", if that is all that is needed. Adding the minimum increment of additional capacity might result in poor utilization and thereby prevent expansion at that site.
- Cell sites are in many separate locations, and each requires establishment and maintenance of an acceptable environment for the equipment. In turn this requires separate visits for any physical operations.
- Micro sites tend to have much lower average utilization than macro sites, but each can experience considerable peaks.
- "Planned obsolescence" occurs, due to ongoing evolution of smartphone capabilities and throughput improvements, as well as introduction of new features and services. It is common practice today to upgrade ("forklift replace") BBUs every 36-60 months.

These factors motivate the centralization of resources where possible. For the O-DU function, we can think of two types of centralization: *simple* centralization and *pooled* centralization.

If the equipment uses O-DU centralization in an Edge Cloud, at any given hour an O-RU will be using a single specific O-DU resource that is assigned to it (e.g. via Kubernetes). On a broad time scale, traffic from any cell site can be rehomed, without any physical work, to use other/additional resources that are available at that Edge Cloud location. This would likely be done infrequently; e.g., about as often as cell sites are expanded.

Centralization can have some additional benefits, such as only having to maintain a single large controlled environment for many cell sites rather than creating and maintaining many distributed locations that might be less controlled (e.g., outside cabinets or huts). Capacity can be added at the central site and assigned to cell sites as needed. Note that *simple* centralization still assigns each O-RU to a single O-DU resource[1], as shown below, and that traffic from one O-RU is not split into subsets that could be assigned to different O-DUs. Also note that a Fronthaul (FH) Gateway (GW) may exist between the cell site and the centralized resources, not only to improve economics but also to enable traffic re-routing when desired.

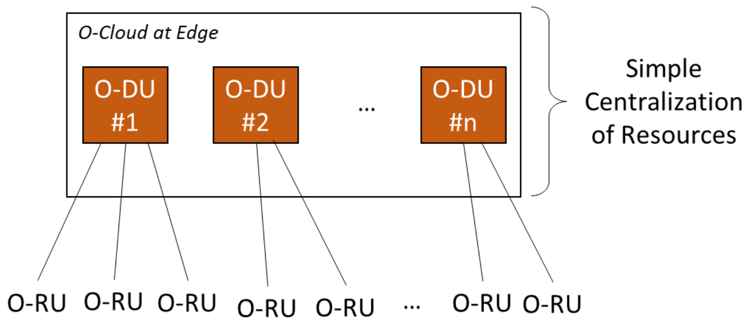


Figure 10: Simple Centralization of O-DU Resources

By comparison, with *pooled* centralization, traffic from an O-RU (or subsets of the O-RU's traffic) can be assigned more dynamically to any of several shared O-DU resources. So if one cell site is mostly idle and another experiences high traffic demand, the traffic can be routed to the appropriate O-DU resources in the shared pool. The total resources of this shared pool can be smaller than resources of distributed locations, because the peak of the sum of the traffic will be markedly lower than the sum of the individual cell site traffic peaks.

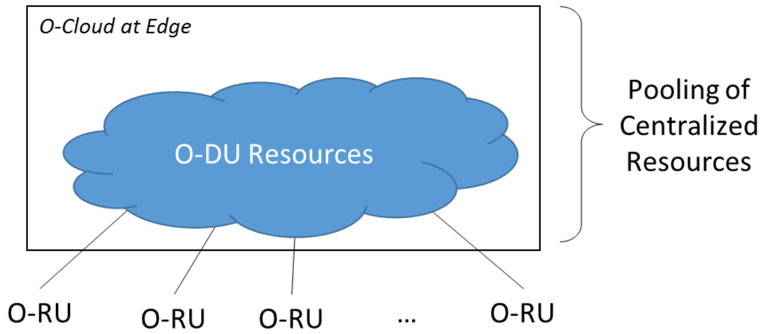


Figure 11: Pooling of Centralized O-DU Resources

We note that being able to share O-DU resources somewhat dynamically is expected to be a solvable problem, although we understand that it is by no means a trivial problem. There are management considerations, among others. There may be incremental steps toward true shared pooling, where rehoming of O-RUs to different O-DUs can be performed more dynamically, based on traffic conditions.

It is noted that O-DU centralization benefits the most dense networks where several cell sites are within the O-RU to O-DU latency limits. Sparsely populated areas most probably will be addressed by vO-CU centralization only.

Figure 12 shows the results of an analysis of a simulated greenfield deployment as an attempt to visualize the relative merit of simple centralization of O-DU ("oDU") vs. pooled centralization of O-DU ("poDU") vs. legacy DU ("BBU"), plotted against the realizable Cell Site pool size.

[1] In this figure, each O-DU block can be thought of as a unit of server resources that includes a hardware accelerator, a GPP, memory and any other associated hardware.

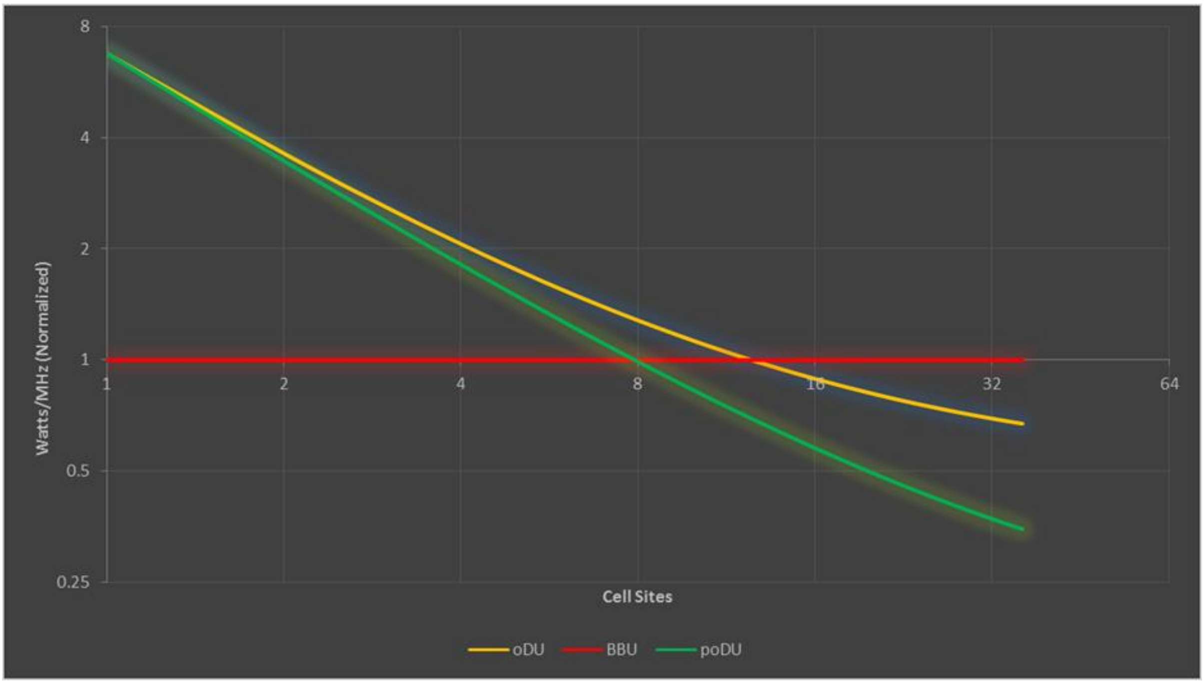


Figure 12: Comparison of Merit of Centralization Options vs. Number of Cell Sites in a Pool

An often-used measure is related to the power required to support a given number of carrier MHz. The lower the power used per carrier, the more efficient is the implementation. In Figure 12, the values of each curve are normalized to the metric of Watts/MHz for distributed legacy BBUs, normalized to equal 1. Please note that in this diagram, a lower value is better. The following assumptions apply to the figure:

- A legacy BBU processes X MHz (for carriers) and consumes Y watts. For example, a specific BBU might process 1600 MHz and consume 160 watts.
- N legacy BBUs will process N x X MHz and consume N x Y watts and have a merit figure of 1, per normalization. If a given site requires less than X MHz, it will still be necessary to deploy an X MHz BBU. For example, we may need only 480 MHz but still deploy a 1600 MHz BBU.
- Simple Centralization (the "oDU" line): In this case, active TRPs are statically mapped to specific VMs and vO-DU tiles[2]. Fewer vO-DU tiles are required to support the same number of TRPs, because MHz per site is not a constant.
    - Independent of resources to support active user traffic, a fixed power level is required to power Ethernet "frontplane" switches and hardware to support management and orchestration processes.
    - In a pool, processing capacity will be added over time as required.
    - Due to mobility traffic behavior, tiles will not be fully utilized, although centralization of resources will improve utilization when compared with a legacy BBU approach.
- Centralization with more dynamic pooling (the "poDU" line): In addition to active load balancing, individual traffic flows (which can last from a few hundreds of msecs to several seconds) will be routed to the least used tile, further optimizing (reducing) vO-DU tile requirements.
    - As in the simple centralization approach above, there is a fixed power level required for hardware that supports switching, management and orchestration processes.

As a final note, any form of centralization requires efficient transport between the O-RU and the O-DU resources. When O-RU functionality is distributed over a relatively large area (e.g., not concentrated in a single large building), the existence of a Fronthaul Gateway is a key enabler.

[2] A "vO-DU tile" refers to a chip or System on Chip (SoC) that provides hardware acceleration for math-intensive functionality such as that required for Digital Signal Processing. With the Option 7.2x split, acceleration of Forward Error Correction (FEC) functionality is required (FEC is optional for e.g. low band.), and other functionality could be considered for acceleration if desired.

5.2 Performance Aspects

Performance requirements drive architectural and design considerations. Performance can include attributes such as delay, packet loss, transmission loss, and delay variation (aka "jitter").

Editor's Note: While all aspects are of interest, delay has the largest impact on network design and will be the focus of this version. Future versions can address other performance aspects if desired and is FFS.

5.2.1 User Plane Delay

This section discusses the framework for discussing delay of user-plane packets[3], and also general delay numbers that it can be agreed that apply across all scenarios. Details relevant to a specific Scenario will be discussed in each Scenario's subsection, as applicable. The purpose of these high-level targets is to act as a baseline for allocating the total latency budget to subsystems that are on the path of each constraint, as required for system engineering and dimensioning calculations, and to assess the impact on the function placement within the specific network site tiers.

The goal is to establish reasonable maximum delay targets, as well as to identify and document the major infrastructure as well as O-RAN NF-specific delay contributing components. For each service or element, minimum delay should be considered to be zero. The implication of this is that any of the elements can be moved towards the Cell Site (e.g. in a fully distributed Cloud RAN configuration, all of O-CU-UP, O-DU and O-RU would be distributed to Cell Site).

In real network deployments, the expectation is that, depending on the operator-specific implementation constraints such as location and fiber availability, deployment area density, etc., deployments result in anything between the fully distributed and maximally centralized configuration. Even on one operator's network, it is common that there are many different sizes of Edge Cloud instances, and combinations of Centralized and Distributed architectures in same network are also common (e.g. network operator may choose to centralize the deployments on dense Metro areas to the extent possible and distribute the configurations on suburban/rural areas with larger cell sizes / cell density that do not translate to pooling benefits from more centralized architecture). However, the maximum centralization within the constraints of latencies that can be tolerable is useful for establishing the basis for dimensioning of the maximum sizes, especially for the Edge and Regional cloud PoPs. Figure 13 below illustrates the relationship among some key delay parameters.

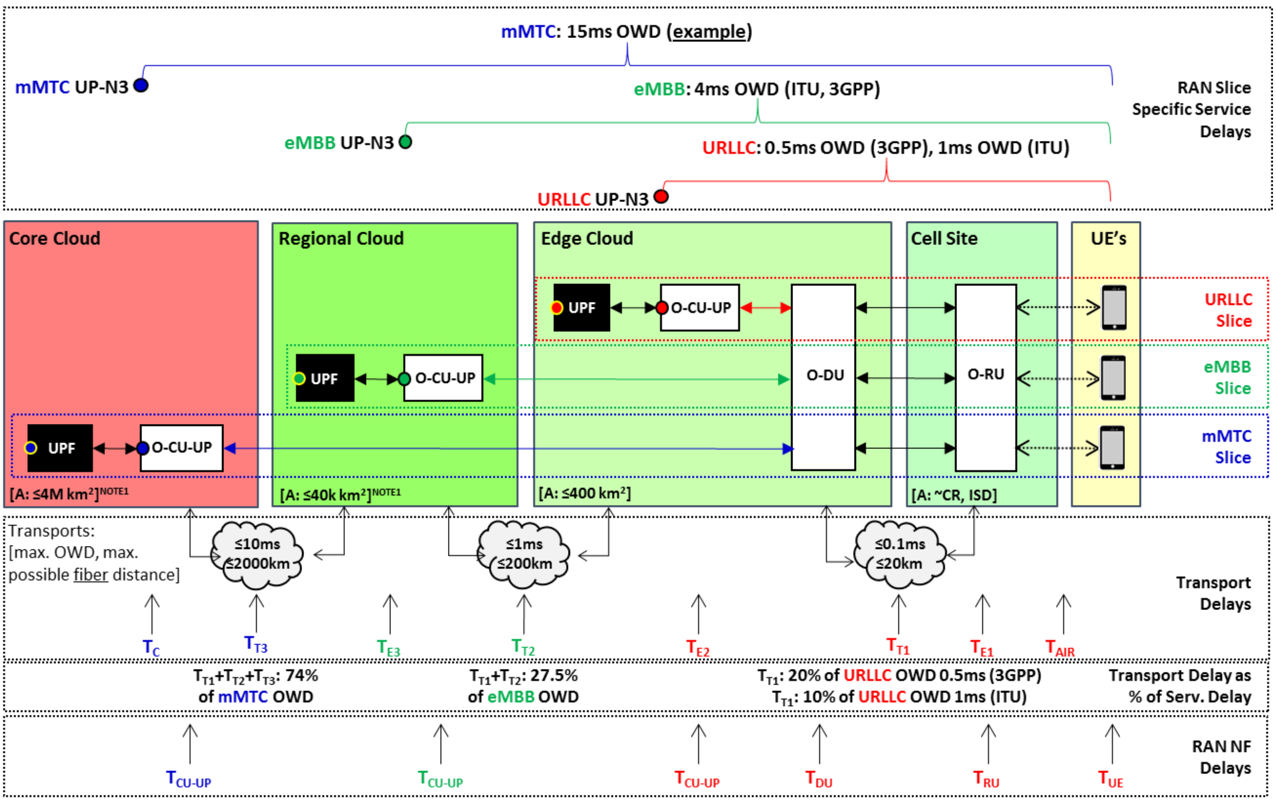


Figure 13: Major User Plane Latency Components, by 5G Service Slice and Function Placement

[3] Delay of control plane or OAM traffic is not considered in this section.

