



# 오픈랜 가상화/지능화 및 AI-RAN 기술 소개

2025. 5. 30.

지능형기지국SW연구실 나지현

# CONTENTS

PART Ⅰ 배경 및 개념

PART Ⅱ 오픈랜 가상화/지능화

PART Ⅲ AI-RAN 기술동향



Chapter

# I

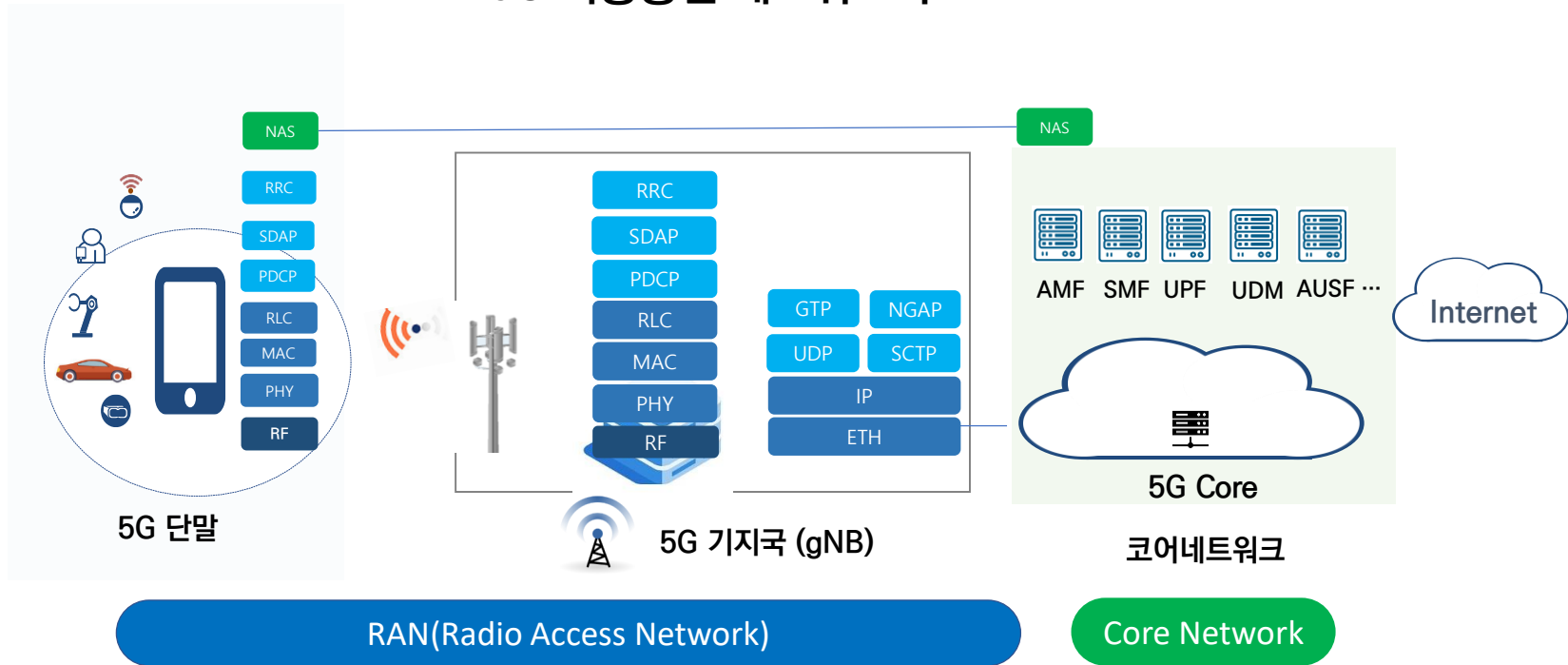
오픈랜 가상화/지능화 및 AI-RAN

## 배경 및 개념



## RAN (Radio Access Network) 이란?

### 5G 이동통신 네트워크 구조

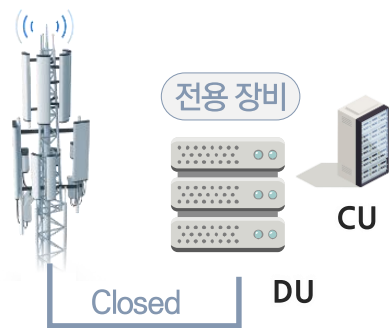




## 가상화/지능화 관점에서 RAN (Radio Access Network) 분류

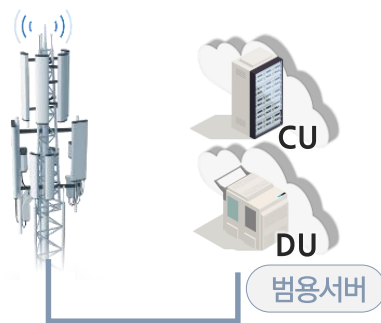
### Legacy RAN

특정구간 기업 전용 인터페이스  
(HW중심, 폐쇄적)



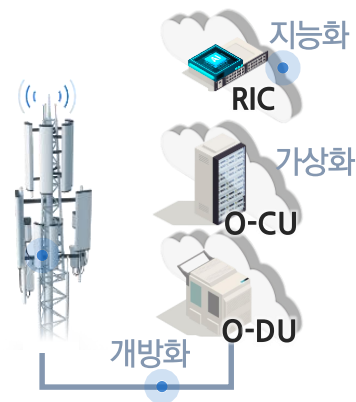
### vRAN

일반 서버 기반 기지국  
(SW중심, 가상화)



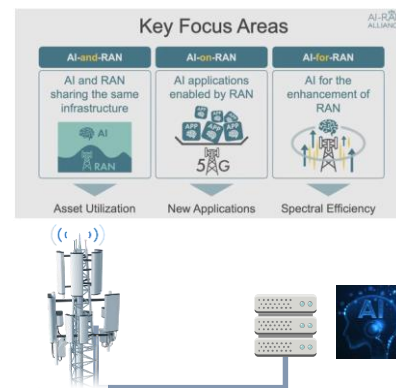
### O-RAN

상호연동이 가능한 개방형 기지국  
(SW중심, 개방화/가상화/지능화)

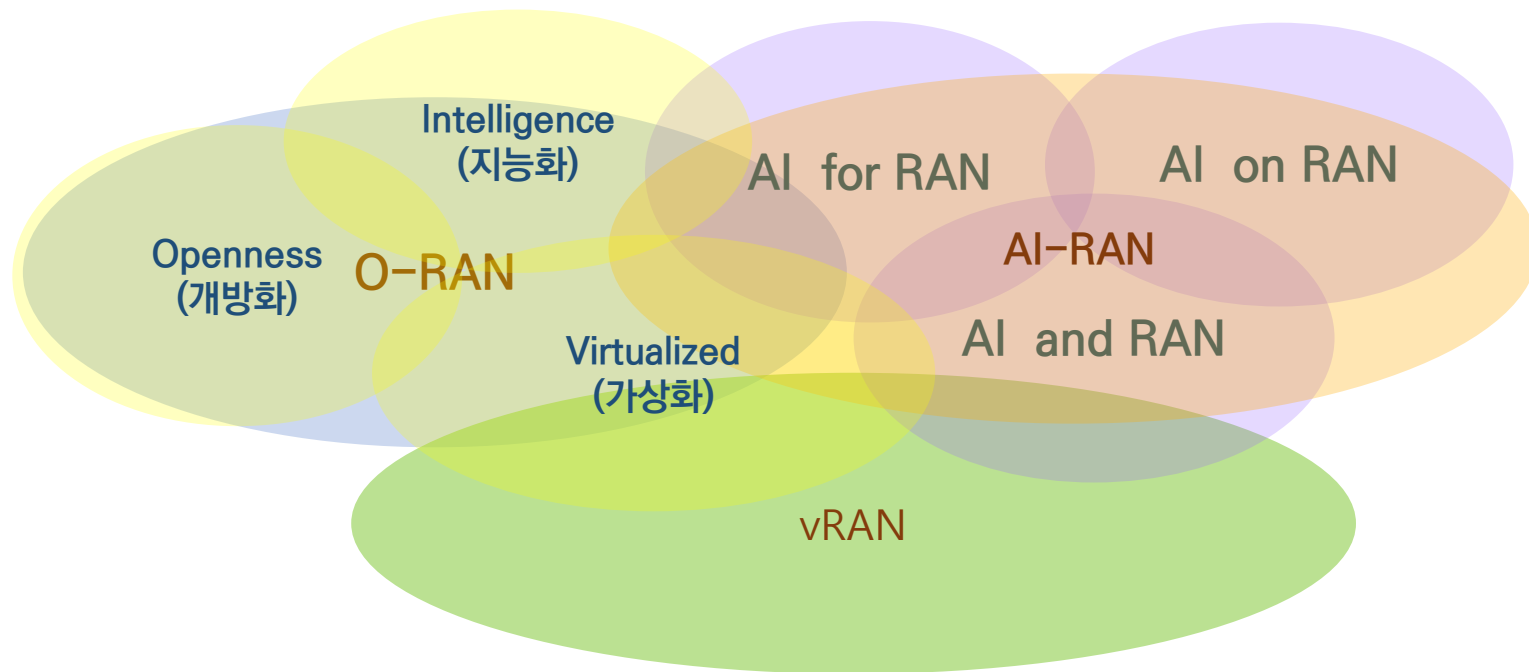


### AI-RAN

RAN 최적화를 위한 AI,  
AI와 RAN의 Infra 공유,  
RAN에 의한 새로운 AI 응용 제공



## vRAN, O-RAN, AI-RAN의 관계는?



\* 본 그림은 개인적인 견해임

Chapter

# II

H ORIA 가상화/지능화 위원회

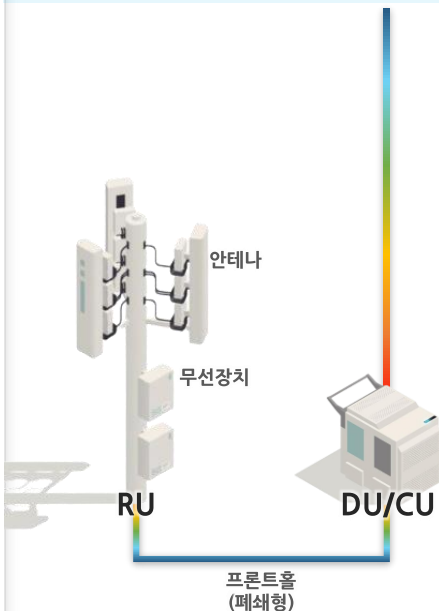
## 오픈랜 가상화/지능화



## 오픈랜(Open RAN, O-RAN) 이란?

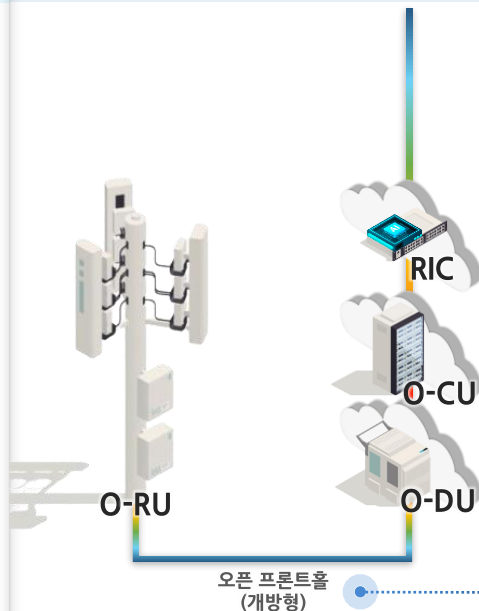
### 기존 기지국

특정구간 기업 전용 인터페이스  
(HW중심, 폐쇄적)



### 오픈랜 기지국

상호연동이 가능한 개방형 인터페이스  
(SW중심, 개방화/가상화/지능화)



### O-RAN Alliance 규격을 따르는 기지국

#### 오픈랜은 vRAN을 포괄하는 개념

※ vRAN : 네트워크 기능 일부 또는 전부가 클라우드 플랫폼의 SW로 구현된 RAN 환경을 의미

#### 특징

연결성	확장성	통합성	신뢰성
효율성	편의성	유연성	안정성

#### 자능화

지능형 컨트롤러  
(RAN Intelligent Controller)

- 인공지능 기반 무선자원 제어/관리
- Near-RT와 비 실시간으로 구분

#### 가상화

vRAN (virtualized RAN)

- HW 모뎀 → SW 모뎀
- RAN 기술의 SW 화

#### 개방화

RAN 노드간 인터페이스 개방

- 오픈 프론트 홀 (RU와 DU간 개방)
- 지능화제어, 가상화를 위한 개방화

# O-RAN Alliance ?

## 조직 구성

O-RAN ALLIANCE's Governing Bodies

총회, 이사회, 집행위, 기술운영위

O-RAN 테크니컬 워킹 그룹(11개)

O-RAN 포커스&리서치 그룹(6개)



## 생태계 지원 프로그램

O-RAN ECOSYSTEM SUPPORT PROGRAM

플러그앤피스트(상호운용성)

개방형 테스트 및 통합센터(시험인증)

온라인 가상 전시관(프로모션)

## 운영자 멤버

(LGU+, KT, SKT 참여)

32개  
통신사



## 기여자 및 학술기여자

(SAMSUNG, SOLID, FRTEK, HFR, Innowireless, ETRI, TTA, KTL, 단국대 등 참여)

282개  
기여자

267개  
기여자

235개  
기여자



# O-RU와 O-DU간 프론트홀 표준화, 가상화/지능화 관련 표준을 지정

WG1 Use Cases and Overall Architecture

WG2 Non-real-time RIC and A1 Interface

WG3 Near-real-time RIC and E2 Interface

WG4 Open Fronthaul Interface

WG5 Open F1/W1/E1/X2/Xn Interface

WG6 Cloudification and Orchestration

WG7 White-Box Hardware

WG8 Stack Reference Design

WG9 Open X-haul Transport

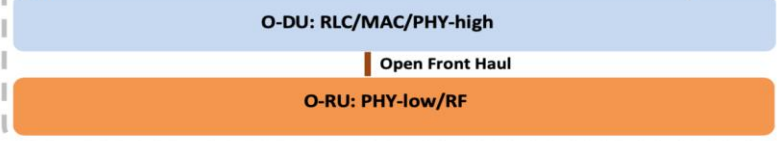
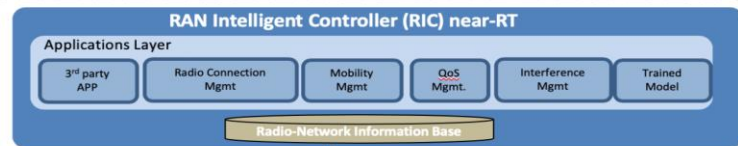
WG10 OAM for O-RAN

