AI-RAN Alliance

배정숙

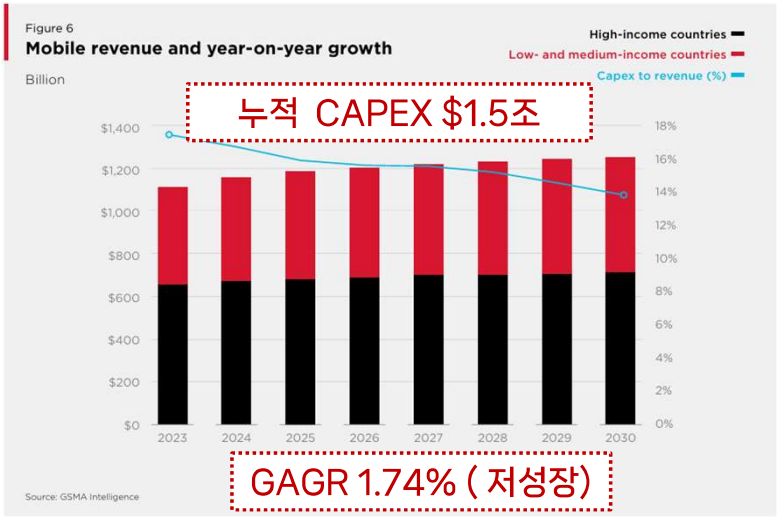
한국전자통신연구원

Why AI-RAN?

통신 산업의 위기 요인

빠르게 고도화되는 기술, 폭증하는 데이터 수요 → 막대한 자본 투자 부담

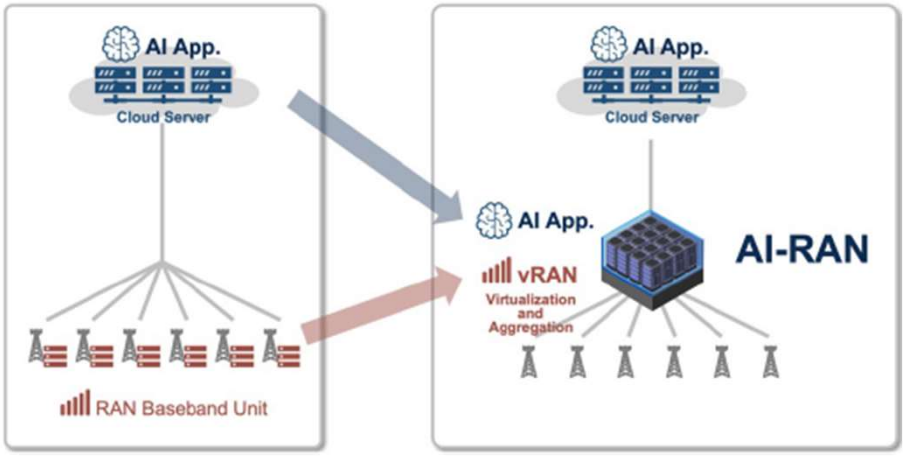
- **5G 투자 비용 급증**
  - 고주파수 대역, Massive MIMO 안테나, 초고밀도 셀
- **네트워크 인프라 부담 증가**
  - 트래픽 증가(LLM 기반 생성형 AI 서비스 확산)
  - '22: 10.2GB → '23: 12.8GB (연평균 23%))
- **소수 벤더 의존 장비 생태계**
  - 장비 가격 하락 / 범용화 저해 → 투자 효율성 하락
- **수익성 약화**
  - 데이터 요금 중심 구조로 CAPEX 회수 불가
  - OTT 사업자 수익 창출 → 추가 수익 없이 망 부담 증가



투자 대비 수익 구조의 비효율성

What AI-RAN?

AI-RAN



- **AI 기술을 무선 접속망(RAN)에 완전히 통합하여 RAN 운영 성능을 혁신적으로 향상**시키고, 새로운 **AI 기반 서비스를 동시에 제공하여 신규 수익 창출** 기회를 여는 **기술 패러다임**

- RAN 기능은 전용 하드웨어에서 동작
- AI 워크로드와 분리 운영

- 동일한 범용 인프라(GPU 기반 클라우드)에서 RAN 기능과 AI 워크로드를 병행 실행

전통적인 RAN 인프라를 다목적 지능형 엣지 인프라로 진화

AI-RAN Alliance

AI의 잠재력을 통해 5G와 다가오는 6G 시대의 RAN을 혁신

Founding Members

arm, DEEPSIG, ERICSSON, Microsoft, NOKIA, Northeastern University, NVIDIA, SAMSUNG, SoftBank, T-Mobile, UTokyo

