2025년 7월 ORIA 가상화/지능화위원회 정례회의

O-RAN nGRG 연구 동향

고남석
가상화 · 지능화위원회 부위원장
ORIA

Agenda

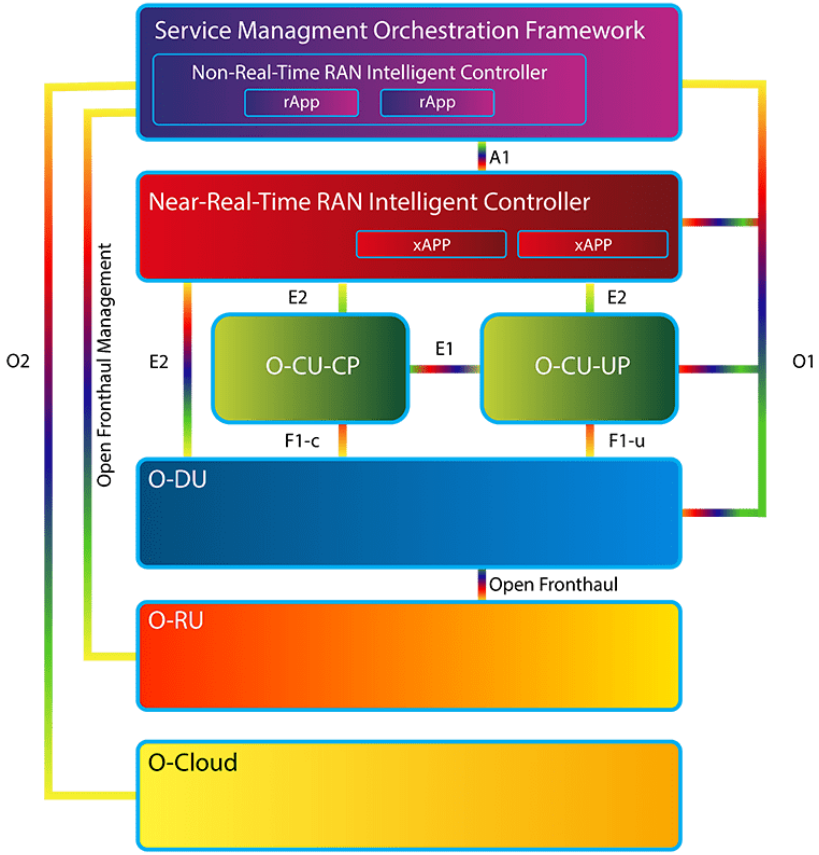
1. O-RAN nGRG 개요
2. AI-Native O-RAN
3. Communication & Computing Integrated O-RAN
4. SBA-based RAN-CN Convergence

O-RAN nGRG 개요

O-RAN Alliance

2018년 출범, 11개 작업반(WG)/5개 포커스 그룹/1개 연구 그룹으로 구성

O-RAN Architecture



O-RAN specifications are available to the general public at https://www.o-ran.org/specifications (100개 이상 문서 발간)

Work Groups

- WG1: Use Cases and Overall Architecture Work Group
- WG2: The **Non-Real-Time RIC and A1** Interface Work Group
- WG3: The **Near-Real-Time RIC and E2** Interface Work Group
- WG4: The **Open Fronthaul Interfaces** Work Group
- WG5: The Open F1/W1/E1/X2/Xn Interface Work Group
- WG6: The Cloudification and Orchestration Work Group
- WG7: The White-box Hardware Work Group
- WG8: Stack Reference Design Work Group
- WG9: Open X-haul Transport Work Group
- WG10: OAM Work Group
- WG11: Security Work Group

Focus and Research Groups

- IEFG: Industry Engagement Focus Group
- OSFG: Open Source Focus Group
- SDFG: Standard Development Focus Group
- SuFG: Sustainability Focus Group
- TIFG: Testing and Integration Focus Group
- **nGRG: next Generation Research Group**

O-RAN Alliance - nGRG Mission

2022년 출범, O-RAN의 6G 연계를 위한 전략적 연구 주도

목적

- O-RAN 기반 6G 연구와 결과 공유를 위한 포럼 제공
- 글로벌 연구 및 표준 동향을 반영해 O-RAN의 6G 대응 방향 수립
- 4G/5G에서 6G로의 O-RAN 지속 가능성 확보
- 산업과 협력하여 O-RAN과 타 표준 간 비호환 방지를 위한 6G 기술 진화 방향 정렬

운영 방식

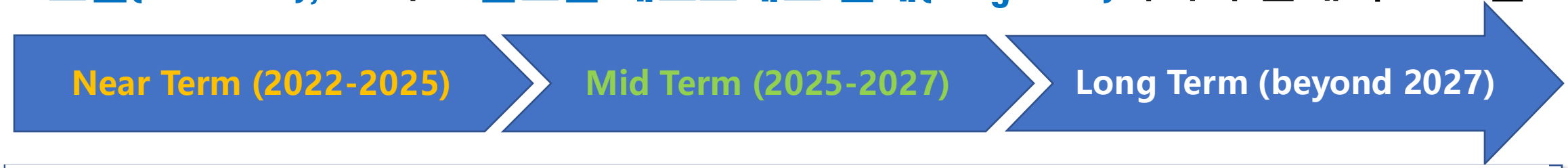
- O-RAN 차세대(nG) 연구 의제와 핵심 우선순위 정의
- 우선순위에 따라 연구 흐름(stream)을 구성하고, 이에 따른 연구 과제 제안 요청
- 연구 흐름의 진행 상황과 성과에 대한 정기적 논의 및 검토
- 다양한 기술 간 O-RAN 솔루션 연동 방안 연구

결과

- 그룹의 연구 성과를 바탕으로 백서 및 연구 보고서 발간 및 대응 권고
- 관련 파트너와 함께 주제별 워크숍, 세미나, 회의 개최

O-RAN Alliance - nGRG Roadmap

6G 기술 동향 분석과 연구 과제 도출(Near Term)에서 시작해, **표준화 기여 및 협력 조율(Mid Term)**, 그리고 **글로벌 테스트베드 연계(Long Term)**까지의 단계적 로드맵



- **Near Term (2022-2025):**

  산업 파트너들과 협력하여 6G 기술 트렌드 및 차세대 IMT에 대한 ITU-R 비전이 O-RAN 아키텍처에 미치는 영향을 분석하고, 연구 우선순위에 기반한 연구 과제를 도출하며, 결과물로 백서 및 연구 보고서를 발간

- **Mid Term (2025-2027):**

  연구 결과를 바탕으로 O-RAN 작업반(WG)/포커스 그룹(FG)에 입력을 제공하고, O-RAN 6G 표준화 연구 준비 및 타 SDO와의 협력 조율

- **Long Term (beyond 2027):**

  O-RAN SDFG를 통해 6G 연구 기관, SDO, 관련 프로그램들과 연계하고, O-RAN TIFG, IEEE, NSF 등과 함께 6G 테스트베드 공동 연구

O-RAN Alliance - nGRG Research Streams

6G 대응을 위한 요구사항, 구조, 보안, AI-RAN 등의 주제를 아우르는 8개의 스트림 기반의 연구 체계

RS01: Requirements

6G use cases and standard gap analysis
- What 6G use cases and requirements must be considered?
- What are 5G standard gaps towards these requirements?

RS02: Architecture

Architecture towards 6G O-RAN
- What are key architecture principles?

Extreme disaggregated RAN
- How does disaggregation play role in future RAN?

RIC evolution
- What are gaps in RIC to meet 6G use cases?

RAN-core-edge harmonization
- What is the trend for RAN-CN NF and edge AF integration?

Architecture support for new features in air interface
- What enhancements are needed to support JCS, CF-MIMO, RIS…?

RS05: Management

Zero touch management and orchestration
- How should O-RAN evolve on network management?
- What are gaps regarding OAM/SMO orchestration and automation?

RS04: Security

Native security, privacy and trust
- What is native security design in O-RAN?
- What are key areas regarding security, privacy and trust in new O-RAN design?

RS03: AI/ML

Native and cross-domain AI
- What is native AI design in 6G O-RAN?
- How to integrate AI cross-domain?

RS06: Cross domain

Softwarization and cloudification
- What are key problems in RAN cloudification?

New sharing paradigms
- How does 6G O-RAN deal with spectrum and infrastructure sharing?

Network exposure and cross-domain integration
- What are trends for cross-domain integration with verticals?

Network migration
- How to migrate from 5G/5Gadv, WiFi to 6G O-RAN?

RS07: Sustainability

Carbon-neutral O-RAN
- Energy efficiency aspects in 6G O-RAN

RS08: nG Research Platforms

O-RAN nGRG Research Reports (https://www.o-ran.org/research-reports)

문서	요약	기술 영역 분류
O-RAN Towards 6G	• 6G 시대 Open RAN의 발전 방향을 모색하기 위해 6G 활용 사례와 요구사항을 분석 • 2022년 O-RAN 회원사를 대상으로 시행한 6G 활용 사례 설문 결과를 토대로 13개 범주의 미래 use case를 정리하고, 이를 통해 도출된 요구사항 대비 현재 O-RAN 아키텍처·표준의 격차를 분석하여 6G 대응을 위한 향후 연구 방향을 제시	6G 유즈케이스, 요구사항
Research Report on Native and Cross-domain AI: State of the art and future outlook	• 차세대 네트워크에서 네이티브 AI와 크로스 도메인 AI의 적용 현황과 전망 제시 • 글로벌 표준 동향(3GPP, O-RAN, ETSI 등)을 소개하고 두 개념을 정의한 뒤, AI 도입이 RAN 아키텍처에 미치는 영향과 계층 간 데이터 통합의 과제를 분석 분산 지능의 신뢰성 확보 중요성을 강조하고, 대규모 분산 AI 구현을 위한 HPC 및 오픈소스 플랫폼, 의도 기반 관리 등 핵심 기술을 논의하며 향후 심화 연구 방향을 제언	AI/ML
Research Report on Quantum Security	• 양자 컴퓨팅으로 인한 미래 통신 보안 위협과 그 대응 전략을 분석 • 향후 양자 공격에 취약해질 기존의 공개키 및 대칭키 암호를 양자 내성 암호(PQC)와 양자 키 분배(QKD) 기술로 대체할 필요성을 강조하고, 이들 양자 보안 기술 도입 시 예상되는 성능 및 상호운용성 문제 등의 도전 과제를 논의	보안
Spectrum Sharing based on Shared O-RUs	• 다수의 운영자가 O-RAN 무선 장치(O-RU)를 공유하여 스펙트럼을 공동 활용하는 프레임워크 제시 • 표준화된 프론트홀 인터페이스를 통해 여러 운영자가 동적으로 주파수를 함께 사용하면서도 품질을 보장할 수 있도록 함으로써, 기존 비면허 대역이나 3GPP RAN 공유 방식의 한계를 극복하고 주파수 이용 효율을 높이는 방안을 모색	스펙트럼 공유 및 활용
O-RAN Native AI Architecture Description	• 6G 시대 O-RAN에 네이티브 AI를 도입하기 위한 아키텍처 방안을 제시 • RAN 지능형 컨트롤러(RIC)를 기반으로 네트워크 품질 향상과 효율 최적화를 위한 6G 네이티브 AI의 요구사항을 분석하고, 네이티브 AI 구현을 위한 원칙과 기능을 논의한 후 미래의 AI 활용 RAN 구조	RAN 아키텍처 및 설계, AI/ML
Architecture principles for a cloud-friendly future 6G RAN architecture	• 미래 6G RAN의 클라우드 친화적 아키텍처 설계를 위한 주요 원칙 정리 • 자동화, 가시성, 근본 원인 분석, 제로트러스트 보안 등 클라우드 네이티브 RAN의 핵심 설계 요소를 정의하고, 의도 기반 API, 라이프 사이클 관리, TEE (신뢰 실행 환경), 원격 attest 등 이를 구현하기 위한 기술들을 제안	RAN 아키텍처 및 설계, 클라우드 네이티브

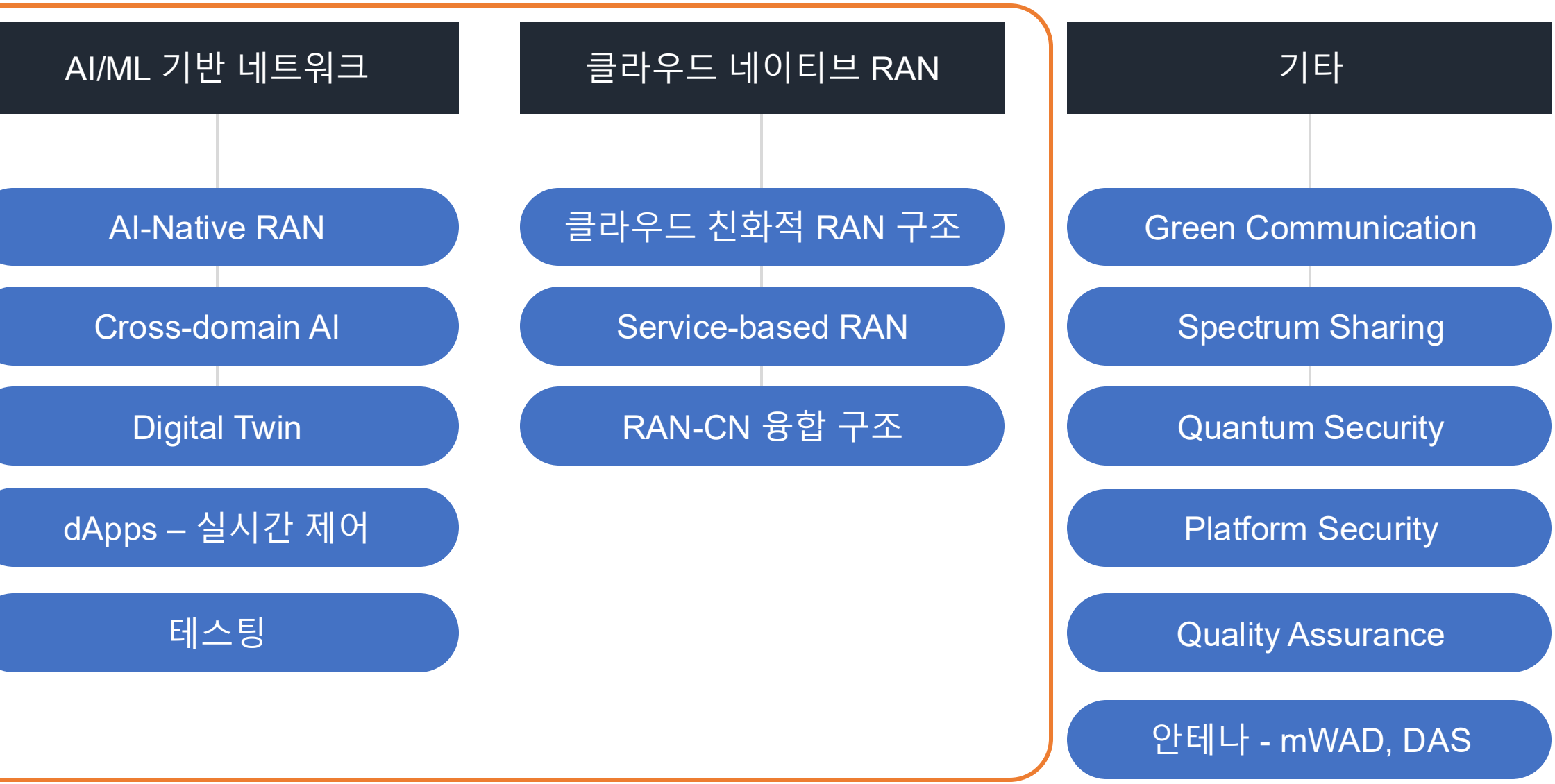
O-RAN nGRG Research Reports (https://www.o-ran.org/research-reports)

문서	요약	기술 영역 분류
Research Report on Cross-domain AI	• 차세대 네트워크에서 크로스 도메인 AI 활용 사례와 기술 방향 제시 • 네트워크 각 분야에서의 AI 적용 현황과 3GPP, O-RAN, ONAP, ETSI 등 표준 동향을 살펴본 뒤, 이종 도메인 간 데이터·연산·모델 협력을 통한 크로스 도메인 AI 기술의 고려사항과 현행 네트워크에 미치는 영향, 향후 협업 방향을 제시	AI/ML
Principles and Methodologies for AI/ML Testing in Next Generation Networks	• 차세대 네트워크에 AI/ML 시스템을 원활히 통합하고 테스트하기 위한 원칙과 방법론 제시 • Open RAN 등 다양한 네트워크 환경에서 AI/ML 알고리즘의 불확실성에 대응하는 새로운 시험 절차와 검증 방안을 논의하고, AI 도입 시 프라이버시·보안·신뢰 확보 방안을 다루며, 예측 분석을 통해 병목을 식별하고 자원을 최적화하는 등의 AI 기반 네트워크 성능 향상 방안을 제안	AI/ML, 보안
Use Case Analysis on mmWave Antenna Distribution (mWAD)	• mmWave 분산 안테나(mWAD) 기술의 잠재력과 과제 분석 • 170개 6G 활용사례 조사 기반 29개 통합 활용사례와 31개 요구사항 추출하고, mWAD 적용이 예상되는 사례들의 요구사항 분석하여 5G의 기술 격차를 식별	무선 기술 및 안테나 시스템
Digital Twin RAN Use Cases	• 디지털 트윈 RAN(DT-RAN)의 주요 활용 사례와 O-RAN 표준 격차 분석 • O-RAN 커뮤니티 설문을 통해 선정된 상위 5개 DT-RAN 활용 사례(모델 학습·테스트, 네트워크 시험 자동화, 네트워크 설계 최적화, 에너지 절감, 사이트 맞춤형 최적화)를 집중 검토하고, 이러한 활용 사례를 지원하기 위해 필요한 표준 개선 사항을 도출	디지털 트윈
Use Case Analysis Related to Green Communication in O-RAN	• O-RAN 네트워크의 에너지 효율 향상을 위한 요구사항과 설계 방안 제시 • 네트워크 기능별 에너지 효율(EE) 요건을 분석하고, 에너지 인지형 프로토콜 및 아키텍처 설계, 혁신적인 절전 기법과 자원 최적화 전략 등을 통해 6G 통신망에서 에너지 소비를 줄이고 친환경 통신을 구현하는 방안을 제시	에너지 효율 및 그린 통신
Research Report on Emerging Indoor Use Cases	• 실내 무선 통신의 새로운 활용 사례들과 그 요구 사항 • 스마트홈, 몰입형 엔터테인먼트, 첨단 업무환경 등 다양한 6G 실내 시나리오별로 독특한 특성과 필요 요건을 제시하고, 차세대 실내 환경에서 6G 기술이 창출할 기회와 극복해야 할 과제를 논의	실내 무선 통신 유즈케이스, 요구사항