Please note the following:

- NOTE 1: If the T2 or/and T3 transport network(s) is/are Packet Transport Network(s), then time allocation for the transport network elements processing and queuing delays will require some portion of maximum latency allocation, and will require reduction of the maximum area accordingly.
- NOTE 2: Site Internal / fabric networks are not shown for clarity, but need some latency allocation (effectively extensions or part of transport delays; per PoP tier designations T_{E1}, T_{E2}, T_{E3} and T_C).
- NOTE 3: To maximize the potential for resource pooling benefits, minimize network function redundancy cost, and minimize the amount of hardware / power in progressively more distributed sites (towards UEs), target design should attempt to maximize the distances and therefore latencies available for transport networks within the service- and RAN-specific time constraints, especially for T_{T1}.
- NOTE 4: UPF, like EC/MEC, is outside of the scope of O-RAN, so UPF shown as a “black box” to illustrate where it needs to be placed in context of specific services to be able to take advantage of the RAN service-specific latency improvements.

Figure 13 represents User Equipment locations on the right, and network tiers towards the left, with increasing latency and increasing maximum area covered per tier towards the left. These Mobile Network Operator’s (MNO’s) Edge tiers are nominated as Cell Site, Edge Cloud, and Regional Cloud, with one additional tier nominated as Core Cloud in the figure.

The summary of the associated latency constraints as well as major latency contributing components as depicted in Figure 13 above is given in Table 1, below.

Table 1: Service Delay Constraints and Major Delay Contributors

RAN Service-Specific User Plane Delay Constraints			
Identifier	**Brief Description**	**Max. OWD (ms)**	**Max. RTT (ms)**
URLLC	Ultra-Reliable Low Latency Communications (3GPP)	0.5	1
URLLC	Ultra-Reliable Low Latency Communications (ITU)	1	2
eMBB	enhanced Mobile Broadband	4	8
mMTC	massive Machine Type Communications	15	30
Transport Specific Delay Components			
T_{AIR}	Transport propagation delay over air interface		
T_{E1}	Cell Site Switch/Router delay		
T_{T1}	Transport delay between Cell Site and Edge Cloud	0.1	0.2
T_{E2}	Edge Cloud Site Fabric delay		
T_{T2}	Transport delay between Edge and Regional Cloud	1	2
T_{E3}	Regional Cloud Site Fabric delay		
T_{T3}	Transport delay between Regional and Core Cloud	10	20
T_C	Core Cloud Site Fabric delay		
Network Function Specific Delay Components			
T_{UE}	Delay Through the UE SW and HW stack		
T_{RU}	Delay Through the O-RU User Plane		
T_{DU}	Delay Through the O-DU User Plane		
T_{CU-UP}	Delay Through the O-CU User Plane		

The transport network delays are specified as maximums, and link speeds are considered to be symmetric for all components with exception of the air interface (T_{AIR}). For the S-Plane services utilizing PTP protocol, it is a requirement that the link lengths, link speeds and forward-reverse path routing for PTP are all symmetric.

Radios (O-RUs) are always located in the Cell Site tier, while O-DU can be located “up to” Edge Cloud tier. It is possible to move any of the user plane NF instances closer towards the cell site, as implicitly they would be inside the target maximum delay, but it is not necessarily possible to move them further away from the Cell Sites while remaining within the RAN internal and/or RAN service-specific timing constraints. A common expected deployment case is one where O-DU instances are moved towards or even to the Cell Site and O-RUs (e.g. in Distributed Cloud-RAN configurations), or in situations where the Edge Cloud needs to be located closer to the Cell Site due to fiber and/or location availability, or other constraints. While this is expected to work well from the delay constraints perspective, the centralization and pooling-related benefits will be potentially reduced or even eliminated in the context of such deployment scenarios.

The maximum transport network latency between the site hosting O-DU(s) and sites hosting associated O-RU(s) is primarily determined by the RAN internal processes time constraints (such as HARQ loop, scheduling, etc., time-sensitive operations). For the purposes of this document, we use 100us latency, which is commonly used as a target maximum latency for this transport segment in related industry specifications for user-plane, specifically “High100” on E-CPRI transport requirements [4] section 4.1.1, as well as “Fronthaul” latency requirement in ITU technical report GSTR-TN5G [6], section 7-2, and IEEE Std 802.1CM-2018 [5], section 6.3.3.1. Based on the 5us/km fiber propagation delay, this implies that in a 2D Manhattan tessellation model, which is a common simple topology model for dense urban area fiber routing, the maximum area that can be covered from a single Edge Cloud tier site hosting O-DUs is up to a 400km^2 area of Cell Sites and associated RUs. Based on the radio inter-site distances, number of bands and other radio network dimensioning specific parameters, this can be used to estimate the maximum number of Cell Sites and cell sectors that can be covered from single Edge Cloud tier location, as well as maximum number of UEs in this coverage area.

The maximum transport network latencies towards the entities located at higher tiers are constrained by the lower of F1 interface latency (max 10 ms as per GSTR-TN5G [6], section 7.2), or alternatively service-specific latency constraints, for the edge-located services that are positioned to take advantage of improved latencies. For eMBB, UE-CU latency target is 4ms one-way delay, while for the URLLC it is 0.5ms as per 3GPP (or 1ms as per ITU requirements). The placement of the O-CU-UP as well as associated UPF, to be able to provide URLLC services would have to be at most at the Edge Cloud tier to satisfy the service latency constraint. For the eMBB services with 4ms OWD target, it is possible to locate O-CU-UP and UPF on next higher latency location tier, i.e. Regional Cloud tier. Note that while not shown in the picture, Edge compute / Multi-Access Edge Compute (MEC) services for a given RAN service type are expected to be collocated with the associated UPF function to take advantage of the associated service latency reduction potential.

For the services that do not have specific low-latency targets, the associated O-CU-UP and UPF can be located on higher tier, similar to deployments in typical LTE network designs. This is designated as Core Cloud tier in the example in Figure 13 above. For eMBB services, if there are no local service instances in the Edge or Regional clouds to take advantage of the 4ms OWD enabled by eMBB service definition, but the associated services are provided from either core clouds, external networks or from other Edge Cloud / RAN instances (in case of user-to-user traffic), the associated non-constrained (i.e. over 4ms from subscriber) eMBB O-CU-UP and UPF instances can be located in Core Cloud sites without perceivable impact to the service user, as in such cases the transport and/or service-specific latencies are dominant latency components.

The intent of this section is not to micromanage the latency budget, but to rather establish a reasonable baseline for dimensioning purposes, particularly to provide basic assessment to enable sizing of the cloud tiers within the context of the service-specific constraints and transport allocations. As such, we get the following “allowances” for the aggregate unspecified elements:

- URLLC_{3GPP}: 0.5ms - 0.1ms (T_{T1}) = 0.4ms $\geq T_{UE} + T_{AIR} + T_{E1} + T_{RU} + 2(T_{E2}) + T_{DU} + T_{CU\text{-}UP}$
- URLLC_{ITU}: 1ms - 0.1ms (T_{T1}) = 0.9ms $\geq T_{UE} + T_{AIR} + T_{E1} + T_{RU} + 2(T_{E2}) + T_{DU} + T_{CU\text{-}UP}$
- eMBB: 4ms - 0.1ms (T_{T1}) - 1ms (T_{T2}) = 2.9ms $\geq T_{UE} + T_{AIR} + T_{E1} + T_{RU} + 2(T_{E2}) + T_{DU} + T_{E3} + T_{CU\text{-}UP}$
- mMTC_{15}: 15ms - 0.1ms (T_{T1}) - 1ms (T_{T2}) - 10ms (T_{T3}) = 3.9ms $\geq T_{UE} + T_{AIR} + T_{E1} + T_{RU} + 2(T_{E2}) + T_{DU} + T_{E3} + T_{CU\text{-}UP} + T_C$

If required, we may provide more specific allocations in later versions of the document, as we gain more implementation experience and associated test data, but at this stage it is considered to be premature to do so. It should also be noted that the URLLC specification is still work in progress at this stage in 3GPP, so likely first implementations will focus on eMBB service, which leaves 2.9ms for combined O-RAN NFs, air interface, UE and cloud fabric latencies.

It is possible that network queuing delays may be the dominant delay contributor for some service classes. However, these delay components should be understood to be in context of the most latency-sensitive services, particularly on RU-DU interfaces, and relevant to the system level dimensioning. It is expected that if we will have multiple QoS classes, then the delay and loss parameters are specified on per-class basis, but such specification is outside of scope of this section.

The delay components in this section are based on presently supported O-RAN splits, i.e. 3GPP reference split configurations 7-2 & 8 for the RU-DU split (as defined in O-RAN), and 3GPP split 2 for F1 (as defined in O-RAN) and associated transport allocations, and constraints are based on the 5G service requirements from ITU & 3GPP.

Other extensions have been approved and included in version 2.0 of the O-RAN Fronthaul specification [7], which allow for so called "non-ideal" Fronthaul. It should be noted that while they allow substantially larger delays (e.g. 10 ms FH splits have been described and implemented outside of O-RAN), they cannot be considered for all possible 5G use cases, as for example it is clearly impossible to meet the 5G service-specification requirements over such large delay values over the FH for URLLC or even 4 ms eMBB services. In addition, in specific scenarios (e.g. high-speed users), adding latency to the fronthaul interface can result in reduced performance, and lower potential benefits, e.g. in Co-Ordinated Multi-Point (CoMP) mechanisms.

5.3 Hardware Acceleration and Acceleration Abstraction Layer (AAL)

As stated in Section 4.3.2, an O-Cloud Node is a collection of CPUs, Memory, Storage, NICs, BIOSes, BMCs, etc., and may include hardware accelerators to offload computational-intense functions with the aim of optimizing the performance of the O-RAN Cloudified NF (e.g., O-RU, O-DU, O-CU-CP, O-CU-UP, near-RT RIC). There are many different types of hardware accelerators, such as FPGA, ASIC, DSP, GPU, and many different types of acceleration functions, such as Low-Density Parity-Check (LDPC), Forward Error Correction (FEC), end-to-end high-PHY for O-DU, security algorithms for O-CU, and Artificial Intelligence for RIC. The combination of hardware accelerator and acceleration function, and indeed the option to use hardware acceleration, is the vendor's choice; however, all types of hardware acceleration on the cloud platform should ensure the decoupling of software from hardware. This decoupling implies the following key objectives:

- Multiple vendors of hardware GPP CPUs and accelerators (e.g., FGPA, ASIC, DSP, or GPU) can be used in O-Cloud platforms (including agreed-upon Acceleration Abstraction Layer as defined in an upcoming specification) from multiple vendors, which in turn can support the software providing RAN functionality.
- A given hardware and cloud platform shall support RAN software (including near-RT RIC, O-CU-CP, O-CU-UP, O-DU, and possibly O-RU functionality in the future) from multiple vendors.

There are different concepts that should be considered for the hardware acceleration abstraction layer on the cloud platform; these are usually the following:

- Accelerator Deployment Model
- Acceleration Abstraction Layer (AAL) Interface (i.e., the APIs used by the NFs)

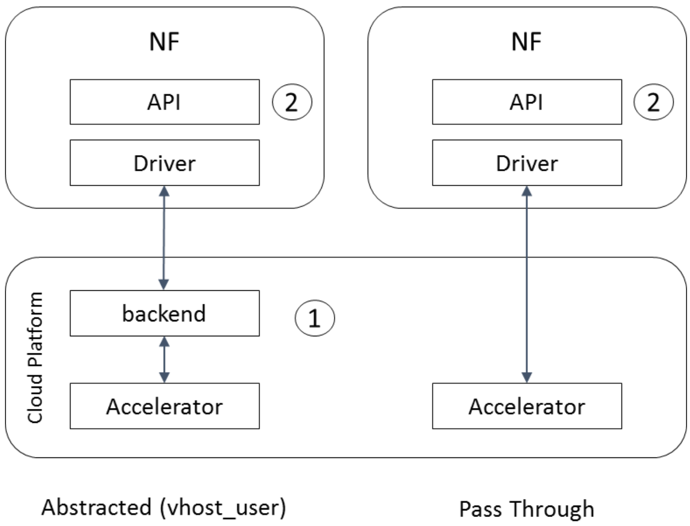


Figure 14: Hardware Abstraction Considerations

5.3.1 Accelerator Deployment Model

Figure 14 above presents two common hardware accelerator deployment models as examples: an abstracted implementation utilizing a vhost_user and virtIO type deployment, and a pass-through model using SR-IOV. While the abstracted model allows a full decoupling of the Network Function (NF) from the hardware accelerator, this model may not suit real-time latency sensitive NFs such as the O-DU. For better acceleration capabilities, SR-IOV pass through may be required, as it is supported in both VM and container environments.

5.3.2 Acceleration Abstraction Layer (AAL) Interface

To allow multiple NF vendors to utilize a given accelerator through its Acceleration Abstraction Layer (AAL) interface, the accelerators must provide an open-sourced API. Likewise, this API shall allow NFs applications to discover, configure, select and use (one or more) acceleration functions provided by a given accelerator on the cloud platform. Moreover, this API shall also support different offload architectures including look aside, inline and any combination of both. Examples of open APIs include DPDK's CryptoDev, EthDev, EventDev, and Base Band Device (BBDEV).

When delivering an NF to an Operator, it is assumed that the supplier of that Network Function will provide not only the Network Function, but it will also package the appropriate Accelerator Driver (possibly provided by a 3rd party) and will indicate the corresponding AAL profile needed in the Operator's O-Cloud. Figure 15 illustrates this for both Container and Virtual Machine (VM) deployments.