General Members

WIND, AIRA, amdocs, boost mobile, Booz Allen, BROADRADIO, CableLabs, CLOUDERA, CloudRAN, cohere, CTONE, DELL Technologies, EDGECORTIX, ETRI, EURECOM, FUJITSU, G REIGNS, ITRI, KEYSIGHT, KOREA UNIVERSITY, kt, LG U+, LotusCom, METANOIA, NEC, PURESTORAGE, Red Hat, FIMEDO, SEOUL NATIONAL UNIVERSITY, SKT telecom, SYNAXG, TECH mahindra, tietoevry, TOYOTA, UNIST, VIAVI, YONSEI UNIVERSITY, zTOUCH

AI-RAN ALLIANCE

History

Feb. 2024.	Founded AI-RAN Alliance @MWC 2024, Barcelona
Aug. 2024.	Dr. Jinsung Choi head up the Alliance as a Chairman
Aug. 2024.	Create Three Working Groups (AI-for-RAN WG, AI-and-RAN WG, AI-on-RAN WG) On-line Meetings
Nov. 2024.	The 1st F2F Member Meeting @Ericsson Inc., Santa Clara, CA
Mar. 2025.	Lab. Demo @MWC2025, Barcelona
June. 2025.	The 2nd F2F Member Meeting, Espoo, Finland

설립과 회원사

2024.02@MWC
11 개 창립 회원사, 80개 일반 회원사

임무

* Transform the network, Transform our business
* Evolving networked sustainability

비전

*Realizing and harnessing the potential of an AI-native RAN

AI-RAN Alliance

비표준화 기반의 실용 중심의 독립 연합체

중점 분야

Reference Design

AI 기능과 RAN 네트워크 요소 간의 통합, 무중단 운영, 성능 최적화 보장 설계 프레임워크 제시

Blueprints

AI 기능을 RAN에 효율적으로 통합하기 위한 기술 구성요소, 단계별 절차, 운영 시나리오의 체계적 정의
→ 확장 가능하고 반복 가능한 아키텍처 구현 지원

.

Benchmarking

다양한 RAN 환경에서의 AI/ML 모델의 성능을 정량적으로 분석 및 평가
→ 산업 전반의 요구사항과 표준화된 기대 수준에 부합하는 성능 기준 마련

AI-RAN Alliance

Committee

- Chair: Kuntal Chowdhury, NVIDIA
- Vice Chair: Dr. Felipe Arraño Scharager, Ericsson

Technical Streering Committee (TSC)

Marketing Steering Committee (MSC)

- Chair: Steph Delvoye, Head of Mobile Networks Marketing, Nokia
- Vice Chair: Yashar Nezami, Global Partnering & Ecosystem Director, Ericsson

Working Group & Task Group

WG-1: AI-for-RAN

- Chair: Chris Dick, Nvidia
- Vice Chair: Dr. Felipe Arraño Scharager, Ericsson

WG-2: AI-and-RAN

- Chair: Tanveer Saad, Nokia
- Vice Chair: Dr. Yuji Sekiya, University of Tokyo

WG-3: AI-on-RAN

- Chair: Dr. Athul Prasad, Samsung Research America
- Vice Chair: TBD

Work Process

1. Work item proposal at the WG
   - The proposal must be as per current WG charter
2. Work item report initiated
   - Work plan defined
   - Lab setup process initiated
3. T&M Task Group defined test and benchmarking criteria, setup
   - Data for AI step as necessary
4. Lab setup and test and benchmarking executed
   - Results discussed at WG
   - TSC review of the results approval
5. Board approval for publication
   - WI report publish (Dataset, and any Notebook if
   - Optionally Blueprint published

AI-RAN Alliance Activity

WG-1: AI-for-RAN WG (Enhancing network performance)

AI-for-RAN

AI for the enhancement of RAN

Spectral Efficiency

목표

- AI를 RAN에 적용하여 성능 향상을 실현하는 기술을 탐색하고 발전시키는 것

추진 분야

- RAN의 효율성, 용량 및 기타 KPI를 향상시키기 위한 AI-네이티브 RAN 솔루션을 발굴/정의/구축 (3GPP 및 기타 산업 단체의 AI/ML 활동을 활용 및 보완)

핵심 기술 영역

- Radio Link Optimization: AI 기반 신호 처리 기술을 통해 링크 품질 및 안정성 개선
- Spectral Efficiency: 주파수 자원 활용을 극대화하며 고품질 사용자 경험을 유지
- Performance: 고신뢰성, 고용량, 광역 커버리지, 고에너지 효율 (혼잡하거나 요구사항이 높은 환경에서도 QoS 유지 가능)