WG11 Security

지능화

가상화

## O-RAN Alliance WG 구성



## WG6. Cloudification &amp; Orchestration

## WG6 Cloudification and Orchestration

## WG6: Cloudification and Orchestration Workgroup

Title	Publication date	Document type	Release	Feature package	Download
<b>O-RAN Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN 8.01</b> O-RAN.WG6.CADS-v08.01	February 2025	Technical Report	R004		<a href="#">Download</a>
This document introduces and examines different scenarios and use cases for O-RAN deployments of Network Functionality into Cloud Platforms, O-RAN Cloudification...					
<b>O-RAN O-Cloud Interface Conformance Test Specification 4.01</b> O-RAN.WG6.TS.O-CLOUD-INTF-CONF-R004-v04.01	February 2025	Technical Specification	R004		<a href="#">Download</a>
This document specifies a test specification for test and validation of the O-Cloud interfaces/APIs. This version includes refinements to the test cases for the O2 D...					
<b>O-RAN O2 Interface General Aspects and Principles 8.01</b> O-RAN.WG6.TS.O2-GA&P-R004-v08.01	February 2025	Technical Specification	R004		<a href="#">Download</a>
This specification defines O-RAN O-Cloud functions and protocols for the O-RAN O2 Interface. This version clarifies the network to which O-Cloud connects.					
<b>O-RAN O-Cloud Information Model 3.0</b> O-RAN.WG6.TS.O-CLOUD-IM.O-R004-v03.00	February 2025	Technical Specification	R004		<a href="#">Download</a>
The O-Cloud Information Model provides the logical model of information elements and their relationships. This release introduces additional namespaces which a...					
<b>O-RAN O-Cloud Interoperability Test (IOT) Specification 2.0</b> O-RAN.WG6.O-CLOUD-IOT-R004-v02.00	February 2025	Technical Specification	R004		<a href="#">Download</a>
This document provides O-Cloud IoT specification, and includes definition of the interoperability testing methodology and a set of tests for the O-Cloud Notification...					
<b>O-RAN O2ims Interface Specification 8.0</b> O-RAN.WG6.TS.O2IMS-INTERFACE-R004-v08.00	February 2025	Technical Specification	R004		<a href="#">Download</a>
This specification defines O-RAN O-Cloud IMS interface functions and protocols for the O-RAN O2 Interface. This release introduces support for additional FM an...					
<b>O-RAN Acceleration Abstraction Layer General Aspects and Principles 11.0</b> O-RAN.WG6.AAL-GA&P-R004-v11.00	February 2025	Technical Specification	R004		<a href="#">Download</a>
This specification defines O-RAN O-Cloud hardware accelerator interface functions and protocols for the O-RAN AAL interface. This release introduces the declar...					
<b>O-RAN Acceleration Abstraction Layer High-PHY Profiles 7.0</b> O-RAN.WG6.AAL-HI-PHY-R004-v07.00	February 2025	Technical Specification	R004		<a href="#">Download</a>
This document specifies how to accelerate the High-PHY functionality of an O-DU where the High-PHY Acceleration Function (AF) is hosted on an AAL-LPU that is ...					
<b>O-RAN Cloudification and Orchestration Use Cases and Requirements for O-RAN Virtualized RAN 12.0</b> O-RAN.WG6.ORCH-USE-CASES-R004-v12.00	February 2025	Technical Specification	R004		<a href="#">Download</a>

<https://specifications.o-ran.org/specifications>

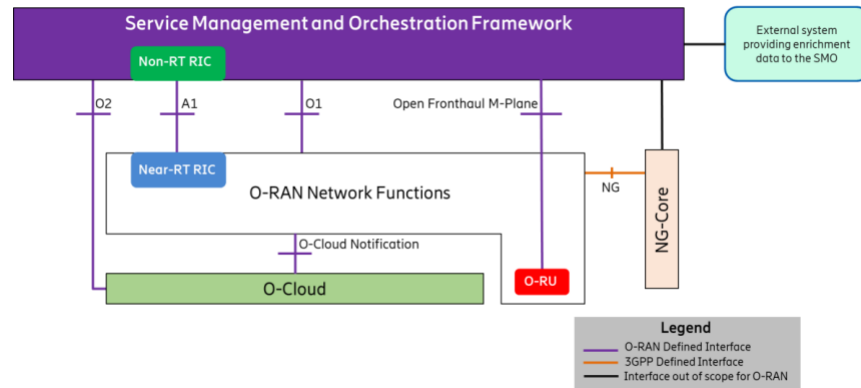
## WG6

1. O2 Interface

2. Cloud Platform Reference Design

3. O-Cloud 구조 및 Deployment 시나리오, Conformance test 등

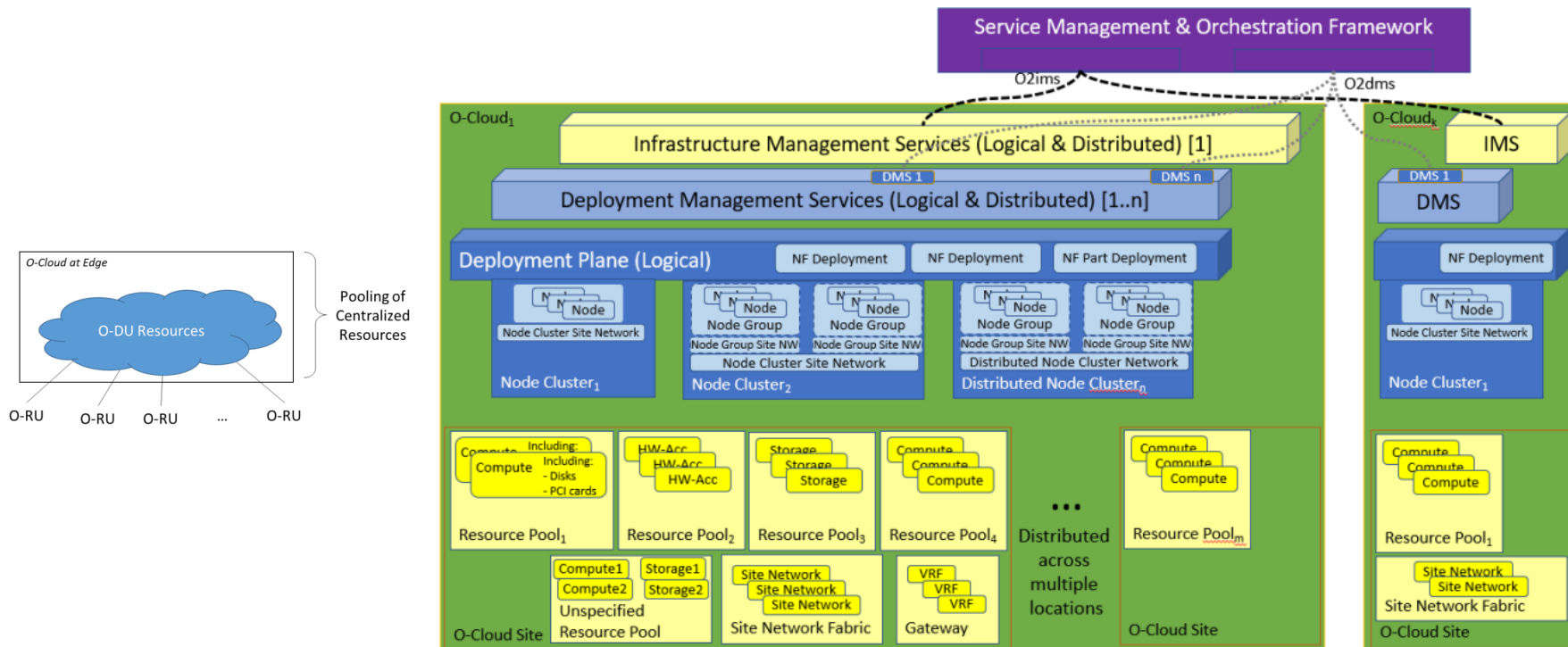
4. O-RAN Acceleration Abstract Layer



출처: O-RAN Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN 8.01

## WG6. Cloudification & Orchestration

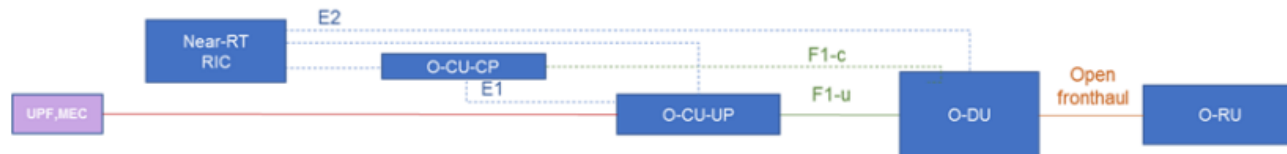
### ● O-Cloud의 key Concept





## WG6. Cloudification &amp; Orchestration

## ● RAN 기능과 Cloud, HW 요구사항 간의 관계



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					✓

## 오픈랜 장비 구성 ( 모뎀 기능을 위한 HW Accelerator)

### Look-Aside 가속화 방식

#### ● Look-Aside Features

- 선택된 기능만 가속기로 전송
- 가속화 작업동안 CPU는 다른 유용한 작업 처리 가능
- CPU와 가속기 사이에 대용량 데이터 전송 필요

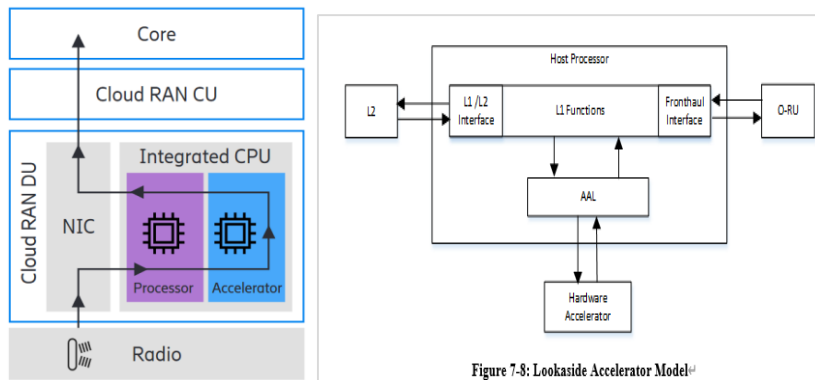


Figure 7-8: Lookaside Accelerator Model<sup>[1]</sup>

### Inline 가속화 방식

#### ● Inline Features

- 데이터 흐름 및 기능 전체가 가속기로 전송
- COTS 서버에 간단한 인터페이스 제공
- 가속기에서 Layer 1(물리계층) 기능 모두 처리

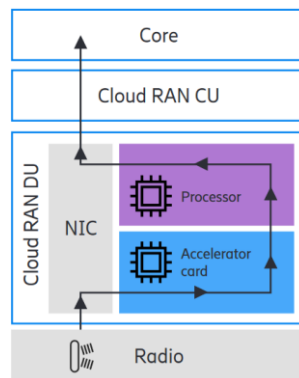


Figure 7-9: Inline Accelerator Model<sup>[1]</sup>

## vRAN HW Accelerator Card

### vRAN HW Accelerator Card 제조사



Genevisio PAC-010 DU Accelerator Card

Inline L1/DU Acceleration

Layerscape LX2160  
(16x Arm A72 Cores)



NVIDIA Converged A30x/A100x Accelerator Card

Inline L1/DU Acceleration

BlueField-2 DPU  
(8x Arm A72 Cores)



Dell Open RAN Accelerator Card

Inline L1 Acceleration

Octeon Fusion 95xx  
(6x Arm v8.2 Cores)



Qualcomm 5G X100 Accelerator Card

Inline L1 Acceleration

Distributed Unit Platform  
(4x A55 Arm Cores)



EdgeQ M Series Acceleration Card

Inline L1 Acceleration

EdgeQ S Series  
(9x E1 Arm Cores)



Xilinx T2 Accelerator Card

Lookaside Acceleration

Zynq UltraScale+ RFSoc  
(4x A53 Arm Cores)



Lookaside Acceleration

# 삼성의 Virtualized RAN 3.0 (1)

## Flexible Deployment and Dynamic Scaling

- ✓ vDU allows an optimal deployment of a network by dynamically allocating resources to various site configurations and traffic demand

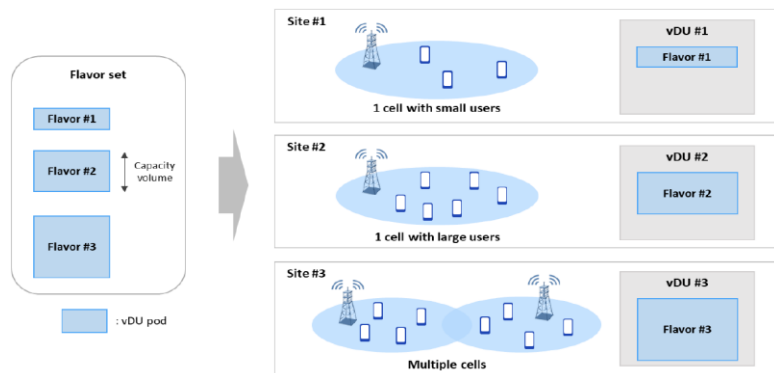


Figure 8. Flexible deployment with multiple flavors

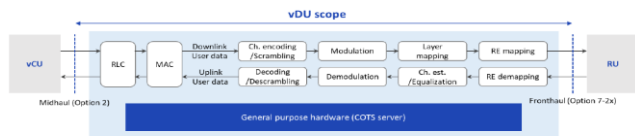


Figure 3. Function split between CU and DU

## Flexible Deployment and Dynamic Scaling

- ✓ **Dynamic scaling** enables flexible management of vDU resources and also enables pooling to efficiently cope with the following network changes

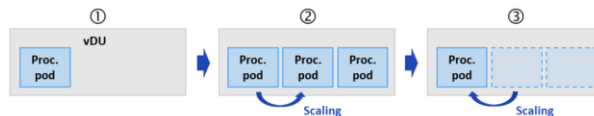
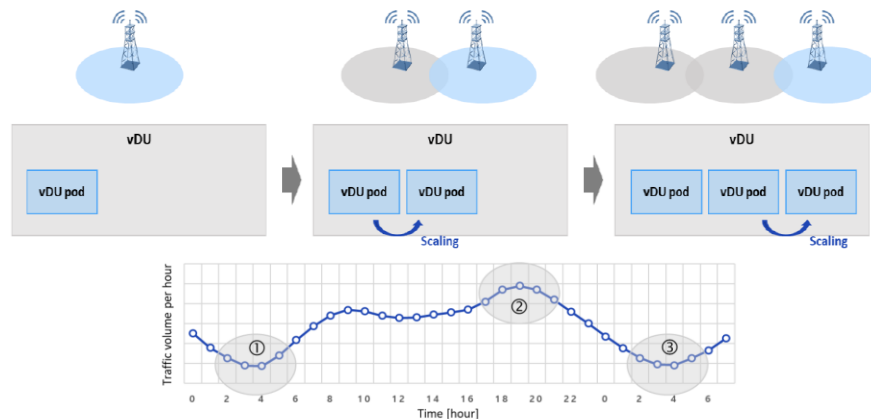


Figure 10. Dynamic scaling with traffic change

## 삼성의 Virtualized RAN 3.0 (2)

### Efficient Resource Utilization via Pooling

- ✓ **vDU pooling** enables single vDU to support multiple cell-site baseband processing by sharing its baseband processing resources within a baseband cloud and allowing other cell sites and radio technologies to use its resources