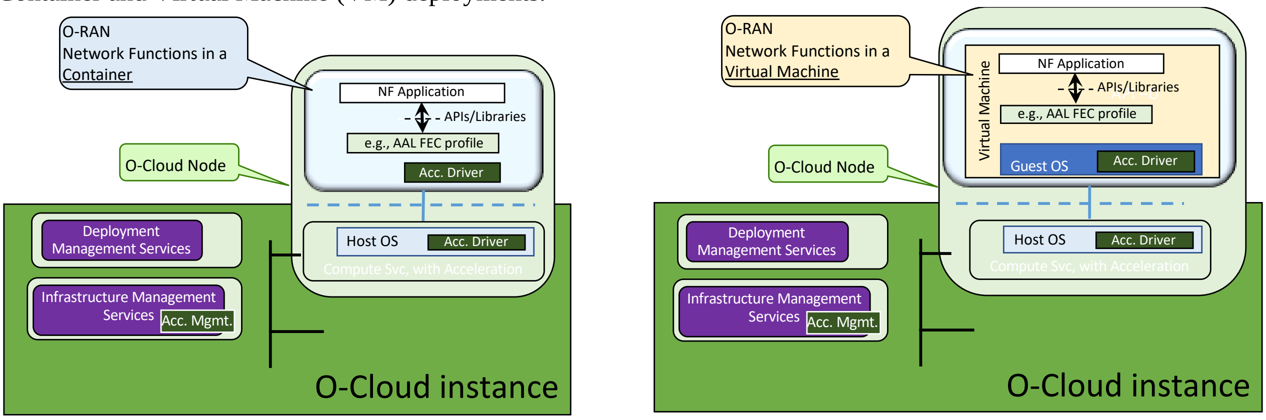


Figure 15: Accelerator APIs/Libraries in Container and Virtual Machine Implementations

5.3.3 Accelerator Management and Orchestration Considerations

Note that Figure 15 shows the APIs/Libraries as used by the NF application running in a Container or a VM, but there are several entities that require management. Accordingly, the figure also shows the Accelerator Management and Accelerator Driver in the O-Cloud. As will be discussed in Section 5.6, these entities (in addition to any hardware accelerator considerations) will be managed via O2, specifically the Infrastructure Management Services. Figure 15 also shows that the Accelerator Driver (e.g., the PMD driver) needs to be supported both by the O-Cloud Platform, by the Guest OS in case of VMs, and by the NF packaged into a container.

In general, the hardware accelerators shall be capable of being managed and orchestrated. In particular, hardware accelerators shall support feature discovery and life cycle management. Existing Open Source solutions may be leveraged for both VMs and containers as defined in an upcomingO2 specification. Examples include OpenStack Nova and Cyborg, while in Kubernetes we can leverage the device plugin framework for vendors to advertise their device and associated resources for the accelerator management.

5.4 Cloud Considerations

In this section we talk about the list of cloud platform capabilities which is expected to be provided by the cloud platform to be able to support the deployment of the scenarios which are covered by this document.

It is assumed that some or all deployment scenarios may be using VM orchestrated/managed by OpenStack and / or Container managed/orchestrated by Kubernetes, and therefore this section will cover both options.

The discussion in most sub-sections of this section is structured into (up to) three parts: (1) Common, (2) Container only, and (3) VM only.

5.4.1 Networking requirements

A Cloud Platform should have the ability to support high performance N – S and E – W networking, with high throughput and low latency.

5.4.1.1 Support for Multiple Networking Interfaces

Common: In the different scenarios, near-RT RIC, vO-CU, and vO-DU all depend on having support for multiple network interfaces. The Cloud Platform is required to support the ability to assign multiple networking interfaces to a single container or VM instance, so that the cloud platform could support successful deployment for the different scenarios.

Container-only: For example, the cloud platform can achieve this by supporting the implementation of Multus Container Networking Interface (CNI) Plugin. For more details, please see https://github.com/intel/multus-cni.

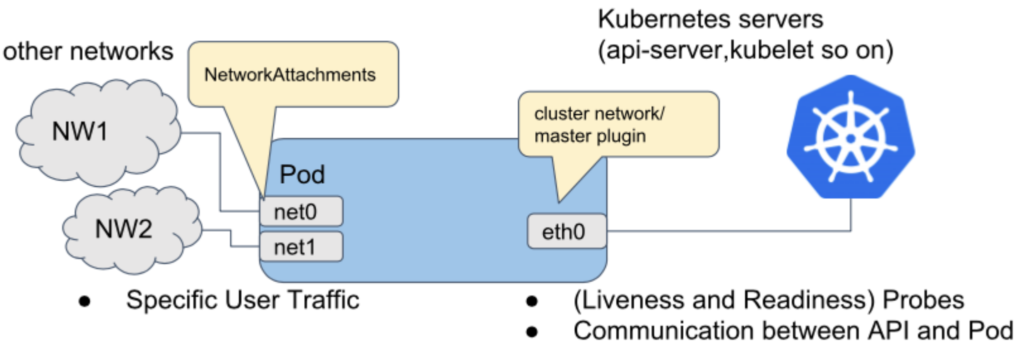


Figure 16: Illustration of the Network Interfaces Attached to a Pod, as Provisioned by Multus CNI

VM-only: OpenStack provides the Neutron component for networking. For more details, please see https://docs.openstack.org/neutron/stein/

5.4.1.2 Support for High Performance N-S Data Plane

Common: The Fronthaul connection between the O-RU/RU and vO-DU requires high performance and low latency. This means handling packets at high speed and low latency. As per the different scenarios covered in this document, multiple vO-DUs may be running on the same physical cloud platform, which will result in the need for sharing the same physical networking interface with multiple functions. Typically, the SR-IOV networking interface is used for this.

The cloud platform will need to provide support for assigning SR-IOV networking interfaces to a container or VM instance, so the instance can use the network interface (physical function or virtual function) directly without using a virtual switch.

If only one container needs to use the networking interface, the PCI pass-through network interface can provide high performance and low latency without using a virtual switch.

In general, the following two items are needed for high performance N-S data throughput:

- Support for SR-IOV; i.e., the ability to assign SR-IOV NIC interfaces to the containers/ VMs
- Support for PCI pass-through for direct access to the NIC by the container/ VM

Container-only: When containers are used, the cloud platform can achieve this by supporting the implementation of SR-IOV Network device plugin for Kubernetes. For more details, please refer to https://github.com/intel/sriov-network-device-plugin

VM-only: OpenStack provides the Neutron component for networking. For more details, please see https://docs.openstack.org/neutron/stein/admin/config-sriov.html .

5.4.1.3 Support for High-Performance E-W Data Plane

Common: High-performance E-W data plane throughput is a requirement for the implementation of the different near-RT RIC, vO-CU, and vO-DU scenarios which are covered in this document.

One of commonly used options for E-W high-performance data plane is the use of a virtual switch which provides basic communication capability for instances deployed at either the same machine or different machines. It provides L2 and L3 network functions.

To get the high performance required, one of the options is to use a Data Plan Development Kit (DPDK)-based virtual switch. Using this method, the packets will not go into Linux kernel space networking, and instead will implement userspace networking which will improve the throughput and latency. To support this, the container or VM instance will need to use DPDK to accelerate packet handling.

The cloud platform will need to provide the mechanism to support the implementation of userspace networking for container(s) / VM(s).

Container-only: As an example, the cloud platform can achieve this by supporting implementation of Userspace CNI Plugin. For more details, please refer to https://github.com/intel/userspace-cni-network-plugin.

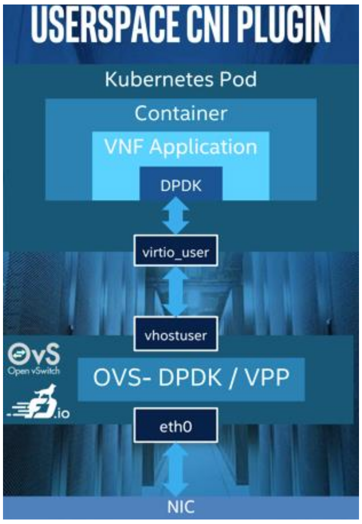


Figure 17: Illustration of the Userspace CNI Plugin

VM-only: OVS DPDK is an example of a Host userspace virtual switch and could provide high performance L2/L3 packet receive and transmit.

5.4.1.4 Support for Service Function Chaining

Common: Support for a Service Function Chaining (SFC) capability requires the ability to create a service function chain between multiple VMs or containers. In the virtualization environment, multiple instances will usually be deployed, and being able to efficiently connect the instances to provide service will be a fundamental requirement.

The ability to dynamically configure traffic flow will provide flexibility to Operators. When the service requirement or flow direction needs to be changed, the Service Function Chaining capability can be used to easily implement it instead of having to restart and reconfigure the services, networking configuration and Containers/VMs.

Container-only: An example of SFC functionality is found at: https://networkservicemesh.io/

VM only: The OpenStack Neutron SFC and OpenFlow-based SFC are examples of solutions that can implement the Service Function Chaining capability.

5.4.2 Assignment of Acceleration Resources

Common: For both container and VM solutions, specific devices such as accelerator (e.g., FPGA, GPU) may be needed. In this case, the cloud platform needs to be able to assign the specified device to container instance or VM instance.

For example, some L1 protocols require an FFT algorithm (to compute the DFT) that could be implemented in an FPGA or GPU, and the vO-DU would need the PCI Pass-Through to assign the accelerator device to the vO-DU for access and use.

5.4.3 Real-time / General Performance Feature Requirements

5.4.3.1 Host Linux OS

5.4.3.1.1 Support for Pre-emptive Scheduling

Support may be required to support Pre-emptive Scheduling (real time Linux uses the preempt_rt patch). Generally, without real time features, it is very difficult for an application to get deterministic response times for events, interrupts and other reasons[4]. In addition, during the housekeeping processes in Linux system, the application also cannot guarantee the running time (CPU cycle), so from the wireless application design perspective, it needs the real time feature. In addition, to support the requirements of high throughput, multiple accesses and low latency, some wireless applications need the priority-based OS environment.

5.4.3.2 Support for Node Feature Discovery

Common: Automated and dynamic placement of Cloud-Native Network Functions (CNFs) / microservices and VMs is needed, based on the hardware requirements imposed on the vO-DU, vO-CU and near-RT RIC functions. This requires the cloud platform to support the ability to discover the hardware capabilities on each node and advertise it via labels vs. nodes, and allows O-RAN Cloudified NFs' descriptions to have hardware requirements via labels. This mechanism is also known as Node Feature Discovery (NFD).

Container-only: For example, the cloud platform can achieve this by supporting implementation of NFD for Kubernetes. For more details, please see https://github.com/kubernetes-sigs/node-feature-discovery.

VM-only: VMs can use OpenStack mechanisms. For example, the OpenStack Nova filter, host aggregates and availability zones can be used to implement the same function.

5.4.3.3 Support for CPU Affinity and Isolation

Common: The vO-DU, vO-CU and even the near-RT RIC are performance sensitive and require the ability to consume a large amount of CPU cycles to work correctly. They depend on the ability of the cloud platform to provide a mechanism to guarantee performance determinism even when there are noisy neighbors.

Container-only: This requires the cloud platform to support using affinity and isolation of cores, so high performance Kubernetes Pod cores also can be dedicated to specified tasks. For example, the cloud platform can achieve this by implementing CPU Manager for Kubernetes. For more details, please refer to https://github.com/intel/CPU-Manager-for-Kubernetes .

VM-only: For example the modern Linux operating system uses the Symmetric MultiProcessing (SMP) mode, so the system process and application will be located at different CPU cores. To run the VM and guarantee the VM performance, the capability to assign the specific CPU cores to a VM is the way to do that. And at the same time, CPU isolation will reduce the inter-core affinity. Please refer to https://docs.openstack.org/senlin/pike/scenarios/affinity.html

5.4.3.4 Support for Dynamic HugePages Allocation

Common: When an application requires high performance and performance determinism, the reduction of paging is very helpful. vO-DU, vO-CU and even near-RT RIC can require performance determinism. The cloud platform needs to be able to support the ability to provide this mechanism to applications that require it.

This requires the cloud platform to support ability to dynamically allocate the necessary amount of the faster memory (a.k.a. HugePages) to the container or VM as necessary, and also to relinquish this memory allocation in the event of unexpected termination.

[4] Other options include things such as Linux signal, softwareirq, and perhaps using a common process. Because the pre-emptive kernel could interrupt the low priority process and occupy the CPU, it will get more chance to run the high priority process. Then through proper application design, it will have guaranteed time/resource and can have deterministic performance.

Container-only: For example, the cloud platform can achieve this by supporting implementation of Manage HugePages in Kubernetes. For more details please refer to https://kubernetes.io/docs/tasks/manage-hugepages/scheduling-hugepages/ .

VM-only: For example, the OpenStack Nova flavor setting can be used to configure the HugePage size for a VM instance. See https://docs.openstack.org/nova/pike/admin/huge-pages.html

5.4.3.5 Support for Topology Manager

Common: Some of the cloud infrastructure which is targeted in the scenarios in this document may have servers which utilize a multiple-socket configuration which comes with multiple memory regions. Each core[5] is connected to a memory region. While each CPU on one socket can access the memory region of the CPUs on another socket of the same board, the access time is significantly slower when crossing socket boundaries, and this will affect performance significantly.

The configuration of hardware with multiple memory regions is also known as Non-Uniform Memory Access (NUMA) regions. To support automated and dynamic placement of CNFs/microservices or VMs based on cloud infrastructure that has multiple NUMA regions and guarantee the response time of the application (especially for vO-DU), it is critical to be able to ensure that all the containers/VMs are associated with core(s) which are connected to the same NUMA region. In addition, if the application relies on access to hardware accelerators and/or I/O which uses memory as a way to interact with the application, it is also critical that those also use the same NUMA region that the application uses.

The cloud platform will need to provide the mechanism to enable managing the NUMA topology to ensure the placement of specified containers/VMs on cores which are on the same NUMA region, as well as making sure that the devices which the application uses are also connected to the same NUMA region.

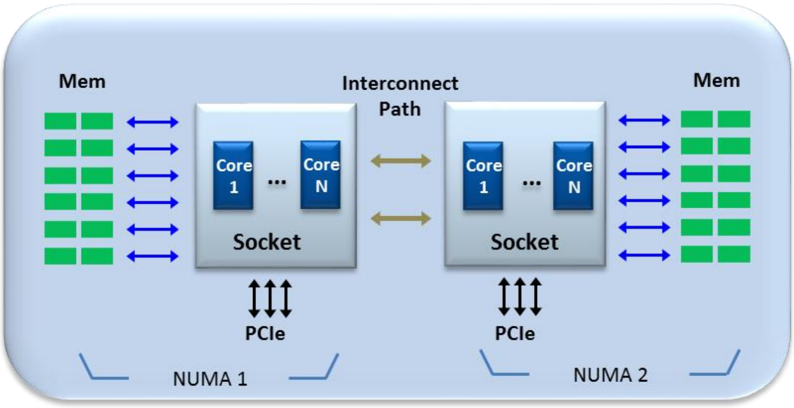


Figure 18: Example Illustration of Two NUMA Regions

5.4.3.6 Support for Scale In/Out

Common: The act of scaling in/out of containers/ VMs can be based on triggers such as CPU load, network load, and storage consumption. The network service usually is not just a single container or VM, and in order to leverage the container/ VM benefit, the network service usually will have multiple containers/ VMs. But if demand is changing dynamically, especially for the O-CU, the service needs to be scaled in/out according to service requirements such as subscriber quantity.