AI-RAN Alliance Activity

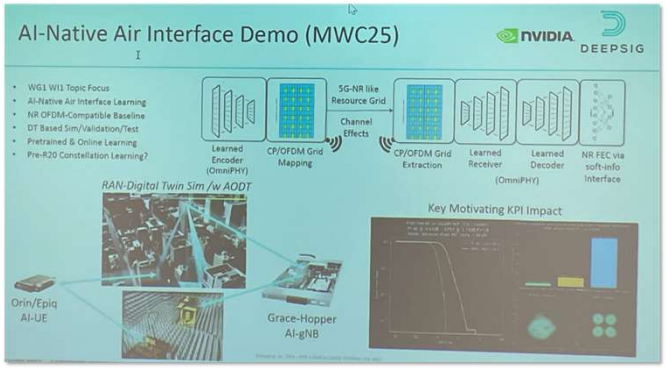
WG1 Work Item

WI #	Title	To be developed ?	WI Report Status
1	Learned Air Interface with Online Learning (DeepSig, NVIDIA)	Yes	초안 완료
3	AI-based PUSCH Channel Estimation (Keysight)	Yes	초안 완료
5	AI-based 5G Beamforming for Mobility-Aware Interference Mitigation and Power Saving (Viavi, SUTD, Yonsei)	TBD	미정
6	AI-based Spectrum Sensing in the RAN (Northeastern University)	Yes	초안 완료
7	Cell energy saving (Ericsson)	Yes	초안 완료
8	Agentic Architecture for AI-native RAN (U of Leeds, NVIDIA)	Yes	초안 완료
9	Multimodal sensing (U of Oulu, NVIDIA)	Yes	초안 완료
10	Neuromorphic transceiver (Viavi)	Yes	SW 완료
11	ML for slicing (NEU)	Yes	초안 완료
12	DeepMIMO (NVIDIA, *Arizona State University*)	TBD	미정
13	SRS prediction (SoftBank, NVIDIA)	TBD	미정
14	AI-for-RIC (Amdocs)	Yes	초안 완료

* 2: Hold, [Softbank, NVIDIA, NEU] Realization of uplink channel interpolation in actual RAN

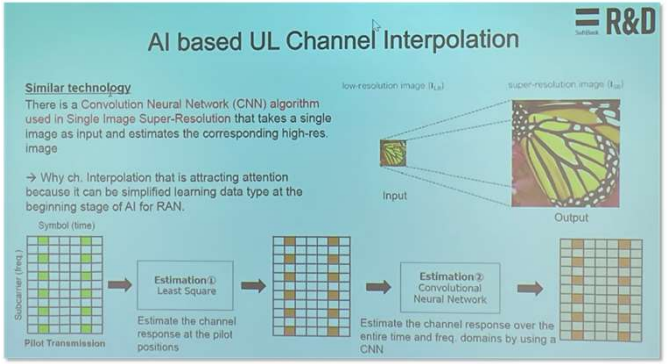
AI-RAN Alliance MWC2025 WG1 demos

Demo1: Learned Air Interface with Online Learning
: Deepsig, NVIDIA



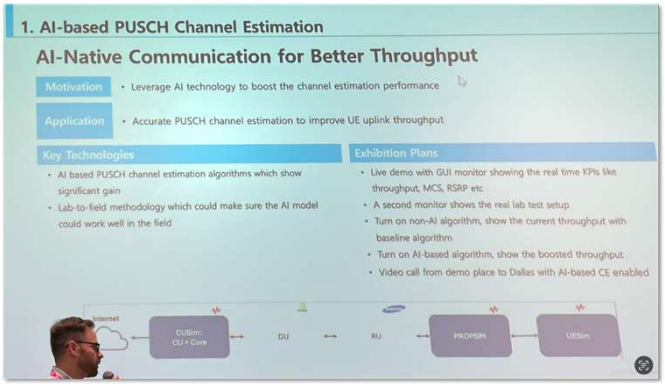
* AI가 최적의 변조 방식과 등화 구조를 직접 학습
* DMRS 없이도 채널 추정 및 복원 가능
* Online으로 환경에 맞춰 학습 업데이트 가능
* 고속 채널 vs 정지형 채널에 따라 적응적 최적화
→ Shannon 한계에 근접하는 스펙트럼 효율 달성

Demo2: Realization of UL Ch Interpolation in Actual RAN : SB, FJ, NVIDIA



* AI를 실제 5G NR 물리 계층에 적용하여 채널 추정 성능 향상
* 딥러닝으로 채널의 시공간분포를 추론하여 전체 그리드 채널 상태 정보 복원
* CNN 기반 UL 채널 보간 기술을 적용해 슬롯당 성능을 3~6% 향상

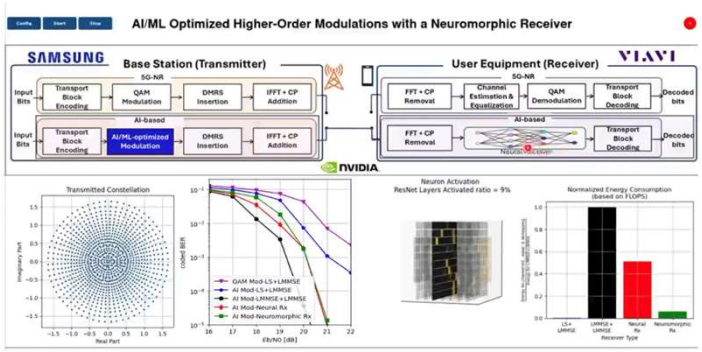
Demo3: AI-based Channel Estimation
: Samsung, Keysight, NVIDIA – Keysight booth



