

# AI-RAN: Unveiling the Future of Telecom with Breakthrough Innovations

Feb 6, 2025 5:00 pm EST / Feb 7, 2025 7:00 am JST





# Today's Speakers



**Ryuji Wakikawa**

Vice President, Head of Research  
Institute of Advanced Technology  
SoftBank Corp.



**Alex Jinsung Choi**

Principal Fellow  
SoftBank Corp.



**Rajeev Koodli**

Principal Fellow  
SoftBank Corp.



**Koichiro Furueda**

Deputy Director, Wireless System  
Development Department, Wireless  
System Innovation Division  
SoftBank Corp.



# Speaker



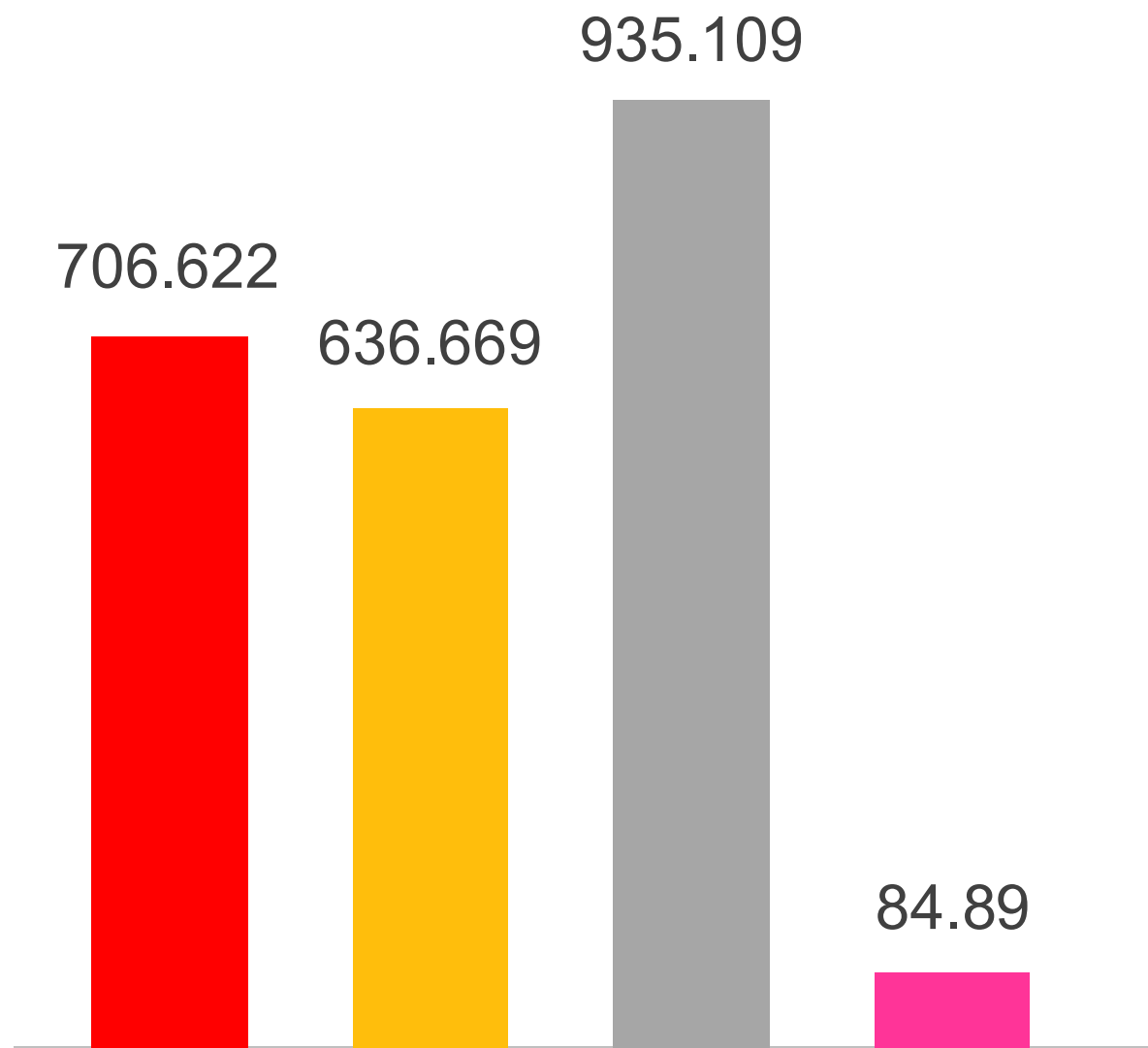