For example, when the number of subscribers increases, the system needs to start more container/ VM instances to ensure the service quality. From the cloud platform perspective, it could monitor the CPU load; if the load reaches a level such as 80%, it needs to scale out. If the CPU load drops 40%, it could then scale in.

Different services can scale in/out depending on different criteria, such as the CPU load, network load and storage consumption. Support for scale in/out can be helpful in implementing on-demand services.

Editor's Note: Support for scale up/down is not discussed at this time, but may be revisited in the future.

---

[5] In this document, we use the terms core and socket in the following way. A socket, or more precisely the multichip platform that fits into a server socket, contains multiple cores, each of which is a separate CPU. Each core in a socket has some dedicated memory, and also some shared memory among other cores of the same socket, which are within the same NUMA zone.

5.4.3.7 Support for Device Plugin

Common: For vO-DU, vO-CU and near-RT RIC applications, hardware accelerators such as SmartNICs, FPGAs and GPUs may be required to meet performance objectives that can't be met by using software only implementations. In other cases, such accelerators can be useful as an option to reduce the consumption of CPU cycles to achieve better cost efficiency.

The cloud platform will need to provide the mechanism to support those accelerators. This in turn requires support the ability to discover, advertise, schedule and manage devices such as SR-IOV, GPU, and FPGA.

Container-only: For example, the cloud platform can achieve this by supporting implementation of Device Plugins in Kubernetes. For more details please check: https://kubernetes.io/docs/concepts/extend-kubernetes/compute-storage-net/device-plugins/.

VM-only: The PCI passthrough feature in OpenStack allows full access and direct control of a physical PCI device in guests. This mechanism is generic for any kind of PCI device, and runs with a Network Interface Card (NIC), Graphics Processing Unit (GPU), or any other devices that can be attached to a PCI bus. Correct driver installation is the only requirement for the guest to properly use the devices.

Some PCI devices provide Single Root I/O Virtualization and Sharing (SR-IOV) capabilities. When SR-IOV is used, a physical device is virtualized and appears as multiple PCI devices. Virtual PCI devices are assigned to the same or different guests. In the case of PCI passthrough, the full physical device is assigned to only one guest and cannot be shared.

See https://wiki.openstack.org/wiki/Cyborg

5.4.3.8 Support for Direct IRQ Assignment

VM-only: The general-purpose platform has many devices that will generate the IRQ to the system. To develop a performance-sensitive application, inclusion of low-latency and deterministic timing features, and assigning the IRQ to a specific CPU core, will reduce the impact of housekeeping processes and decrease the response time to desired IRQs.

5.4.3.9 Support for No Over Commit CPU

VM-only: The "No Over Commit CPU" VM creation option is able to guarantee VM performance with a "dedicated CPU" model.

In traditional telecom equipment design, this will maintain the level of CPU utilization to avoid burst and congestion situations. In a virtualization environment, performance-sensitive applications such as vO-DU, vO-CU, and near-RT RIC will need the platform to provide a mechanism to secure the CPU resource.

5.4.3.10 Support for Specifying CPU Model

VM-only: OpenStack can use the CPU model setting to configure the vCPU for a VM. For example, QEMU allows the CPU options to be "Nehalem", "Westmere", "SandyBridge" or "IvyBridge", or alternatively it could be configured as "host-passthrough". This allows VMs to leverage advanced features of selected CPU architectures. For the vO-CU and vO-DU design and implementation, there will be some algorithm and computing functions that can leverage host CPU instructions to realize some benefits such as performance. The cloud platform needs to provide this capability to VMs.

5.4.4 Storage Requirements

The storage requirements are the same for both VM and Container based implementations.

For O-RAN components, the O-RAN Cloudified NF needs storage for the image and for the O-RAN Cloudified NF itself. It should support different scale, e.g., for a Regional Cloud vs. an Edge Cloud. The cloud platform needs to support a large-scale storage solution with redundancy, medium and small-scale storage solutions for two or more servers, and a very small-scale solution for a single server.

5.5 Sync Architecture

Synchronization mechanisms and options are receiving significant attention in the industry.

Editor's Note: O-RAN Working Groups 4 and 5 are addressing some aspects of synchronization, and more discussion of Sync is expected in future versions of this document.

Version 2 of the Control, User and Synchronization (CUS) Plane Specification [7] discusses, in chapter 9.2.2, "Clock Model and Synchronization Topology", four topology configuration options Lower Layer Split Control Plane 1 – 4 (LLS-C1 – LLS-C4) that are required to support different O-RAN deployment scenarios. Configuration LLS-C3 is seen as the most likely initial option for deployment and is discussed below. This section will provide a summary of what is required to support the LLS-C3 synchronization topology from the cloud platform perspective.

Note that in chapter 6 "Deployment Scenarios and Implementation Considerations" of this document, we call the site which runs the O-vDU the "Edge Cloud", while the Control, User and Synchronization (CUS) Plane Specification [7] calls it the "Central Site". However, the meaning is the same.

5.5.1 Cloud Platform Time Synchronization Architecture

The Time Sync deployment architecture which is described below relies on usage of Precision Time Protocol (PTP) IEEE 1588-2008 (a.k.a. IEEE 1588 Version 2) to synchronize clocks throughout the Edge Cloud site.

For LLS-C3 in the CUS specification [7], vO-DU may act Telecom Slave Clock (T-TSC) and select the time source the same SyncE and PTP distribution from fronthaul as O-RU. For vO-DU, only the ITU-T G.8275.2 type T-TSC will be addressed; others are For Further Study.

5.5.1.1 Edge Cloud Site Level – LLS-C3 Synchronization Topology

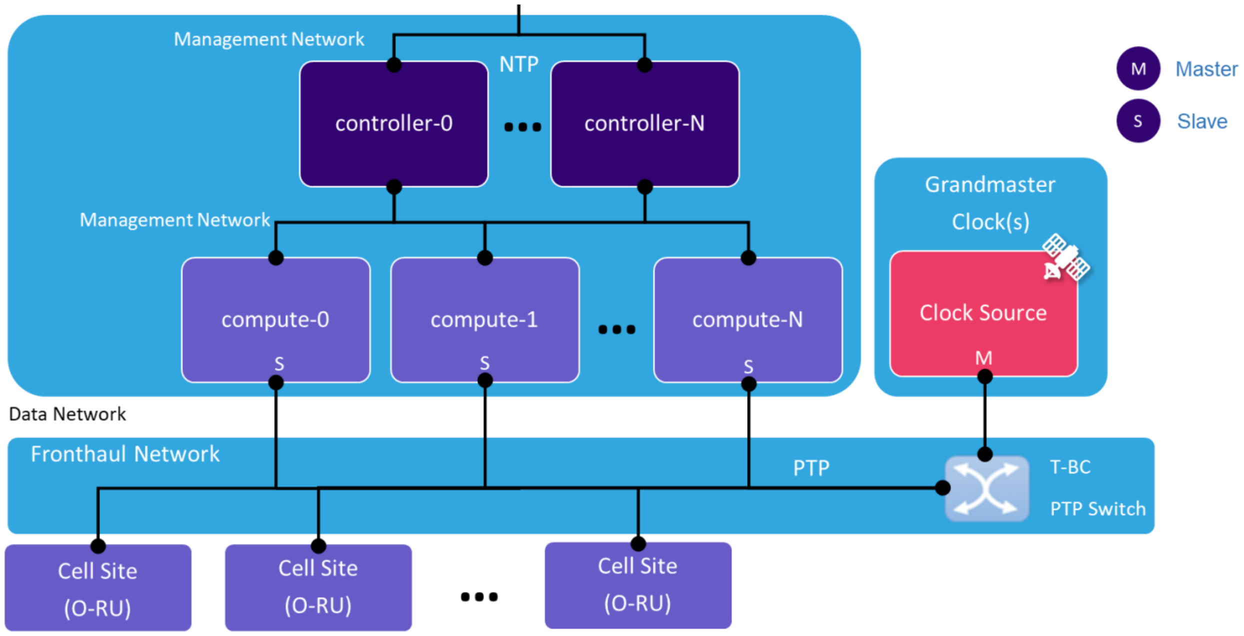
This section outlines what the time synchronization architecture should be from the cloud platform perspective, and identifies requirements that the Cloud Platform and Edge Site need to support in order to support the O-RAN deployment scenarios that use the LLS-C3 synchronization topology described in CUS specification [7].

5.5.1.1.1 LLS-C3 Synchronization Topology Edge Site Time Synchronization Architecture

The deployment architecture at the Edge Cloud site level includes:

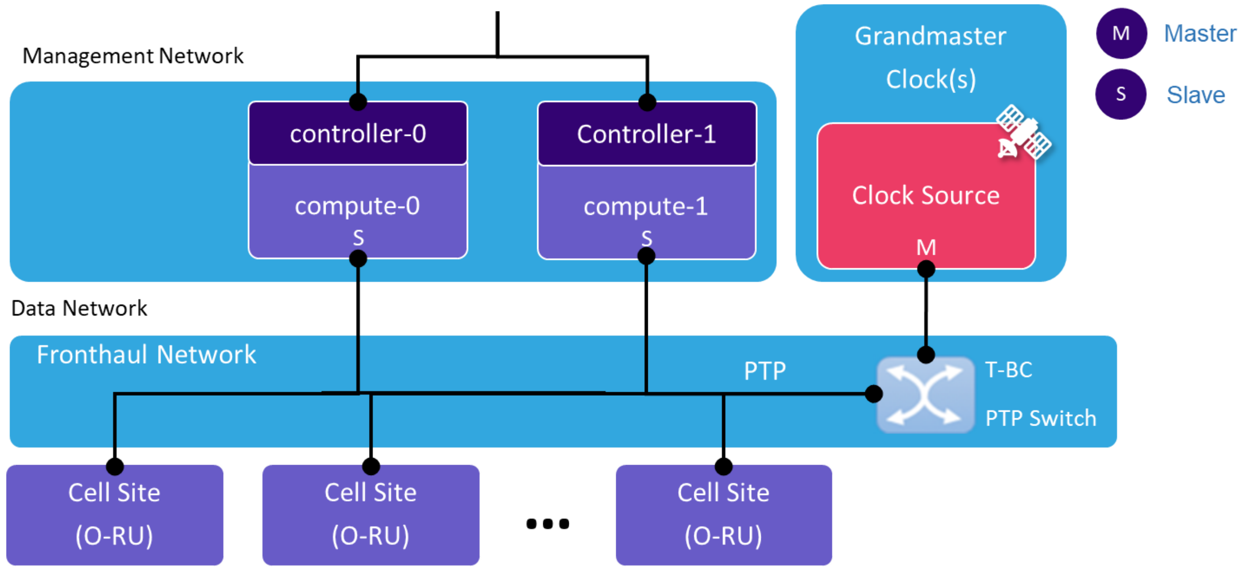
- Primary Reference Time Clock (PRTC)-traceable time source (i.e., Grandmaster Clocks):
  - External precision time source for the PTP networks, usually based on Global Navigation Satellite System/Global Positioning System (GNSS/GPS)
- Compute Nodes:
  - Compute Nodes synchronize their clocks to a Grandmaster Clock via the Fronthaul Network
- Controller Nodes:
  - Controller Nodes synchronize their clocks to the Network Time Protocol (NTP) via the Management Network

Figure 19 illustrates the relationship of these entities where the Controller functions are hosted on separate nodes from the Compute nodes. Figure 20 illustrates the relationships where each Compute node also includes the Controller functions (i.e., the hyper-converged case).



1045

1046 **Figure 19: Edge Cloud Site Time Sync Architecture for LLS-C3**



1047

1048 **Figure 20: Hyperconverged Edge Cloud Time Sync Architecture for LLS-C3**

1049 ##### 5.5.1.1.2 LLS-C3 Synchronization Topology Edge Site Requirements

1050 To support time synchronization at the Edge site, the cloud platform (O-Cloud) used at the Edge site needs to support
1051 implementation of the PTP IEEE 1588-2008 (a.k.a. IEEE 1588 Version 2) standard. The following software and
1052 hardware capabilities are required:

1053 ###### 5.5.1.1.2.1 Software

1054 Support for PTP will be needed in all the Edge Site O-Cloud nodes that support compute roles and will run vO-DU
1055 service operating as a Slave Clock. The following PTP configuration options should be provided:

1056 - Network Transport – G.8275.1 sync over Ethernet (Layer 2)
1057 - Delay Measurement Mechanism – utilize E2E or P2P to measure the delay
1058 - Time Stamping – support for hardware time stamping

1059

For example: in the case when an O-Cloud is based on the Linux OS, this will require support for Linux PTP (see http://linuxptp.sourceforge.net) with the following:

- ptp4l – implementation of PTP (Ordinary Clock, Boundary Clock), HW / SW timestamping, Delay request-response / Peer delay mechanism, and IEEE 802.3 (Ethernet) / UDP IPv4 / UDP IPv6 network transport
- phc2sys – Synchronization of two clocks, PHC and system clock (Linux clock) when using HW timestamping

5.5.1.1.2.2 Hardware

Use of High speed, low latency Network Interface Card (NIC) with support for PTP Hardware Clock (PHC) subsystem for the data interface (fronthaul) on all the compute node(s) that will run the O-vDU function.

5.6 Operations and Maintenance Considerations

Management of cloudified RAN Network Functions introduces some new management considerations, because the mapping between Network Functionality and physical hardware can be done in multiple ways, depending on the Scenario that is chosen. Thus, management of aspects that are related to physical aspects rather than logical aspects need to be designed with flexibility in mind from the start. For example, logging of physical functions, scale out actions, and survivability considerations are affected.

The O-RAN Alliance has defined key fundamentals of the OAM framework (see [8] and [9], and refer to Figure 1). Given the number of deployment scenario options and possible variations of O-RAN Managed Functions (MFs) being mapped into Managed Elements (MEs) in different ways, it is important for all MEs to support a consistent level of visibility and control of their contained Managed Functions to the Service Management & Orchestration Framework. This consistency will be enabled by support of the common OAM Interface Specification [9] for Fault Configuration Accounting Performance Security (FCAPS) and Life Cycle Management (LCM) functionality, and a common Information Modelling Framework that will provide underlying information models used for the MEs and MFs in a particular deployment.