*고차원 무선 채널 정보(frequency-time-space)를 AI로 처리
*DMRS 신호를 Denoising하여 더 정밀한 채널 추정
*전통적 MMSE 기반 UL MAC 처리량에 비해 2배 이상 throughput 향상

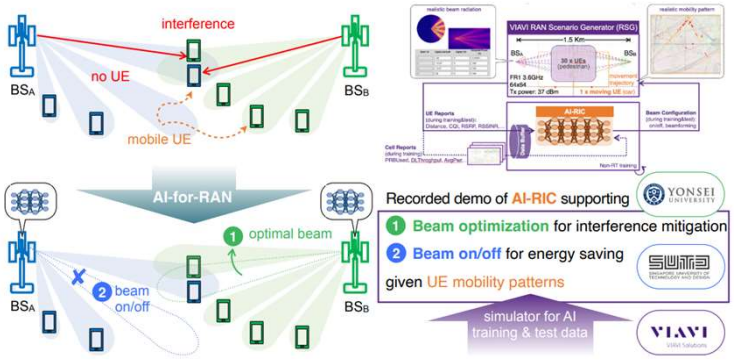
AI-RAN Alliance MWC2025 WG1 demos

Demo4: AI/ML Optimized Higher-Order Modulations with a Neuromorphic Receiver, Samsung, Viavi, NVIDIA



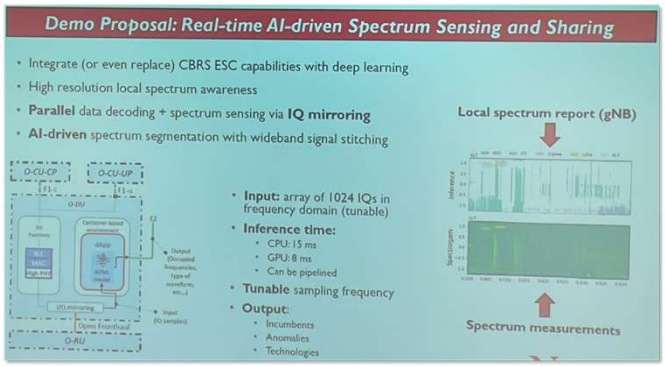
* AI/ML 기반 고차 non-uniform 변조 방식
 - 256QAM, 1024QAM 대비 0.7dB, 1.2dB 성능 향상
* Neuromorphic Receiver 기반 AI 복조기
 - 기존 Neural Rx 대비 전력 소모 9배 절감
 - 5G 수신기보다 더 높은 전송률 실현

Demo5: AI-based 5G Beamforming for Mobility-Aware Interference Mitigation and Power Saving, Viavi, SUTD, Yonsei



* AI-RIC를 도입하여 사용자 이동성과 무관하게 간섭을 줄이고, 불필요한 빔 송신을 방지
* 전력 절감 최대 20%
* DL 처리량 최대 30% 증가

Demo6: AI-based Spectrum Sensing in the RAN, Northeastern



*RAN에 AI 기반 스펙트럼 센싱 기능을 통합
*Uplink 대역 주기적 감지→ AI 기반 분류기로 신호 감지 및 파형 분류 → 해당 주파수 대역 차단 또는 PRB 재할당 → 간섭 감소

AI-RAN Alliance Activity

AI-and-RAN WG (Shared Infrastructure)