O-RAN nGRG Research Reports (https://www.o-ran.org/research-reports)

문서	요약	기술 영역 분류
Use Case and Gap Analysis for Radio Quality Assurance	• 6G 시대 IoT 및 로봇 도입으로 중요해진 무선 품질(QoS) 보장을 위한 사례와 과제 분석 • 무선 채널 품질(SINR) 변동으로 인한 QoS 저하 문제에 대응하기 위해 사용자 중심 RAN의 필요성을 강조하고, 로봇·미션크리티컬·UAV 등 첨단 응용을 지원하는 데 있어서 현행 O-RAN 표준의 격차와 기술적 도전 과제 식별	무선 기술 및 안테나 시스템
Digital Twin RAN: Key Enablers	• 디지털 트윈 RAN(DT-RAN)의 실현을 위한 주요 기술 요소 정리 • 물리 네트워크 데이터 수집, 고급 모델링 기법, 디지털 트윈을 위한 인터페이스 등 고충실도 DT-RAN 구현에 필요한 요소들을 분석하고, 디지털 영역과 물리 영역 간 원활한 동기화와 상호작용을 위한 기술들을 논의	디지털 트윈
dApps for Real-Time RAN Control: Use Cases and Requirements	• 경량 분산 애플리케이션(dApp)의 실시간 RAN 제어 활용 사례 분석 • 스펙트럼 관리, 스케줄링, 에너지 효율, 트래픽 분류 등 10가지 dApp 활용 시나리오를 제시하고 이를 기반으로 dApp 아키텍처 요구사항을 도출 독립형 dApp과 실시간 RIC 기반 제어 방식을 비교 분석하여 차세대 RAN 실시간 제어 및 최적화 방향을 논의논의	RAN 아키텍처 및 설계, 실시간 제어 및 최적화
Service-based RAN for 6G Network	• 6G 시대의 서비스 기반 RAN 도입 필요성과 요구사항을 분석하고, 모듈화 및 재사용 가능한 서비스 정의 원칙과 제어/사용자 평면 측면의 구현 방향을 제시	RAN 아키텍처 및 설계
Platform Security in Next-Gen Mobile Networks	• 6G 개방형 Disaggregated 네트워크 환경에서의 보안 위협을 분석하고, 보안 통제 및 대응 방안을 제시하여 안전한 네트워크 구현 전략 설명	보안
Research Report on RAN-CN converged Architecture	• RAN과 Core 기능의 융합을 통해 성능, 효율성, 보안성을 향상시키고, 6G 네트워크의 지연 및 오버헤드를 줄이기 위한 아키텍처와 기술 과제를 분석	RAN 아키텍처 및 설계
Requirements and Design Principles on RAN Architecture for Next Generation Antenna Distribution System	• 2040년 대 트래픽 대응을 위한 차세대 DAS 요구사항과 A-RoF 기술 기반 설계 방안 제시	무선 기술 및 안테나 시스템

O-RAN nGRG Research Reports (https://www.o-ran.org/research-reports)

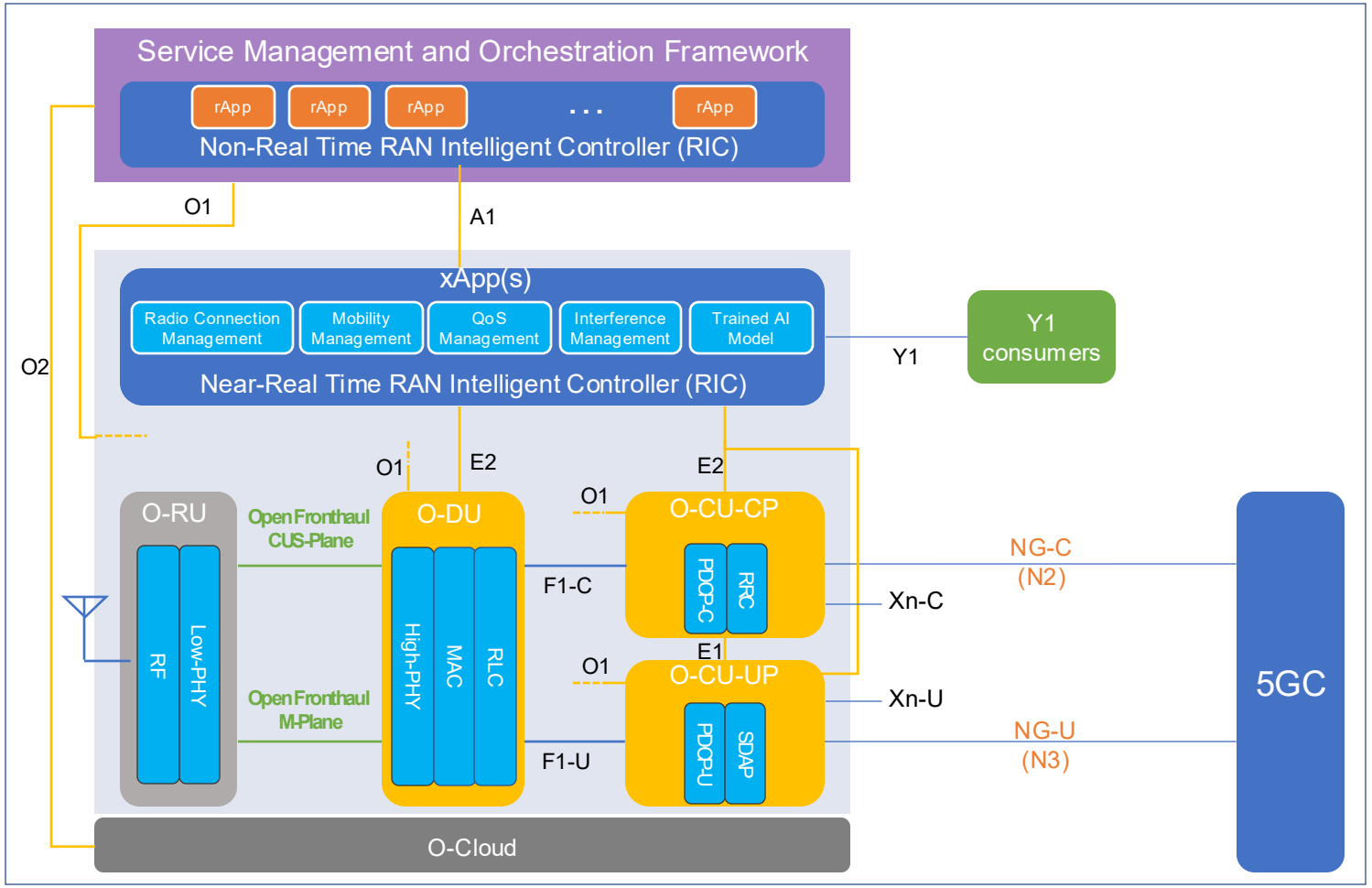


AI-Native O-RAN – Part I

AI-Native, Cross-domain AI, Digital Twin, dApp

RAN의 진화 – O-RAN

개방성, 지능화, 가상화를 핵심 목표로 하는 개방형 무선 액세스 네트워크



rApp(s)
(Non-Real Time RIC)

- 제어 대기 시간: > 1 s
- RAN 정책 관리, 네트워크 트래픽 패턴, 단말 이동성 패턴, 서비스 유형, 서비스 품질(QoS) 예측 패턴 등의 **빅데이터 분석 및 기계학습(Machine Learning)을 통한 인공지능 기반 관리**를 수행

xApp(s)
(Near-RT RIC)

- 제어 대기 시간: 10 ms ~ 1 s
- **단말 단위 부하 조정, 서비스 품질 관리, 핸드오버 등 준실시간 무선 자원 관리 기능**

※ dApp(s) : distributed App(s)의 약자로 O-RAN Alliance의 RR-2024-10 (dApps for Real-Time RAN Control: Use Cases and Requirements)에 포함됨

* CUS: Control User Synchronization, RU: Radio Unit, DU: Distributed Unit, CU: Central Unit

6G를 위한 AI Native 해결 과제

6G에서 AI를 네이티브하게 통합하기 위해서는 도메인 간 협력 부족, 실시간성 한계, 신뢰성과 보안 문제 등 세 가지 주요 과제를 해결해야 함

(1) AI 기능 간 협력 부족 (Lack of Coordination Across Network Domains)

- 5G까지는 RAN, CN, 관리 도메인별로 AI를 독립적으로 논의
- 데이터, 모델, 연산 자원이 도메인 간 공유되지 않음 → E2E AI 불가
- 6G에서는 클라우드-네트워크-단말 간 협력적 AI 인프라 필요

(2) 실시간 AI 성능의 한계 (Insufficient Real-Time AI at the Edge)

- 엣지에서 요구되는 초저지연 응답성 충족 어려움
- 대규모 모델은 연산이 무겁고, 소형 모델은 정확도 부족
- 모델 최적화 + 엣지 AI 배치 전략이 필요

(3) AI의 신뢰성과 보안 확보 과제 (AI Trustworthiness and Security Challenges)

- AI는 불확실성, 비가시성(설명력 부족) 등의 문제 존재
- 데이터 수집 → 처리 → 모델 실행 전 단계에 걸친 보안·검증 체계 필요
- 디지털 트윈을 통한 사전 시뮬레이션 기반 검증이 핵심 해법

Key Technologies – Cross-domain AI Collaboration

6G에서는 AI가 다양한 네트워크 도메인(RAN, CN, 클라우드, 엣지, 단말)에 분산되므로, 도메인 간 AI 기능, 데이터, 자원의 통합 관리와 협업이 필수

(1)E2E AI management and orchestration

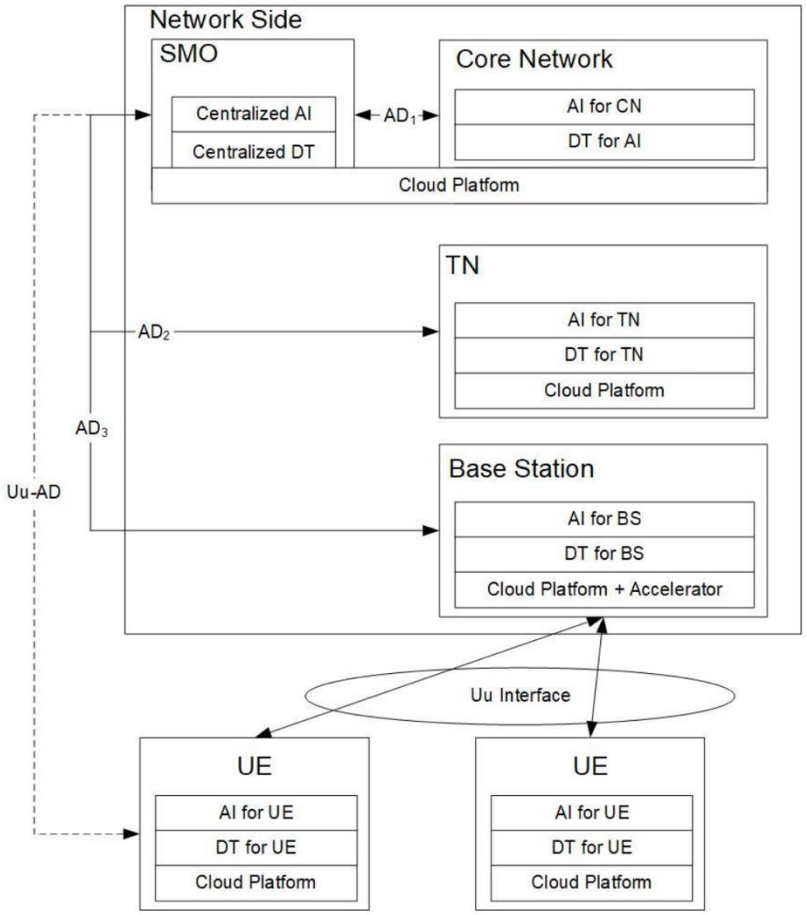
- 네트워크 전 영역에 분산된 AI 자원(데이터, AI 모델, 컴퓨팅 자원)을 통합 관리·오케스트레이션
- Non-RT RIC을 넘어선 상위 제어 계층이 필요하며, AI 서비스의 유연한 배치와 실행 지원

(2)Data and model collaboration across domains

- AI 모델과 데이터가 다양한 도메인 간에 효율적으로 교환되고 벤더간 호환성이 제공될 수 있는 구조가 필요
- 기존 통신 인터페이스는 대규모 데이터 전달 및 모델 전송에 효율적이지 않기 때문에, 별도의 전용 프로토콜과 모델 식별 체계가 필요

(3)Collaboration between large models and other models

- 대형 모델은 클라우드나 중앙 서버에 배치되어, 풍부한 연산 자원을 활용해 범용 지식 학습 및 전역적 의사결정 수행하고 다양한 소형 모델의 기초 모델(Foundation) 역할 제공
- 소형 모델 (Small Models)은 엣지 노드나 단말에 배치되어, 특정 용도에 맞게 경량화·최적화되어 빠르게 추론 수행



Key Technologies – Service-based RAN

RAN과 CN 간의 신호 부하를 줄이고, AI를 활용한 자율 최적화 및 지능형 네트워크 운영을 가능하게 하는 핵심 요소

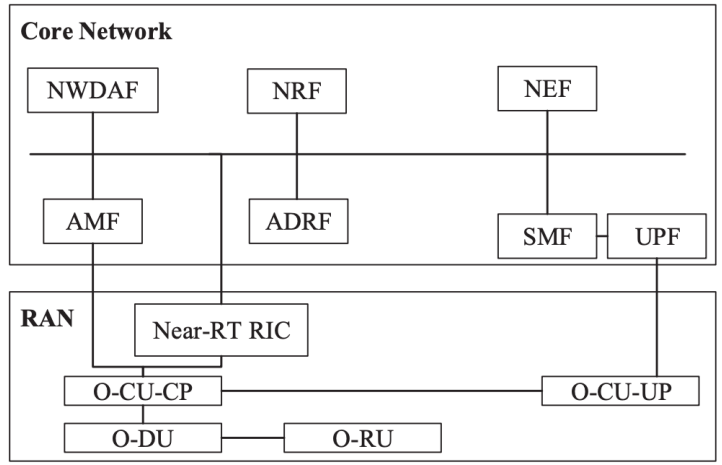
기존 RAN의 한계점

- 구조가 모놀리식(monolithic) → 유연한 서비스 대응 어려움
- 기능 간 결합이 강해 업그레이드, AI 기능 추가에 제약
- CN은 이미 서비스 기반 구조 도입 완료 → RAN도 변화를 요구받음

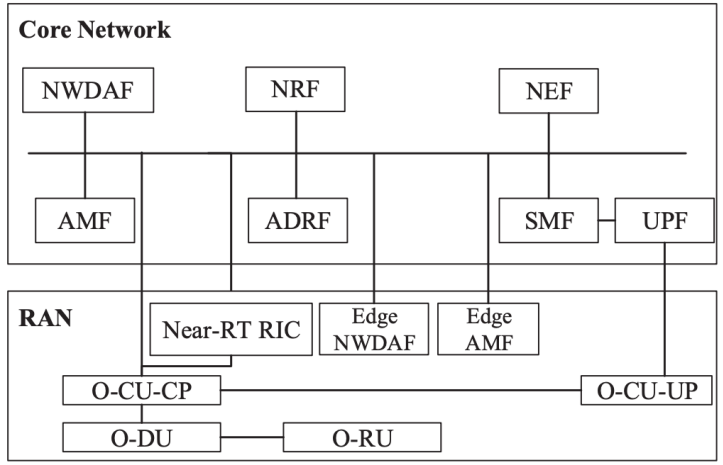
Service-based RAN 구조의 장점

- RAN 기능을 모듈화하여 서비스 단위로 제공
- 유연한 확장과 AI 기반 최적화 가능
- CN과의 연동 효율성 향상, 이기종 서비스에 적응 용이

※ NWDAF가 네트워크 엣지에서 Service-based RAN에 통합 가능



Option 1 - Near-RT RIC만 서비스 버스에 연결



Option 2 - RIC + CU-CP 모두 연결해 더 높은 유연성 확보

Key Technologies – Digital Twin RAN

Digital Twin RAN은 실제 무선망을 가상공간에 실시간으로 복제하여, 실제 RAN 운영에 영향을 주지 않고 예측·최적화·검증이 가능

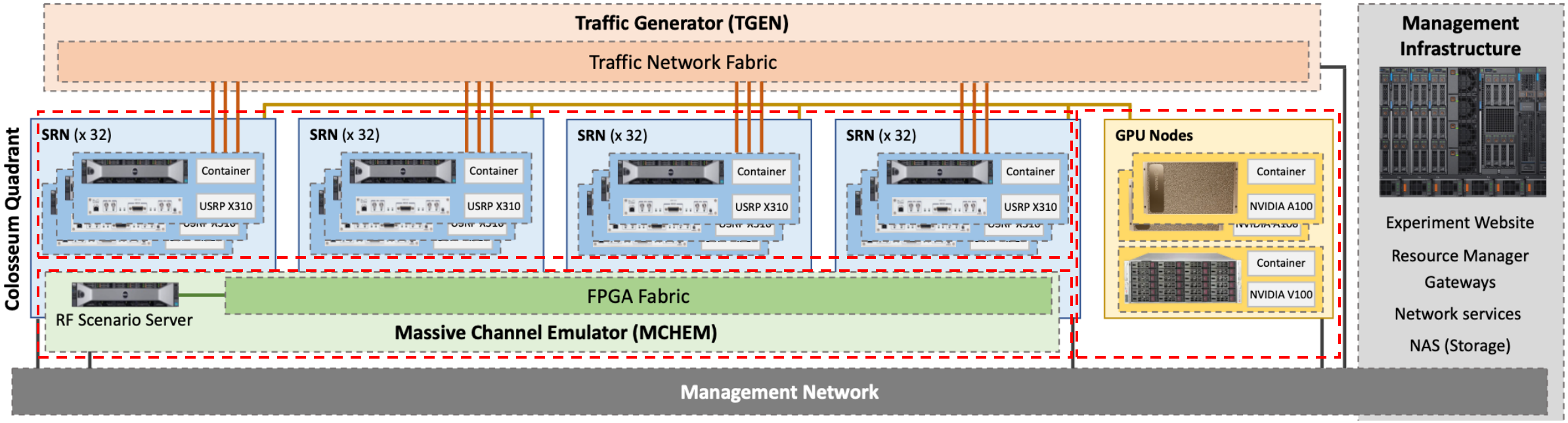
주요 기능 및 역할

1. 실시간 상태 모사
   - 실제 RAN 상태(트래픽, 설정, 품질 등)를 가상 환경에 동기화
2. AI 전략 사전 검증
   - AI가 생성한 최적화 정책을 실제 적용 전 디지털 트윈에서 안전하게 시험 가능
3. 예측 기반 운영
   - 이상 감지, 자원 병목 예측, 서비스 품질 향상에 활용
4. 지속적 네트워크 최적화
   - 자원 활용 효율 향상 → 성능 개선 + 운영 비용 절감