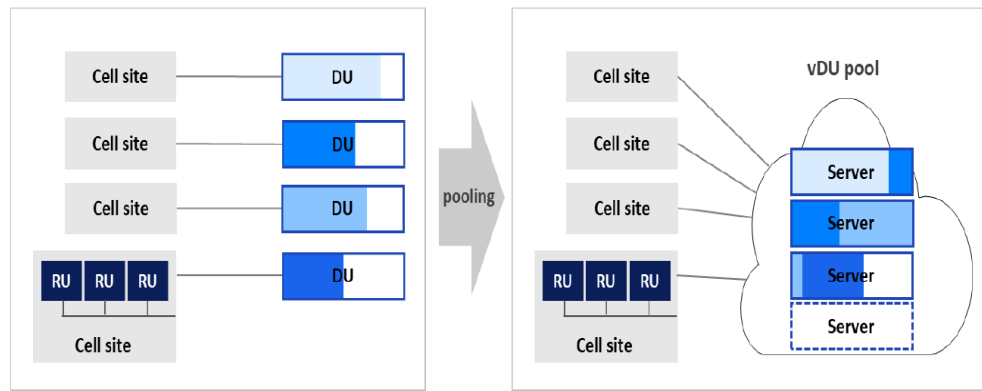


Figure 11. Resource efficiency via vDU pooling

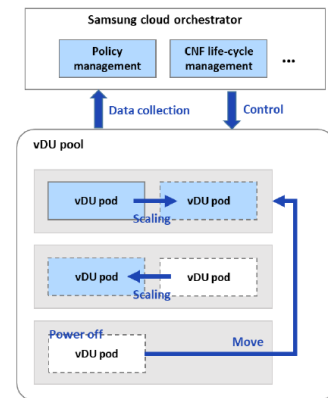


Figure 13. Evolution of pooling with Samsung Cloud Orchestrator

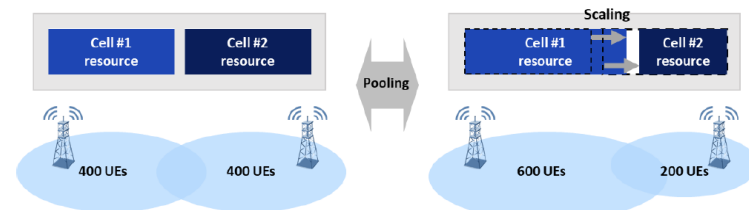


Figure 12. Resource pooling when load imbalance occurs

## WG2. Non-real-time RIC and A1 Interface

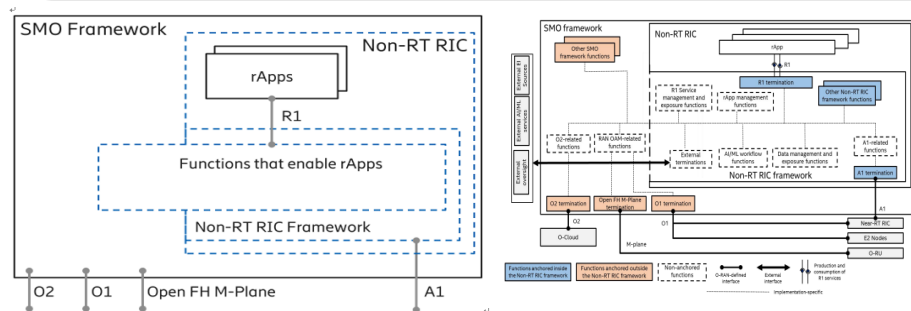
## WG2 Non-real-time RIC and A1 Interface

Title	Date	Document type	Release	Work item
<b>O-RAN A1 Interface: General Aspects and Principles 5.0</b> O-RAN.WG2.TS.A1GAP-R004-v05.00	February 2025	Technical Specification	R004	
This document specifies the general aspects and principles of the A1 Interface. This version brings updated A1 service architecture, and A1 ML functions and relate...				
<b>O-RAN A1 Interface: Transport Protocol 3.04</b> O-RAN.WG2.TS.A1TP-R004-v03.04	February 2025	Technical Specification	R004	
This document describes the transport protocol of the O-RAN A1 Interface. This version brings Editorial enhancement of references.				
<b>O-RAN A1 Interface: Application Protocol 4.04</b> O-RAN.WG2.TS.A1AP-R004-v04.04	February 2025	Technical Specification	R004	
This document specifies the application protocol of the A1 Interface. This version brings an update to the general referencing and A1 service architecture.				
<b>O-RAN A1 Interface: Test Specification 4.03</b> O-RAN.WG2.TS.A1TS-R004-v04.03	February 2025	Technical Specification	R004	
This document specifies test cases for conformance testing and interoperability testing of the Non-RT RIC and the Near-RT RIC over the A1 Interface. This version...				
<b>O-RAN R1 Interface: General Aspects and Principles 10.0</b> O-RAN.WG2.TS.R1GAP-R004-v10.00	February 2025	Technical Specification	R004	
The O-RAN R1GAP Specification summarizes the R1 Interface specification objectives and specifies the principles and procedures related to the O-RAN R1 interfa...				
<b>O-RAN R1 Interface: Type Definitions for R1 Services 4.0</b> O-RAN.WG2.TS.R1TD-R004-v04.00	February 2025	Technical Specification	R004	
This document specifies the Type Definitions for R1 Services. It is part of a TS-family covering the R1 Interface specifications. This release includes the new DME ty...				
<b>O-RAN R1 Interface: Test Specification 2.0</b> O-RAN.WG2.TS.R1TS-R004-v02.00	February 2025	Technical Specification	R004	
This document specifies test cases for conformance testing and interoperability testing of the rApps and R1 services over R1 Interface. This release includes the co...				
<b>O-RAN R1 Interface: Use Cases and Requirements 9.0</b> O-RAN.WG2.TS.R1UCR-R004-v09.00	February 2025	Technical Specification	R004	
This document describes use cases and requirements for the O-RAN R1 Interface. This version adds use cases for AIML inference capability information Query.				
<b>O-RAN R1 Interface: Application Protocols for R1 Services 7.0</b> O-RAN.WG2.TS.R1AP-R004-v07.00	February 2025	Technical Specification	R004	
This document contains a realization for the procedures identified in O-RAN R1 Interface: General Aspects and Principles. This version updated the specification by...				
<b>O-RAN A1 Interface: Type Definitions 10.0</b> O-RAN.WG2.TS.A1TD-R004-v10.00	February 2025	Technical Specification	R004	
This specification defines the data types for A1 Policies and A1 Enrichment Information in a reusable and extensible way. It allows new policy types to be created in...				
<b>O-RAN A1 Interface: Use Cases and Requirements 2.0</b> O-RAN.WG2.TS.A1UCR-R004-v02.00	February 2025	Technical Specification	R004	
This document describes use cases for the O-RAN A1 Interface. This version brings Use cases for Authorization of service access requests and AI/ML model training.				

<https://specifications.o-ran.org/specifications>

## WG2 Non-real-time RIC and A1 interface

1. Non-RT RIC 요구사항 및 Use Case, 구조
2. A1 인터페이스 (Type, 일반, Transport, Test spec. , Application Protocol)
3. rAPP을 위한 R1 인터페이스
4. AI/ML Work Flow
5. Non-RT RIC 기능 구조



## WG3. Near-real-time RIC and E2 Interface

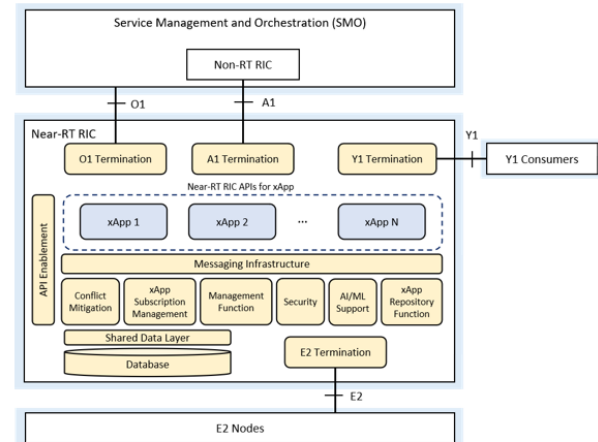
## WG3 Near RT RIC and E2 Interface

## WG3: Near-real-time RIC and E2 Interface Workgroup

Title	Publication date	Document type	Release	Feature package	Download
<b>O-RAN E2 General Aspects and Principles (E2GAP) 7.0</b> O-RAN.WG3.TS.E2GAP-R004-v07.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
E2 General Aspects and Principles (E2GAP) provides a stage 2 description of the E2 Interface between Near-RT RIC and an E2 Node. This version includes support f...					
<b>O-RAN E2 Application Protocol (E2AP) 7.0</b> O-RAN.WG3.TS.E2AP-R004-v07.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
E2 Application Protocol (E2AP) provides a stage 3 description of the E2 Interface between Near-RT RIC and an E2 Node. This version includes support for new feat...					
<b>O-RAN E2 Service Model (E2SM) 7.0</b> O-RAN.WG3.TS.E2SM-R004-v07.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
E2 Service Model (E2SM) provides describes the O-RAN specified RAN Function-specific Service Models supported over E2 (E2SM) and specifies the common elem...					
<b>O-RAN E2 Service Model (E2SM), Lower Layers Control 1.0</b> O-RAN.WG3.TS.E2SM-LLC-R004-v01.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
This document specifies the capabilities exposed over E2 Interface to enable efficient control of lower layers of the RAN, including collecting channel and traffic info...					
<b>O-RAN E2 Service Model (E2SM) KPM 6.0</b> O-RAN.WG3.TS.E2SM-KPM-R004-v06.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
E2 Service Model KPM (E2SM-KPM) provides RAN Function-specific Service Models of KPM (Key Performance Measurement) for use in xApps. This version of E2...					
<b>O-RAN E2 Service Model (E2SM) Cell Configuration and Control 5.0</b> O-RAN.WG3.TS.E2SM-CCC-R004-v05.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
This document specifies the E2 Service Model (E2SM) to enable exposure of services for node and cell level configuration and control over the E2 Interface. This ver...					
<b>O-RAN E2 Service Model (E2SM), RAN Control 7.0</b> O-RAN.WG3.TS.E2SM-RC-R004-v07.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
This document specifies the E2 Service Model (E2SM) for UE and Cell level services over E2 Interface. This version of E2SM-RC includes support for Massive MIM...					
<b>O-RAN Near-RT RIC Architecture 7.0</b> O-RAN.WG3.TS.RICARCH-R004-v07.00	February 2025	Technical Specification	R004		<a href="#">DOWNLOAD</a>
The O-RAN Near-RT RIC Architecture document specifies the Near-RT RIC internal architecture and functionalities, and the stage 2 definitions of Near-RT RIC API...					

## WG3

1. E2: 서비스 모델, 일반, E2AP, E2SM, E2SM KPM
2. Near-RT RIC 구조, Use Case and Requirement, API
3. Y1 인터페이스 (일반, Use Case)
4. Cell Configuration and Control
5. O1 Interface



## O-RAN Use Cases (1)

## O-RAN Use Case

RIC을 통한 AI for RAN에 Focus

	Use Case
1	Context-Based Dynamic HO Management for V2X
2	Flight Path Based Dynamic UAV Radio Resource Allocation
3	Radio Resource Allocation for UAV Application Scenario
4	QoE Optimization
5	Traffic Steering
6	Massive MIMO Beamforming Optimization
7	RAN Sharing
8	QoS Based Resource Optimization
9	RAN Slice SLA Assurance
10	Multi-vendor Slices
11	Dynamic Spectrum Sharing
12	NSSI Resource Allocation Optimization
13	Local Indoor Positioning in RAN

	Use Case
14	Massive MIMO SU/MU-MIMO Grouping Optimization
15	O-RAN Signalling Storm Protection
16	Congestion Prediction & Management
17	Industrial IoT Optimization
18	BBU Pooling to achieve RAN Elasticity
19	Integrated SON Function
20	Shared O-RU
21	Energy Saving
22	MU-MIMO Optimization
23	Sharing Non-RT RIC Data with the Core
24	Industrial vision SLA Assurance
25	Non-Public Network (NPN) RAN-Sharing via Midhaul for Multi-Operator Coverage
26	Interference Detection and Optimization



## O-RAN Use Cases (2)

## O-RAN Use Case

RIC을 통한 AI for RAN에 Focus

분야	Use cases (WG1 spec.)	관련 세부 기술	추가 필요 기능	성능 지표
MIMO	Use case 6: Massive MIMO Beamforming Optimization	Beamforming control	Massive MIMO	User throughput
MIMO	Use case 14: Massive MIMO SU/MU-MIMO Grouping Optimization	Beamforming control	SU-MIMO/MU-MIMO	User throughput
MIMO	Use case 22: MU-MIMO Optimization	Beamforming control	MU-MIMO	User throughput
<b>Traffic Steering</b>	Use case 5: Traffic Steering	Resource management	Multi-RAT, Dual connectivity	User experience
<b>QoS</b>	Use case 8: QoS Based Resource Optimization	Resource management	Emergency service, Slice	User experience
<b>QoE</b>	Use case 4: QoE Optimization	Resource management		User experience
SON	Use case 19: Integrated SON Function	Resource management	PCI allocation	Cell throughput
Energy Saving	Use case 21: Energy Saving	Resource management	Traffic steering, SON	Energy consumption
RAN Sharing	Use case 7: RAN Sharing	O-CU/DU sharing	Virtual network functions (VNF)	Multi-vendor operation
Sharing	Use case 20: Shared O-RU	O-RU sharing	Multi O-DU	Multi-operator operation
Slice	Use case 9: RAN Slice SLA Assurance	Service Level Agreement (SLA)	Network slicing	RAN slice performance
Slice	Use case 10: Multi-vendor Slices	Sharing	O-RU sharing	Multi-vendor operation
Slice	Use case 12: NSSI Resource Allocation Optimization	Resource management	Slice	User throughput
Spectrum Sharing	Use case 11: Dynamic Spectrum Sharing (DSS)	Low band/high band sharing	4G, 5G	Cell throughput
Congestion	Use case 16: Congestion Prediction & Management	Cell load balancing		Cell throughput
Security	Use case 15: O-RAN Signalling Storm Protection	DDoS attack	Abnormal activity detection	System security
IIoT	Use case 17: Industrial IoT Optimization	IIoT reliability	PDCCP duplication	Reliability
V2X	Use case 1: Context-Based Dynamic HO Management for V2X	Handover	V2X Application Server	Handover performance
BBU management	Use case 18: BBU Pooling to achieve RAN Elasticity	Cloud architecture		Resource utilization
Positioning	Use case 13: Local Indoor Positioning in RAN	Location Server	Positioning protocol	Positioning accuracy
UAV	Use case 2: Flight Path Based Dynamic UAV Radio Resource Allocation	Resource management	UTM (Unmanned Traffic Management)	User throughput
UAV	Use case 3: Radio Resource Allocation for UAV Application Scenario	Resource management	UAV control vehicle	Low latency

## O-RAN Use Cases (3)

### O-RAN Use Case 21 : Energy Saving

#### Background

- 5G Network에서의 에너지 절감이 사업자들에게 매우 중요한 부분
- Traffic Load, User mobility의 다양으로 RAN 에너지 절감은 매우 복잡
- Traffic이 없거나 매우 적을때도 기지국 장비를 켜서 소모되는 에너지량 심각

#### Motivation

- 기존 기지국의 에너지 절감은 수동으로 하거나 SON에 의하여 가능
- ORAN의 AI/ML을 활용한 경우 효율적인 에너지 절감 가능

#### Proposed Solution

- Carrier and Cell Switch Off/On ES
- RF Channel Switch off/On ES
- Advanced Sleep Mode ES

#### Benefits of O-RAN

- AI/ML Assisted 솔루션들을 제공 가능 : Non-RT RIC, Near RT-RIC
- O-RAN Open Interface를 활용하여 Multi-vendor disaggregated RAN 지원