AI-and-RAN

AI and RAN sharing the same infrastructure

Asset Utilization

목표

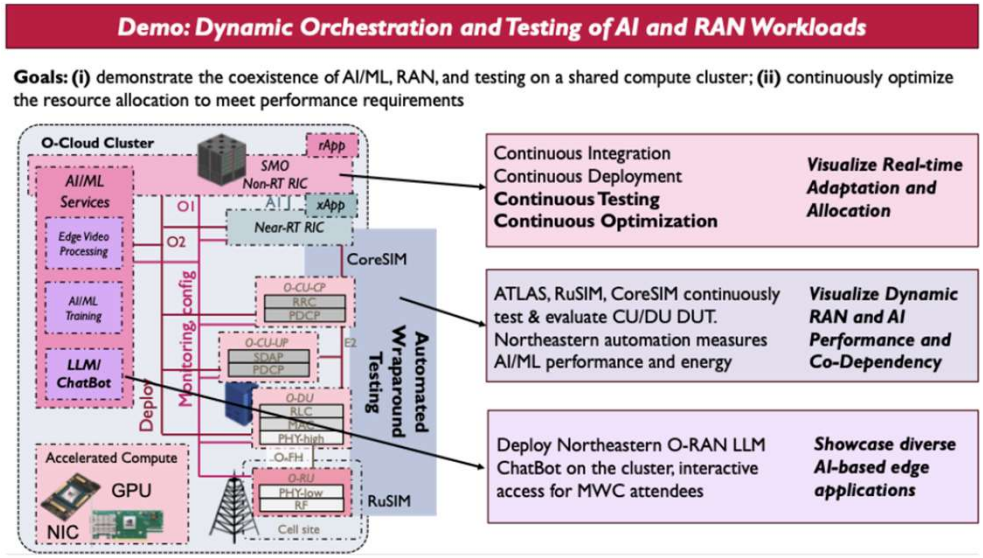
- AI/GenAI 애플리케이션과 RAN 워크로드를 하나의 통합 인프라 상에서 동시에 실행하기 위한 기술 탐색컴퓨팅 자원의 효율적 활용 및 AI 서비스 수익화 기회 확보

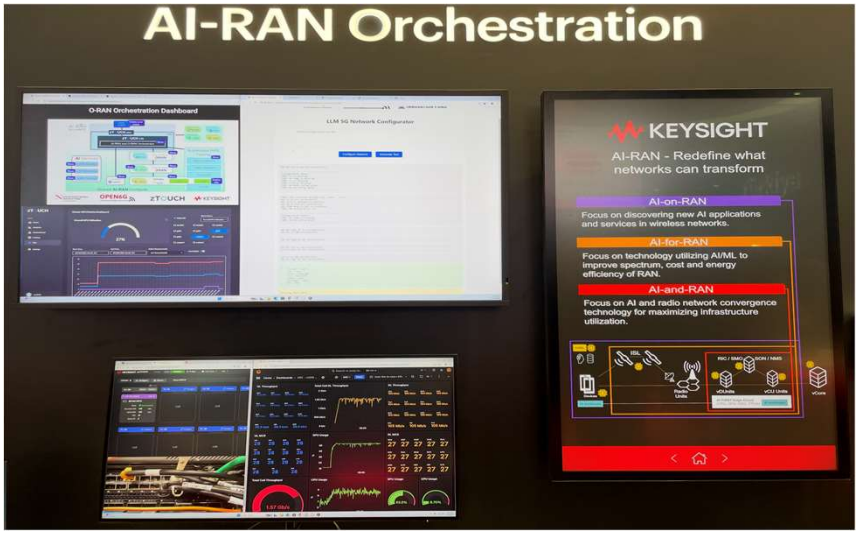
핵심 기술 영역

- **통신+컴퓨팅 융합 인프라 활용 모델 연구:** RAN과 AI/GenAI 워크로드의 병렬 실행을 통한 자원 활용률 극대화, 특히 멀티테넌시 환경에서 리소스 분할·공유 기술을 단계별로 실증(사전 구성 기반 정적 분할, 규칙 기반 동적 분할, AI 기반 동적 자원 관리)
- **RAN 가상화와 엣지 집약형 처리 모델 연구:** 가상화된 RAN (DU/CU/UPF 등)을 엣지에 배치, 통계적 트래픽 평균화를 통한 인프라 최적 활용

AI-RAN Alliance MWC2025 WG2 Demo

WG2 Demo: AI-RAN Orchestration
: NEU, Keysight

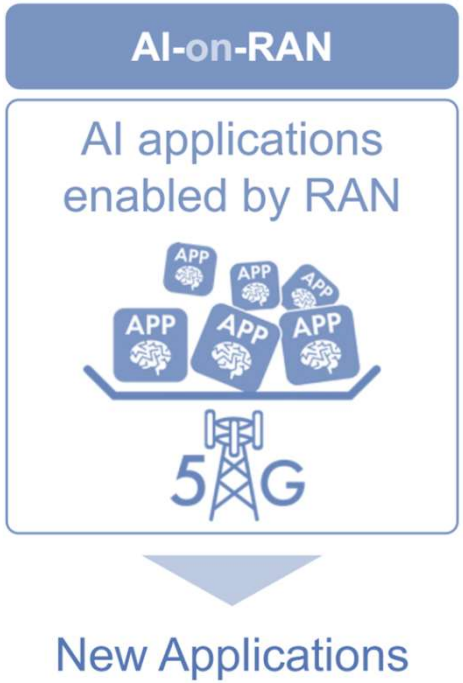




* 이기종 워크로드(RAN + AI)를 동일한 컴퓨팅 인프라에서 통합 운영하며,자동화된 테스트 및 실시간 성능 분석 → AI-RAN 환경의 공존성 실증
* AI 챗봇과 실시간 상호작용, CNN 및 기타 AI 애플리케이션 실행
* 다양한 RAN 트래픽 조건 변화 시 자동 성능 측정
* 워크로드 간 자원 충돌/공존의 영향 시각화