Key Technologies – Digital Twin RAN (Colosseum)

DT-RAN for Network Testing Automation

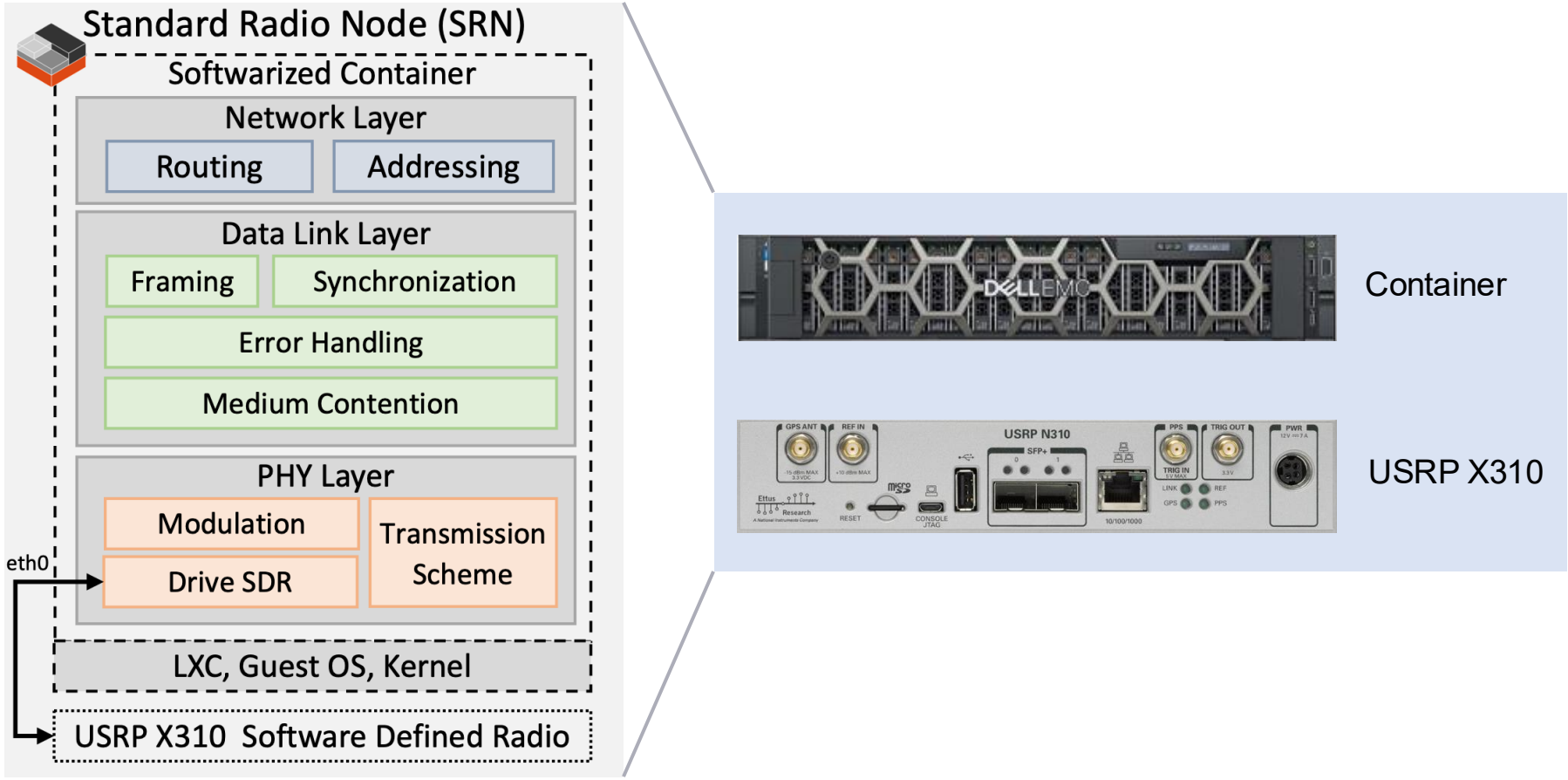
- 총 128쌍의 SDR(Software Defined Radio)과 서버로 구성된 SRN(Standard Radio Node)
- 실시간 채널 에뮬레이션 시스템 (MCHEM, Massive Channel Emulator) : 4개의 ATCA 섀시
- 고성능 AI/ML 인프라



Key Technologies – Digital Twin RAN (Colosseum)

SRN (Standard Radio Node)

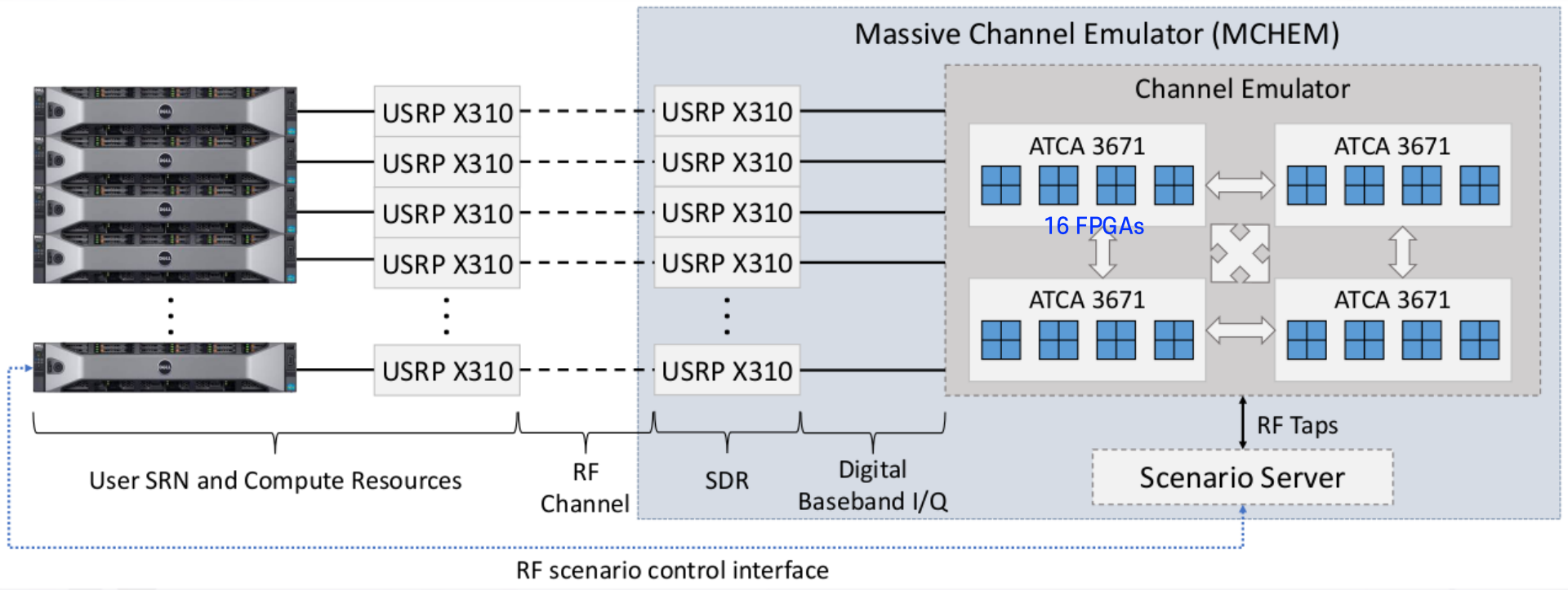
- Dell 서버 + NI/Ettus USRP X310 (10 Gbps 링크)
- 사용자 지정 혹은 사전 구성 프로토콜 스택 배포 (LXC 컨테이너 기반)



Key Technologies – Digital Twin RAN (Colosseum)

Massive Channel Emulator

- 4개 사분면 구성, 각 사분면 당 32개의 SDR & 하나의 ATCA 섀시(16개 FPGAs) → 총 128 SDRs + 64 FPGAs
  ※ 하나의 섀시가 32개의 SDR에 대한 채널 에뮬레이션 수행
- 각 SDR 트랜시버 간에는 실시간 다중경로 무선 채널 모델링 기법(Tapped Delay Line, TDL)로 모델링
  ※ 무선 통신 채널에서 발생하는 '다중경로 현상(multipath propagation)'을 수학적으로 표현하기 위한 모델
  ※ 최대 4개의 multipath 구성 요소 (1ms 단위 CIR)



CIR (Channel Impulse Response) : **무선 채널이 신호를 얼마나, 어떻게 왜곡시키는지를 나타내는 특성**

Key Technologies – Digital Twin RAN (Colosseum)

실제 무선 환경을 디지털 트윈으로 변환하는 과정

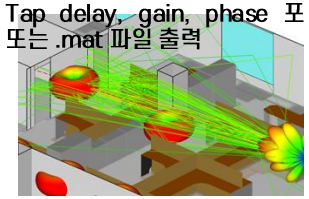
- 무선 환경의 3D 모델을 확보하여 레이 트레이서에 로드
- 송수신 노드의 위치와 이동 경로를 설정하고, 채널 특성을 시간별로 샘플링한 뒤 결과를 파싱
- 파싱된 데이터를 채널 에뮬레이터에 적합한 형식으로 변환하여 디지털 시나리오로 완성

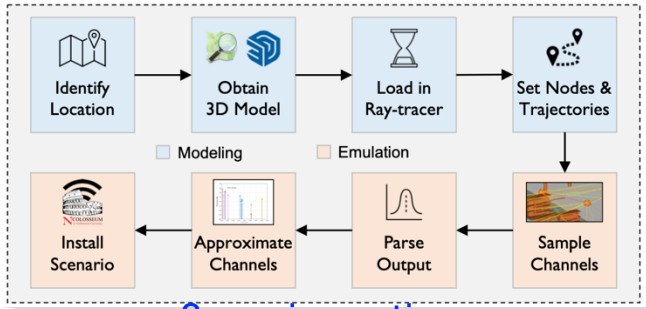
- 송신 노드는 미리 정의된 코드 시퀀스를 전송하고, 수신 노드는 이 신호를 수신하여 원래 신호와 Cross-correlation을 통해 실제 채널 응답(CIR)을 추정
- 추정된 채널 응답이 설계된 시나리오와 얼마나 일치하는지를 비교함으로써, 디지털 트윈 시나리오의 정확성과 신뢰성을 검증

1.도시/건물의 3D 모델 확보
- 드론 스캔, LiDAR, 공개 GIS 기반 모델 활용

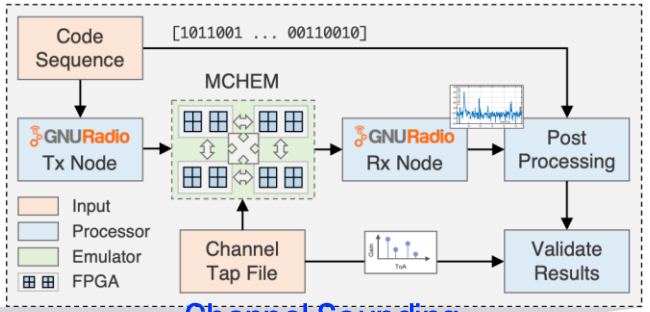
2.사용자 이동 궤적 반영한 채널 생성
- 시뮬레이션 시간에 따라 동적으로 채널 추출

3.CIR 데이터 → 에뮬레이터용 파일로 저장
- Tap delay, gain, phase 포함된 .tap 또는 .mat 파일 출력

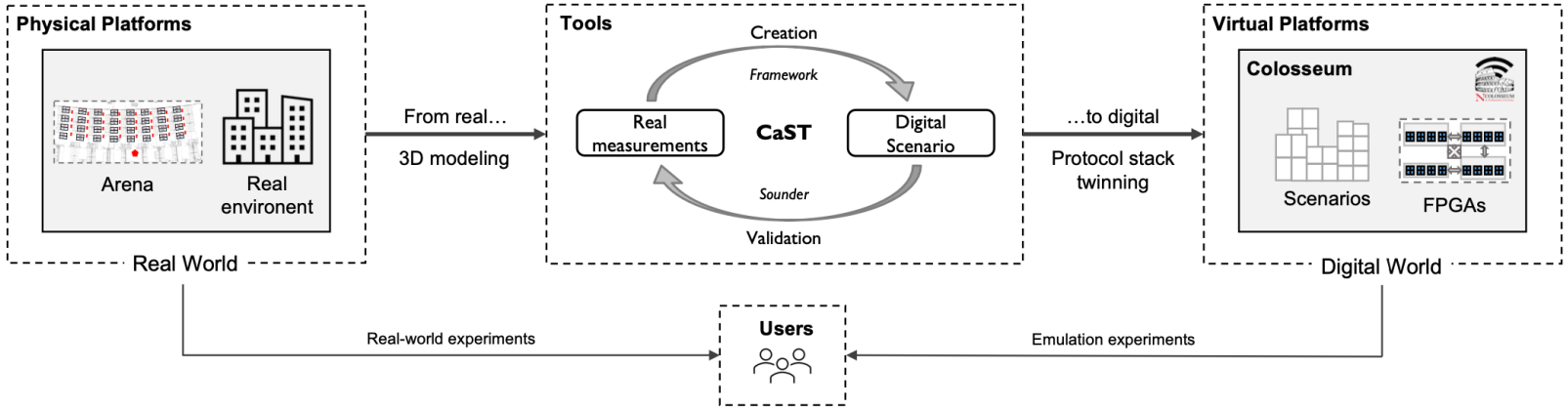




Scenario creation



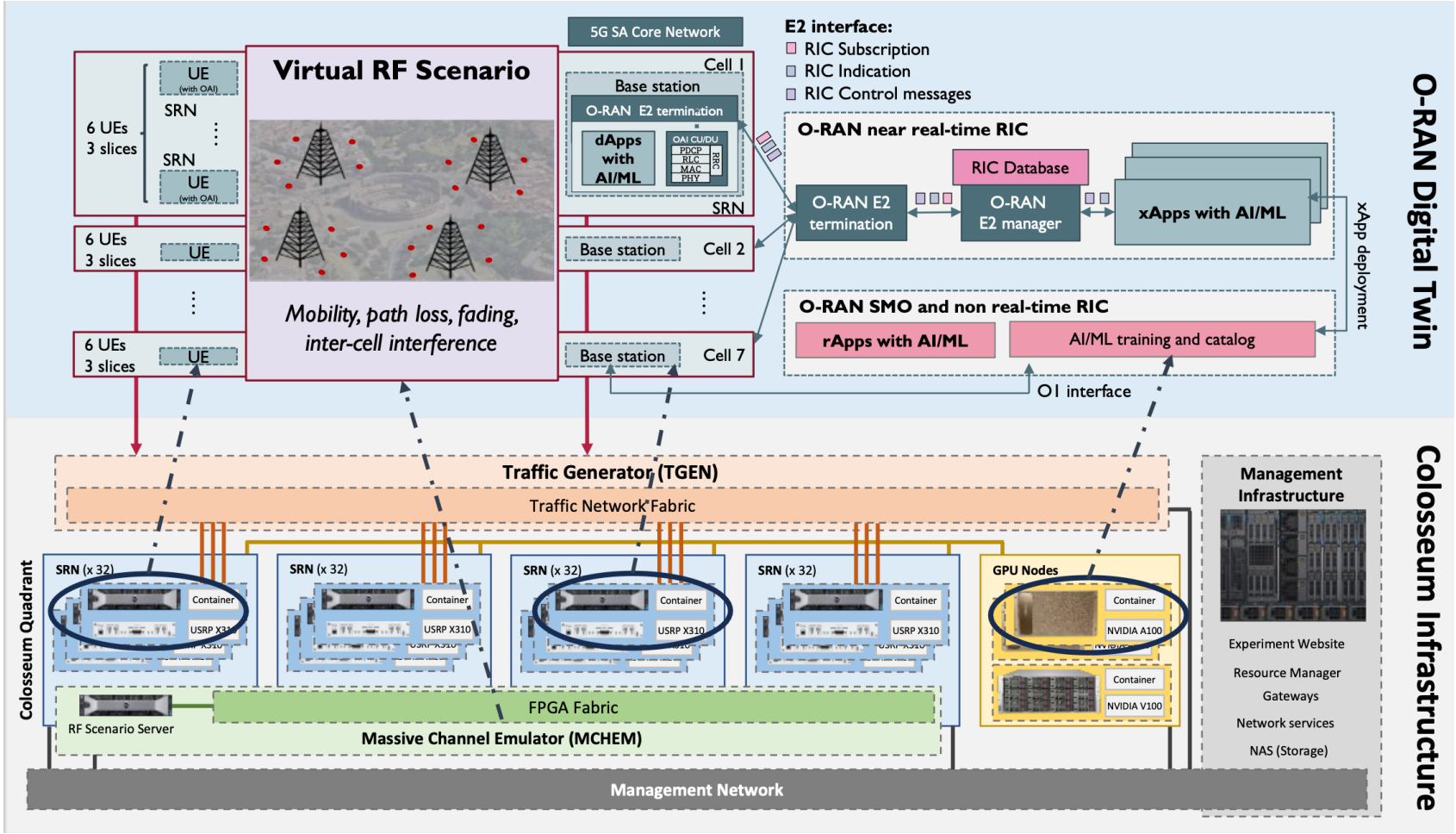
Channel Sounding



High-level representation of a DT

Key Technologies – Digital Twin RAN (Colosseum)