5.6.1 The O1 Interface

As described in [8], the O1 is an interface between management entities in Service Management and Orchestration Framework and O-RAN managed elements, for operation and management, by which FCAPS management, Software management, File management shall be achieved.

5.6.2 The O2 Interface

The O2 Interface is a collection of services and their associated interfaces that are provided by the O-Cloud platform to the SMO. The services are categorized into two logical groups:

- **Infrastructure Management Services**: which include the subset of O2 functions that are responsible for deploying and managing cloud infrastructure.

- **Deployment Management Services:** which include the subset of O2 functions that are responsible for managing the lifecycle of virtualized/containerized deployments on the cloud infrastructure.

The O2 services and their associated interfaces shall be specified in the upcoming O2 specification. Any definitions of SMO functional elements needed to consume these services shall be described in OAM architecture. O2 interface would also address the management of hardware acceleration and supporting software in the O-Cloud platform.

5.7 Transport Network Architecture

While a Transport Network is a necessary foundation upon which to build any O-RAN deployment, a great many of the aspects of transport do not have to be addressed or specified in O-RAN Alliance documents. For example, any location with cloud servers will be connected by layer 2 or layer 3 switches, but we do not need to specify much if anything about them in this document.

The transport media used, particularly for fronthaul, can have an effect on aspects such as performance. However, in the current version of this document we have been assuming that fiber transport is used.

Editor's Note: Other transport technologies (e.g., microwave) are also possible, and could be addressed at a later date.

That said, the use of an (optional) Fronthaul Gateway (FH GW) will have noteworthy effects on any O-RAN deployment that uses it.

5.7.1 Fronthaul Gateways

In the deployment scenarios that follow, when the O-DU and O-RU functions are not implemented in the same physical node, a Fronthaul Gateway is shown as an *optional* element between them. A Fronthaul Gateway can be motivated by different factors depending on a carrier's deployment, and may perform different functions.

The O-RAN Alliance does not currently have a single definition of a Fronthaul Gateway, and this document does not attempt to define one. However, the Fronthaul Gateway is included in the diagrams as an optional implementation to acknowledge the fact that carriers are considering Fronthaul Gateways in their plans. Below are some examples of the functionality that could be provided:

- A FH GW can convert CPRI connections to the node supporting the O-RU function to eCPRI connections to the node that provides O-DU functionality.
    - Note that when there is no FH GW, it is assumed that the Open Fronthaul interface between the O-RU and O-DU uses Option 7-2, as mentioned earlier in Section 4.1. When there is a FH GW, it may have an Option 7-2 interface to both the O-DU and the O-RU, but it is also possible for the FH GW to have a different interface to the O-RU/RU; for example, where CPRI is supported.
- A FH GW can support the aggregation of fiber pairs.
- A FH GW must support the following forwarding functions:
    - Downlink: Transport the traffic from O-DU to each O-RU (and cascading FH GW, if present)
    - Uplink: Summation of traffic from O-RUs
- A FH GW can provide power to the NEs supporting the O-RU function, e.g. via Power over Ethernet (PoE) or hybrid cable/fibers

5.8 Overview of Deployment Scenarios

The description of logical functionality in O-RAN includes the definition of key interfaces E2, F1, and Open Fronthaul. However, as noted earlier, this does not mean that each Network Function block must be implemented in a separate O-RAN Physical NF/O-RAN Cloudified NF. Multiple logical functions can be implemented in a single O-RAN Physical NF/O-RAN Cloudified NF (for example O-DU and O-RU may be packaged as a single appliance).

We assume that when Network Functions are implemented as different O-RAN Physical NFs/O-RAN Cloudified NFs, the interfaces between them must conform to the O-RAN specifications. However, when multiple Network Functions are implemented by a single O-RAN Physical NF/O-RAN Cloudified NF, it is up to the operator to decide whether to enforce the O-RAN interfaces between the embedded Network Functions. However, note that the OAM requirements for each separate Network Function will still need to be met.

The current deployment scenarios for discussion are summarized in the figure below. This includes options that are deployable in both the short and long term. Each will be discussed in some detail in the following sections, followed by a summary of which one or ones are candidates for initial focus. Please note that, to help ease the high-level depiction of functionality, a single O-CU box is shown with an F1 interface, but in detailed discussions of specific scenarios, this will need to be discussed properly as composed of an O-CU-CP function with an F1-c interface and an O-CU-UP function with an F1-u interface. Furthermore, there would in general be an unequal number of O-CU-CP and O-CU-UP instances.

Figure 21 below shows the Network Functions at the top, and each identified scenario shows how these Network Functions are deployed as O-RAN Physical NFs or as O-RAN Cloudified NFs running on an O-RAN compliant O-Cloud. The term O-Cloud is defined in Section 4. Please note that the requirements for an O-Cloud are driven by the Network Functions that need to be supported by the hardware, so for instance an O-Cloud that supports an O-RU function would be different from an O-Cloud that supports O-CU functionality.

Finally, note that in the high-level figure below, the User Plane (UP) traffic is shown being delivered to the UPF. As will be discussed, in specific scenarios it is sometimes possible for UP traffic to be delivered to edge applications that are supported by Mobile Edge Computing (MEC). However, note that the specification of MEC itself is out of scope of this document.

Note that vendors are not required to support all scenarios – it is a business decision to be made by each vendor. Similarly, each operator will decide which scenarios it wishes to deploy.

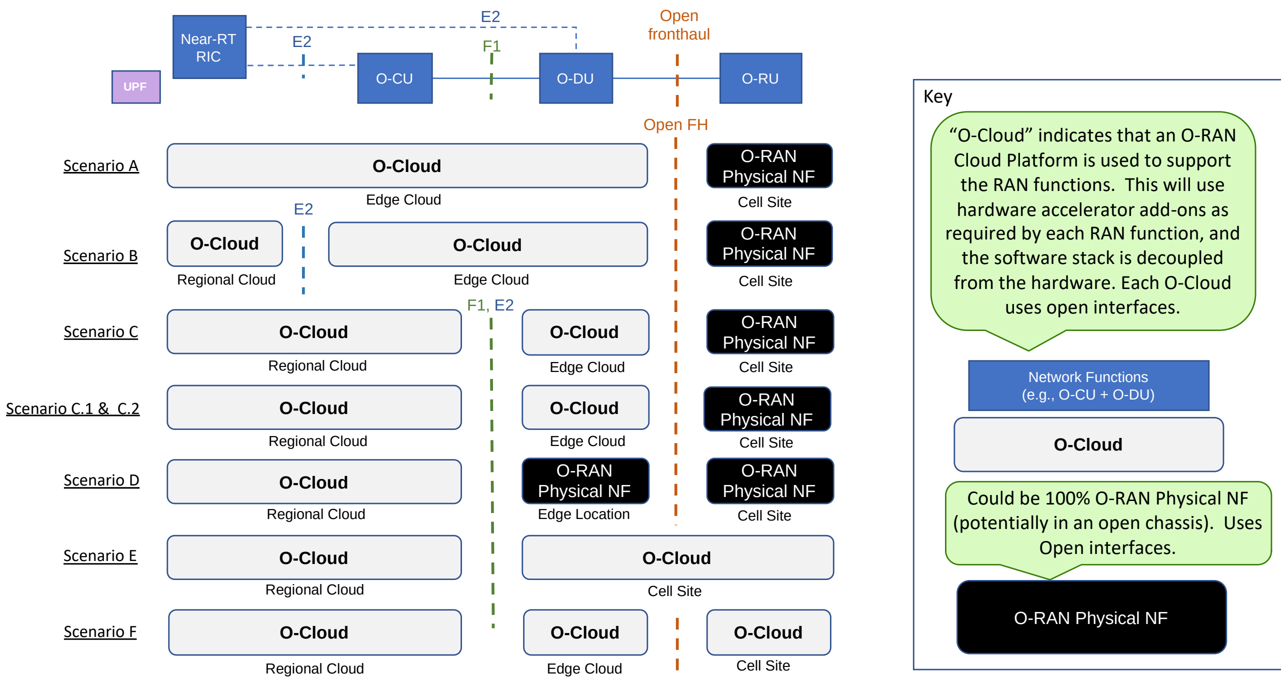


Figure 21: High-Level Comparison of Scenarios

Each scenario is discussed in the next section.

6 Deployment Scenarios and Implementation Considerations

This section reviews each of the deployment scenarios in turn. For a given scenario, the requirements that apply to the O-RAN Physical NFs, O-RAN Cloudified NFs or O-Cloud platforms may become more specific and unique, while many of the logical Network Function requirements will remain the same.

Please note that in all of the scenario figures of this section, the interfaces are logical interfaces (e.g., F1, E2, etc.). This has a couple of implications. First, the two functions on each side of an interface could be on different devices separated by physical transport connections (e.g., fiber or Ethernet transport connections), could be on different devices within the same cloud platform, or could even exist within the same server. Second, the functions on each side of an interface could be from the same vendor or different vendors.

In addition, please note that all User Plane interfaces are shown with a solid lines, and all Control Plane interfaces use dashed lines.

Editor's note: The terms vO-CU and vO-DU represent virtualized or containerized O-CU and O-DU, and are used interchangeably with O-CU and O-DU in these scenarios (with the exception when the O-DU is explicitly stated as a non-virtualized O-DU).

6.1 Scenario A

In this scenario, the near-RT RIC, O-CU, and O-DU functions are all virtualized on the same cloud platform, and interfaces between those functions are within the same cloud platform.

This scenario supports deployments in dense urban areas with an abundance of fronthaul capacity that allows BBU functionality to be pooled in a central location with sufficiently low latency to meet the O-DU latency requirements. Therefore, it does not attempt to centralize the near-RT RIC more than the limit that O-DU functionality can be centralized.

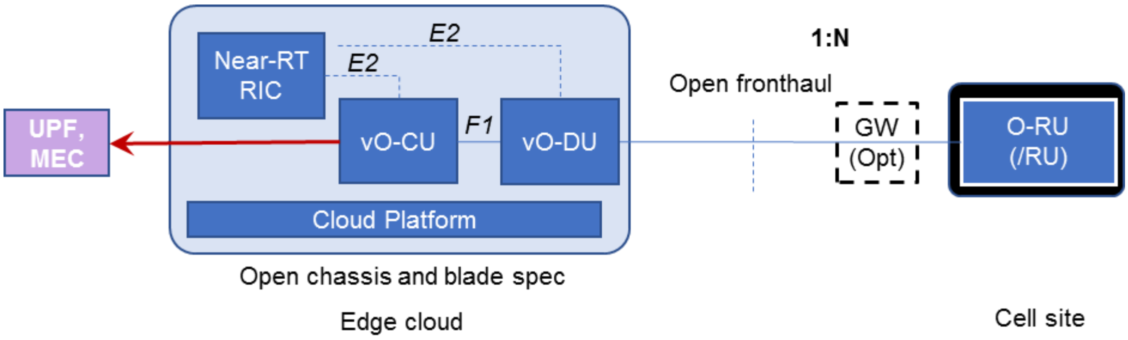


Figure 22: Scenario A

Also please note that if the optional FH GW is present, the interface between it and the Radio Unit might not meet the O-RAN Fronthaul requirements (e.g., it might be an Option 8 interface), in which case the Radio Unit could be referred to as an "RU", not an "O-RU". However, if FH GWs are defined to support an interface such as Option 8, it could be argued that the O-RU definition at that time will support Option 8.

6.1.1 Key Use Cases and Drivers

Editor's Note: This section is FFS.

6.2 Scenario B

In this scenario, the near-RT RIC Network Function is virtualized on a Regional Cloud Platform, and the O-CU and O-DU functions are virtualized on an Edge Cloud hardware platform that in general will be at a different location. The interface between the Regional Cloud and the Edge cloud is E2. Interfaces between the O-CU and O-DU Network Functions are within the same Cloud Platform.

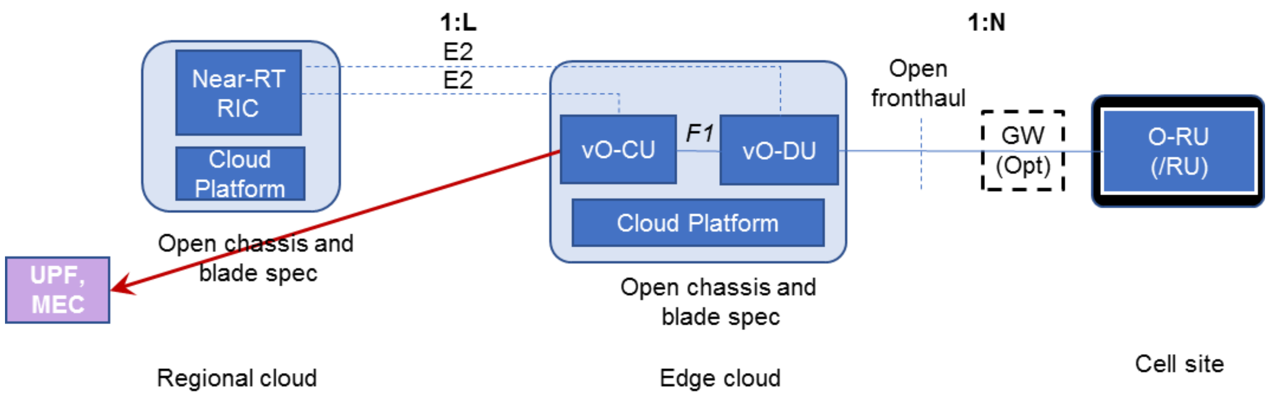


Figure 23: Scenario B

This scenario is to support deployments in locations with limited remote fronthaul capacity and O-RUs spread out in an area that limits the number of O-RUs that can be supported by pooled vO-CU/vO-DU functionality while still meeting the O-DU latency requirements. The use of a FH GW in the architecture allows significant savings in providing transport between the O-RU and vO-DU functionality.

As discussed earlier in Section 5.1.3, the O-CU and O-DU functions can be virtualized using either simple centralization or pooled centralization. The desire is to have support for pooled centralization, although we need to understand what needs to be developed to enable such sharing. Perhaps pooling will be a later feature, but any initial solution should not preclude a future path to a pooled solution.

6.2.1 Key Use Cases and Drivers

In this case, there are multiple O-RUs distributed in an area served by a centralized vO-DU functionality that can meet the latency requirements. Depending on the concentration of the O-RUs, N could vary, but in general is expected to be engineered to support < 64 TRPs per O-DU.[6] The near-RT RIC is centralized further to allow for optimization based on a more global view (e.g., a single large metropolitan area), and to reduce the number of separate near-RT RIC instances that need to be managed.