AI-RAN Alliance Activity

AI-on-RAN WG(Enabling Edge Services)



목표

- AI 및 생성형 AI(GenAI) 응용을 RAN 상에서 실행하기 위한 무선 인터페이스 요구사항 정의
- 5G 환경에서 응용 성능 벤치마크 → 향후 6G 시스템에 필요한 신규 요구사항 도출
- 지연(Latency), 처리량(Throughput), 지터(Jitter), 패킷 지연(Packet Delay), 암호화 요구사항

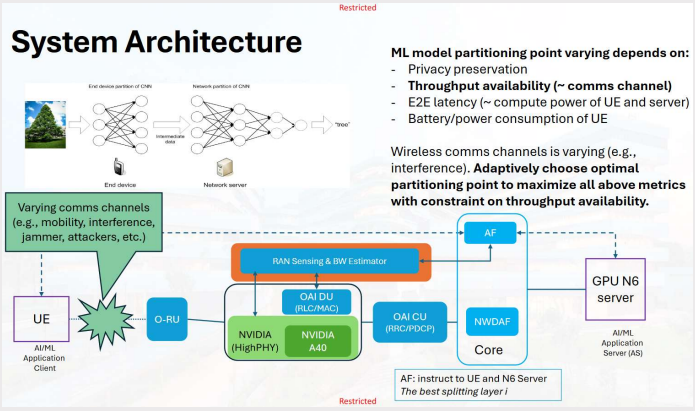
핵심 기술 영역

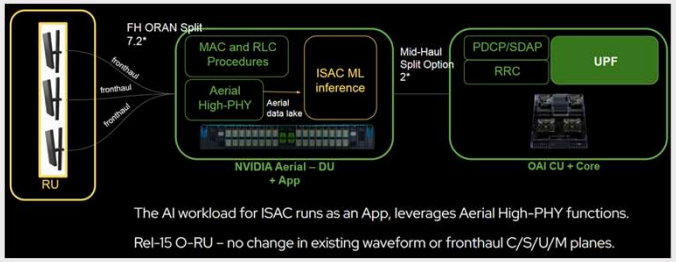
- 현재의 AI/ML 및 GenAI 기술을 검토하고, 기술적 난제를 파악하며, 활용 사례를 정의하고, 성능 테스트용 랩 시스템 개발

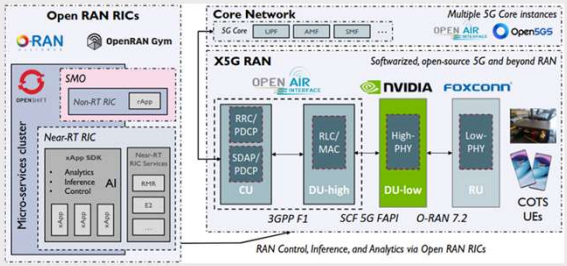
AI-RAN Alliance Activity Fields

WG3 Items

- **WI #1: AI enabled split compute.**
  - AI 기반 애플리케이션을 위한 Split Compute(분산 연산) 구조 연구
- **WI #2: AI-enabled Critical Applications (Joint Communication and Sensing)**
  - RAN 내에 AI를 통합하여 통신과 센싱 기능이 결합된 중요 애플리케이션을 지원하는 방법 정의

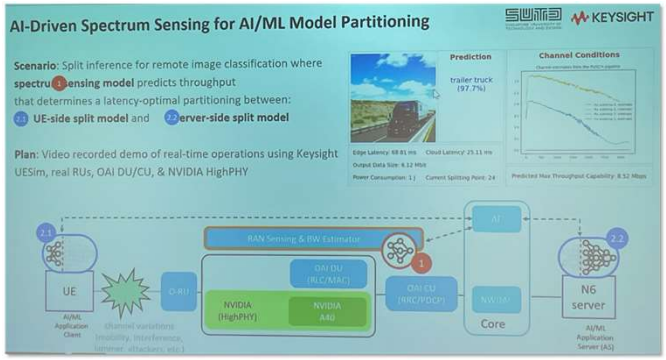






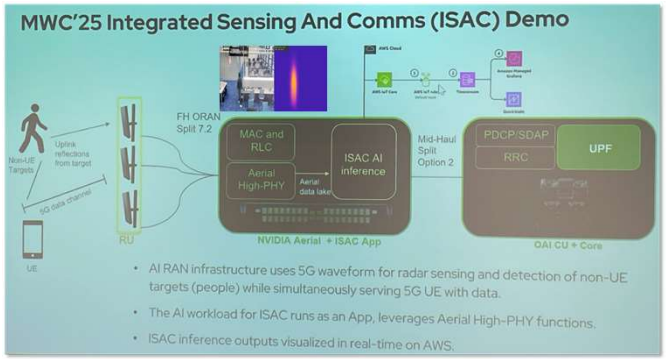
AI-RAN Alliance MWC2025 WG3 demos

WG3 Demo 1: AI-Powered Spectrum Sensing with NVIDIA Aerial 5G for AI/ML Model Partitioning: Privacy-Focused Image Processing: SUTD, Keysight, NeuroRAN, LITEON