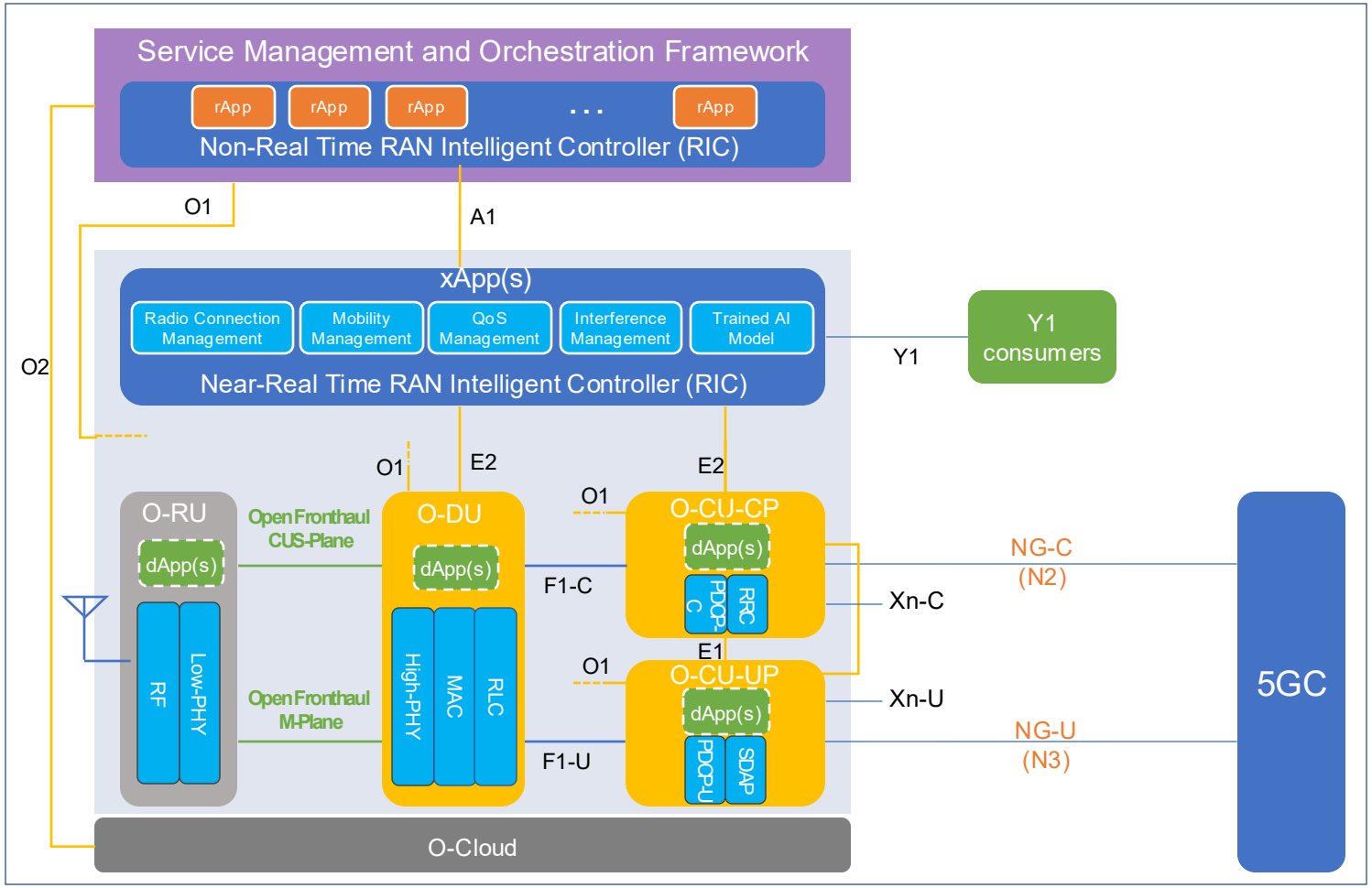
SRN, MCHEM, TGEN, GPU 노드를 활용해 실제 무선 채널 환경을 디지털 트윈으로 재현



기지국 단에서 실시간 제어를 수행하는 분산형 AI 애플리케이션(dApp)부터 준실시간 RIC(xApp) 및 비실시간 RIC(rApp)에 이르기까지 다양한 계층의 AI/ML 실험을 안전하고 확장 가능하게 지원

Key Technologies – Realtime Control

실시간 추론 및 제어 기능을 위한 O-RAN 확장: dApp



rApp(s)
(Non-Real Time RIC)

- 제어 대기 시간: > 1 s
- RAN 정책 관리, 네트워크 트래픽 패턴, 단말 이동성 패턴, 서비스 유형, 서비스 품질(QoS) 예측 패턴 등의 빅데이터 분석 및 기계학습(Machine Learning)을 통한 인공지능 기반 관리를 수행

xApp(s)
(Near-RT RIC)

- 제어 대기 시간: 10 ms ~ 1 s
- 단말 단위 부하 조정, 서비스 품질 관리, 핸드오버 등 준실시간 무선 자원 관리 기능

dApp(s)

- 제어 대기 시간: 1 ms ~ 10 ms
- 자원 스케줄링, 빔포밍 등 실시간 무선 자원 제어 기능

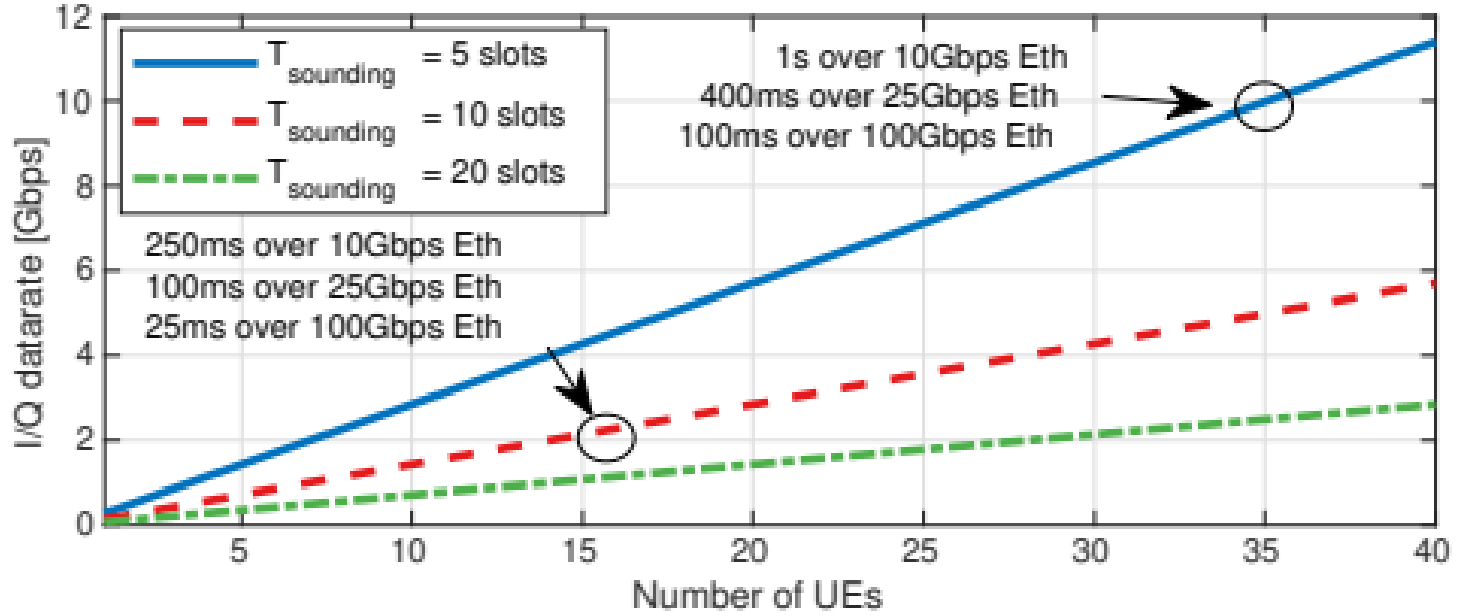
※ dApp(s) : distributed App(s)의 약자로 O-RAN Alliance의 RR-2024-10 (dApps for Real-Time RAN Control: Use Cases and Requirements)에 포함됨

* CUS: Control User Synchronization, RU: Radio Unit, DU: Distributed Unit, CU: Central Unit

Key Technologies – Realtime Control

실시간 추론 및 제어를 위한 기존 O-RAN xApp/rApp 아키텍처의 한계

- 주로 제어 평면(Control Plane) 에만 초점을 맞추고 있어, **사용자 평면(User Plane) 에 기반한 데이터 기반 제어/추론 불가능**
- I/Q 샘플이나 사용자 평면 데이터(PDUs 등)에 접근시 보안 및 개인정보 보호(privacy) 이슈
- Near-RT RIC가 RAN 외부(엣지 클라우드 등)에 위치할 경우, **E2 인터페이스를 통한 제어/측정 지연 발생**

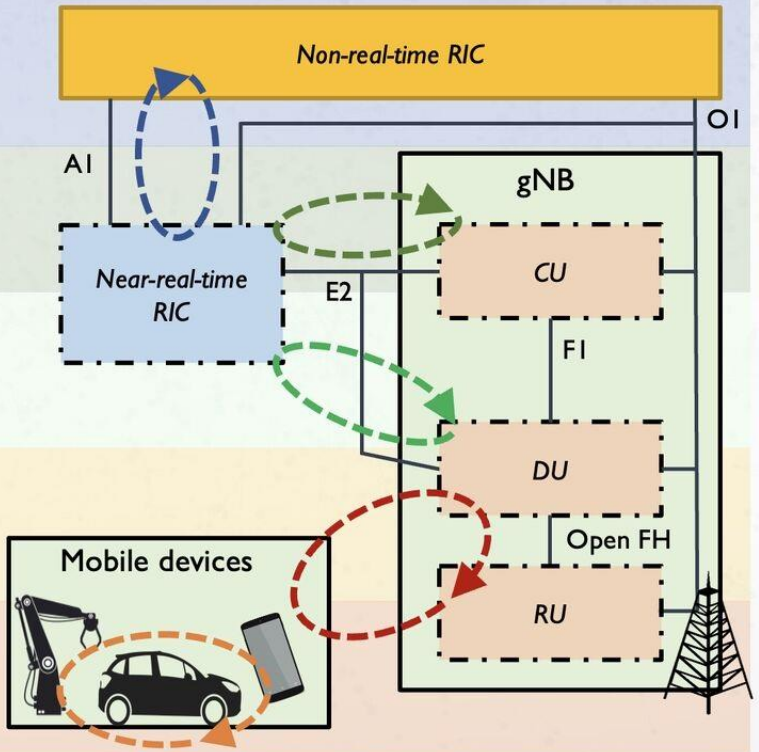


E2 인터페이스를 통한 I/Q 샘플 전송은 10ms 이하의 실시간 요구사항을 충족하지 못함

Key Technologies – Realtime Control

실시간 추론 및 제어 기능을 위한 O-RAN 확장: dApp

	Objective	Scale	Input data	Timescale
rApps	Policies, models, slicing	> 1000 devices	Infrastructure-level KPIs	> 1 s Non real-time
xApps	User Session Management e.g., load balancing, handover	> 100 devices	CU-level KPIs e.g., number of sessions, PDCP traffic	10-1000 ms
xApps	Medium Access Management e.g., scheduling policy, RAN slicing	> 100 devices	MAC-level KPIs e.g., PRB utilization, buffering	10-1000 ms Near real-time
dApps	Radio Management e.g., resource scheduling, beamforming	~10 devices	MAC/PHY-level KPIs e.g., PRB utilization, channel estimation	< 10 ms
dApps	Device DL/UL Management e.g., modulation, interference, blockage detection	1 device	I/Q samples	< 1 ms Real-time



dAPP은 xApps/rApps를 보완하여 CU/DU에 AI를 도입하고, RIC보다 더 짧은 시간 척도에서 하위 계층 기능에 대한 실시간 추론 및 제어를 지원하는 분산 애플리케이션

Key Technologies – Realtime Control

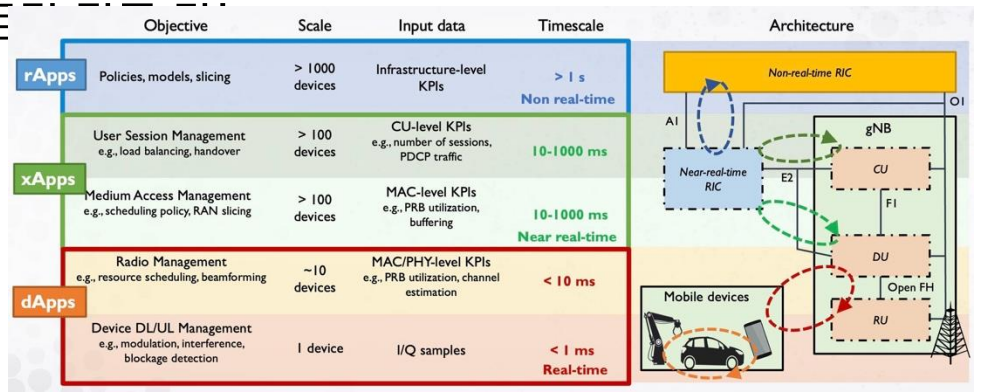
dApp의 주요 특징

1. **지연 및 오버헤드 감소 (Reduced Latency and Overhead)**
   - dApp은 엣지의 CU/DU에서 직접 실행 가능
   - 기존 대비 약 3.57배 오버헤드 감소 효과
   - 중앙 집중형 구조 대비 빠른 반응성과 경량화된 운영 가능

2. **MAC 및 PHY 계층 제어 (Controlling MAC- and PHY-Layer Functionalities)**
   - 빔포밍, 변조 방식, RF 핑거프린팅 등 물리계층 기능을 소프트웨어적으로 제어
   - I/Q 샘플 기반의 세밀한 무선 최적화 가능

3. **CU/DU 데이터 및 기능의 실시간 접근 (Access DU/CU Data and Functionalities in Real Time)**
   - I/Q 샘플, 데이터 패킷, UE 핸드오버 정보 등 세부 네트워크 메트
   - 이를 활용한 실시간 네트워크 조정 및 예측 제어

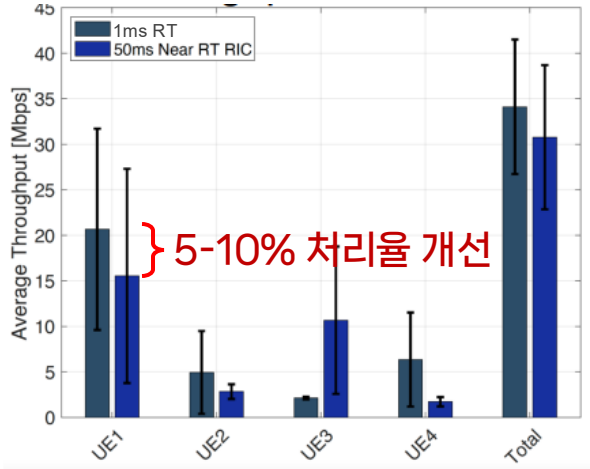
4. **확장성 및 재구성 (Extensibility and Reconfigurability)**
   - O-RAN 표준 인터페이스를 기반으로 하여 모듈화된 구조 지원
   - 하드웨어 가속기와의 연계 가능
   - DU/RU 내 AI 통합을 위한 전환점 역할 수행



Key Technologies – Realtime Control

dApp 유즈케이스

- **물리 계층 보안**
  - O-DU 내 dApp이 실시간으로 I/Q 샘플을 분석하여 이상 탐지, 재밍, 스푸핑 시도를 무선 계층에서 직접 감지
- **스펙트럼 센싱**
  - 실시간 스펙트럼 데이터를 활용해 간섭이나 스펙트럼 공백을 감지하고, 전력 조정, PRB 블랭킹, 빔포밍 등 즉각적인 대응 수행
- **실시간 스케줄링 재구성**
  - MAC/PHY 계층 데이터를 바탕으로 O-DU의 스케줄러 설정을 TTI 단위로 동적으로 조정하여 네트워크 상황에 최적화된 스케줄링 수행



실시간 Resource Block 스케줄링 (<1ms latency) vs. 준실시간 스케줄링 (50ms latency)

Key Technologies – Realtime Control

dApp 유즈케이스

- **실시간 스케줄링 가속화**
  - 복잡한 MAC 스케줄링 연산(예: UE 그룹핑, PRB 할당)을 GPU/CPU에서 병렬 처리하여 실시간 계산을 가속화
- **실시간 스케줄링 조율(Coordination)**
  - 여러 사업자가 자원을 공유할 수 있도록 각 dApp 간 실시간 자원 사용 정보를 교환하여 간섭 없이 스펙트럼을 효율적으로 조율
- **O-DU 에너지 절감**
  - 실시간 트래픽 부하를 분석하여 필요한 CPU 코어 수를 예측하고, 남는 코어는 자동으로 비활성화해 에너지를 절감
- **AI/ML 기반 CSI 피드백 및 채널 최적화**
  - TTI 단위로 CSI 압축, 채널 추정 등을 위한 AI 모델을 학습·추론하여 링크 성능을 실시간으로 향상
- **원거리 간섭 탐지 및 완화**
  - 원거리 간섭을 감지하고 가드 구간 확장 등 즉각적인 대응을 수행하며, 필요시 xApp과 협력하여 외부 기지국과 간섭 완화를 조율
- **업링크 처리량 최적화**
  - 채널 품질과 UE 이동성 정보를 기반으로 최적 스트림 수를 결정하여 업링크 압축률을 조정함으로써 처리량을 개선