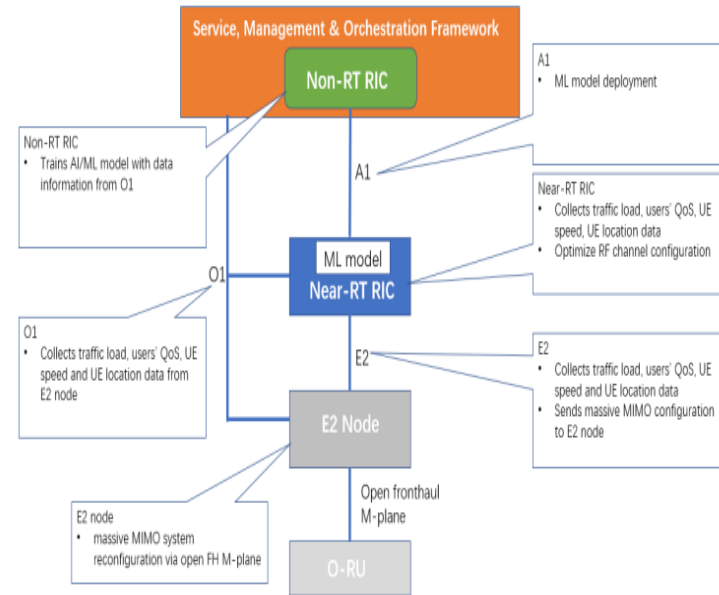


Figure 4.21.3.2-2: RF channel switch off/on ES sub-use case: Near-real-time implementation

## 오픈랜 (Open Radio Access Network) Use Case

## O-RAN Use Case 21 : Energy Saving

## Background

- 5G Network에서의 에너지 절감이 사업자들에게 매우 중요한 부분
- Traffic Load, User mobility의 다양으로 RAN 에너지 절감은 매우 복잡
- Traffic이 없거나 매우 적을때도 기지국 장비를 켜서 소모되는 에너지량 심각

## Motivation

- 기존 기지국의 에너지 절감은 수동으로 하거나 SON에 의하여 가능
- ORAN의 AI/ML을 활용한 경우 효율적인 에너지 절감 가능

## Proposed Solution

- Carrier and Cell Switch Off/On ES
- RF Channel Switch off/On ES
- Advanced Sleep Mode ES

## Benefits of O-RAN

- AI/ML Assisted 솔루션들을 제공 가능 : Non-RT RIC, Near RT-RIC
- O-RAN Open Interface를 활용하여 Multi-vendor disaggregated RAN 지원

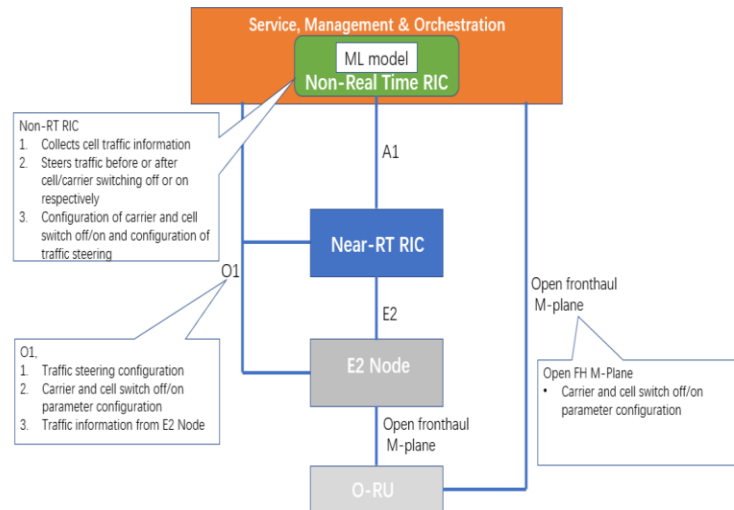


Figure 4.21.3.1-1: Carrier and cell switch off/on ES sub-use case.

Chapter

# III

ORIA 가상화/지능화 위원회

## AI-RAN 기술 동향



## AI-RAN ? "AI 내장하여 RAN에 통합하는 개념으로 AI-RAN Alliance 등장'으로 주목



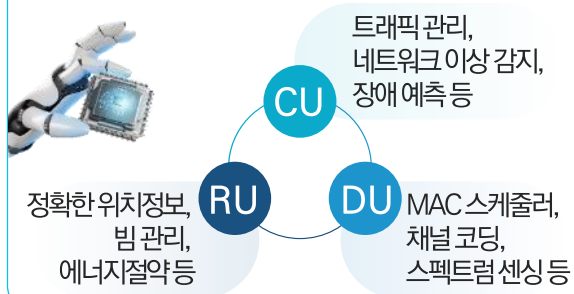
미국 기업을 중심으로 AI-RAN Alliance 출범(2024.02)이후  
회원사 급격히 증가 (현재 73개)

## AI-RAN Alliance WG

### WG 1: AI for RAN

- AI-native air Interface and signaling processing
- Positioning and beam management
- Radio resource management and scheduling
- Energy and spectrum efficiency
- Network Optimization and anomaly detection

#### • AI 기반 RAN 최적화 자동화 기술



### WG2: AI and RAN

- Design architecture and components of multi-tenant system
- Lifecycle management of AI and RAN workloads
- Validating the multitenancy scenarios
- Data center optimization for AI and RAN workloads

#### 동일한 인프라 내 AI와 RAN 융합 기술



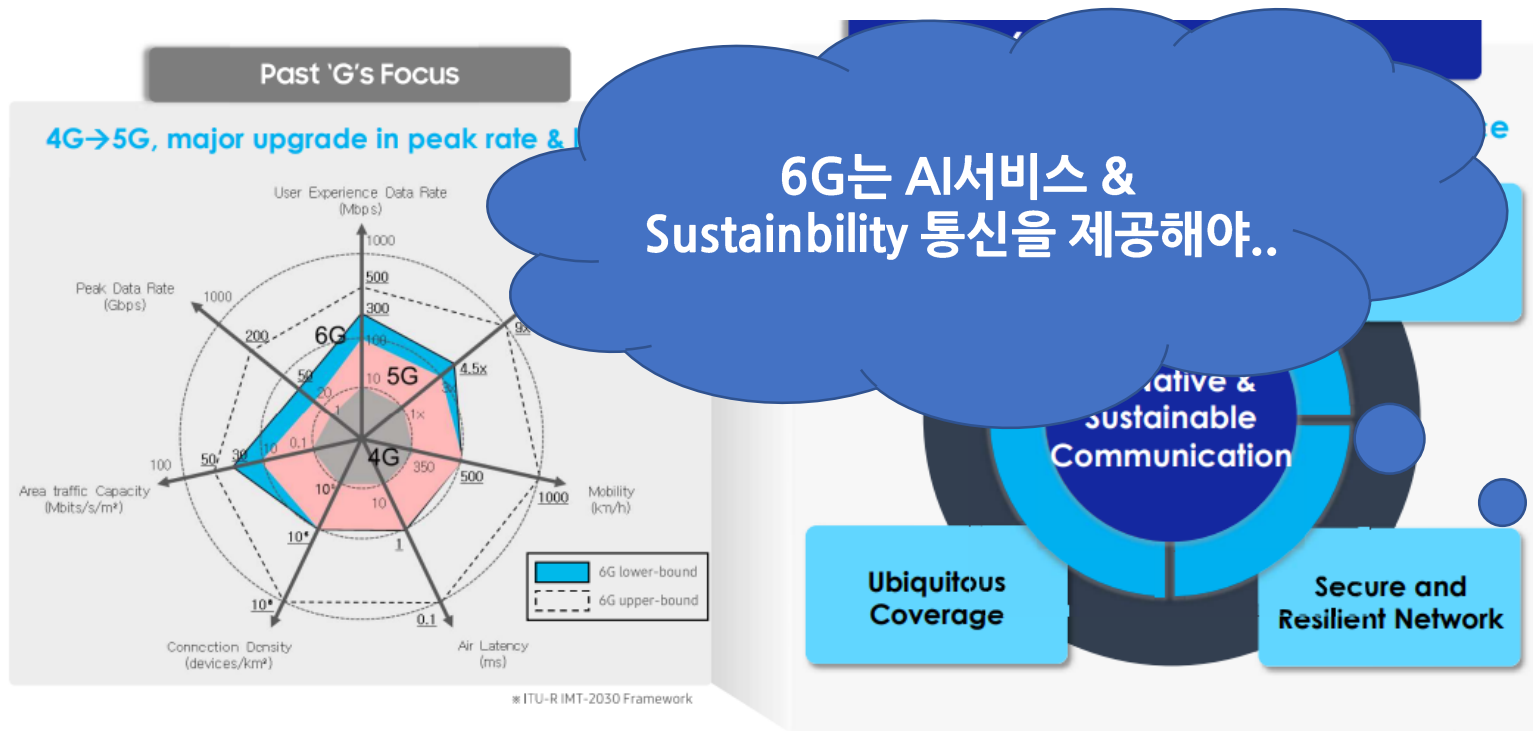
### WG3: AI on RAN

- AI-based multimedia applications
- AI-based security and critical applications,
- AI-based automation and industrial applications
- GenAI/AI-based network service
- Efficient AI/ML model splitting

#### AI Service를 RAN으로 제공하는 기술



## 3GPP 6G Workshop을 통한 AI 를 대비한 R&D 이슈



## 3GPP 6G Workshop을 통한 AI 를 대비한 R&D 이슈

### "Service Plane" for 6G Services

- 6G architecture should be designed to support new characteristics (AI/ML) which are different than traditional telecommunication

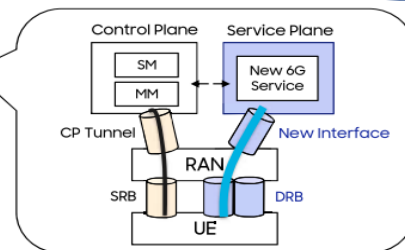
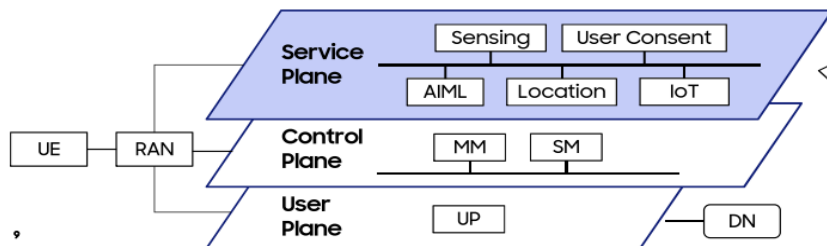
#### Motivation

- New 6G services require large data transmission for which current design of control plane is not suitable
- Make it easy for customers to determine essential features for operations and services
- Keep control plane compact and simple

#### Areas

- NFs for managing
- NFs for various services can be managed in a
- Service Plane is a general framework for types of operator services

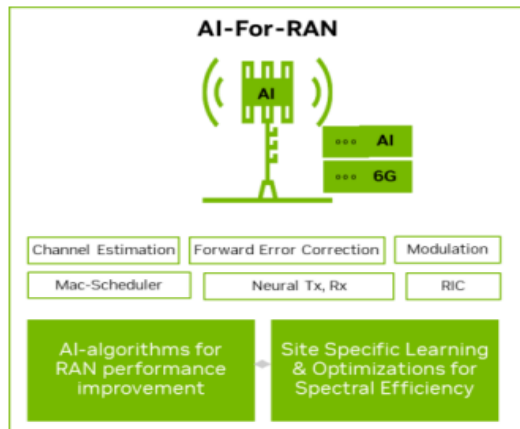
AI 서비스 제공을 위한  
새로운 구조/인터페이스 필요  
(AI on 6G?)..



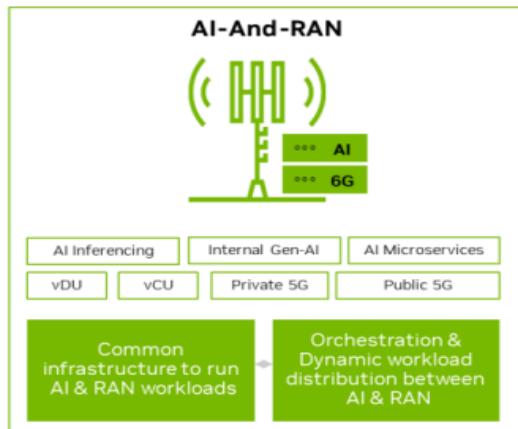


# What is AI-RAN (Artificial Intelligence - Radio Access Network)?

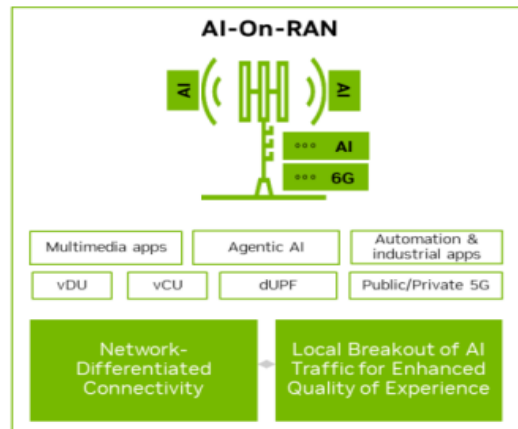
- **(AI와 RAN의 통합)** Technology that enables **full integration of AI in to RAN infrastructure (HW & SW)** to enable **new AI services and monetization opportunities**, in addition to the transformative gains in network utilization, spectral efficiency, and performance.
- **(RAN의 지능화 플랫폼)** Transforming traditional RAN into **intelligent, adaptive, and revenue-generating platforms**



AI for RAN: **advancing RAN capabilities through embedding AI/ML models**, algorithms and neural networks into the radio signal processing layer to improve spectral efficiency, radio coverage, capacity and performance.



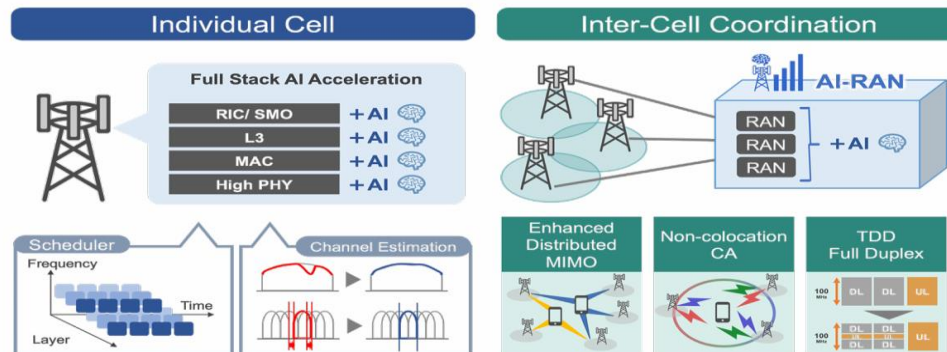
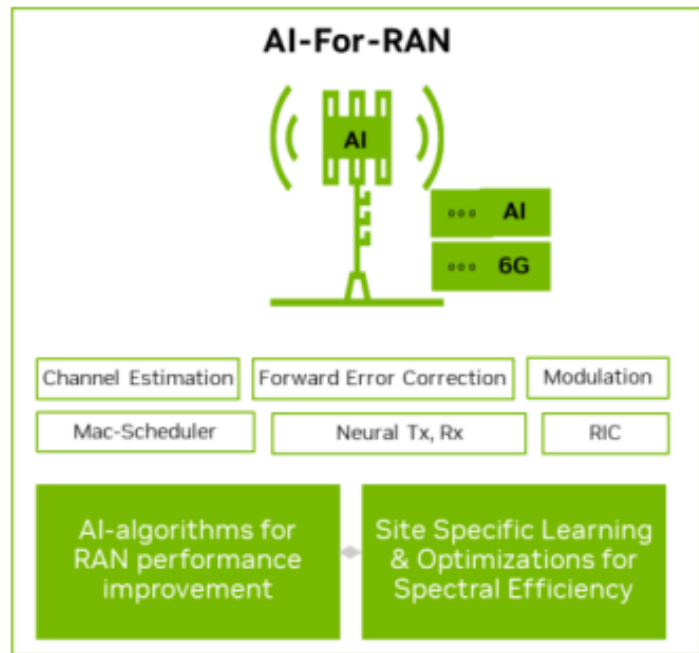
AI and RAN: using a common shared infrastructure to **run both AI and workloads**, with the goal to maximize utilization, **lower Total Cost of Ownership (TCO)** and generate **new AI-driven revenue opportunities**.