* AI-RAN 환경에서 실시간 AI 기반 스펙트럼 감지를 바탕으로 한 적응형 머신러닝 모델 분할을 통해 ,개인정보 보호, 지연 시간, 에너지 효율, 처리량 등 주요 성능 지표를 실시간으로 동적 최적화

WG3 Demo 2: 5G Integrated Sensing and Communications (ISAC): Tiami Networks



* 상용 5G 인프라 기반, AI 활용 환경내 사람의 존재를 감지 및 추적
* UL SRS로 환경 감지
* Sensing 기능을 AI App 형태로 RAN 내에 통합

WG3 Demo 3: AI-on-RAN Object Detection: ARM, Tannera, Phluido & Effnet

* 지연에 민감한 AI 응용 프로그램(산업용 모니터링, 영상 분석 등)을 Private 5G 네트워크 상에서 안정적으로 동작시킬수 있는 상용 수준의 통합 솔루션 제시
AI 추론과 RAN 기능을 같은 서버(Arm Neoverse-N1 기반)에서 병렬 실행

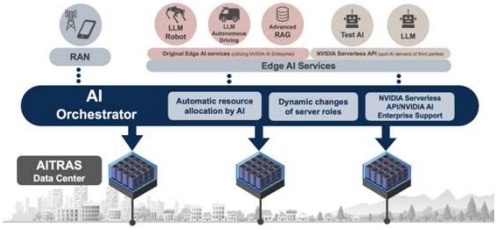
AI-RAN 플랫폼 사례: Softbank AI-RAN

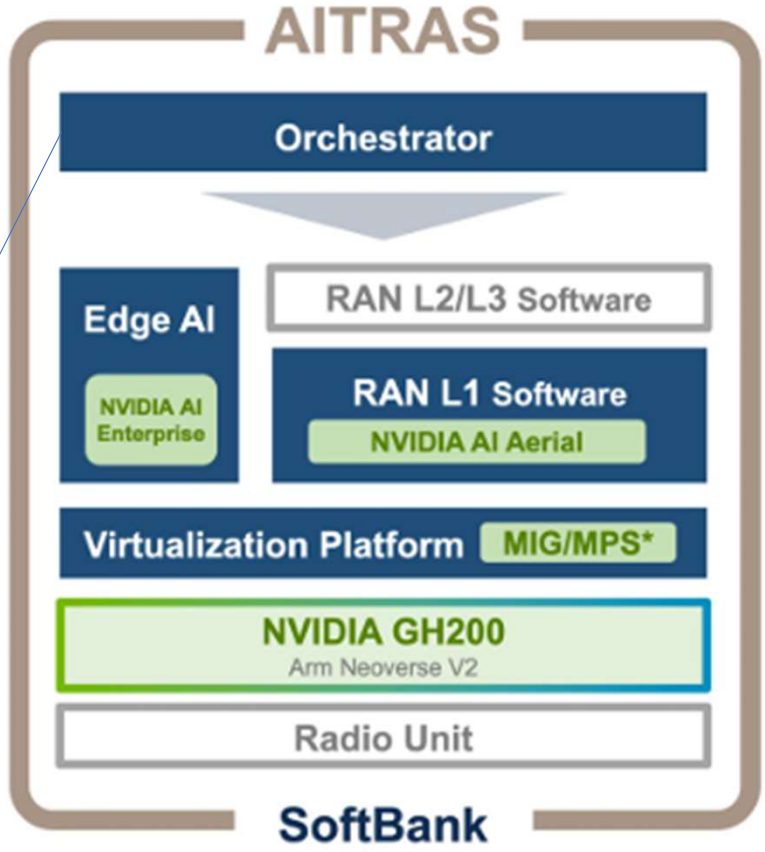
AITRAS

- gRAN 을 기반으로 구축된 차세대 무선 접속망으로,AI 워크로드와 통신 연산을 통합 처리할 수 있는 구조를 통해 멀티테넌시 운영과 고도화된 서비스 제공