Key Technologies – Realtime Control

dApp 유즈케이스

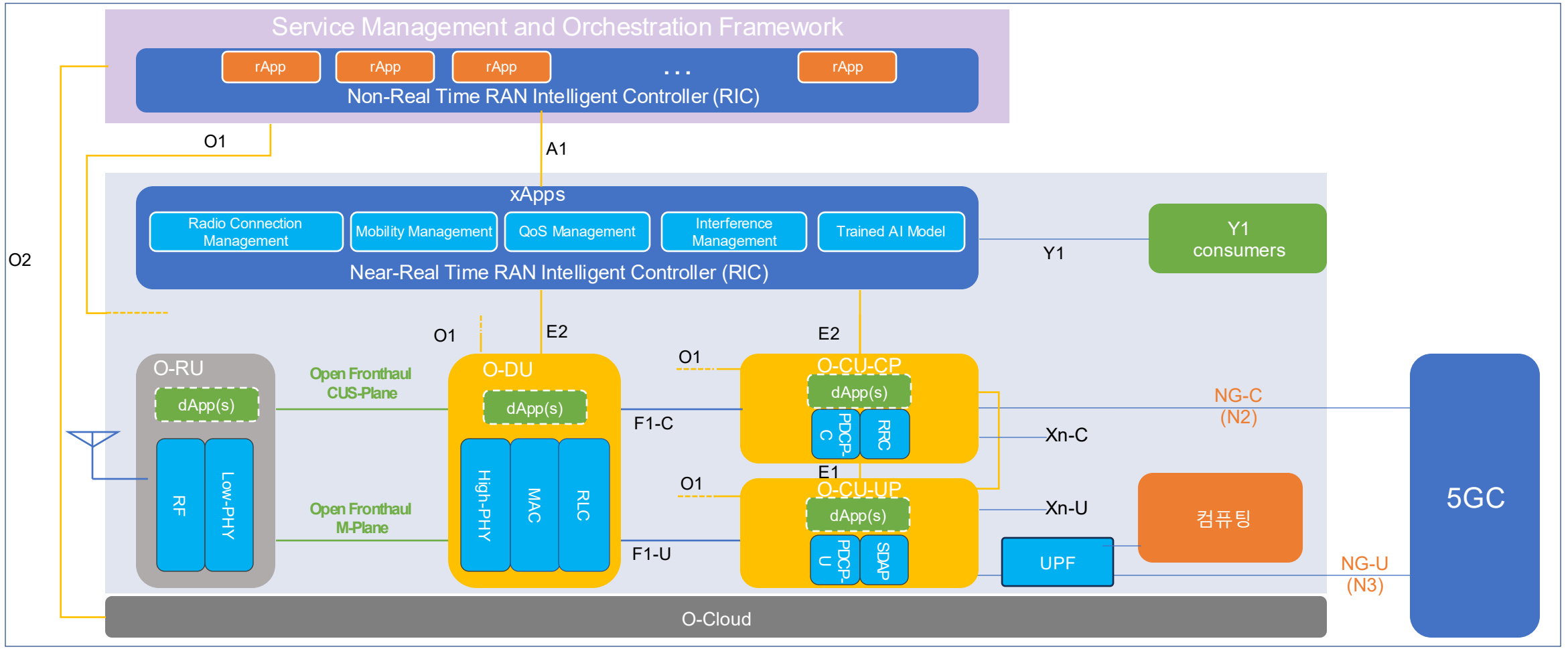
- **빔 관리 및 최적화**
  - O-DU 내에서 직접 빔 선택 및 사운딩 신호 구성을 제어하고, 실시간으로 precoding weight를 계산해 성능과 간섭을 최적화
- **통합 통신 및 센싱(ISAC)**
  - 레퍼런스 신호나 I/Q 샘플을 활용해 환경 정보를 추출하고, 이를 기반으로 CSI 추정 정확도를 향상시키고 오버헤드를 줄임
- **트래픽 분석 및 분류**
  - 사용자 패킷을 실시간으로 분석해 애플리케이션 유형을 식별하고 적절한 슬라이스로 동적으로 재매핑하여 QoS를 정밀하게 제어

AI-Native O-RAN – Part II

Communication & Computing Integrated RAN

Communication & Computing Integrated RAN

동일 인프라 내에서 RAN Communication과 Computing을 동시에 수행하기 위한 확장



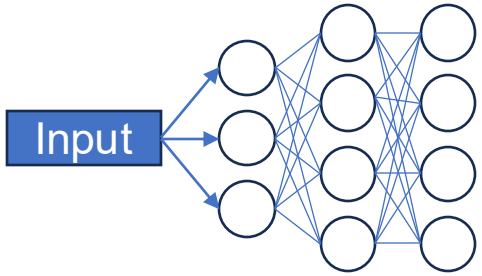
* CUS: Control User Synchronization, RU: Radio Unit, DU: Distributed Unit, CU: Central Unit

유즈케이스: 단말-RAN AI/ML 협력 추론

협력 AI 추론을 위한 End-to-End Latency : T1 + T2 + T3 + T4

Model layer 1 to 10 are computed at mobile device

Computing latency: T1

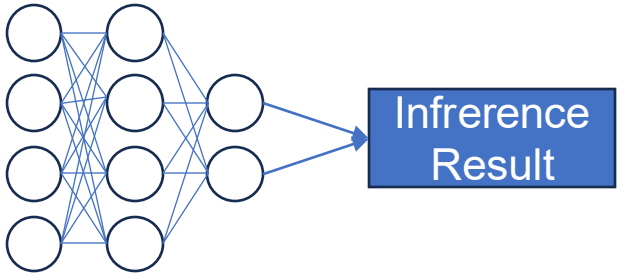


AI/ML model intermediate outcomes transmission latency: T2

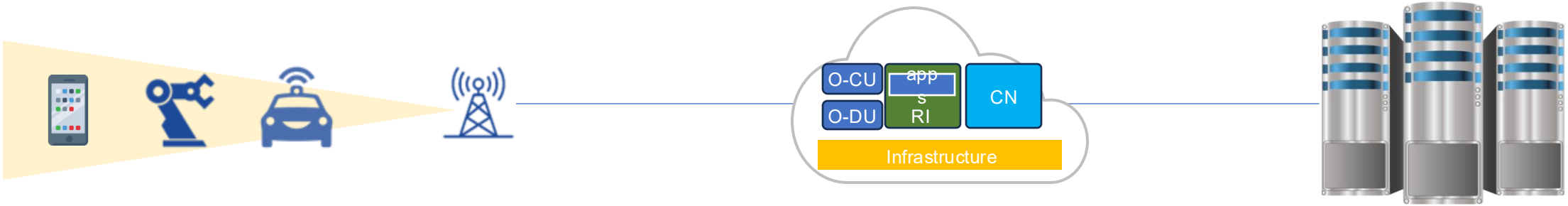
AI Inference result transmission latency: T4

Model layer 11 to output are computed at edge/cloud server

Computing latency: T3



단말 처리 지연 | **전달 지연** | **에지/클라우드 컴퓨팅 지연**



유즈케이스: 단말-RAN AI/ML 협력 추론

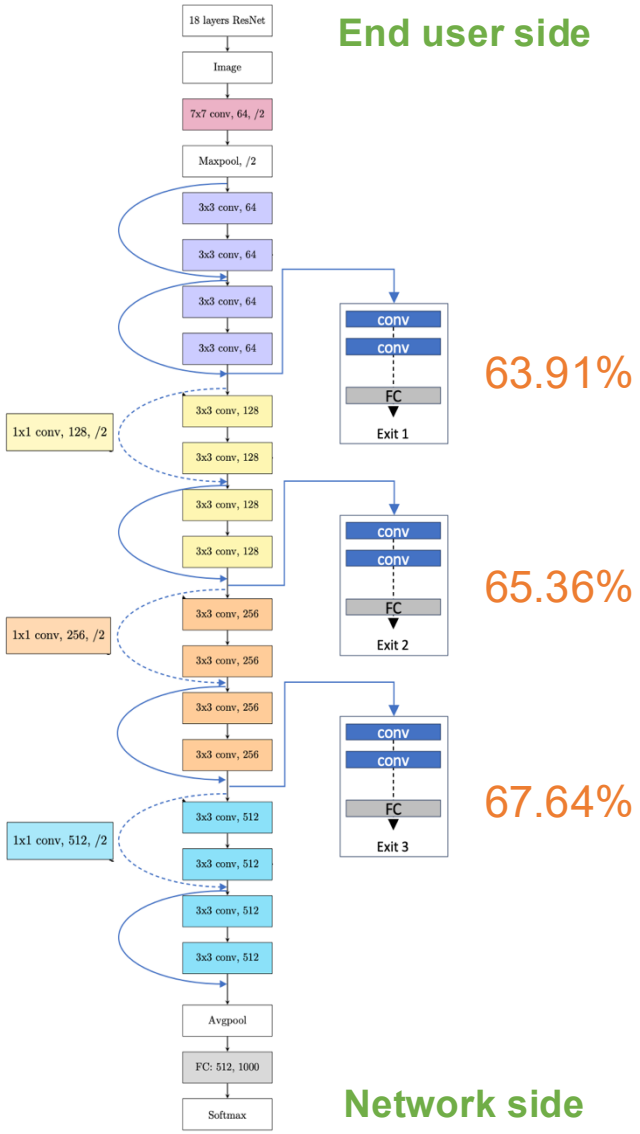
제한된 컴퓨팅 자원을 갖는 단말의 클라우드/에지컴퓨팅에 AI/ML 추론 오프로딩

Mobile devices, Robotics, Cameras 등

AI/ML edge servers

Cloud servers

Application servers



End user side

63.91%

65.36%

67.64%

Network side

Multi-branch ResNet-18 model

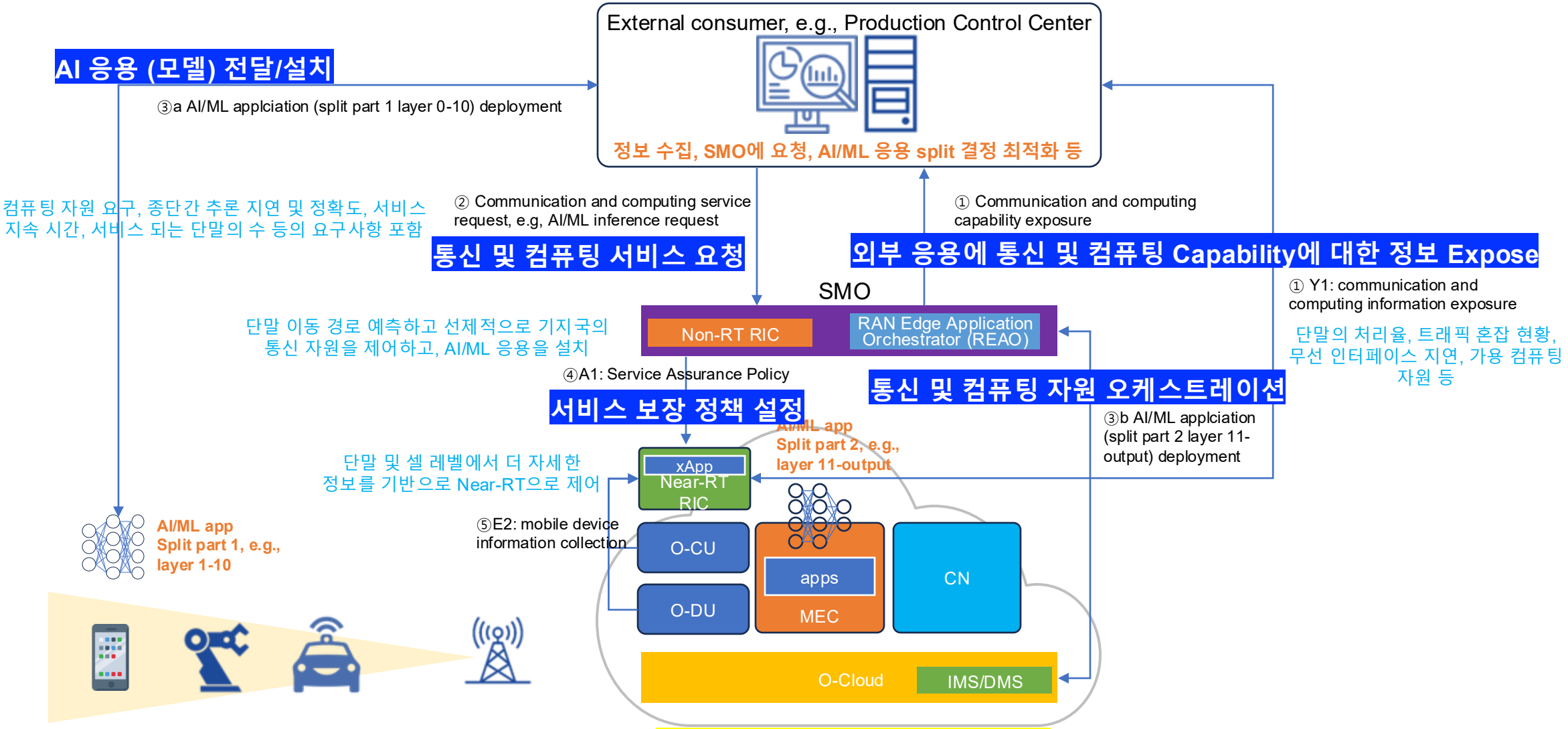
중간 결과 데이터 크기

Layer ID	Type	Data size (bit)	Computation load (FLOPs)	Accuracy of exit point
1	conv1	25690112	118013952	
	maxpool	6422528	7225344	
2	conv2-1	6422528	115605504	
3	conv2-2	6422528	115605504	
4	conv2-3	6422528	115605504	
5 (Exit1)	conv2-4	6422528	115605504	63.91%
6	conv3-1	3211264	57802752	
7	conv3-2	3211264	115605504	
8	conv3-3	3211264	115605504	
9 (Exit2)	conv3-4	3211264	115605504	65.36%
10	conv4-1	1605632	57802752	
11	conv4-2	1605632	115605504	
12	conv4-3	1605632	115605504	
13 (Exit3)	conv4-4	1605632	115605504	67.64%
14	conv5-1	802816	57802752	
15	conv5-2	802816	115605504	
16	conv5-3	802816	115605504	
17	conv5-4	802816	115605504	
	avg pool	16384	1229312	
18 (Exit4)	fc	32000	1023000	68.96%

Data size and computation load of ResNet18

유즈케이스: 단말-RAN AI/ML 협력 추론

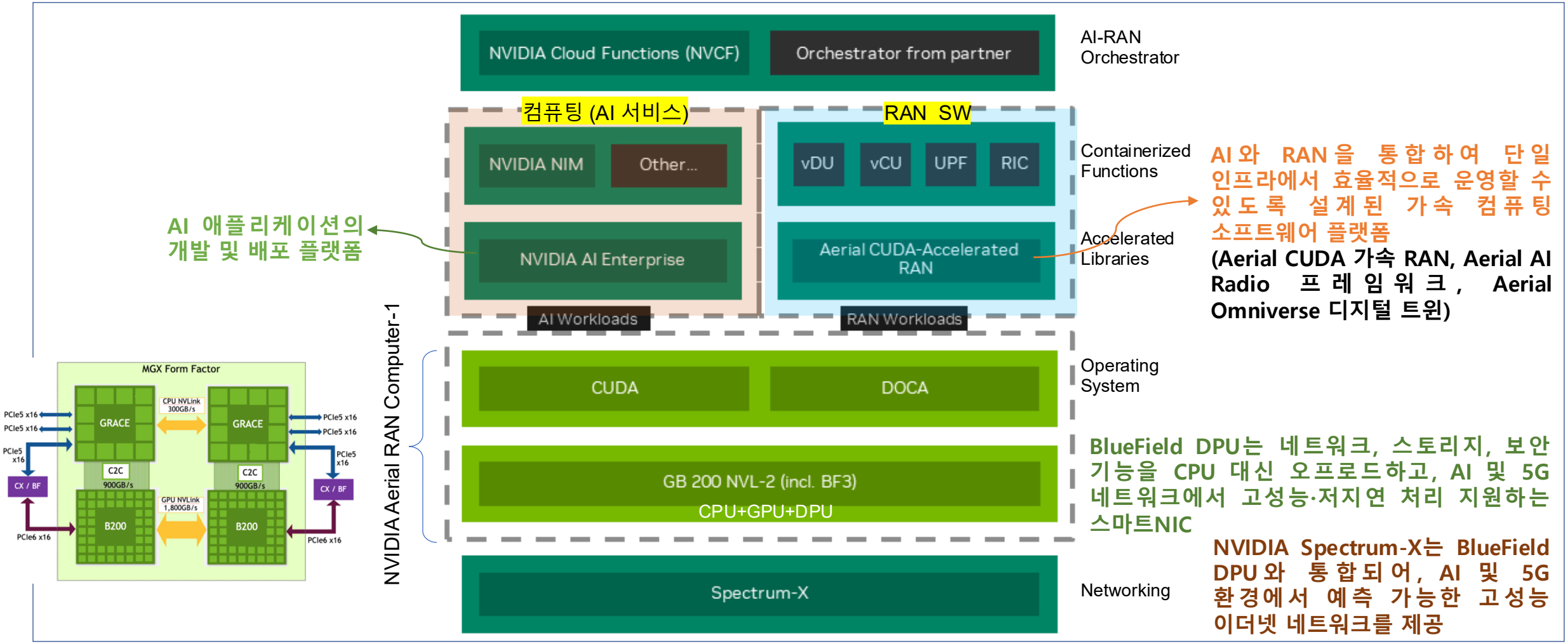
별도의 Edge/Cloud 서버가 아닌 O-Cloud의 컴퓨팅 자원을 이용한 AI/ML 협력 추론



* IMS: Infrastructure Management Service, DMS: Deployment Management Services

Platforms

NVIDIA Grace Hopper/Blackwell 수퍼칩 및 Aerial SDK를 활용하여 L1-L3 RAN 프로세싱과 AI/ML 모델을 통합



NVIDIA Aerial RAN Compute-1 (ARC-1)