## Ryuji Wakikawa

Vice President, Head of Research Institute of  
Advanced Technology  
SoftBank Corp.



# SoftBank's AI-RAN Initiative

# Monthly Aggregate Network Traffic of All Users (Unit : PB)



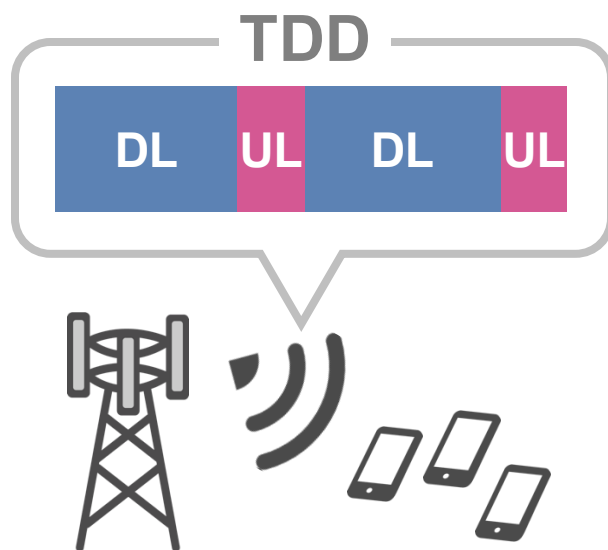
**SoftBank is the operator with the highest network traffic in Japan**

※By group (3G, 4G/5G, Advanced BWA + 5G)

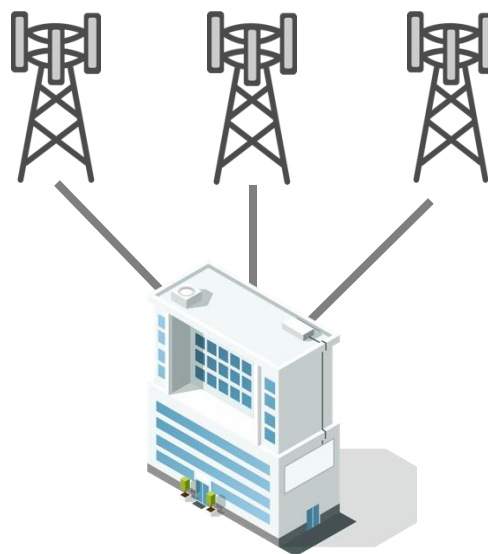
Source: created based on "Summary of the Results of the FY2023 Survey on Actual Radio Spectrum Usage on Mobile Phones and Nationwide BWA"