** gRAN(GPU 기반 RAN):실시간 데이터 분석, 저지연 처리, 대규모 AI 애플리케이션 등 AI-네이티브 기능을 지원하기 위해 GPU를 활용한 병렬 처리 구조를 사용하는 진화된 RAN*

RAN과 엣지 AI 서비스 전반에 걸쳐 RAN 및 AI 워크로드를 관리하고 최적화하기 위한 중앙 제어 메커니즘

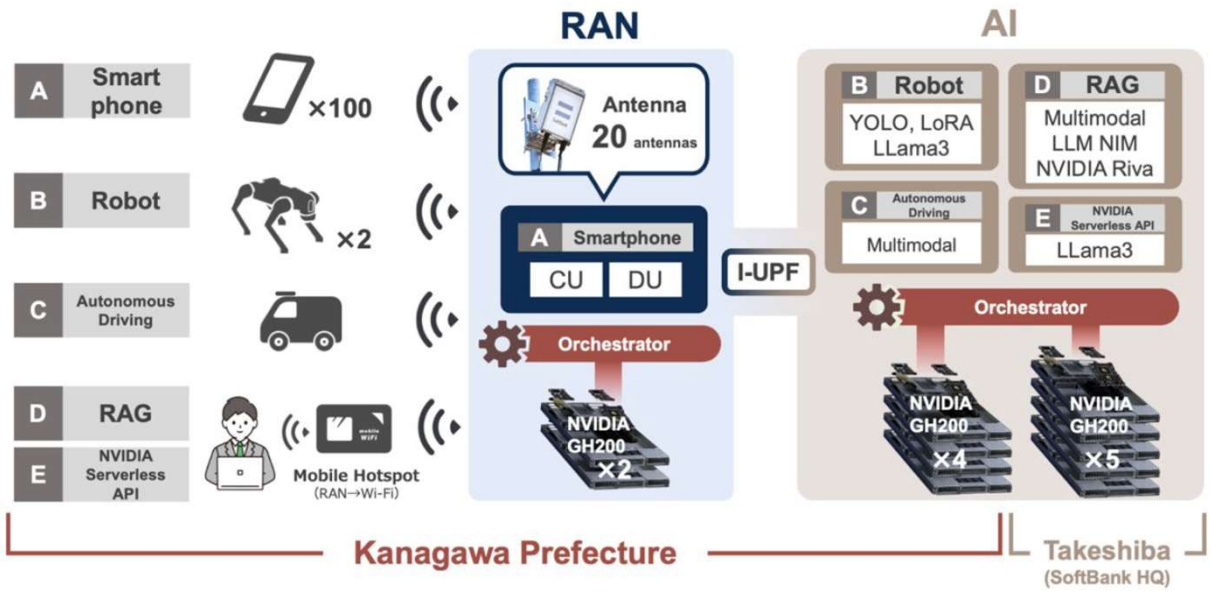


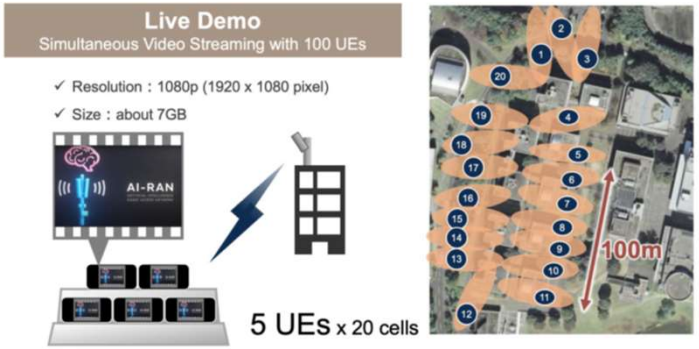


AI-RAN 플랫폼 사례: Softbank AI-RAN

Outdoor Testbed

- 100미터 구간에 고르게 분포된 20개의 5G 셀 구성(최대 4계층 MIMO 및 100MHz 대역폭 지원)
- 주파수 대역: 4.8GHz ~ 4.9GHz (n79 대역)





각 20개 위치에 5개의 단말(UE)을 배치, 동시 영상 스트리밍 환경에서도 기지국 안정적인 동작이 확인

핵심 정리

- **AI는 RAN의 본질적 혁신을 이끈다**

  기존 RAN의 한계를 뛰어넘는 자율 최적화(Self-X), 효율성, 확장성 확보

- **AI-RAN은 단순 기술이 아닌 전략적 인프라 전환**

  투자 효율, 에너지 절감, 유연한 서비스 제공을 동시에 실현

- **AI-RAN Alliance를 통한 글로벌 협력이 가속 중**

  산업계·학계·연구기관이 함께 MWC 데모로 검증한 실현 가능 기술

- **AI-RAN은 6G와 엣지 지능형 서비스의 핵심 플랫폼**

  자율주행, 스마트로봇, 산업용 AI 등 실시간·초지능 서비스의 기반 플랫폼으로 확장 중

- **지금이 AI-RAN 생태계 조성의 골든타임**

  디지털 트윈, 에너지 효율, AI-native 아키텍처 실증은 이미 시작됨
  산업·학계·연구기관의 전략적 연합으로 **데이터, 인재, 파트너십의 조화로운 생태계 구축 필요**

감사합니다.

Thank you.

JungSook Bae(jsbae@etri.re.kr)