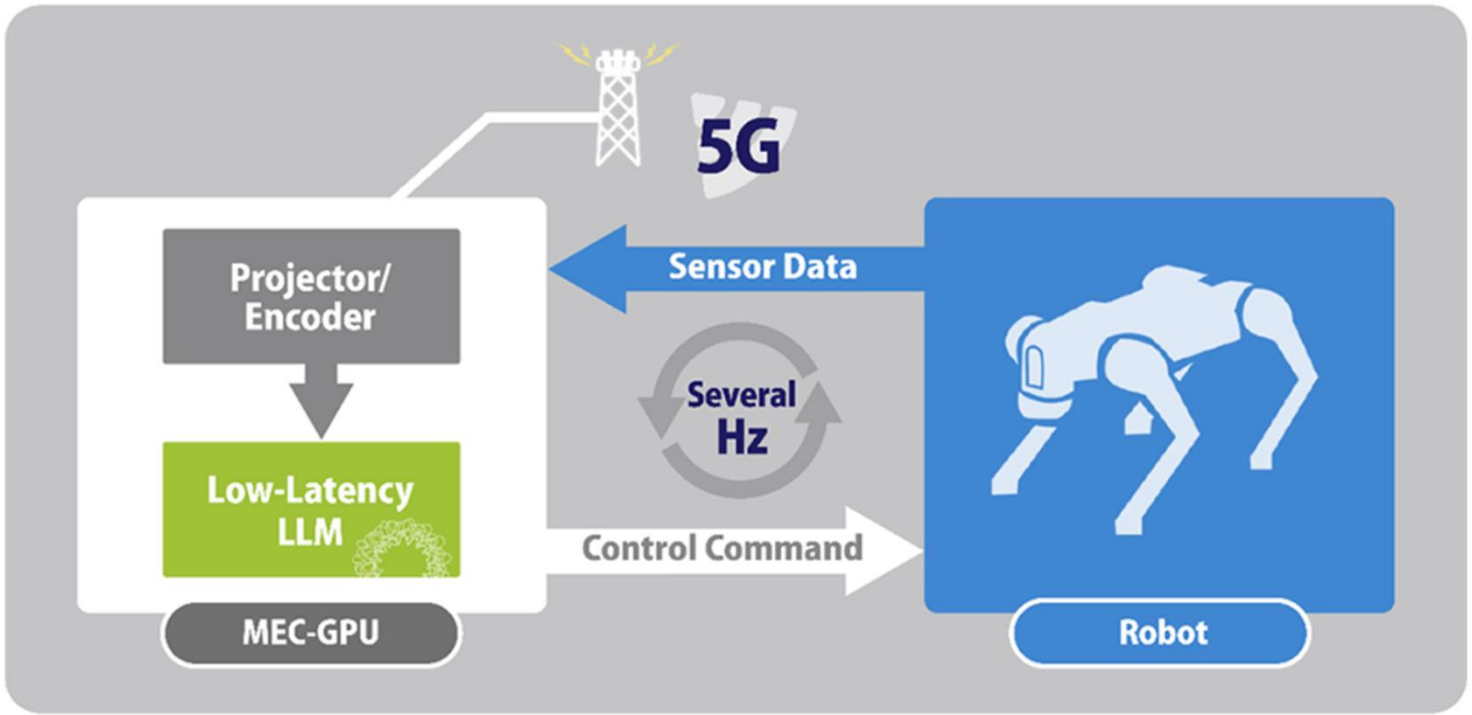
유즈케이스 분석 기반 AI-RAN 이해

AI-on-RAN 중심

초저지연 LLM

실시간 로봇 동작 생성을 위한 초저지연 LLM

※ LLM 모델은 엣지 AI 서버에서 동작하며, 로봇의 센서 데이터를 받아서 제어 로봇 제어



Cloud Robot Utilizing Ultra-low-latency LLMs

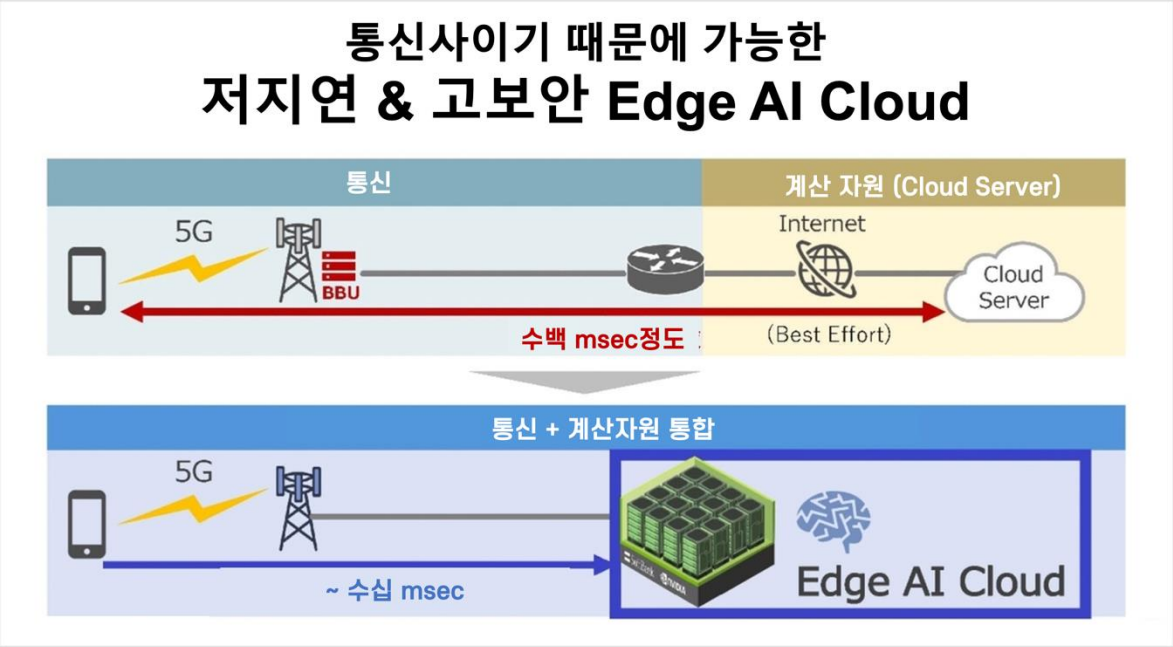
- 지연 민감(real-time) 로봇 제어에 LLM을 클라우드 대신 MEC에서 처리
- 기존의 단순 inference가 아닌, 복합 reasoning 기반의 로봇 제어 가능성 확보
- 향후 산업 로봇, 서비스 로봇, 재난 대응 드론 등에 확장 가능

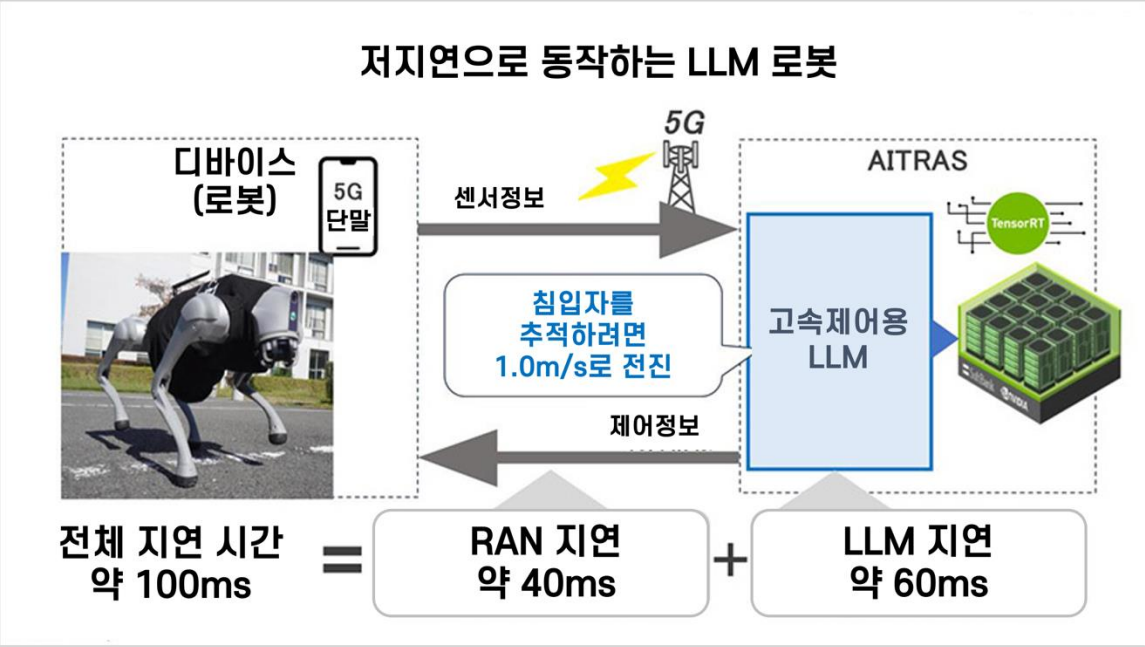
https://www.softbank.jp/en/corp/news/press/sbkk/2024/20241113_01/

초저지연 LLM

실시간 로봇 동작 생성을 위한 초저지연 LLM

※ LLM 모델은 엣지 AI 서버에서 동작하며, 로봇의 센서 데이터를 받아서 제어 로봇 제어

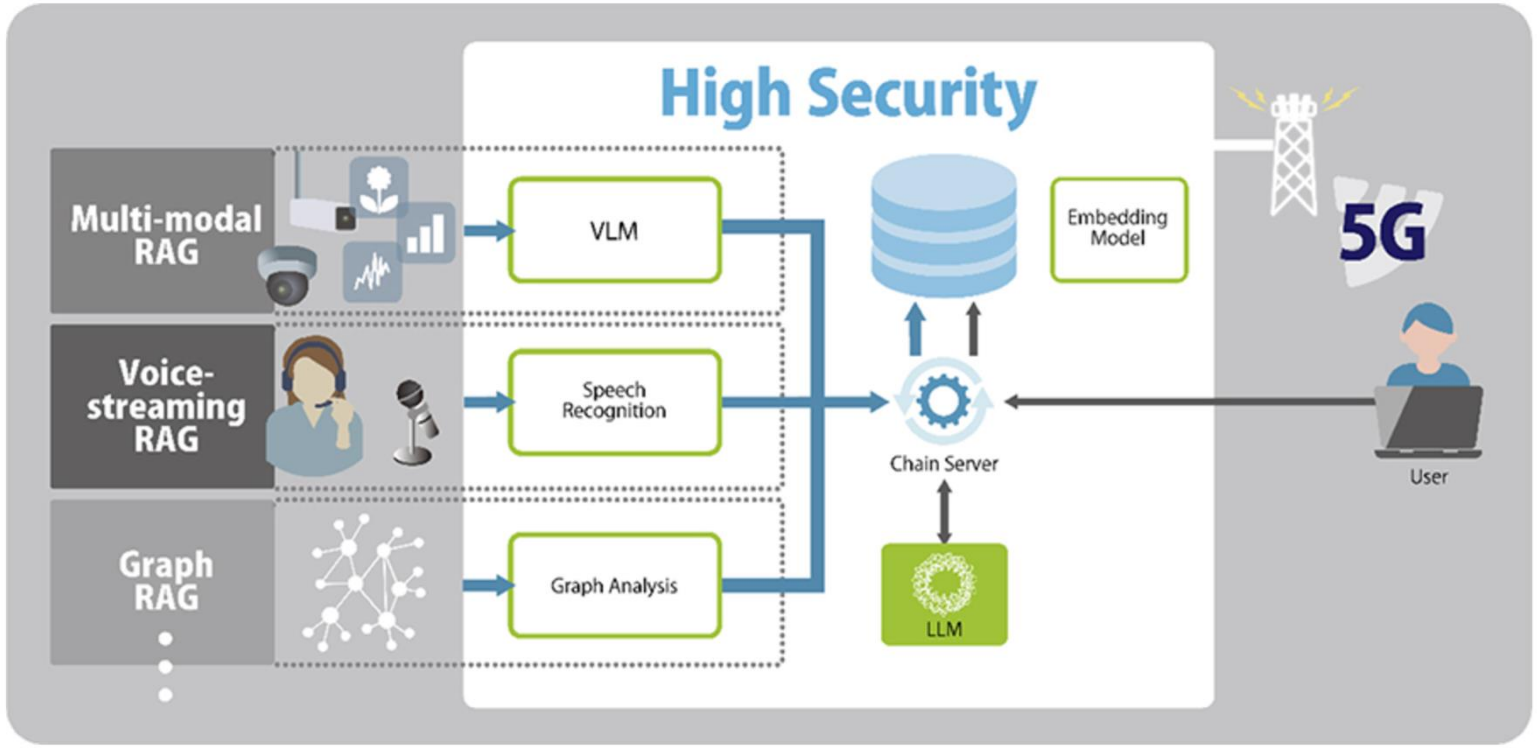




RAG(Retrieval-Augmented Generation)

RAG는 외부 데이터베이스 또는 문서에서 필요한 정보를 검색(Retrieval)하여, 이를 생성형 AI 모델(LLM)에 입력하여 결과를 생성하는 방식

※ LLM의 기존 학습 데이터만 사용하는 것이 아니라, 최신 정보나 특정 도메인의 정보를 실시간으로 반영 가능



1) **Knowledge Base 준비**
   - ✓ 문서(텍스트)를 Embedding Model을 통해 벡터화되어 고속 검색 데이터베이스로 저장
2) **Query Embedding**
   - ✓ 사용자가 질의를 입력하면, 입력에 맞는 전처리 동작(Speech Recognition, VLM, Graph Analysis)
   - ✓ 텍스트화된 사용자 입력을 동일한 임베딩 모델로 벡터화
3) **Similarity Search**
   - ✓ 쿼리 벡터와 DB에 저장된 문서 벡터 간의 가장 가까운 문서 검색
4) **Retrieved 문서 + Query → LLM 입력**
   - ✓ 선택된 문서가 LLM의 컨텍스트로 들어가서 답변 생성

- **Multi-modal RAG:** 이미지, 센서, 카메라 등 다양한 데이터 → VLM(Vision-Language Model)로 처리
- **Voice-streaming RAG:** 음성 스트림 → Speech Recognition을 통해 텍스트로 변환
- **Graph RAG:** 관계망, 트랜잭션 등 → Graph 분석으로 구조화된 정보 도출

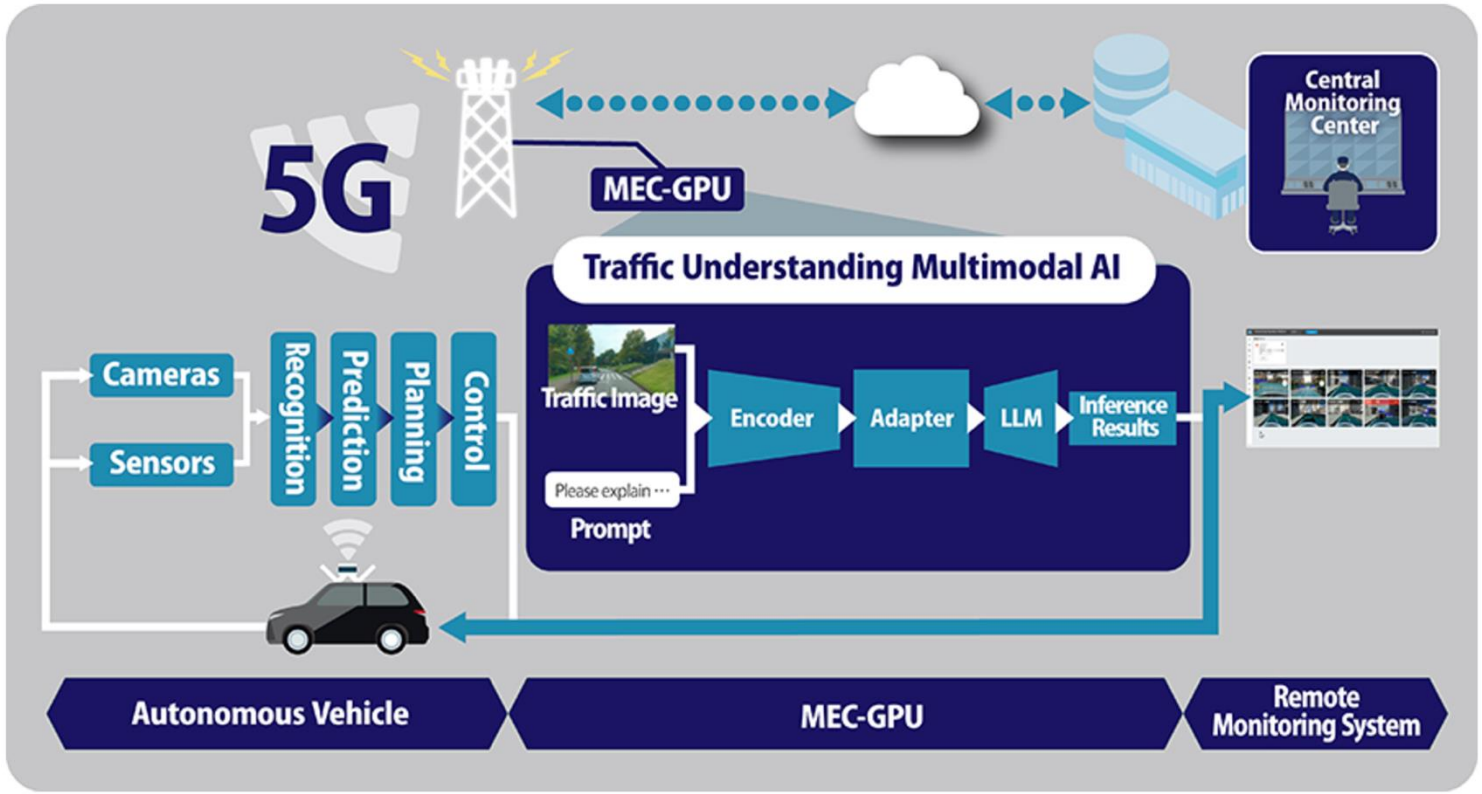
<의료 Graph RAG>

사용자 질문: "당뇨병 치료에 효과적인 약은?"
1.임베딩을 통해 KG에서 관련 노드 검색
2."당뇨병" → 관련 질병 노드, "메트포르민", "인슐린" 등 약물 노드 추출
3.그래프 서브셋을 기반으로 LLM에 컨텍스트 제공
4.GPT가 "당뇨병 치료에는 메트포르민이 1차 약제로 많이 사용되며, …"와 같이 응답 생성

Traffic Understanding Multimodal AI

현재의 “교통 상황”, “운전 위험”, “위험 완화를 위한 권장 조치” 에 대한 실시간 정보를 생성하여, 자율 주행 차량에 직접 지시를 내림

- 차량은 **카메라 및 센서 데이터**를 기반으로 Recognition → Prediction → Planning → Control을 수행
- 동시에 **도로 상황 이미지**와 “프롬프트(설명 요청 등)”를 5G를 통해 MEC로 전송