AI on RAN: enabling **AI services on RAN** at the network **edge to increase operational efficiency and offer new services** to mobile users. This turns the RAN from a cost centre to a revenue source.

### AI for RAN

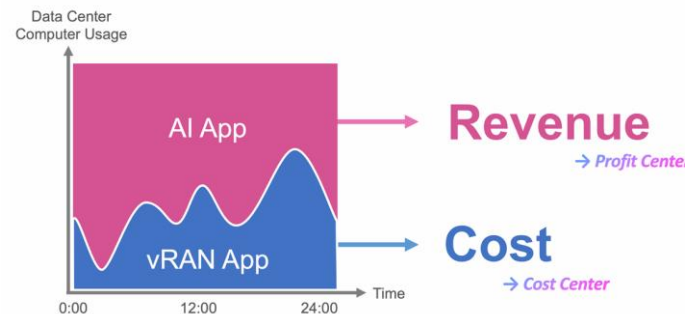
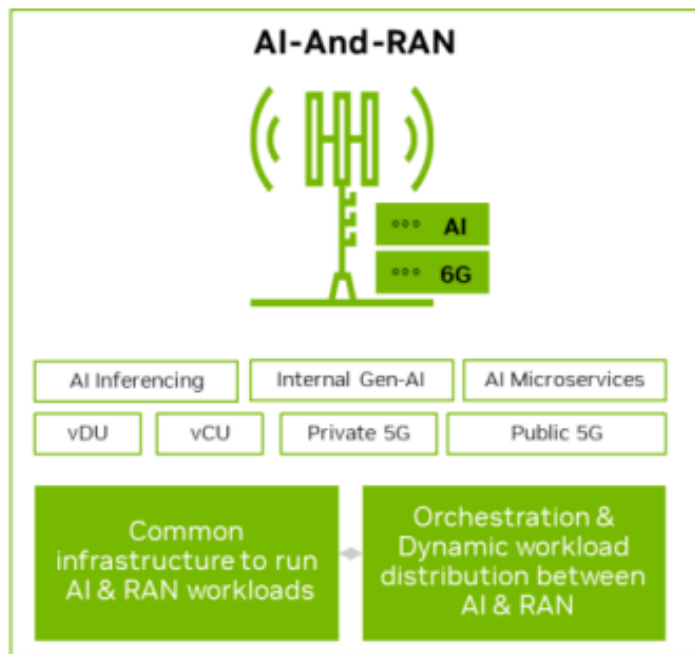
- (AI for RAN) RAN의 성능향상 및 최적화를 위하여 AI 활용



**(AI for RAN)** advancing RAN capabilities through embedding AI/ML models, algorithms and neural networks into the radio signal processing layer to improve spectral efficiency, radio coverage, capacity and performance.

### AI and RAN

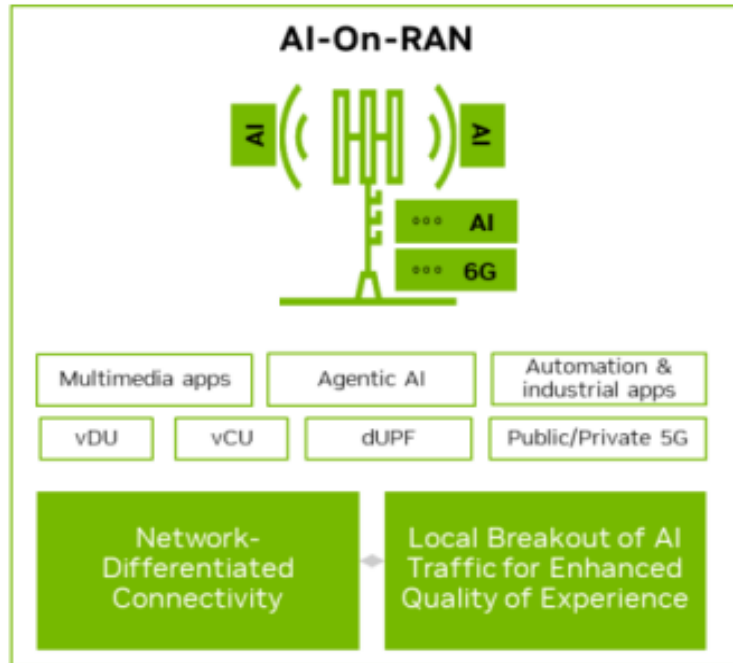
- (AI and RAN) AI와 RAN의 컴퓨팅 자원 공유



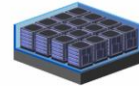
(AI and RAN) using a common shared infrastructure to **run both AI and workloads**, with the goal to maximize utilization, **lower Total Cost of Ownership (TCO)** and generate **new AI-driven revenue opportunities**.

### AI on RAN

- **(AI on RAN)** 무선접속망(RAN) 인프라에서 AI 및 생성형 AI(GenAI) 애플리케이션을 실행하는 기술적·운영적 프레임워크



### AI Inference



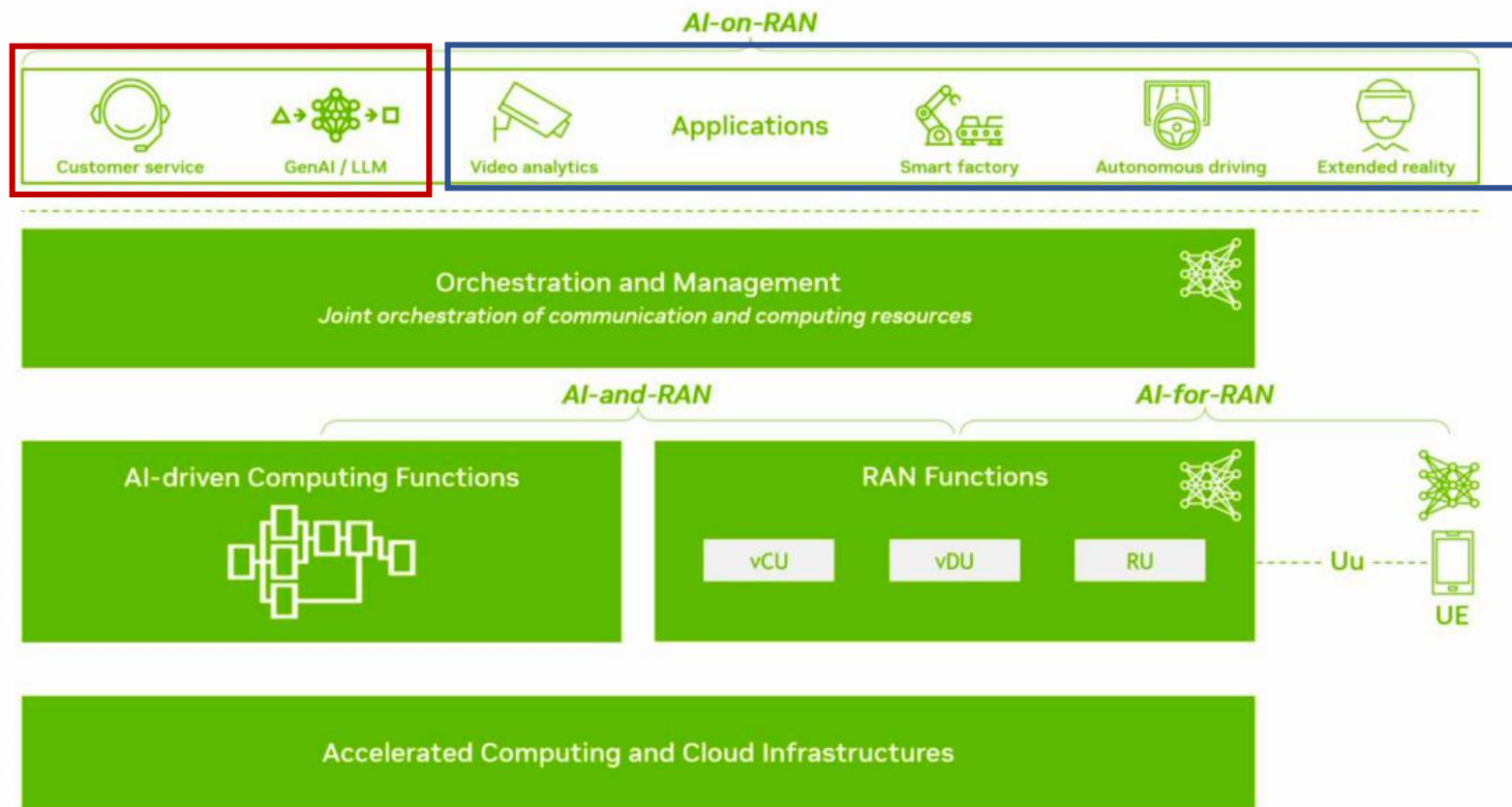
- Low latency
- High bandwidth
- High security
- Distributed compute



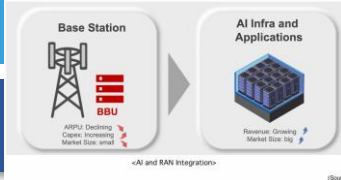
(Source: AI RAN Alliance)

**(AI on RAN)** enabling **AI services on RAN** at the network edge to increase operational efficiency and offer new services to mobile users. This turns the RAN from a cost center to a revenue source.

# AI-RAN의 High Level Overview by nvidia



# Why AI-RAN ?



- **(AI 기술을 RAN 에 통합)** AI-RAN lays the technology foundation for the telecommunications industry **to integrate the rapid advancements in AI technologies into the cellular telecommunications roadmap.**
- **(Edge에서의 AI 추론)** The surge in AI and generative AI applications is creating increased demands on cellular networks, driving demand **for AI inferencing at the edge** and **necessitating new approaches to handle these workloads.**
- **(AI 기반 무선 신호처리 효율증가)** At the same time, advances in AI-based radio signal processing techniques are showing compelling results versus traditional techniques, and **promising transformative gains in radio efficiency and performance**
- **(범용서버기반의 유연한 구조)** As the industry begins its 6G journey, **AI-RAN built with general purpose Commercial Off-The-Shelf (COTS) servers and software defined acceleration**, provides enhanced capabilities to process increased AI and non-AI traffic efficiently, compared to traditional RAN systems that are based on purpose-built hardware, whether it be custom Application-Specific Integrated Circuit (ASICs) or System on Chips (SoCs) with embedded accelerators.
- **(새로운 수익화 모델 창출 및 운영 자동화)** AI-RAN creates new revenue opportunities from hosting AI workloads and enables AI to be integrated into the operations of the RAN to optimize network performance, automate management tasks, and enhance overall user experience

### What are the key networking considerations for AI-for-RAN?

- **(Spectrum 효율 향상을 위한 AI 기회)** There are many opportunities to utilize AI to improve spectral efficiency of RAN such as **Channel Estimation/Prediction, Interference management, Beamforming, Deep Reinforcement Learning (DRL) based Modulation and Coding Scheme (MCS) selection** and more
- **(L1가속기의 유연성 필요)** These can only be achieved with embedded accelerated hardware and software computing capability at Layer 1, that is fully programmable, as with NVIDIA CUDA accelerated libraries for radio signal processing, under the NVIDIA AI Aerial platform
- **(AI-for-RAN의 MWC 사례)**
  - ✓ **(AI 로 성능향상)** SoftBank Demonstrates Performance Improvement in RAN Using AI with NVIDIA, Fujitsu
  - ✓ **(6G용 AI-Native 인터페이스)** DeepSig shows AI-Native Air Interface for 6G, using NVIDIA platforms
  - ✓ **(NVIDIA플랫폼 활용)** Keysight, Samsung, NVIDIA Advance AI-For-RAN, using NVIDIA platforms
- **(기존 가속기의 한계: 반복적인 AI 혁신 주기를 따라가기 어려움)** Purpose built RAN accelerators cannot support these continuous innovations as these accelerators are not programmable for the integration of such new techniques and are also outpacing multi-year cycles for developing custom hardware.

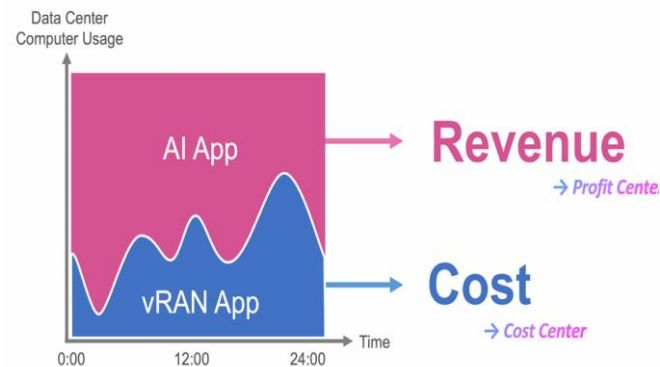
### What are the key networking considerations for AI-and-RAN?

- **(TCO 절감 및 AI 수익화)** NVIDIA's NCP Telco Reference Architecture built with MGX GH200 servers and BF3 DPUs allows the 5G RAN, dUPF and the AI applications to be deployed on the same platform managed by Kubernetes. This brings tremendous TCO benefits as the platform resources (CPU, GPU, DPU/NIC) **are dynamically allocated to RAN and AI functions** thereby increasing their utilization and unlocking new AI monetization.
- **(NVIDIA Spectrum-X + BlueField-3 DPU 효과)**
  - ✓ **(우선순위 및 QoS 보장)** Leveraging AI-driven traffic management to prioritize latency-sensitive RAN traffic and ensure high-priority AI workloads.
  - ✓ **(밴드폭 활용율 증가)** From 50–60% to over 97%, speeding up data transfer for inference workloads.
  - ✓ **(Latency 감소)** Advanced congestion control minimizes bottlenecks, ensuring real-time responsiveness.
  - ✓ **(GPU 활용도 개선)** Efficient network management maximizes GPU use for AI and RAN tasks. This includes software defined fronthaul.
  - ✓ **(Lower inter-token latency)** The increased bandwidth and optimized storage performance provided by Spectrum-X result in lower inter-token latency.
  - ✓ **(Accelerated storage access)** Spectrum-X improves read bandwidth by up to 48% and write bandwidth by up to 41% compared to traditional RoCE v2 protocols. This enhancement speeds up data retrieval and storage operations critical for inference tasks, particularly for techniques like retrieval-augmented generation (RAG). These can only be achieved with embedded accelerated hardware and software computing capability at Layer 1, that is fully programmable, as with NVIDIA CUDA accelerated libraries for radio signal processing, under the NVIDIA AI Aerial platform
- **(기존 가속기의 한계)** Purpose built RAN accelerators and NICs lack these critical capabilities.