The driving use case for this is to support an outdoor deployment of a mix of Small Cells and Macro cells in a relatively dense urban setting. This can support mmWave as well as Sub-6 deployments.

In this scenario, a given "virtual BBU" supports both vO-CU and vO-DU functions, and can connect many O-RUs. Current studies show that savings from pooling are significant but level off once more than 64 Transmission Reception Points (TRPs) are pooled. This would imply N would be around 32-64. This deployment should support tens of thousands of O-RUs per near-RT RIC, so L could easily exceed 100.

Below is a summary of the cardinality requirements assumed for this scenario.

Table 2: Cardinality and Delay Performance for Scenario B

	Attribute	RIC – O-CU	O-CU – O-DU	O-DU – O-RU/RU
	Example Cardinality	L = 100+	M=1	N = 1-64

6.3 Scenario C

In this scenario, the near-RT RIC and O-CU Network Functions are virtualized on a Regional Cloud Platform with a general server hardware platform, and the O-DU Network Functions are virtualized on an Edge Cloud hardware platform that is expected to include significant hardware accelerator capabilities. Interfaces between the near-RT RIC and the O-CU network functions are within the same Cloud Platform. The interface between the Regional Cloud and the Edge cloud is F1, and an E2 interface from the near-RT RIC to the O-DU must also be supported.

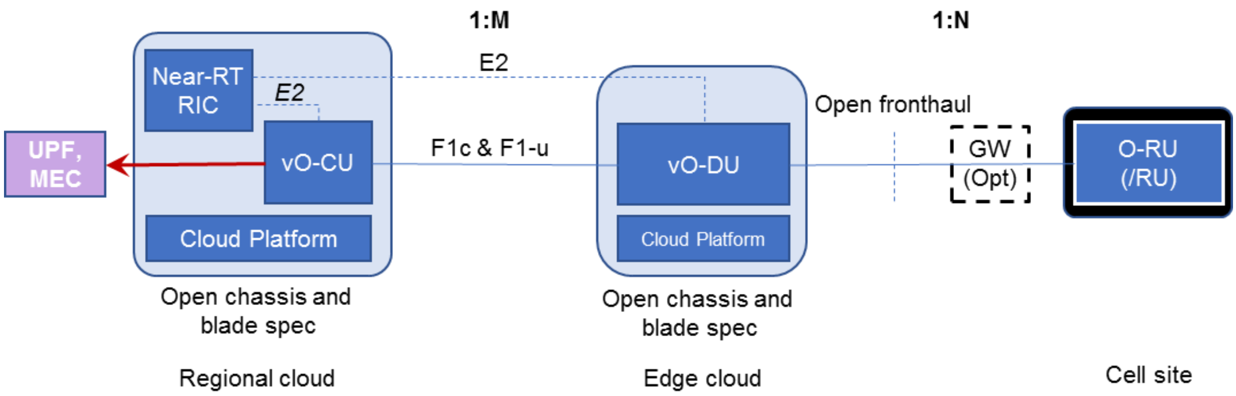


Figure 24: Scenario C

This scenario is to support deployments in locations with limited remote Fronthaul capacity and O-RUs spread out in an area that limits the number of O-RUs that can be pooled while still meeting the O-DU latency requirements. The O-CU

[6] It is assumed that one O-RU is associated with one TRP. For example, if a cell site has three sectors, then each sector would have at least one TRP and hence at least three O-RUs.

Network Function is further pooled to increase the efficiency of the hardware platform which it shares with the near-RT RIC Network Function.

However, note that if a service type has tighter O-CU delay requirements than other services, then that may either severely limit the number of O-RUs supported by the Regional cloud, or a method will be needed to separate the processing of such services. This will be discussed further in the following C.1 and C.2 Scenarios.

The use of a FH GW in the architecture allows significant savings in providing transport between the O-RU and vO-DU functionality.

6.3.1 Key Use Cases and Drivers

In this case, there are multiple O-RUs distributed in an area where each O-RU can meet the latency requirement for the pooled vO-DU function. The near-RT RIC and O-CU Network Functions are further centralized to realize additional efficiencies.

A use case for this is to support an outdoor deployment of a mix of Small Cells and Macro cells in a relatively dense urban setting. This can support mmWave as well as Sub-6 deployments.

In this scenario, as in Scenario B, the Edge Cloud is expected to support roughly 32-64 O-RUs. This deployment should support tens of thousands of O-RUs per near-RT RIC.

Below is a summary of the cardinality and the distance/delay requirements assumed for this scenario.

Table 3: Cardinality and Delay Performance for Scenario C

	Attribute	RIC – O-CU	O-CU – O-DU	O-DU – O-RU/RU
	Example Cardinality	L= 1	M=100+	N=Roughly 32-64

6.3.2 Scenario C.1, and Use Case and Drivers

This is a variation of Scenario C, driven by the fact that different types of traffic (network slices) have different latency requirements. In particular, URLLC has more demanding user-plane latency requirements, and Figure 25 below shows how the vO-CU User Part (vO-CU-UP) could be terminated in different places for different network slices. Below, network slice 3 is terminated in the Edge Cloud. This scenario is also suitable in case there isn't enough space or power supply to install all vO-CUs and vO-DUs in one Edge Cloud site.

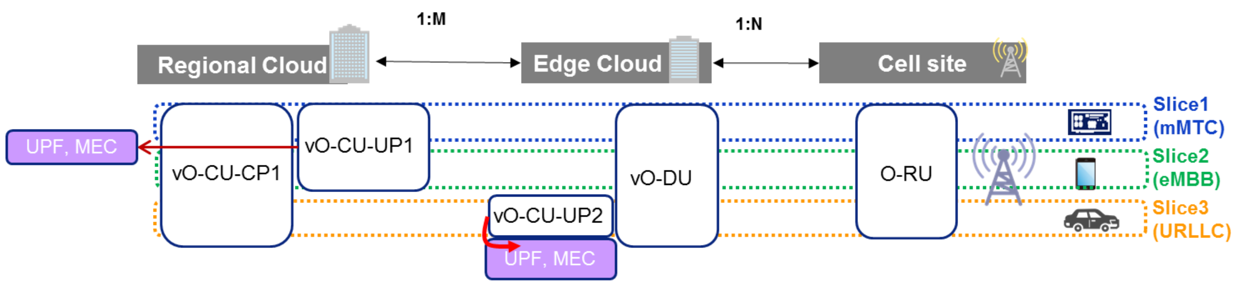


Figure 25: Treatment of Network Slices: MEC for URLLC at Edge Cloud, Centralized Control, Single vO-DU

In Scenario C.1, all O-CU control is placed in the Regional Cloud, and there is a single vO-DU for all Network Slices. Only the placement of the vO-CU-CP differs, depending on the network slice. Below is the diagram of this scenario, using the common diagram conventions of all scenarios.

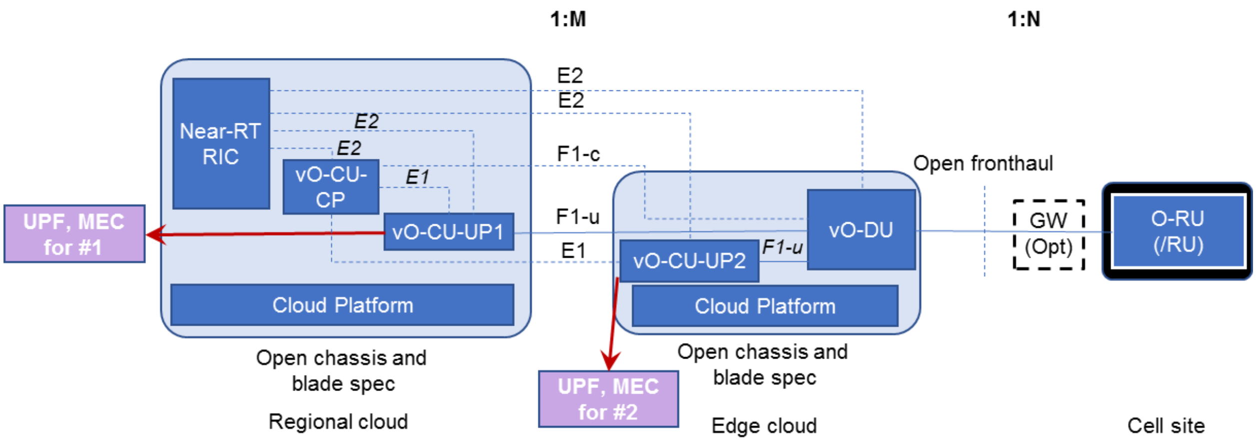


Figure 26: Scenario C.1

Below is a summary of the cardinality and the distance/delay requirements assumed for this scenario. The URLLC user plane requirements are what drive the placement of the vO-CU-UP function to be in the Edge cloud.

Table 4: Cardinality and Delay Performance for Scenario C.1

	Attribute	RIC – O-CU	O-CU – O-DU	O-DU – O-RU/RU
	Example Cardinality	L= 1	M=320	N=100
Delay Max 1-way (distance)	*mMTC*	NA	625 µs (125 km)	100 µs (20 km)
	eMBB	NA	625 µs (125 km)	100 µs (20 km)
	*URLLC (**user**/control)*	NA	**100 µs (20 km)**/625 µs (125 km)	100 µs (20 km)

6.3.3 Scenario C.2, and Use Case and Drivers

This is a second variation of Scenario C, which utilizes the same method of placing some vO-CU user plane functionality in the Edge Cloud, and some in the Regional Cloud. However, instead of having one vO-DU for all network slices, there are different vO-DU instances in the Edge Cloud.

It is driven by factors including the following two use cases:

- One driver is RAN (O-RU) sharing among operators. In this use case, any operator can flexibly launch vO-CU and vO-DU instances at Edge or Regional Cloud site. For example, as shown in Figure 27, Operator #1 wants to launch the vO-CU1 instance in the Regional Cloud, and the vO-DU1 instance at subtending Edge Cloud sites. On the other hand, Operator #2 wants to install both the vO-CU2 and vO-DU2 instances at the same Regional Cloud site. Note that both operators will share the O-RU).
- Another driver is that, even within a single operator, that operator can customize scheduler functions depending on the network slice types, and can place the vO-CU and vO-DU instances depending on the network slice types. For example, an operator may launch both vO-CU and vO-DU at the edge cloud site (see Operator #2 below) to provide a URLLC service.

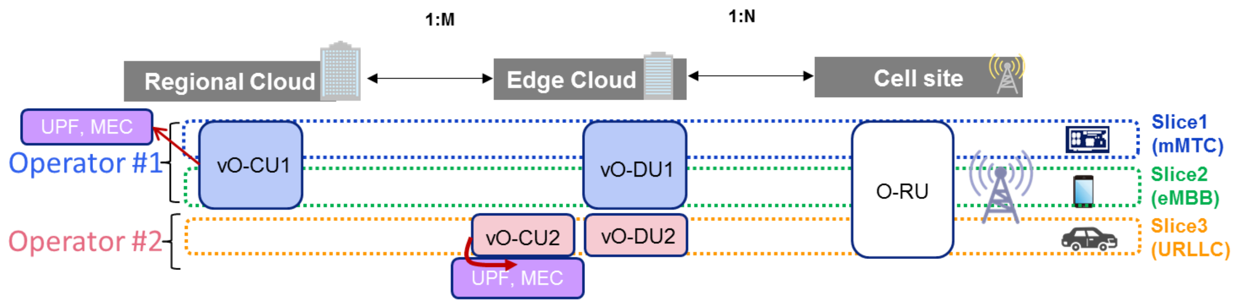


Figure 27: Treatment of Network Slice: MEC for URLLC at Edge Cloud, Separate vO-DUs

The multi-Operator use case has the following pros and cons:

Pros:

- O-RU sharing can reduce TCO
- Flexible CU/DU location allows deployments to consider not only service requirements but also limitations of space or power in each site

Cons:

- Allowing multiple operators to share O-RU resources is expected to require changes to the Open Fronthaul interface (especially the handshake among more than one vO-DU and a given O-RU).
- This change seems likely to have M-plane specification impact. Therefore, this approach would need O-RAN buy-in and approval.

Figure 28 below illustrates how different Component Carriers can be allocated to different operators, at the same O-RU *at the same time*. Note that some updates of not only M-plane but also CUS-plane specifications will be required when considering frequency resource sharing among DUs.

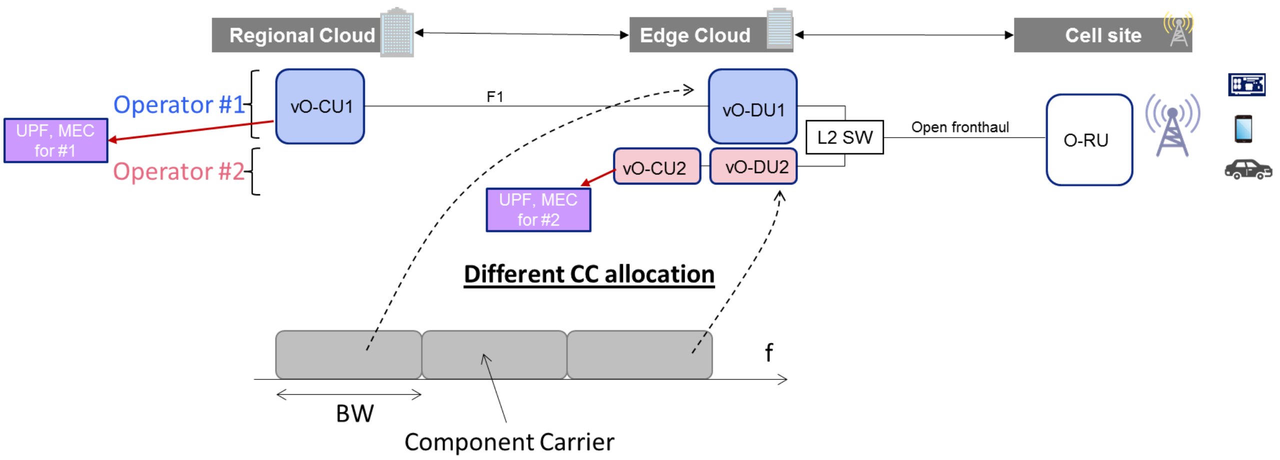


Figure 28: Single O-RU Being Shared by More than One Operator

The diagram of how Network Functions map to Networks Elements for Scenario C.2 is shown below.