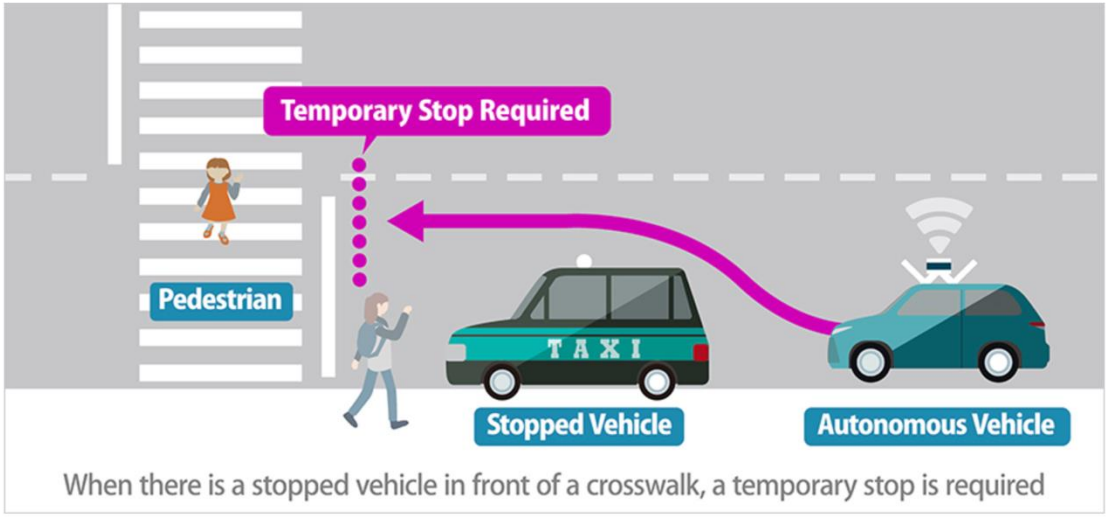
- **Traffic Image + Prompt**를 멀티모달 입력으로 수신
- **LLM**: 상황 설명, 위험 요소 판단, 주행 판단 보조 등 자연어 이해 기반 추론 수행
- **Inference Results**: 결과는 차량 제어뿐 아니라 **원격 모니터링 시스템**과 **중앙 모니터링 센터**로 전달

Traffic Understanding Multimodal AI

Traffic Understanding Multimodal AI

자율주행에 대한 외부 원격 지원 필드 시험

자율주행 차량이 정차 차량과 횡단보도에 접근하며 위험도가 높아지는 상황에서, 시스템이 “횡단보도 앞에 차량이 정차해 있으며 보행자가 갑자기 나타날 수 있으므로 완전히 정지하라”는 지시를 생성

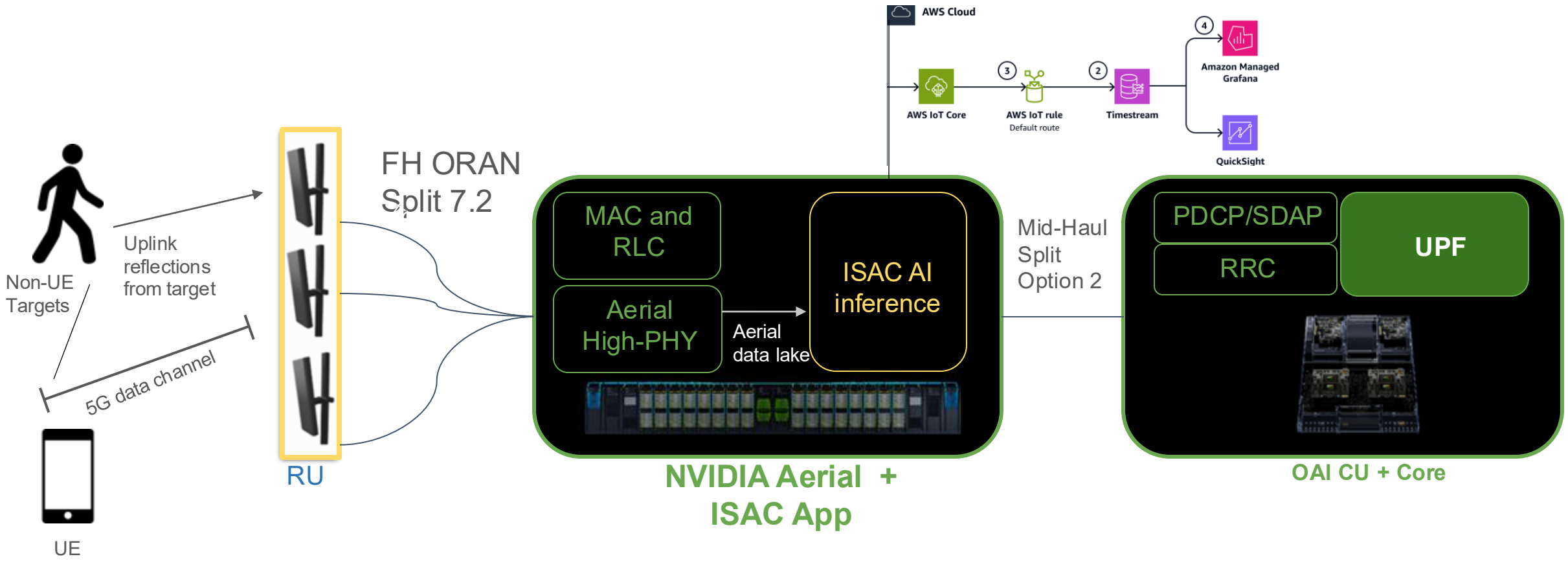






Traffic Understanding Multimodal AI

MWC'25 Integrated Sensing And Comms (ISAC) Demo

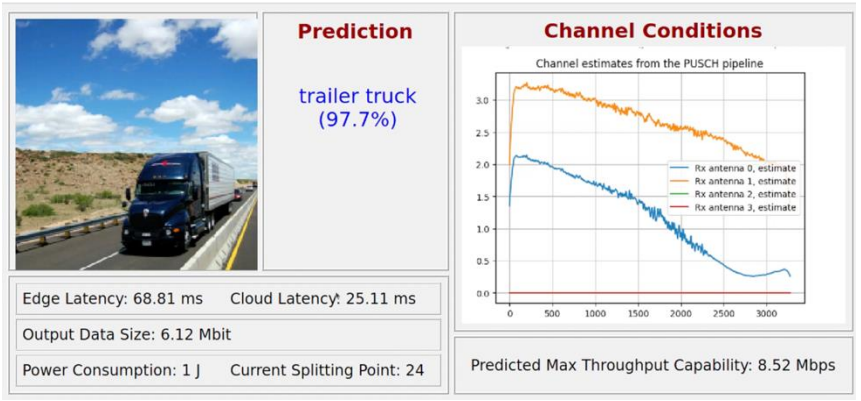


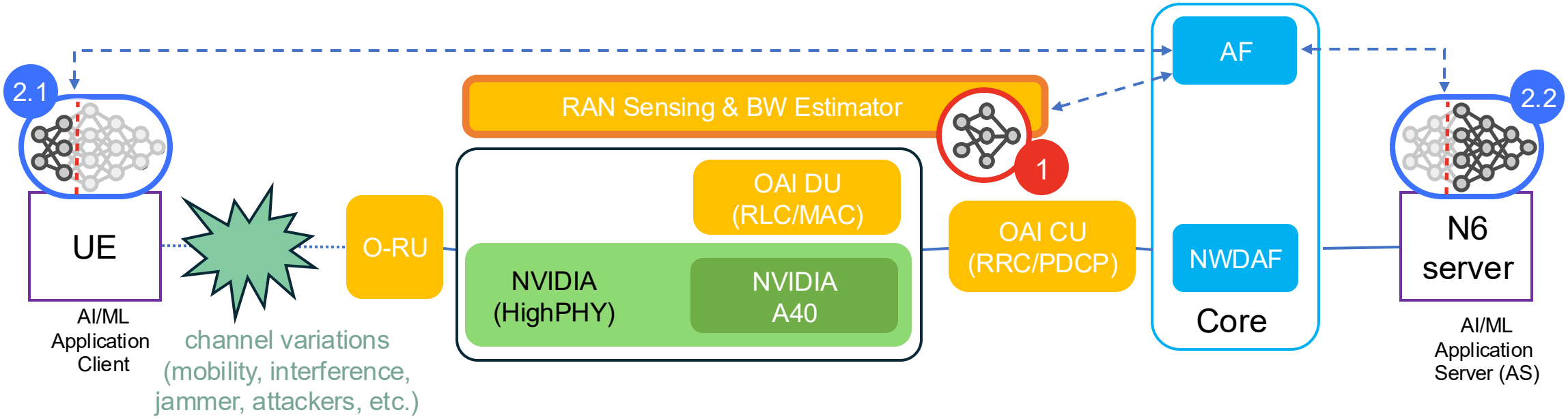
- AI RAN infrastructure uses 5G waveform for radar sensing and detection of non-UE targets (people) while simultaneously serving 5G UE with data.
- The AI workload for ISAC runs as an App, leverages Aerial High-PHY functions.
- ISAC inference outputs visualized in real-time on AWS.

ISAC = Integrated Sensing and Communications

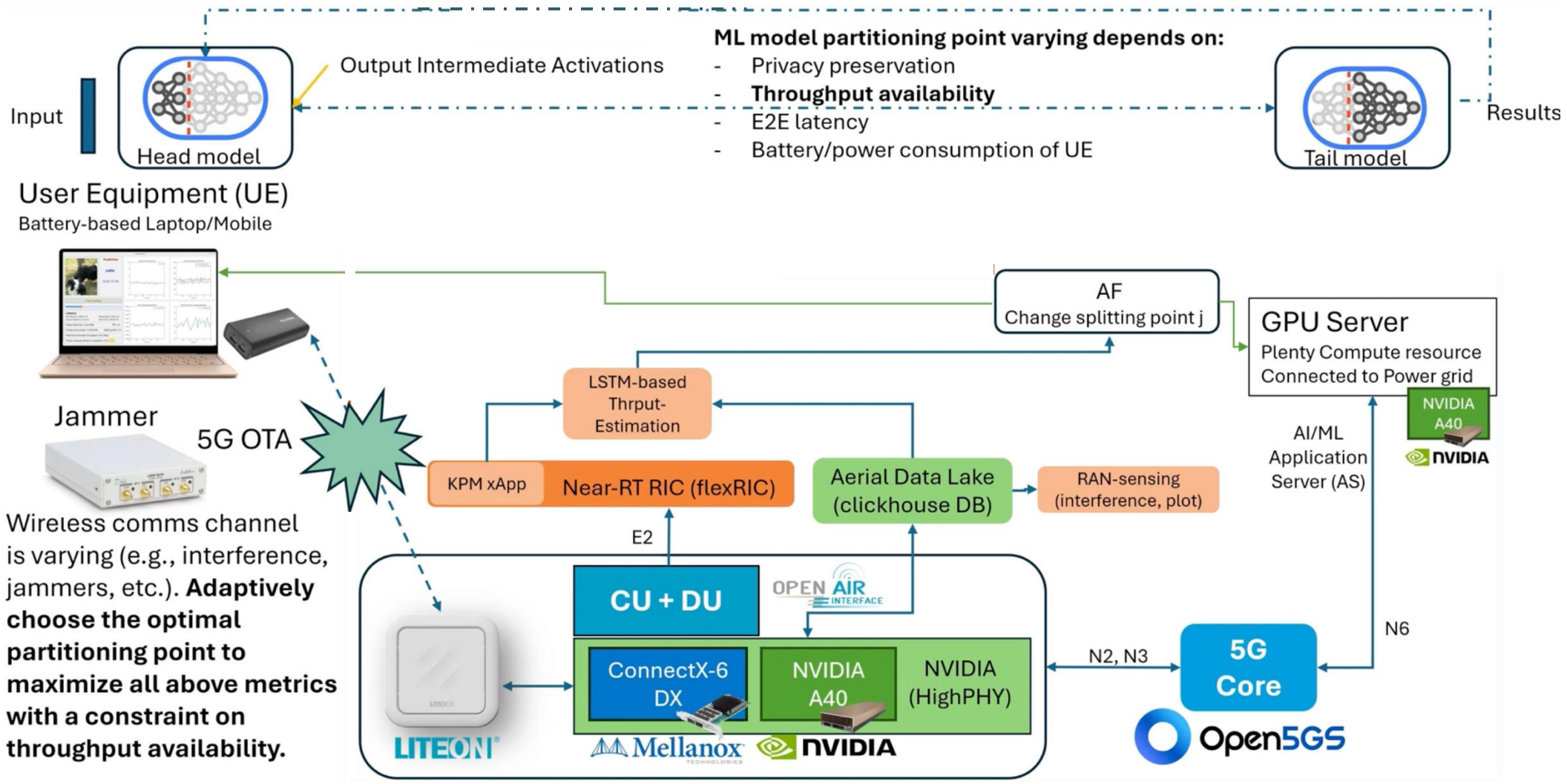
AI-Driven Spectrum Sensing for AI/ML Model Partitioning

Scenario: Split inference for remote image classification where (1) **spectrum sensing model** predicts throughput that determines a latency-optimal partitioning between: (2.1) **UE-side split model** and (2.2) **server-side split model**





AI-Driven Spectrum Sensing for AI/ML Model Partitioning

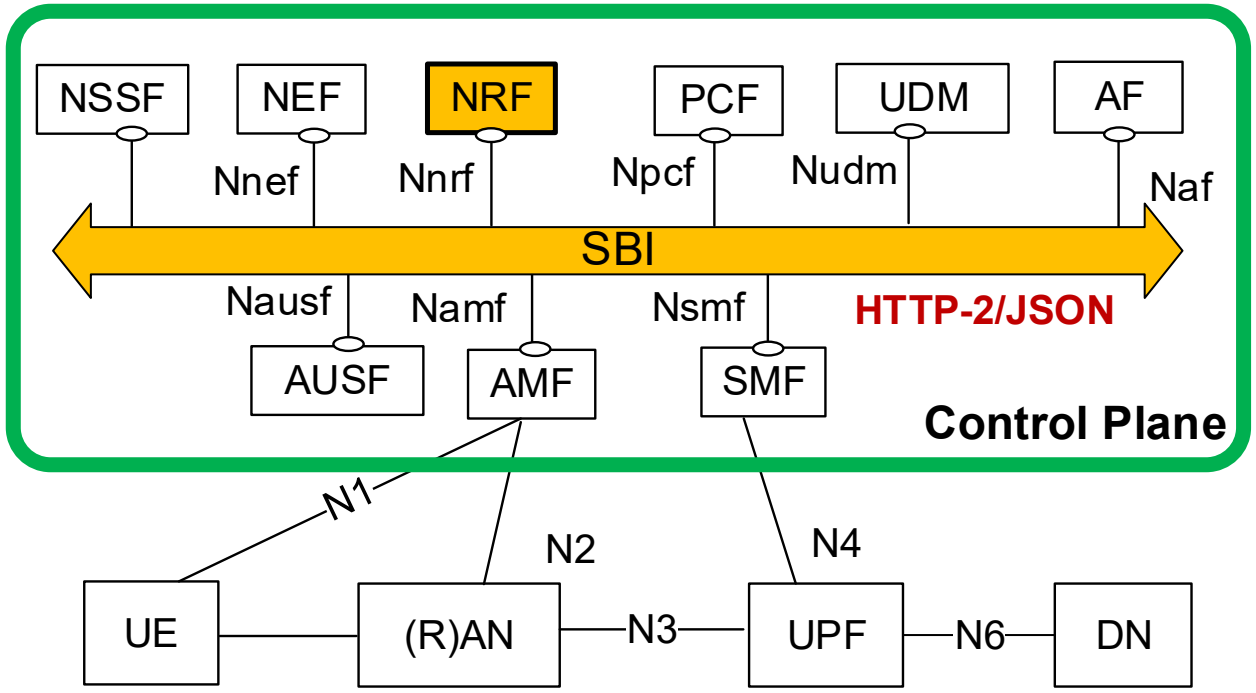


Cloud-Native RAN

Service-based RAN, RAN-CN Converged Architecture

SBA-based RAN-CN Convergence

Service-Based Architecture



5G System Architecture with serviced-based interface [TS 23.501]

- Network Slice Selection Function (NSSF)
- Data Network (DN)
- Network Exposure Function (NEF)
- Network Repository Function (NRF)
- Unified Data Management (UDM)
- Application Function (AF)
- Policy Control Function (PCF)
- Authentication Server Function (AUSF)
- Access and Mobility Management Function (AMF)
- Session Management Function (SMF)
- User Plane Function (UPF)
- User Equipment (UE)
- (Radio) Access Network ((R)AN)

코어 네트워크는 서비스 간 상호 연결을 Restful API 기반의 마이크로서비스 구조로 설계하여 유연성과 확장성을 극대화

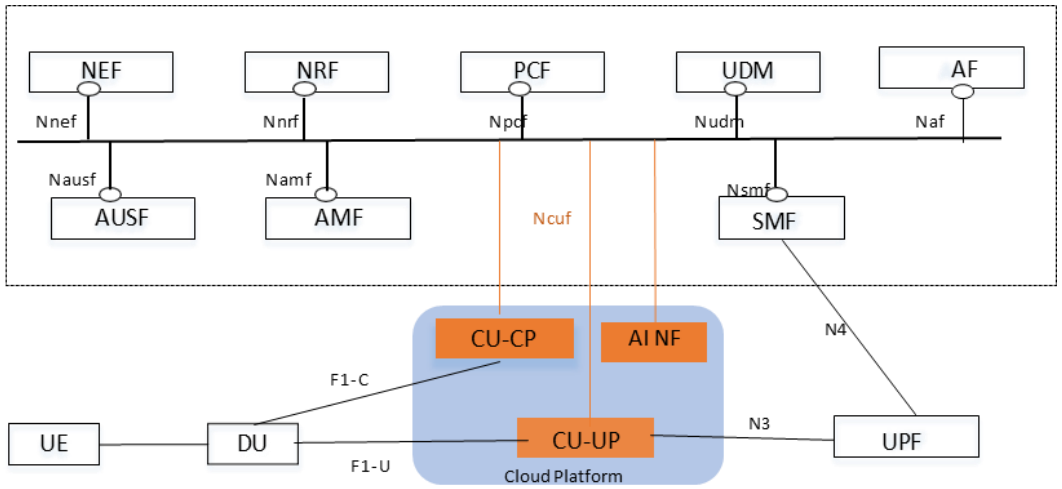
SBA-based RAN-CN Convergence

현재 RAN-CN 구조의 한계점 분석

- 클라우드 네이티브 지원 부족
  - 현재 N2/N3 인터페이스는 점대점(point-to-point) 구조
  - 클라우드 네이티브의 장점을 충분히 활용하지 못함
- 기능 중복
  - RAN과 코어망(CN) 간 역할이 중복됨 : 이동성 / 세션 / 페이징 / 상태 관리 등
  - 결과적으로 복잡성 증가 및 비효율성 초래
- 네트워크 기능 간 과도한 의존성 (tight inter-NF dependencies)
  - RAN과 CN 기능 간 복잡한 시그널링 필요
  - 확장성 저하 및 오류 가능성 증가
- 단일 지점 포워딩 병목점 (Single-point forwarding bottlenecks)
  - UE ⇄ SMF 제어가 반드시 CU-CP와 AMF를 거쳐야 함
  - 이로 인해 지연 증가, CPU 부하, 시그널링 비용 증가