# nVIDIA입장에서의 AI and RAN을 위한 고려사항

- (AI and RAN) AI와 RAN의 컴퓨팅 자원 공유



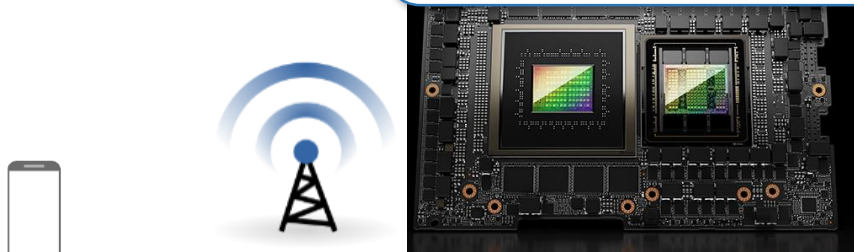
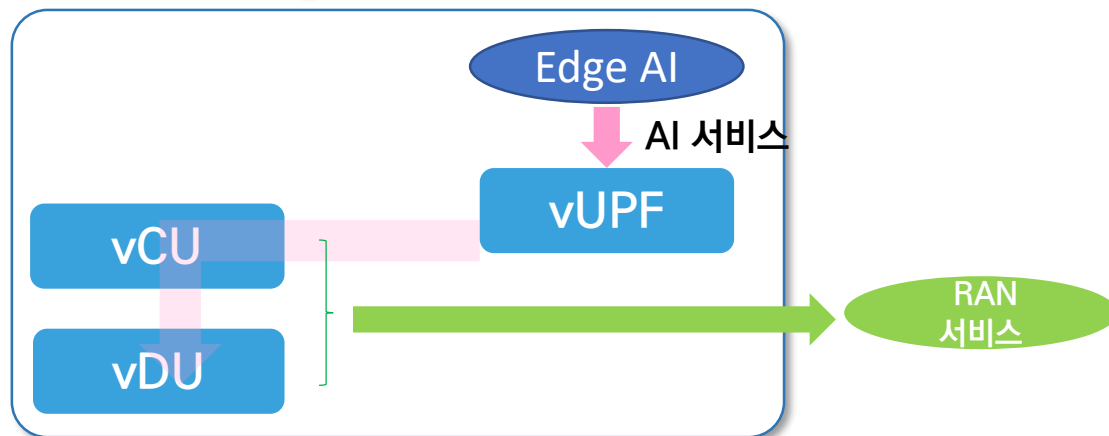
### What are the key networking considerations for AI-on-RAN?

- **(AI 트래픽과 dUPF)** As enterprise applications integrate more AI capabilities and increasingly run on mobile networks, efficient processing of 'AI traffic' in distributed telco datacenters is critical to deliver the best quality and user experience. In this architecture , a dUPF is used to identify and bridge the AI traffic to the AI Inference software such as NVIDIA NIM. A purpose-built RAN accelerator card does not have the features and flexibility for an efficient dUPF (GTP tunnel encap/decap, Packet Classification, Receive Side Scaling (RSS), and QoS (Metering/Marking/Policing).
- **(Agentic AI의 부상)** AI Agents are the next frontier for both consumer and enterprise applications. Agentic AI workloads require optimizations in the accelerated computing hardware and software stack, **such that the compute latency for reasoning tasks is minimized.**
- **(기존 가속기의 한계)** These Agentic AI optimizations are not possible with purpose-built RAN accelerators and NICs as these are not built for AI workloads.

# nVIDIA입장에서의 AI on RAN을 위한 고려사항

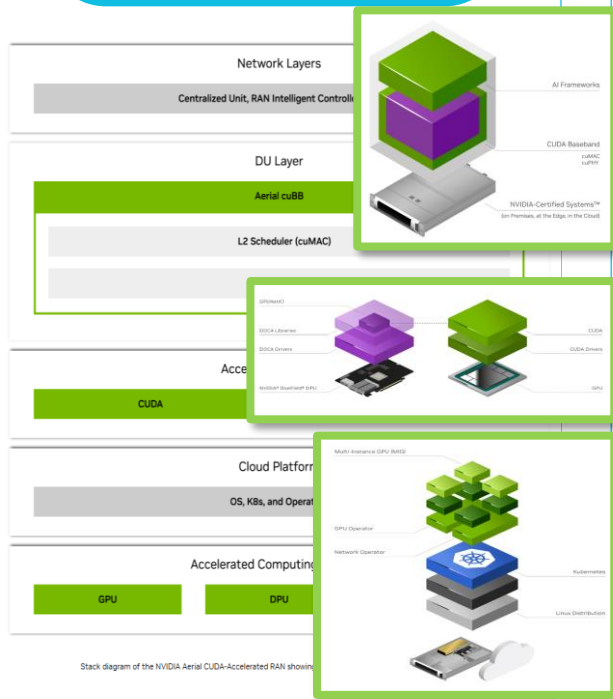
- (AI for RAN) RAN의 성능향상 및 최적화를 위하여 AI 활용

## Edge AI + RAN

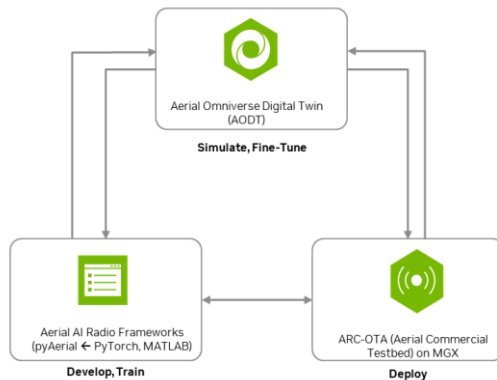


# NVIDIA AI Aerial

## Aerial CUDA-Accelerated RAN



## Aerial Omniverse Digital Twin



## Aerial AI Radio Frameworks

The frameworks include [pyAerial](#), [Aerial Data Lake](#), [Aerial RAN Colab Over-The Air \(ARC-OTA\)](#) and [NVIDIA Sionna](#), the leading link-level research tool for AI/ML-based wireless simulations.

### pyAerial

pyAerial provides a Python interface to the NVIDIA cuBB layer-1 data plane functions. Used with the Aerial Data Lake capture platform to produce training data for layer-1 functions, pyAerial can also be used to evaluate the end-to-end performance of neural network physical layer functions.

### Aerial Data Lake

Generate over-the-air training data with the Aerial Data Lake data collection platform. A data collection app runs on the distributed unit (DU), writing radio frequency (RF) samples to the database. Aerial Data Lake provides APIs to access the data. Used in conjunction with pyAerial, it generates datasets for intermediate nodes in the cuBB layer-1 signal processing pipeline to train neural networks for channel estimation, equalization, soft-demapping, and more.

### Sionna

Sionna is a GPU-accelerated open-source library for link-level simulations. It enables rapid prototyping of complex communication system architectures and provides native support for the integration of machine learning in 6G signal processing.

### ARC-OTA

Aerial RAN Colab Over-The Air (ARC-OTA) is a 3GPP Release 15 compliant and OTA operational campus 5G wireless stack, with all the network elements from RAN and 5G Core. It leverages disaggregated and off-the-shelf hardware and software components to offer full-stack programmability, with complete access to source code to onboard any experiments, quick turnaround validation, and benchmarking results.

# NVIDIA AI Aerial.. Sionna

## Aerial AI Radio Frameworks

### Sionna

Sionna™ is a GPU-accelerated **open-source library for link-level simulations based on TensorFlow and Keras.**

#### • (Principles)

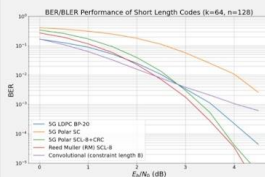
- Modularity
- Extensibility
- Differentiability

#### • (Modules)

- **Sionna RT**: A lightning-fast stand-alone ray tracer for radio propagation modeling
- **Sionna PHY**: A link-level simulator for wireless and optical communication systems
- **Sionna SYS**: System-level simulation functionalities based on physical-layer abstraction

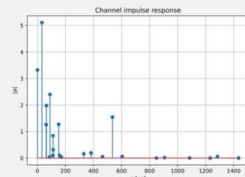
#### FORWARD ERROR CORRECTION

- 5G LDPC and Polar codes incl. rate matching
- Reed-Muller & Conv. Codes
- BP, SC, SCL, SCL-CRC, Viterbi
- Interleaving & Scrambling



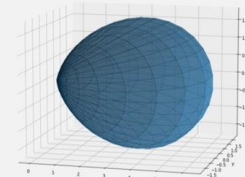
#### CHANNEL MODELS

- 3GPP 38.901 models: TDL, CDL, UMa, UMi, RMa
- AWGN, Rayleigh block fading
- Full time convolution or frequency domain processing



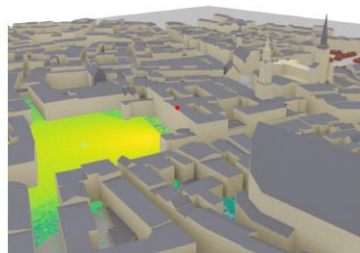
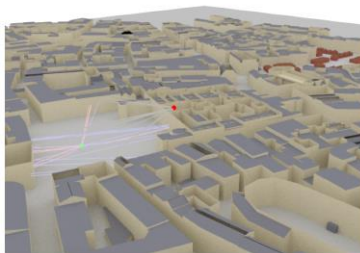
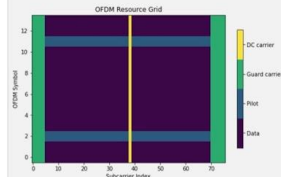
#### MULTIUSER MIMO

- 3GPP 38.901 & user-defined antenna arrays/patters
- ZF Precoding
- LMMSE Equalization
- Multicell support

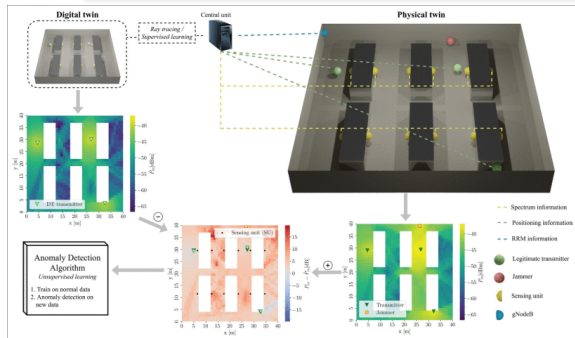


#### OFDM

- Flexible 5G slot-like frame structure
- Arbitrary pilot patterns
- LS channel estimation with nearest-neighbour interpolation

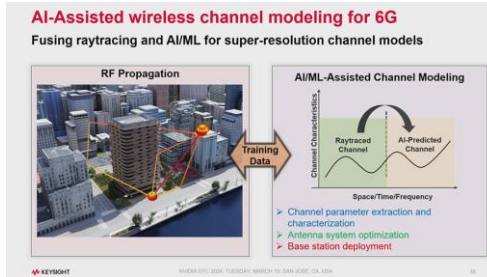


# NVIDIA AI Aerial.. AODT

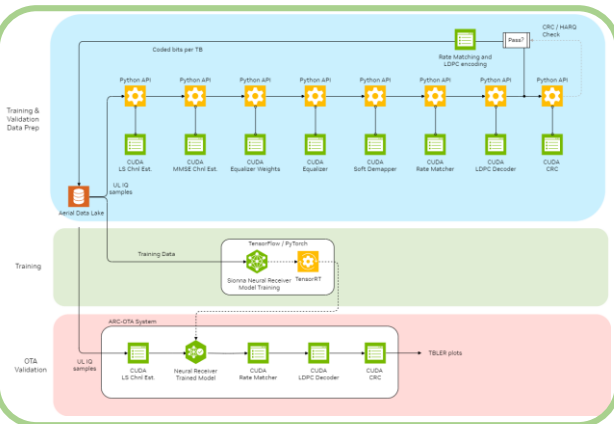


Spectrum anomaly detection  
[25 IEEE Comm. Mag.]

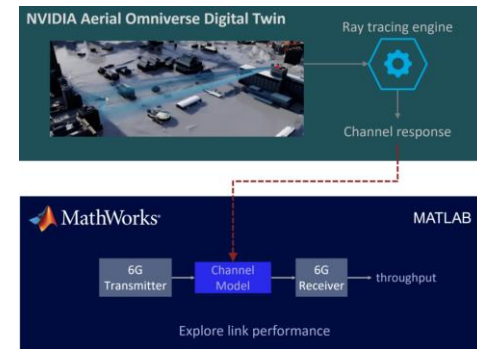
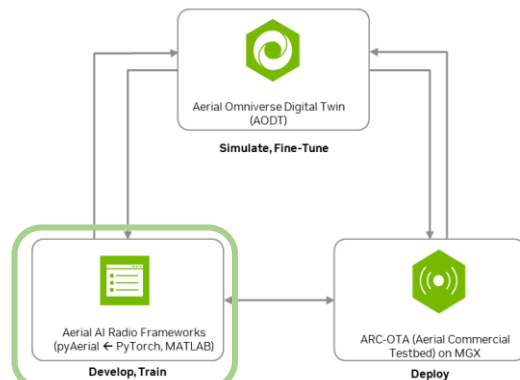
## Aerial Omniverse Digital Twin



AODT-based channel modeling  
[↑ Keysight / ↓ Mathworks]