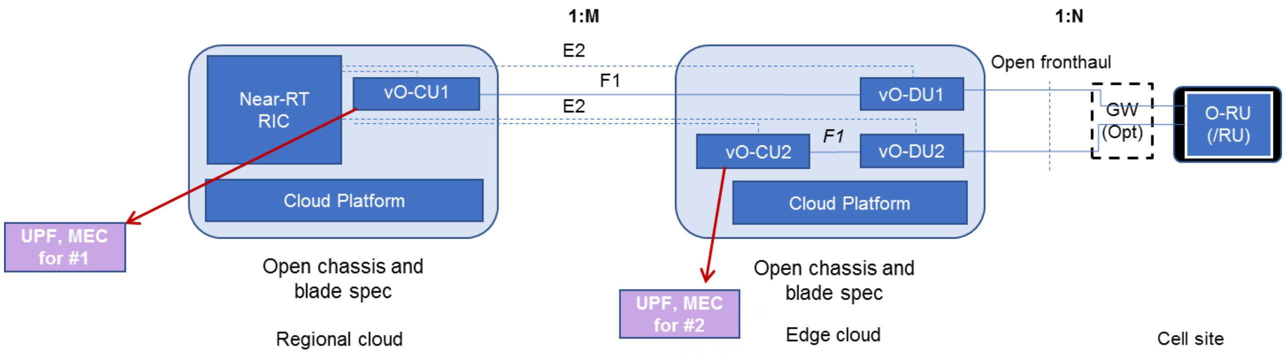


Figure 29: Scenario C.2

The performance requirements are the same as those discussed earlier for Scenario C.1 in Section 6.3.2.

6.4 Scenario D

This scenario is a variation on Scenario C, but in this case the O-DU functionality is supported by an O-RAN Physical NF rather than an O-Cloud.

The general assumption is that Scenario D has the same use cases and performance requirements as Scenario C, and the primary difference is in the business decision of how the O-RAN Physical NF based solution compares with the O-RAN compliant O-Cloud solution. Implementation considerations (discussed in Section 5.1) could lead a carrier to decide that an acceptable O-Cloud solution is not available in a deployment's timeframe.

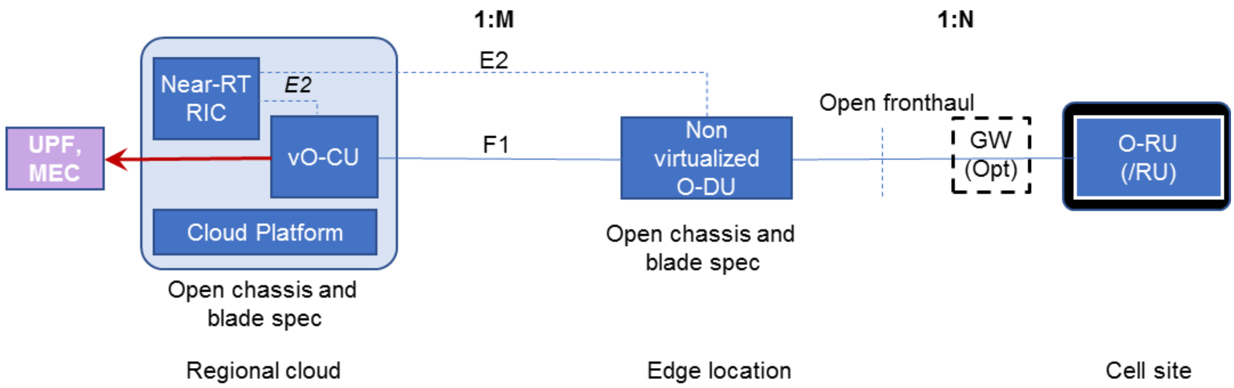


Figure 30: Scenario D

6.5 Scenario E

In contrast to Scenario D, this scenario assumes that not only can the O-DU be virtualized as in Scenario C, but that the O-RU can also be successfully virtualized. Furthermore, the O-RU and O-DU would be implemented in the same O-Cloud, which has acceleration hardware required by both the O-RU and O-DU.

Note, this seems to be a future scenario, and is not part of our initial focus.

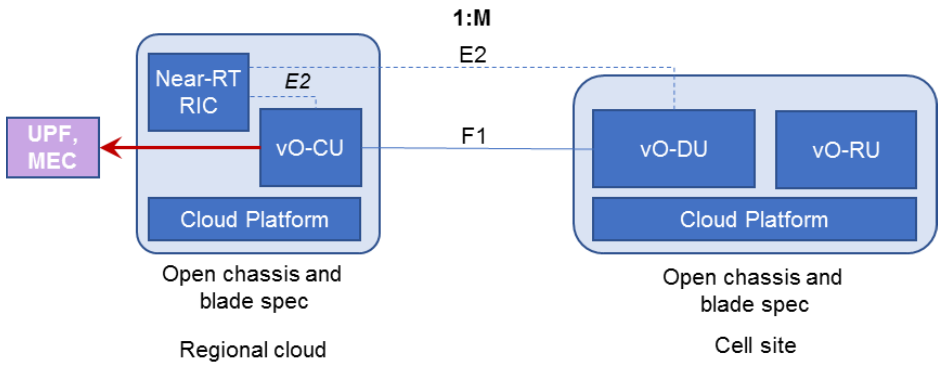


Figure 31: Scenario E

6.5.1 Key Use Cases and Drivers

Because the O-DU and O-RU are implemented in the same O-Cloud in this Scenario, it seems that the O-DU implementation must meet the environmental and accessibility requirements typically associated with an O-RU. Therefore, an indoor use case seems most appropriate.

6.6 Scenario F

This is a variation on Scenario E in which the O-DU and O-RU are both virtualized, but in different O-Clouds. This means that:

- The O-DU function can be placed in a more convenient location in terms of accessibility for maintenance and upgrades.
- The O-DU function can be placed in an environment that is semi-controlled or controlled, which reduces some of the implementation complexity.

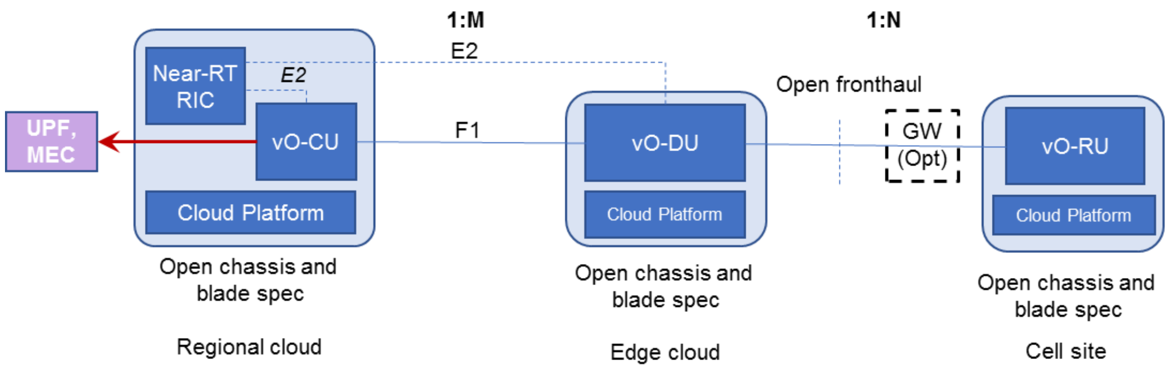


Figure 32: Scenario F

6.6.1 Key Use Cases and Drivers

Because this assumes that the O-RU is virtualized, this is a future use case.

This use case seems to be better suited for outdoor deployments (e.g., pole mounted) than Scenario E.

6.7 Scenarios of Initial Interest

More scenarios have been identified than can be addressed in the initial release of this document. Scenario B has been selected as the one to address initially, and to be the subject of detailed treatment in a Scenario document (refer back to Figure 1). Other scenarios are expected to be addressed in later work.

7 Appendix A (informative): Extensions to Current Deployment Scenarios to Include NSA

In this appendix, some extensions to (some of) the current deployment scenarios are proposed with the aim of introducing Non-Standalone (NSA) in the pictures, consistently with the scope O-RAN cloud architecture. These extensions will be the basis of the discussion for next version of the present document. In the following charts the subscript 'N' is indicating blocks related to NR, while the subscript 'E' is indicating blocks related to E-UTRA.[7] For E-UTRA, the W1 interface is indicated. Its definition is ongoing in a 3GPP work item.

7.1 Scenario A

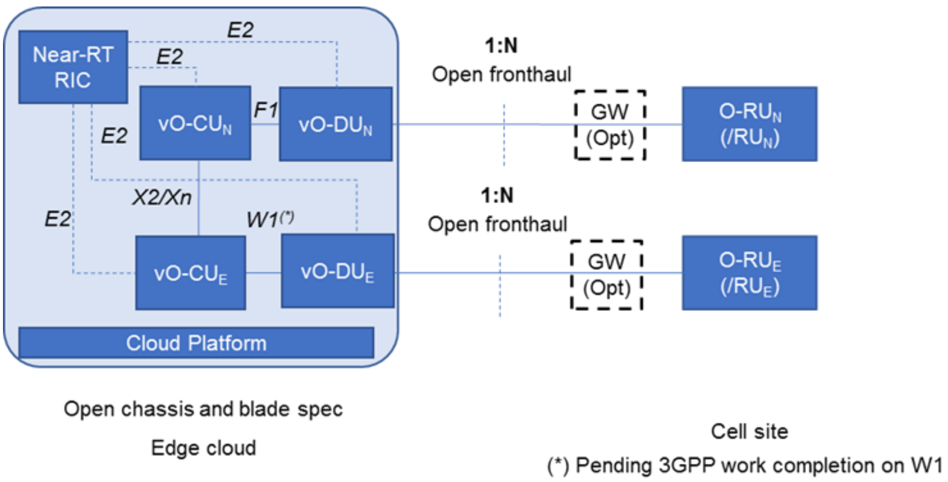


Figure 33: Scenario A, Including NSA

7.2 Scenario B

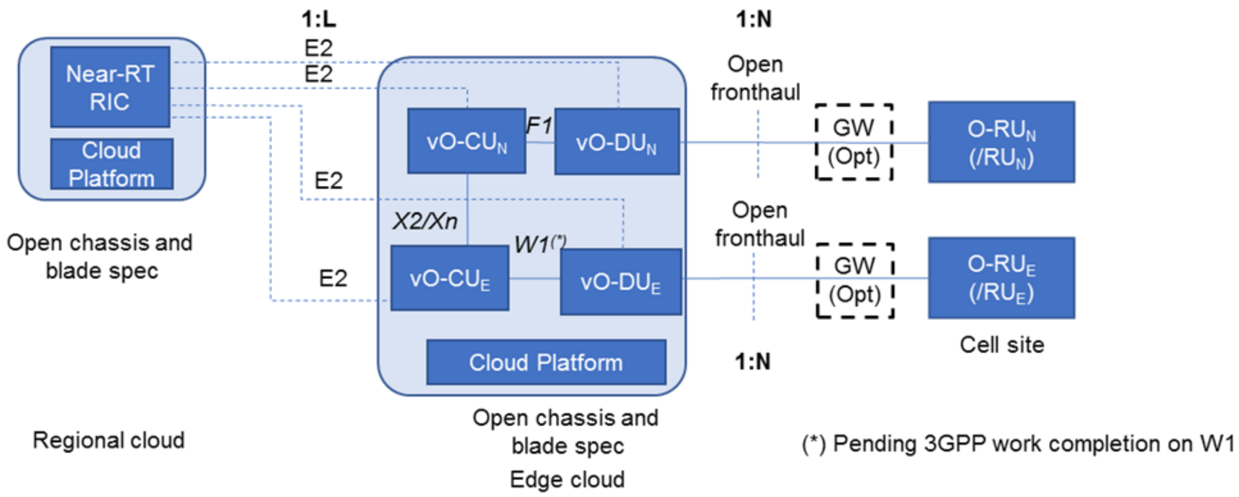


Figure 34: Scenario B, Including NSA

[7] No UPF or MEC blocks are explicitly indicated in the figures of this appendix, as the focus of this appendix is on the radio part.

7.3 Scenario C

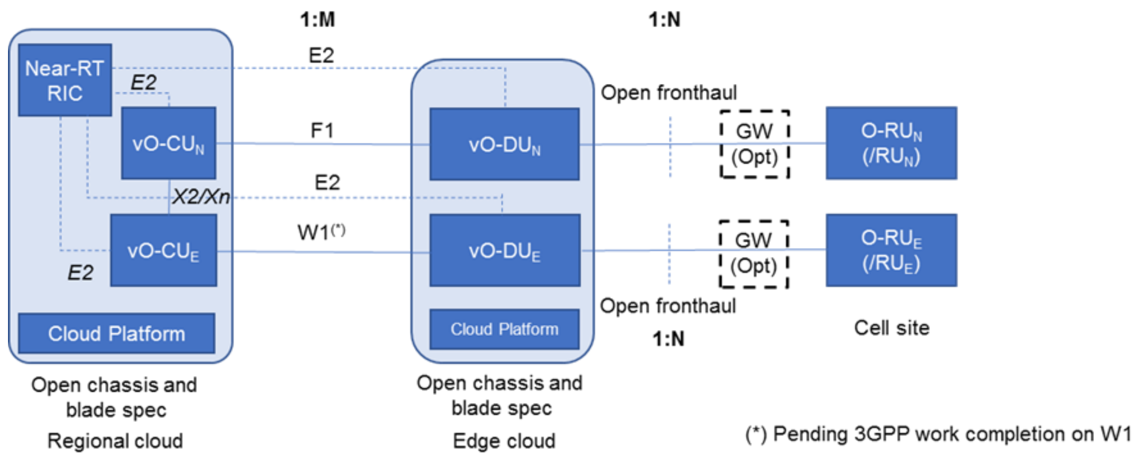


Figure 35: Scenario C, Including NSA

7.4 Scenario C.2

The scenario addresses both the single and multi-operator cases. To reduce the complexity in the figure the multi operator case is considered, so no X2/Xn interface is present between CU_N1 and CU_E2 or between CU_E1 and CU_N2.