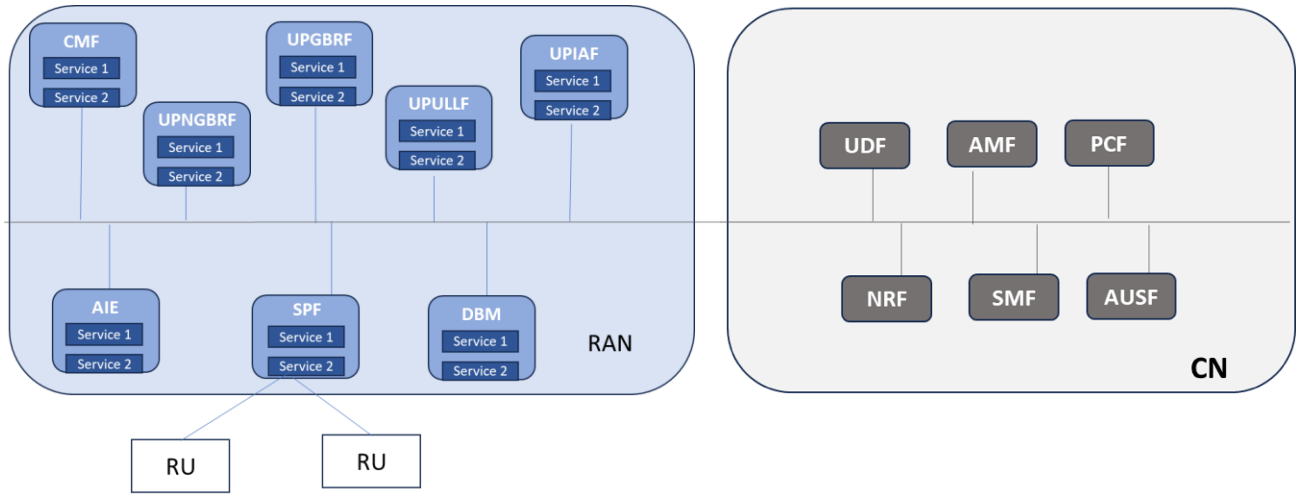
SBA-based RAN-CN Convergence

Interface-Level Service-Based RAN



- 기존 5G RAN(CU/DU) 구조와 **역호환성 확보**
- **CU-CP와 CU-UP을 부분적으로 SBA화**
  - CU-CP
    - SBI 통해 CN NFs(AMF, SMF 등) 및 CU-UP과 연결
    - 연결 제어, 이동성, 세션/베어러 관리, NAS 전달, 위치/추적 등
  - CU-UP
    - E1 인터페이스만 서비스화, P2P 방식 병행
    - SMF 및 NWDAF, NEF와의 데이터 관련 연동 서비스 포함
  - AI, Sensing 기능을 **새로운 NFs로 CU와 함께 배치**

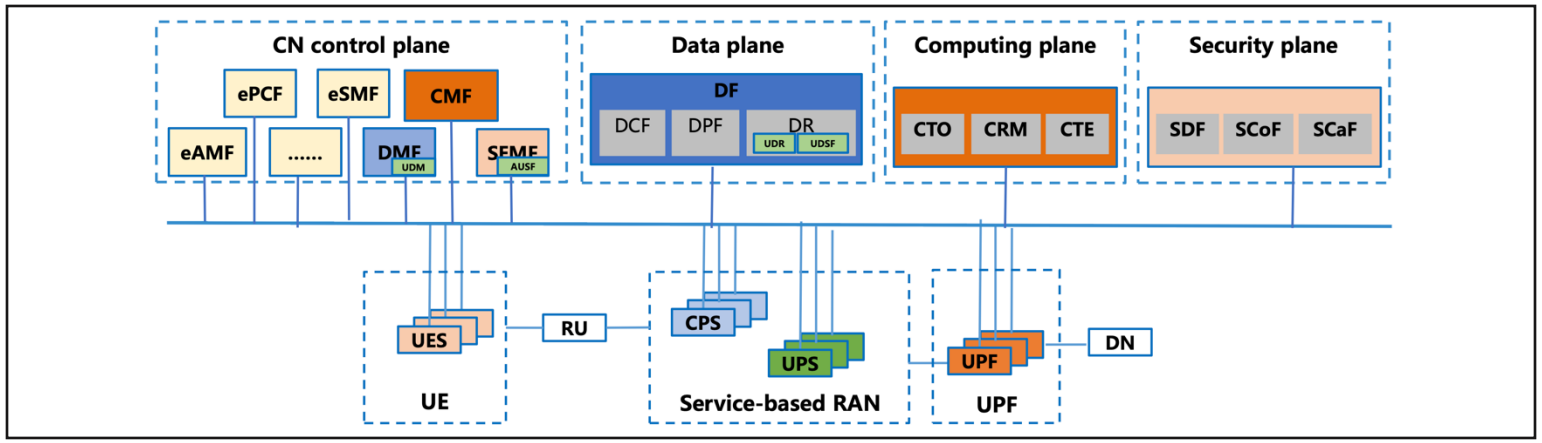
Function-Level Service-Based RAN



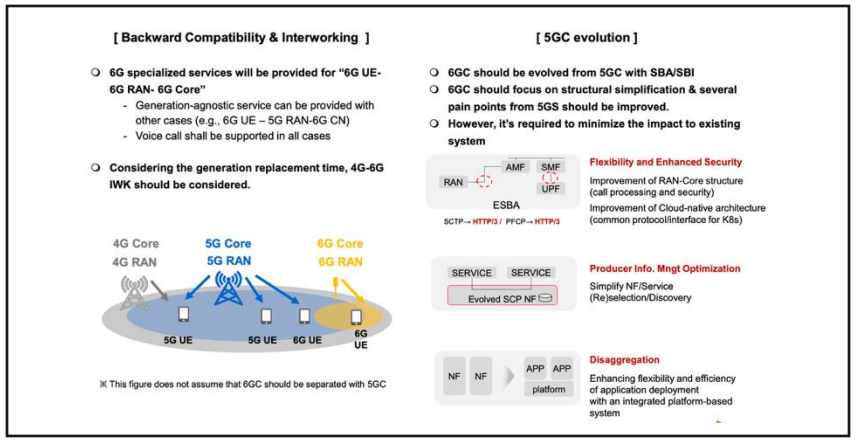
- CU-CP가 서비스로 제공하는 기능
  - 연결 제어, 이동성, 세션/베어러 관리
  - NAS 전달, 경고 메시지 전달, 위치/추적 등
- CU-CP는 SBI 기반으로 AMF, SMF 등과 직접 상호작용
- CU-UP은 SMF 및 NWDAF, NEF와의 데이터 관련 연동 서비스 포함

SBA-based RAN-CN Convergence: 3GPP 6G Workshop 발표

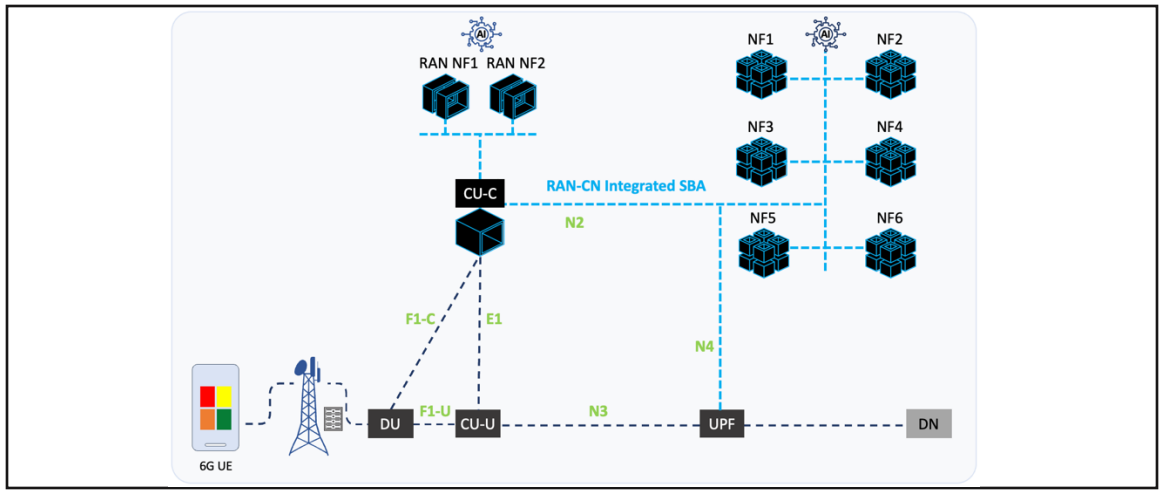
Service-Based Architecture 진화 이슈



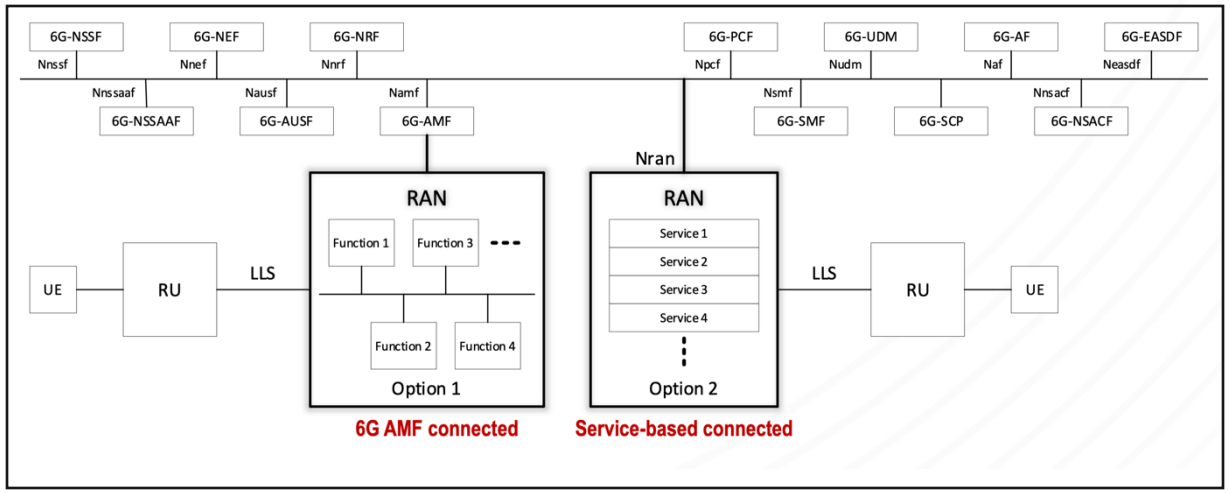
6GWS-250143 (CMCC)



6GWS-250030 (SK Telecom)



6GWS-250047 (ETRI)



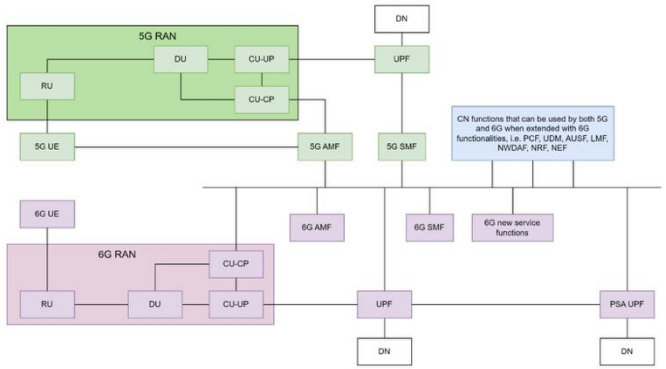
6GWS-250064 (LGE)

6G 코어는 RAN의 SBA화, CN의 SBA 진화, 그리고 RAN-CN 간 SBA 통합을 통해 전 구간 서비스 기반 아키텍처

SBA-based RAN-CN Convergence: 3GPP 6G Workshop 발표

Service-Based Architecture 진화 이슈

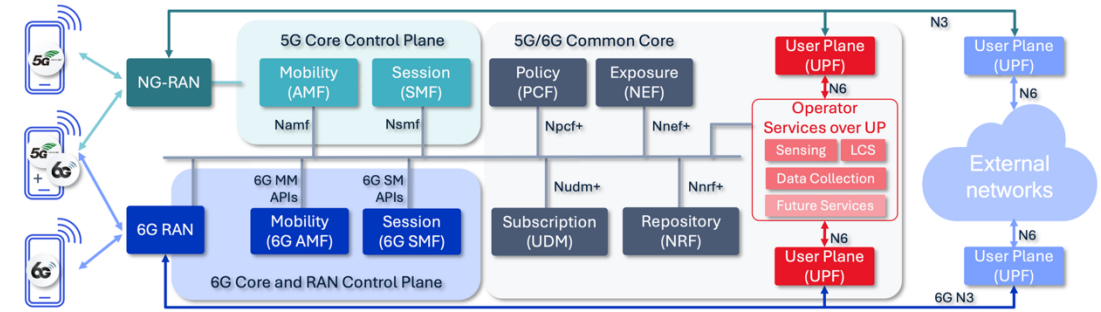
- Preference to have no clean slate design for 6GC
  - Enable overlay deployments of 6G core network by continuing to use the underlying SBA framework.
  - Most of core network functions can be extended to accommodate 6G functionalities, i.e., AUSF, PCF, UDM, NRF, NEF, LMF, NWDAF
  - Interworking can be achieved via the common PCF, UDM, and PSA UPF that are shared between 5G and 6G
  - New service functions can be added to provide 6G value-add services
- Non backward compatible changes should be considered, e.g.
  - Evolve 6G RAN CP to core network interface ("N2") to SBI
  - Evolve 6G UPF to the core network control plane for session management ("N4") to SBI
  - NAS optimizations -
    - Move NAS relay functionality from AMF to RAN
    - Move UE/NF security context storage to a security anchor (can be SEAF as part of 6G AMF)
  - Network simplification
    - Evaluate complexity of existing 5G design, e.g., Network Slicing.



New CP NF examples:
- Computing service discovery and chaining Functions
- Integrated Communication and Sensing Functions

6GWS-250042 (Intel)

Evolve 5GS to a service delivery-friendly and efficient to operate system ready for 6G radio.



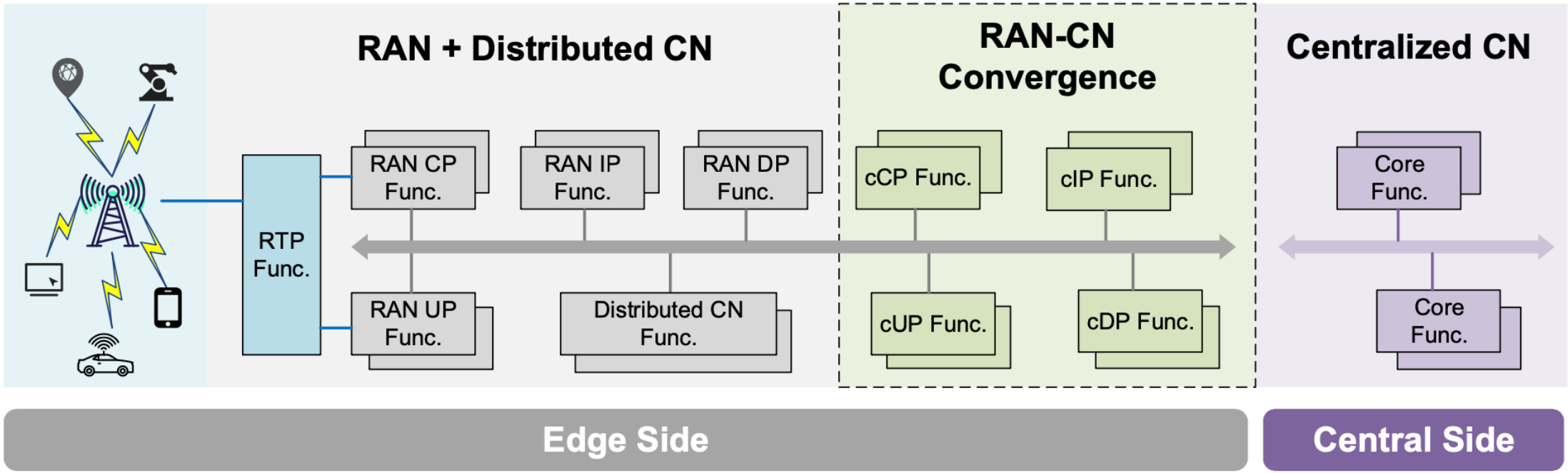
- Relocate operator services to the user-plane for rapid and RAT-agnostic service delivery.
- Replace: Distribute NAS termination for independent evolution of control plane functionalities.
- Replace: Simplify NW Slicing (separate allocation of session resources from CN instance selection)
- Replace: Simplify RAN-CN functional split for efficient network operation (aim for one connected mode mobility solution and one UE tracking/paging solution; store UE radio capabilities at RAN).
- Reuse SBI framework, roaming architecture, exposure APIs & most 5G network functions.

6GWS-250069 (Qualcomm)

6G 코어는 RAN의 SBA화, CN의 SBA 진화, 그리고 RAN-CN 간 SBA 통합을 통해 전 구간 서비스 기반 아키텍처

SBA-based RAN-CN Convergence

RAN과 CN 기능이 “제어, 사용자, 데이터, 지능”의 4개 평면에서 Service-Based Architecture를 통해 융합되어, 엣지 중심의 지능형 네트워크 실현



- RTP: Real-Time Processing
- cCP: converged Control Plane
- cUP: converged User Plane
- cIP: converged Intelligence Plane
- cDP: converged Data Plane

Thank you!

nsko@etri.re.kr
고남석

실장/한국전자통신연구원
부위원장/가상화・지능화위원회 ORIA