Aerial AI Radio Frameworks with pyAerial



# NVIDIA AI Aerial.. CUDA Accelerated-RAN (NV ACR)

## Aerial CUDA- Accelerated RAN

### O-RAN

- 3GPP 및 O-RAN Alliance의 O-RAN 표준화

### 5G & 6G 준비

- 고성능 SW 및 AI 특화 네트워크가 요구됨

### 비용 효율성 및 유연성

- 기존 HW 기반 네트워크의 높은 비용과 낮은 유연성
- 6G로 간편한 업그레이드 지원



**NVIDIA**

Aerial CUDA-  
Accelerated RAN

### AI-RAN 혁신

- AI-RAN을 통한 네트워크 성능 최적화 및 새로운 수익창출 기회 제공

### AI 트래픽 증가

- 각종 디바이스에서 생성되는 AI 트래픽 급증을 지원할 수 있는 통신 네트워크가 요구됨

### 산업 협력 및 생태계 확장

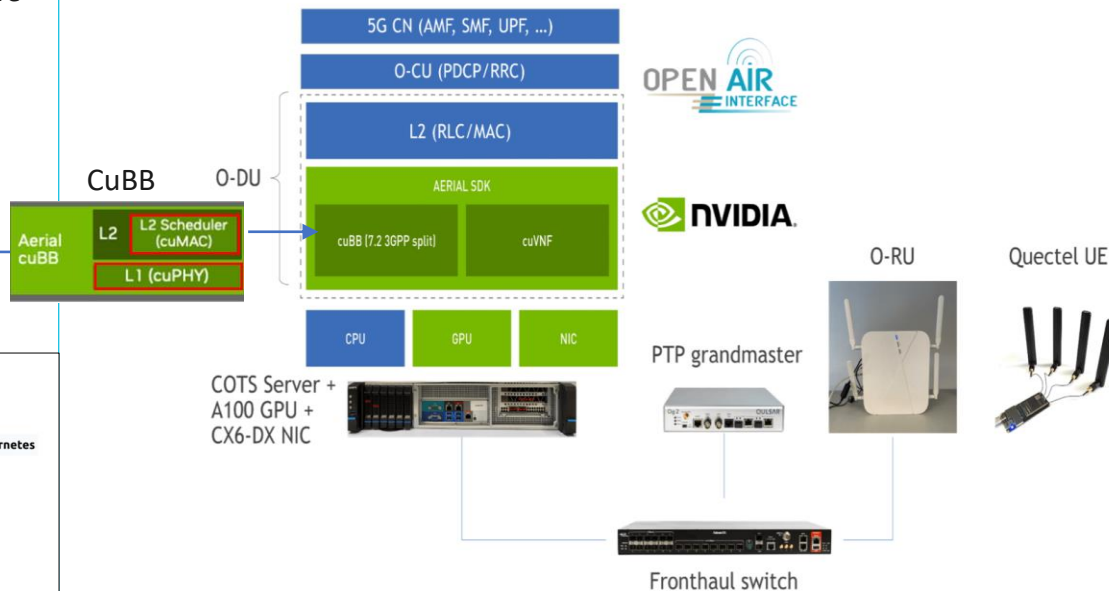
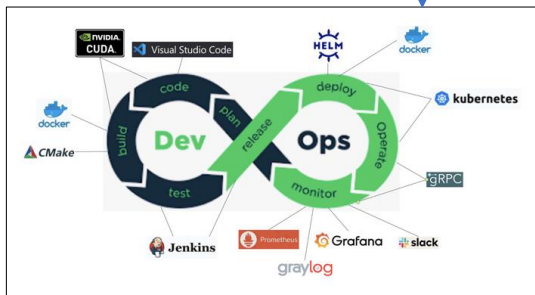
- 주요 통신 업체와 협력하여 AI-RAN 기술 발전을 촉진함
- 더 많은 기업이 해당 플랫폼을 활용하도록 생태계를 확장함

# NVIDIA AI Aerial.. CUDA Accelerated-RAN (NV ACR)

## Aerial Architecture

- **(SW-defined RAN)** GPU accelerated, cloud-native
- **(fully in-line GPU accelerated)** L1/L2 (for 5G/6G)
- **(Fully SW based Solution)**
  - Scalable, Modular, Programmable
  - Kubernetes based Cloud-native
  - SW 업데이트만으로 6G 업그레이드 지원

CuBB OAM: DevOps Workflow지원



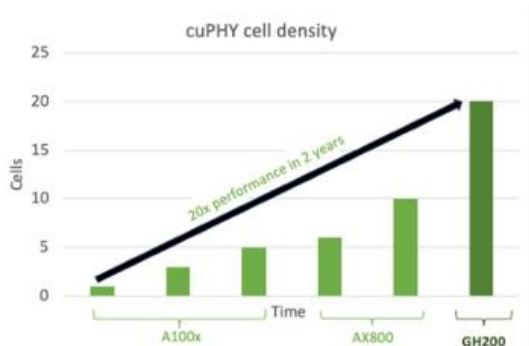


# NVIDIA AI Aerial.. CUDA Accelerated-RAN (NV ACR)

## cuBB



- 5G RAN의 PHY layer 지원 CUDA GPU SW Tool  
- 3GPP Rel.15 PHY 지원
- NVIDIA : A100x, AX800, GH200



[cuPHY peak cell density improvements over time]

## Docker

### cuBB (CUDA Baseband) - Container

- ✓ 5G/6G의 L1/L2 layer를 GPU 내부에서 처리하도록 지원하는 CUDA 기반 프레임워크
- ✓ 고성능 GPU 기반의 fully in-line PHY layer 프로세싱 기능을 SW로 구현
- ✓ CPU ↔ GPU 간의 데이터 이동 최소화를 통한 지연 감소 및 처리 효율 극대화

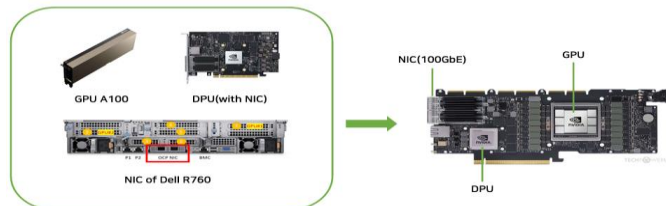
### cuPHY(L1)

- ✓ GPU accelerated L1 PHY
- ✓ 신호 처리 pipeline [FFT, Chnl Est., Eqlzr, De-mapper, LDPC 등]
- ✓ CPU의 중간 개입 없이 L1의 전체 흐름을 구성
- ✓ GPU의 병렬 처리를 활용해 MIMO, mMIMO 등 고 복잡도 처리 과정을 효율적으로 수행
- ✓ 셀/사용자/대역폭 증가에도 유연한 대응

Fully in-line  
L1-L2 IF

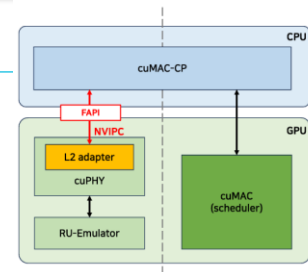
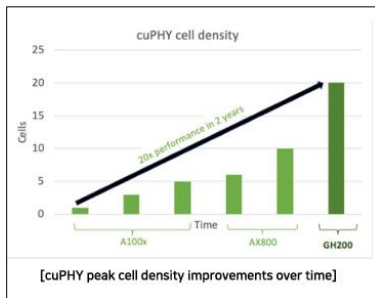
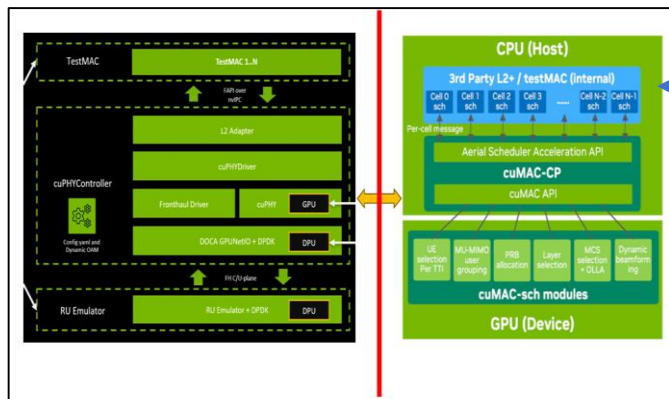
### cuMAC(L2)





- ✓ GPU accelerated L2 MAC
- ✓ 실시간 스케줄링 및 자원 제어 [PRB 할당, MCS 선택, UE grouping 등]
- ✓ SW-defined MAC scheduler
- ✓ GPU의 병렬 처리를 활용해 다중 사용자 관리 및 스케줄링 연산을 동시에 진행
- ✓ cuPHY의 신호 처리 결과를 기반으로 자원 스케줄링 결정



## NVIDIA AI Aerial.. CUDA Accelerated-RAN (NV ACR)

## NV ACR SW

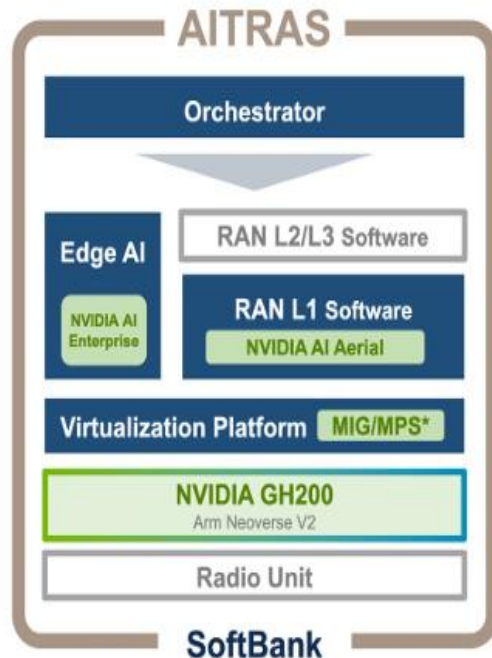


Version	Rel. 24-1 (24.6)	Rel. 24-3
DU	 Dell R760	 NVIDIA Grace Hopper
RU-emulator	 Dell R760	 Dell R760
비고	<ul style="list-style-type: none"> <li>cuPHY와 cuMAC이 독립적으로 개발됨</li> <li>x86 기반 DU는 최신 버전(24-3부터)의 Aerial에서 지원되지 않음</li> </ul>	<ul style="list-style-type: none"> <li>NVIDIA Grace Hopper(MGX)를 Host로 활용함</li> <li>cuPHY-cuMAC의 연동이 지원됨</li> </ul>

# SoftBank (AI-RAN)

## AITRAS

- **(AITRAS) gRAN(GPU based RAN) 기반 소프트뱅크의 AI-RAN 결과물**
- **(AITRAS 주요 특징)**
  - Multi-tenancy for AI-and-RAN with AI-native orchestration..
  - Support for the development, deployment, and monetization of various AI applications
  - Carrier-grade RAN performance,
  - AI-driven enhancements in spectral efficiency and energy efficiency
- **(AITRAS 주요 구성 요소)**
  - 물리적 시스템: **NVIDIA GH200 Grace Hopper , Radio Units, and network switches.**
  - 논리적 시스템 구조
    - . 가상화 플랫폼
    - . RAN 기능 ( L1, L2, L3)
    - . Edge AI: AI 응용 지원
    - . 동적인 할당이 가능한 오케스트레이터 : AI와 RAN Resource
  - 자원 관리 : AI-driven resource control
  - RAN function : fully software-defined
  - RAN에서의 AI/ML 모델 : 다양한 AI 모델로 Spectral Efficiency와 Energy 효율을 높임



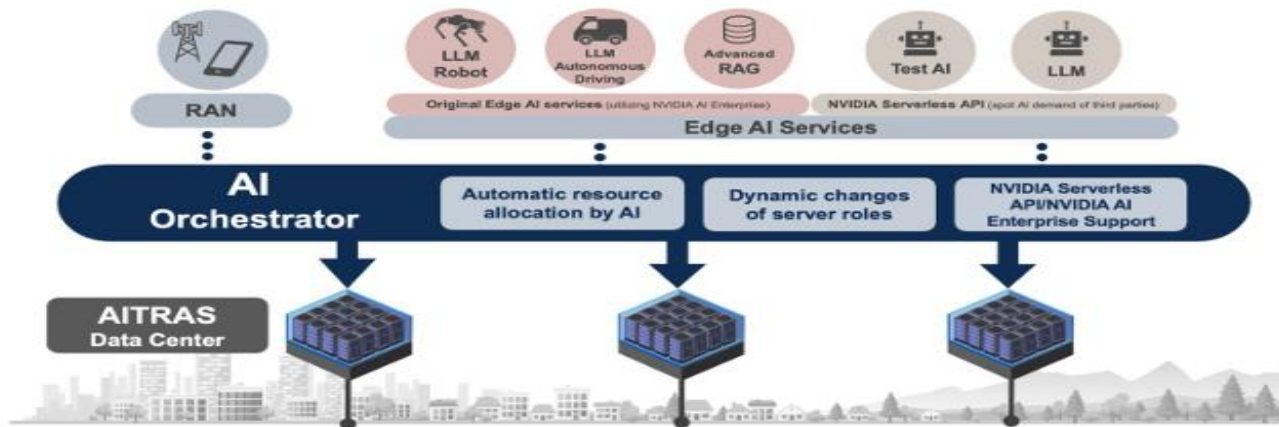
\*MIG: Multi Instance GPU  
MPS: Multi Process Service

# SoftBank (AI-RAN)

## AITRAS

### • (AI-Native Orchestration)

- (Automatic Resource Allocation by AI) AI와 RAN의 workload 요구에 맞추어 동적으로 컴퓨팅 자원할당, RAN과 AI 서비스와 같은 특정요구 수용
- (Dynamic Changes of Server Roles) AI와 RAN 간의 스위칭, Edge AI 서비스 지원
- (Integration with NVIDIA AI Platform) , NVIDIA serverless API와 AI Enterprise SW 간의 leverage,
- (Edge AI Services Management) Edge AI 서비스관리와 배치의 백본역할, LLM 로봇, 자율주행솔루션과 같은 서비스
- (AI-Driven Adaptability) 시스템성능을 계속해서 감시하고 최적화를 위한 AI 프로세스 적용, Latency, 높은 성능, 높은 가용성 제공

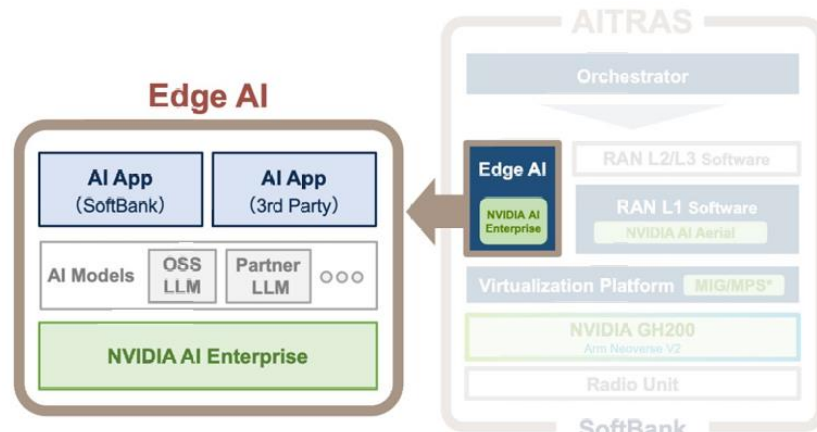


### SoftBank (AI-RAN)

#### AITRAS

##### • (Edge AI)

- 5G기반의 고속, 저지연 통신을 보장하고 데이터 보안 및 Local 처리에 따른 데이터 주권 보장
- (적용 사례)
  - . Multimodal AI 기반 자율주행 원격지원
  - . 기업전용 RAG 지원: 오피스, 공장, 건설현장등에 맞는 기업데이터 입력 및 업무자동화
  - . 실시간 로봇 제어

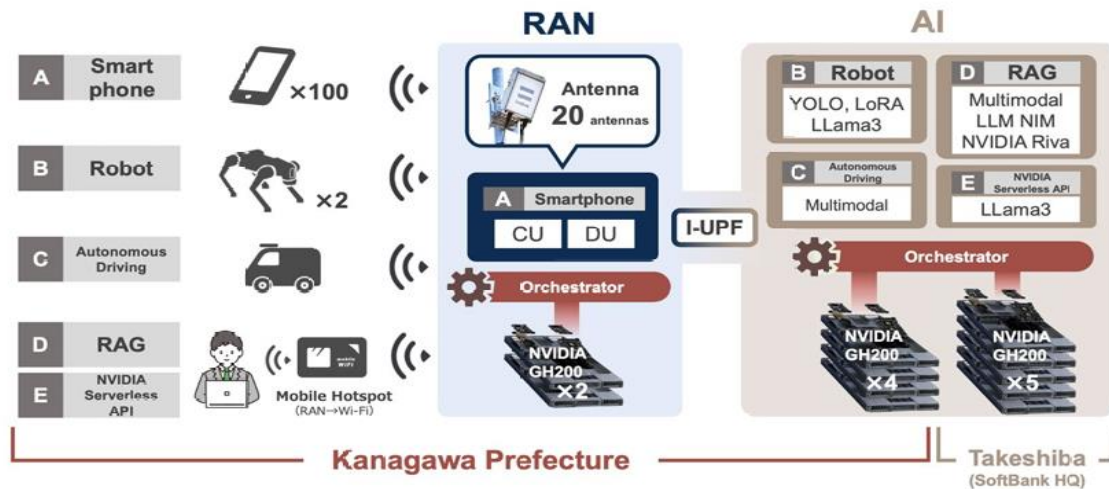


# SoftBank (AI-RAN)

## AITRAS

### • (Outdoor Testbed for AITRAS): kanagawa 현에 설치

- (Background and Objective)
  - . Carrier-grade 안정성 평가 : 5개의 셀(각각 4개의 안테나)을 100m 간격으로 균등배치
- (Testbed Setup)
  - . NVIDIA GH200 Grace Hopper : 한서버에 20개의 5G Cell 동작,
  - . 4 layer MIMO, n79(4.8~4.9GHz)