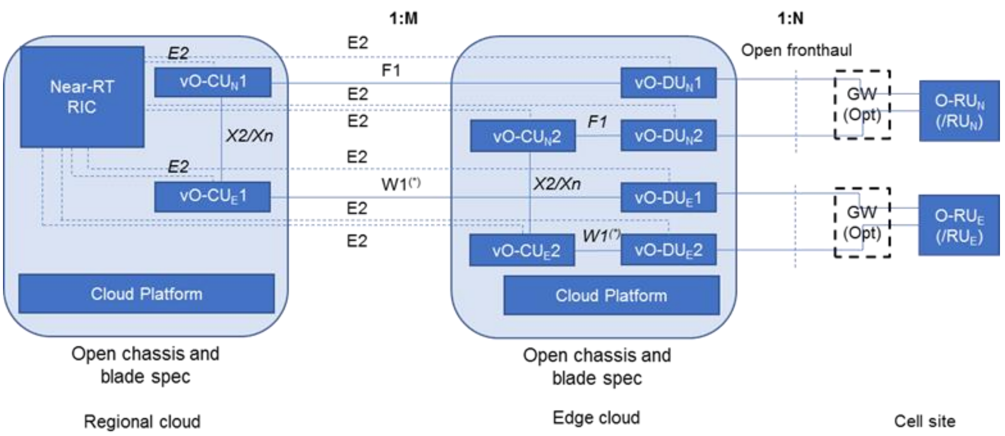


Figure 36: Scenario C.2, Including NSA

7.5 Scenario D

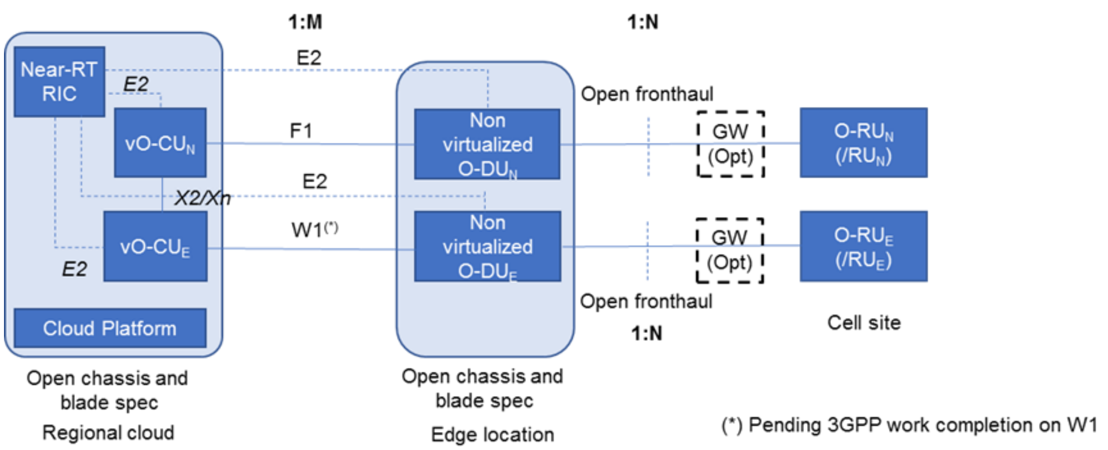


Figure 37: Scenario D, Including NSA

Annex ZZZ: O-RAN Adopter License Agreement

BY DOWNLOADING, USING OR OTHERWISE ACCESSING ANY O-RAN SPECIFICATION, ADOPTER AGREES TO THE TERMS OF THIS AGREEMENT.

This O-RAN Adopter License Agreement (the "Agreement") is made by and between the O-RAN Alliance and the entity that downloads, uses or otherwise accesses any O-RAN Specification, including its Affiliates (the "Adopter").

This is a license agreement for entities who wish to adopt any O-RAN Specification.

SECTION 1: DEFINITIONS

1.1 "Affiliate" means an entity that directly or indirectly controls, is controlled by, or is under common control with another entity, so long as such control exists. For the purpose of this Section, "Control" means beneficial ownership of fifty (50%) percent or more of the voting stock or equity in an entity.

1.2 "Compliant Portion" means only those specific portions of products (hardware, software or combinations thereof) that implement any O-RAN Specification.

1.3 "Adopter(s)" means all entities, who are not Members, Contributors or Academic Contributors, including their Affiliates, who wish to download, use or otherwise access O-RAN Specifications.

1.4 "Minor Update" means an update or revision to an O-RAN Specification published by O-RAN Alliance that does not add any significant new features or functionality and remains interoperable with the prior version of an O-RAN Specification. The term "O-RAN Specifications" includes Minor Updates.

1.5 "Necessary Claims" means those claims of all present and future patents and patent applications, other than design patents and design registrations, throughout the world, which (i) are owned or otherwise licensable by a Member, Contributor or Academic Contributor during the term of its Member, Contributor or Academic Contributorship; (ii) such Member, Contributor or Academic Contributor has the right to grant a license without the payment of consideration to a third party; and (iii) are necessarily infringed by implementation of a Final Specification (without considering any Contributions not included in the Final Specification). A claim is necessarily infringed only when it is not possible on technical (but not commercial) grounds, taking into account normal technical practice and the state of the art generally available at the date any Final Specification was published by the O-RAN Alliance or the date the patent claim first came into existence, whichever last occurred, to make, sell, lease, otherwise dispose of, repair, use or operate an implementation which complies with a Final Specification without infringing that claim. For the avoidance of doubt in exceptional cases where a Final Specification can only be implemented by technical solutions, all of which infringe patent claims, all such patent claims shall be considered Necessary Claims.

1.6 "Defensive Suspension" means for the purposes of any license grant pursuant to Section 3, Member, Contributor, Academic Contributor, Adopter, or any of their Affiliates, may have the discretion to include in their license a term allowing the licensor to suspend the license against a licensee who brings a patent infringement suit against the licensing Member, Contributor, Academic Contributor, Adopter, or any of their Affiliates.

SECTION 2: COPYRIGHT LICENSE

2.1 Subject to the terms and conditions of this Agreement, O-RAN Alliance hereby grants to Adopter a nonexclusive, nontransferable, irrevocable, non-sublicensable, worldwide copyright license to obtain, use and modify O-RAN Specifications, but not to further distribute such O-RAN Specification in any modified or unmodified way, solely in furtherance of implementations of an O-RAN Specification.

2.2 Adopter shall not use O-RAN Specifications except as expressly set forth in this Agreement or in a separate written agreement with O-RAN Alliance.

SECTION 3: FRAND LICENSE

3.1 Members, Contributors and Academic Contributors and their Affiliates are prepared to grant based on a separate Patent License Agreement to each Adopter under Fair, Reasonable And Non-Discriminatory (FRAND) terms and conditions with or without compensation (royalties) a nonexclusive, non-transferable, irrevocable (but subject to Defensive Suspension), non-sublicensable, worldwide license under their Necessary Claims to make, have made, use, import, offer to sell, lease, sell and otherwise distribute Compliant Portions; provided, however, that such license shall not extend: (a) to any part or function of a product in which a Compliant Portion is incorporated that is not itself part of the Compliant Portion; or (b) to any Adopter if that Adopter is not making a reciprocal grant to Members, Contributors and Academic Contributors, as set forth in Section 3.3. For the avoidance of doubt, the foregoing license includes the distribution by the Adopter's distributors and the use by the Adopter's customers of such licensed Compliant Portions.

3.2 Notwithstanding the above, if any Member, Contributor or Academic Contributor, Adopter or their Affiliates has reserved the right to charge a FRAND royalty or other fee for its license of Necessary Claims to Adopter, then Adopter is entitled to charge a FRAND royalty or other fee to such Member, Contributor or Academic Contributor, Adopter and its Affiliates for its license of Necessary Claims to its licensees.

3.3 Adopter, on behalf of itself and its Affiliates, shall be prepared to grant based on a separate Patent License Agreement to each Members, Contributors, Academic Contributors, Adopters and their Affiliates under FRAND terms and conditions with or without compensation (royalties) a nonexclusive, non-transferable, irrevocable (but subject to Defensive Suspension), non-sublicensable, worldwide license under their Necessary Claims to make, have made, use, import, offer to sell, lease, sell and otherwise distribute Compliant Portions; provided, however, that such license will not extend: (a) to any part or function of a product in which a Compliant Portion is incorporated that is not itself part of the Compliant Portion; or (b) to any Members, Contributors, Academic Contributors, Adopters and their Affiliates that is not making a reciprocal grant to Adopter, as set forth in Section 3.1. For the avoidance of doubt, the foregoing license includes the distribution by the Members', Contributors', Academic Contributors', Adopters' and their Affiliates' distributors and the use by the Members', Contributors', Academic Contributors', Adopters' and their Affiliates' customers of such licensed Compliant Portions.

SECTION 4: TERM AND TERMINATION

4.1 This Agreement shall remain in force, unless early terminated according to this Section 4.

4.2 O-RAN Alliance on behalf of its Members, Contributors and Academic Contributors may terminate this Agreement if Adopter materially breaches this Agreement and does not cure or is not capable of curing such breach within thirty (30) days after being given notice specifying the breach.

4.3 Sections 1, 3, 5 - 11 of this Agreement shall survive any termination of this Agreement. Under surviving Section 3, after termination of this Agreement, Adopter will continue to grant licenses (a) to entities who become Adopters after the date of termination; and (b) for future versions of O-RAN

Specifications that are backwards compatible with the version that was current as of the date of termination.

SECTION 5: CONFIDENTIALITY

Adopter will use the same care and discretion to avoid disclosure, publication, and dissemination of O-RAN Specifications to third parties, as Adopter employs with its own confidential information, but no less than reasonable care. Any disclosure by Adopter to its Affiliates, contractors and consultants should be subject to an obligation of confidentiality at least as restrictive as those contained in this Section. The foregoing obligation shall not apply to any information which is: (1) rightfully known by Adopter without any limitation on use or disclosure prior to disclosure; (2) publicly available through no fault of Adopter; (3) rightfully received without a duty of confidentiality; (4) disclosed by O-RAN Alliance or a Member, Contributor or Academic Contributor to a third party without a duty of confidentiality on such third party; (5) independently developed by Adopter; (6) disclosed pursuant to the order of a court or other authorized governmental body, or as required by law, provided that Adopter provides reasonable prior written notice to O-RAN Alliance, and cooperates with O-RAN Alliance and/or the applicable Member, Contributor or Academic Contributor to have the opportunity to oppose any such order; or (7) disclosed by Adopter with O-RAN Alliance's prior written approval.

SECTION 6: INDEMNIFICATION

Adopter shall indemnify, defend, and hold harmless the O-RAN Alliance, its Members, Contributors or Academic Contributors, and their employees, and agents and their respective successors, heirs and assigns (the "Indemnitees"), against any liability, damage, loss, or expense (including reasonable attorneys' fees and expenses) incurred by or imposed upon any of the Indemnitees in connection with any claims, suits, investigations, actions, demands or judgments arising out of Adopter's use of the licensed O-RAN Specifications or Adopter's commercialization of products that comply with O-RAN Specifications.

SECTION 7: LIMITATIONS ON LIABILITY; NO WARRANTY

EXCEPT FOR BREACH OF CONFIDENTIALITY, ADOPTER'S BREACH OF SECTION 3, AND ADOPTER'S INDEMNIFICATION OBLIGATIONS, IN NO EVENT SHALL ANY PARTY BE LIABLE TO ANY OTHER PARTY OR THIRD PARTY FOR ANY INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE OR CONSEQUENTIAL DAMAGES RESULTING FROM ITS PERFORMANCE OR NON-PERFORMANCE UNDER THIS AGREEMENT, IN EACH CASE WHETHER UNDER CONTRACT, TORT, WARRANTY, OR OTHERWISE, AND WHETHER OR NOT SUCH PARTY HAD ADVANCE NOTICE OF THE POSSIBILITY OF SUCH DAMAGES.

O-RAN SPECIFICATIONS ARE PROVIDED "AS IS" WITH NO WARRANTIES OR CONDITIONS WHATSOEVER, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHERWISE. THE O-RAN ALLIANCE AND THE MEMBERS, CONTRIBUTORS OR ACADEMIC CONTRIBUTORS EXPRESSLY DISCLAIM ANY WARRANTY OR CONDITION OF MERCHANTABILITY, SECURITY, SATISFACTORY QUALITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, ERROR-FREE OPERATION, OR ANY WARRANTY OR CONDITION FOR O-RAN SPECIFICATIONS.

SECTION 8: ASSIGNMENT

Adopter may not assign the Agreement or any of its rights or obligations under this Agreement or make any grants or other sublicenses to this Agreement, except as expressly authorized hereunder, without having first received the prior, written consent of the O-RAN Alliance, which consent may be withheld in O-RAN Alliance's sole discretion. O-RAN Alliance may freely assign this Agreement.

SECTION 9: THIRD-PARTY BENEFICIARY RIGHTS

Adopter acknowledges and agrees that Members, Contributors and Academic Contributors (including future Members, Contributors and Academic Contributors) are entitled to rights as a third-party beneficiary under this Agreement, including as licensees under Section 3.

SECTION 10: BINDING ON AFFILIATES

Execution of this Agreement by Adopter in its capacity as a legal entity or association constitutes that legal entity's or association's agreement that its Affiliates are likewise bound to the obligations that are applicable to Adopter hereunder and are also entitled to the benefits of the rights of Adopter hereunder.

SECTION 11: GENERAL

This Agreement is governed by the laws of Germany without regard to its conflict or choice of law provisions.

This Agreement constitutes the entire agreement between the parties as to its express subject matter and expressly supersedes and replaces any prior or contemporaneous agreements between the parties, whether written or oral, relating to the subject matter of this Agreement.

Adopter, on behalf of itself and its Affiliates, agrees to comply at all times with all applicable laws, rules and regulations with respect to its and its Affiliates' performance under this Agreement, including without limitation, export control and antitrust laws. Without limiting the generality of the foregoing, Adopter acknowledges that this Agreement prohibits any communication that would violate the antitrust laws.

By execution hereof, no form of any partnership, joint venture or other special relationship is created between Adopter, or O-RAN Alliance or its Members, Contributors or Academic Contributors. Except as expressly set forth in this Agreement, no party is authorized to make any commitment on behalf of Adopter, or O-RAN Alliance or its Members, Contributors or Academic Contributors.

In the event that any provision of this Agreement conflicts with governing law or if any provision is held to be null, void or otherwise ineffective or invalid by a court of competent jurisdiction, (i) such provisions will be deemed stricken from the contract, and (ii) the remaining terms, provisions, covenants and restrictions of this Agreement will remain in full force and effect.

Any failure by a party or third party beneficiary to insist upon or enforce performance by another party of any of the provisions of this Agreement or to exercise any rights or remedies under this Agreement or otherwise by law shall not be construed as a waiver or relinquishment to any extent of the other parties' or third party beneficiary's right to assert or rely upon any such provision, right or remedy in that or any other instance; rather the same shall be and remain in full force and effect.