### SoftBank (AI-RAN)

#### AITRAS

- **(Trial Result):** 100UE (Cell 당 5개씩) 의 동시 비디오 스트리밍 안정적,
- **(Performance Evaluation)** Power consumption: 500W( 셀당 25W), UE의 수가 증가해도 C-plane 로드안정적, DL only, UL only, DL/UL 동시 트래픽시 에도 안정적임

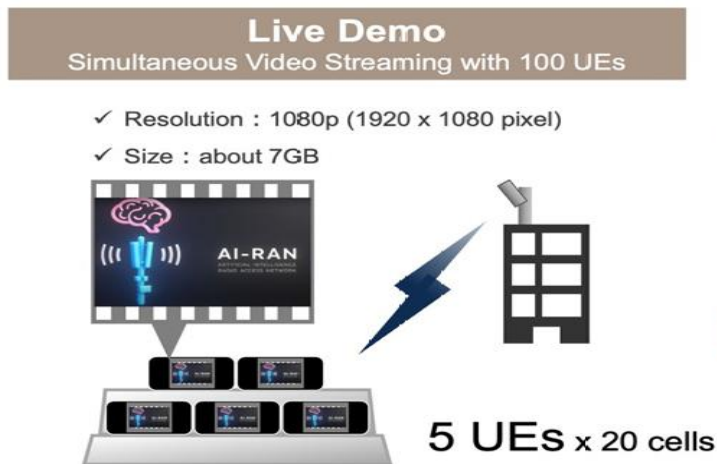


Figure 10. Live demo of simultaneous video streaming with 100 UEs

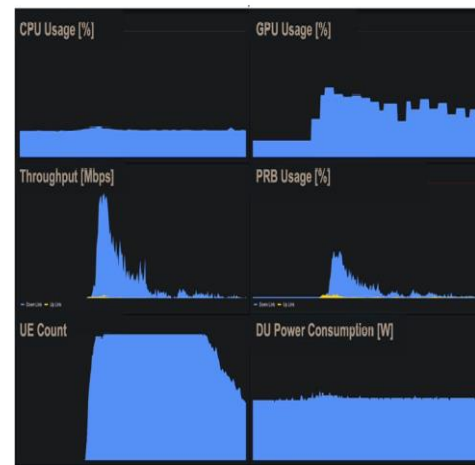


Figure 11. Resource monitoring of simultaneous video streaming with 100 UEs (traffic load and computational resources)

# SoftBank (AI-RAN)

## AITRAS

- **(L1 enhancement):** NVIDIA AERIAL 플랫폼을 활용하여 안정적으로 자체 개발,
  - 향후, 소프트뱅크는 L1의 성능을 AI를 활용하여 높이고 전력소모 줄일 예정,
  - L2/L3에도 AI 적용 예정
- **(전체 RAN 성능 개선):**
  - 셀 용량증가, Throughput 향상, 전력소모 감소
  - L1의 경우, GPU 가속기를 통한 성능저하없이 더 많은 가입자수 수용
  - 실시간 빔포밍 최적화
  - 실시간 네트워크 수요에 맞춘 자원관리, 효율적인 전력사용
  - 기존 RAN과의 차별성

Feature	Traditional RAN	gRAN (AITRAS)
Processing Type	CPU-based	GPU-accelerated (parallel processing)
AI Handling	Limited, sequential	Multiple AI algorithms run <b>concurrently</b>
Latency	Higher	<b>Lower</b> , due to real-time parallelism
Energy Efficiency	Lower	<b>Improved</b> , through optimized workload management
Scalability	Constrained	<b>Highly scalable</b> , suitable for growing 5G/6G demands
Future-readiness	Less adaptable	Optimized for <b>AI-native, next-gen networks</b>

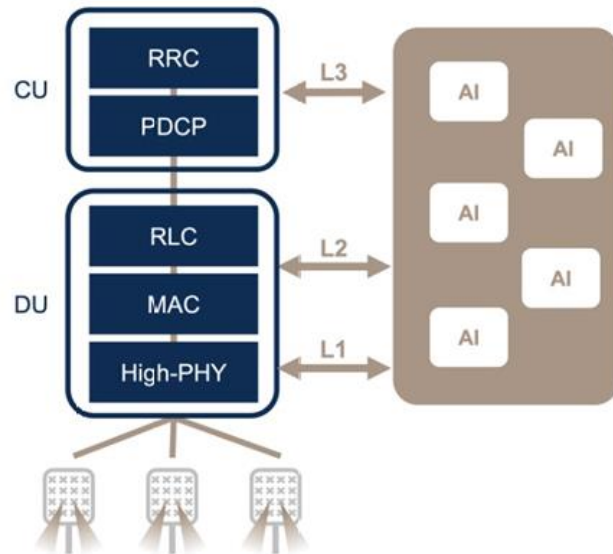


Figure 12. Full-layer optimization



### SoftBank (AI-RAN)

#### AI- and-RAN in ATRAS

- **(소프트뱅크의 AI and RAN ):**

- 가상화 구조는 Management Cluster와 Workload Cluster로 구성
- 추론을 포함한 AI Workload의 증가하는 계산 수요 대응해야 하고 AI 와 RAN 이 모두 동작 가능하도록 구성

- **(하드웨어와 자원관리):**

- 셀 용량증가, Throughput 향상, 전력소모 감소
- L1의 경우, GPU 가속기를 통한 성능저하없이 더 많은 가입자수 수용
- 실시간 빔포밍 최적화
- 실시간 네트워크 수요에 맞춘 자원관리, 효율적인 전력사용
- 기존 RAN과의 차별성

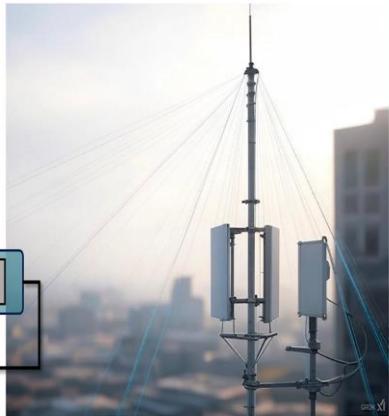
# AI-Native 6G Air Interface Prototype on NVIDIA AI Aerial Platform

## Demo 1: “Learned Air Interface with Online Learning”, AI-for-RAN

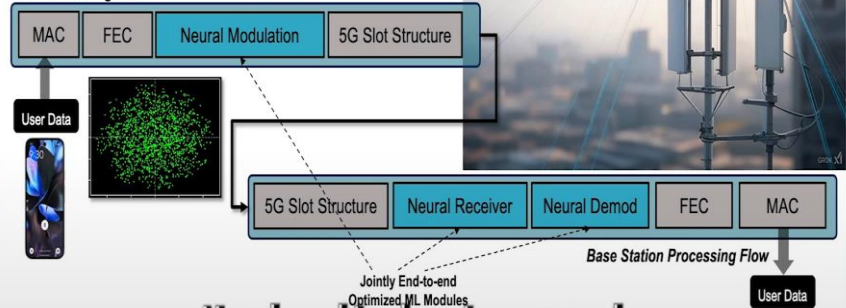
### AI-Native Air Interface for future 6G PHY



- AI-Native Air Interface allows base station and mobile device to jointly optimize processing and signal design for real world conditions.
- Neural encoder, receiver, and decoder replace conventional pilots & modulation while retaining 5G-NR compatible CP/DFT-OFDM slot structure



UE Processing Flow



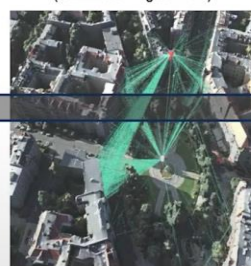
### Prototyping 6G on AI Aerial Platform



- Aerial Omniverse Digital Twin (AODT) provides scalable Ray Tracing Simulation for more accurate propagation conditions reflecting real world deployment geometry.
- We leverage AODT channel emulation for realistic fine-tuning and validation
- End-to-end optimization for realistic site-specific propagation conditions
- The path towards optimizing the whole RAN for digital twin environment



NVIDIA AI Aerial Platform (Base Station) /w OmniPHY Modem



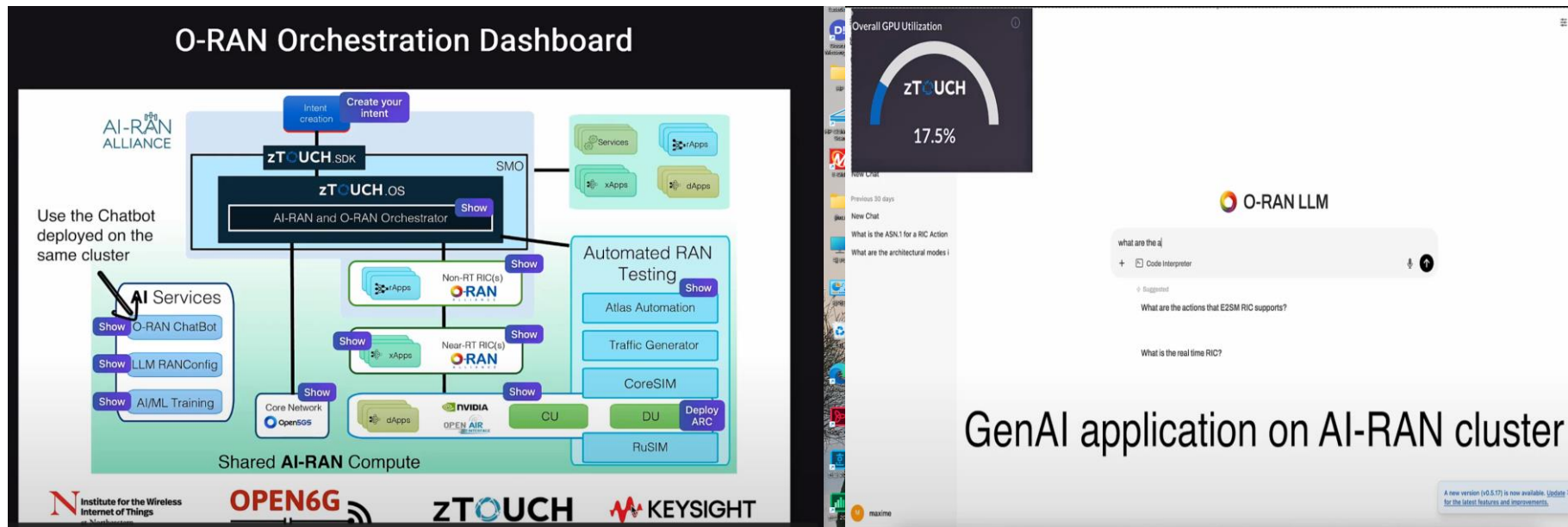
NVIDIA AI Aerial Platform (Omniverse Digital Twin)



Orin NX AI-Native PHY Mobile Device /w OmniPHY Modem

## Demo 7 : AI-RAN Orchestration, AI and RAN

### Demo 7: “AI-RAN Orchestration”, AI-and-RAN

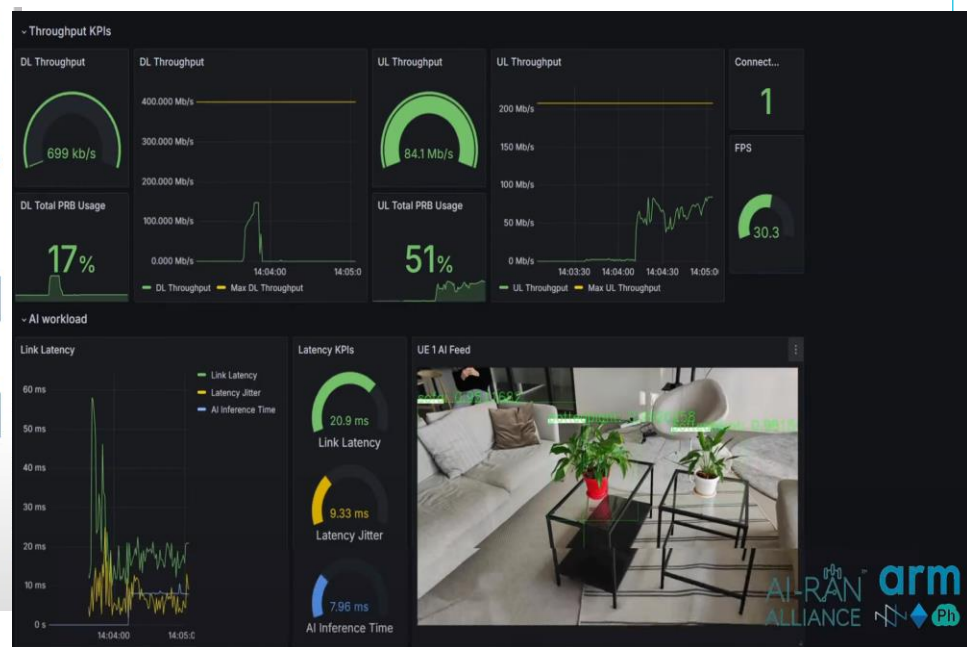
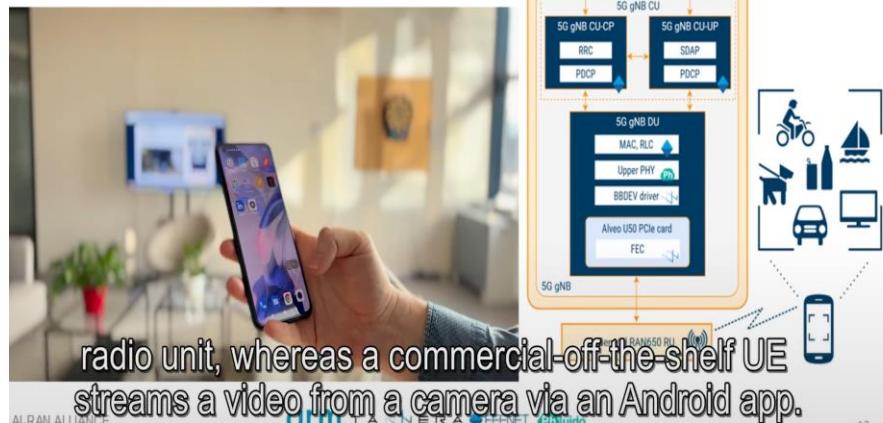


# Object Detection on RAN , AI-on-RAN

## Demo 10: Object Detection on RAN , AI-on-RAN

### Solution Overview and Demo Setup

- Processor: Ampere Altra Q80-30 (Arm Neoverse-N1 cores)
- RU: Benetel RAN650, 100 MHz



# THANK YOU