# SoftBank's Measures on Network Capacity Expansion

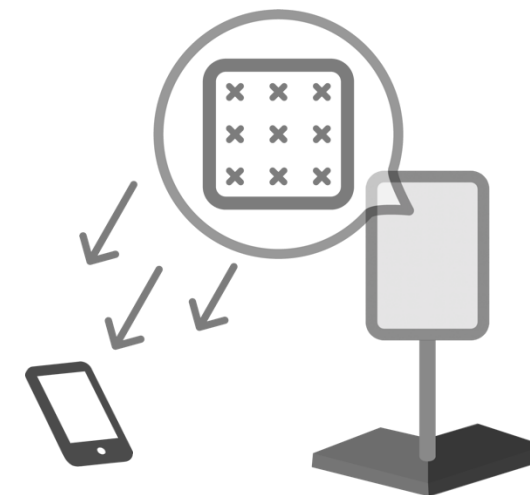
## Implementation of TD-LTE



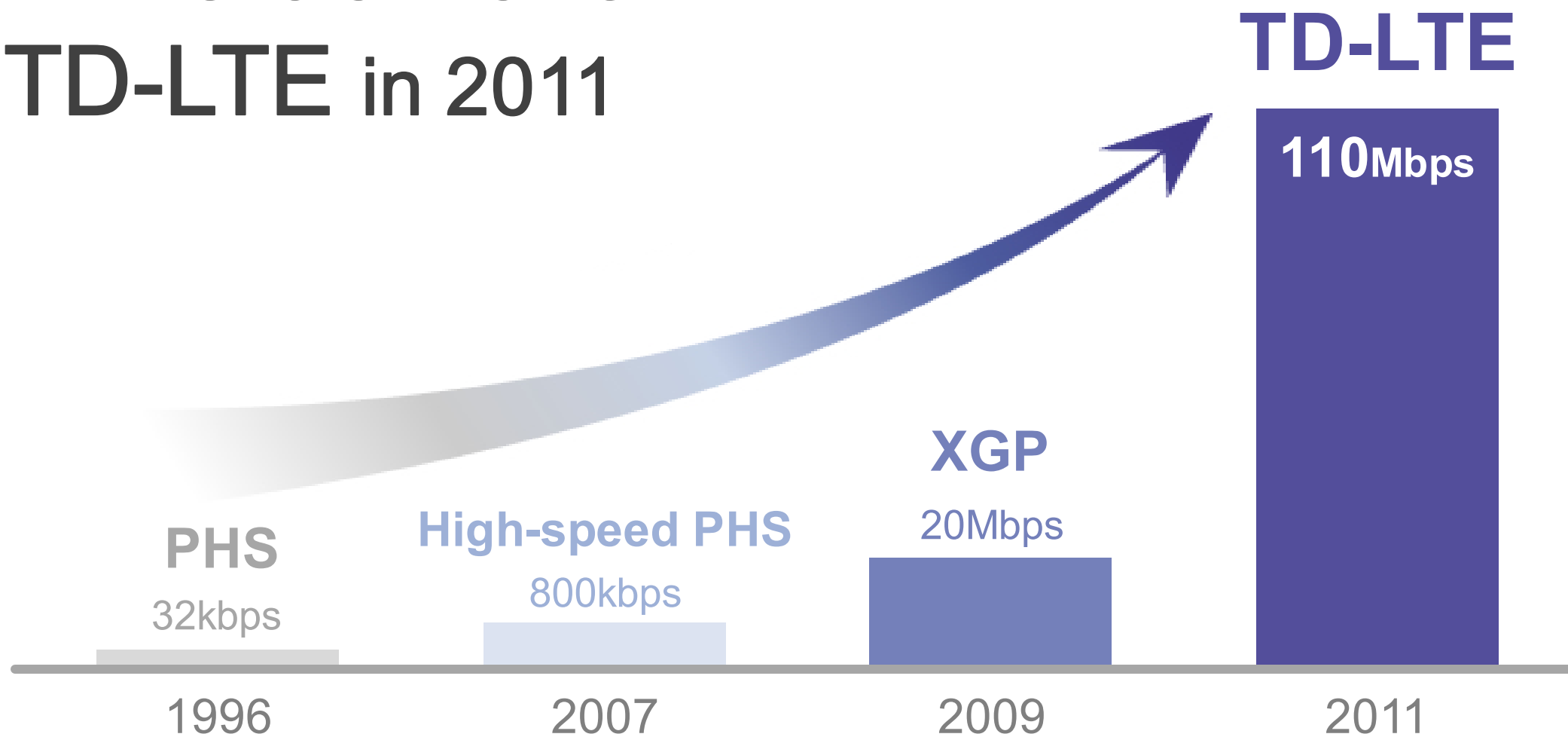
## Implementation of C-RAN



## Implementation of Massive MIMO



# World's First Commercialization of TD-LTE in 2011





Difficulty of  
interference control in  
high-density areas

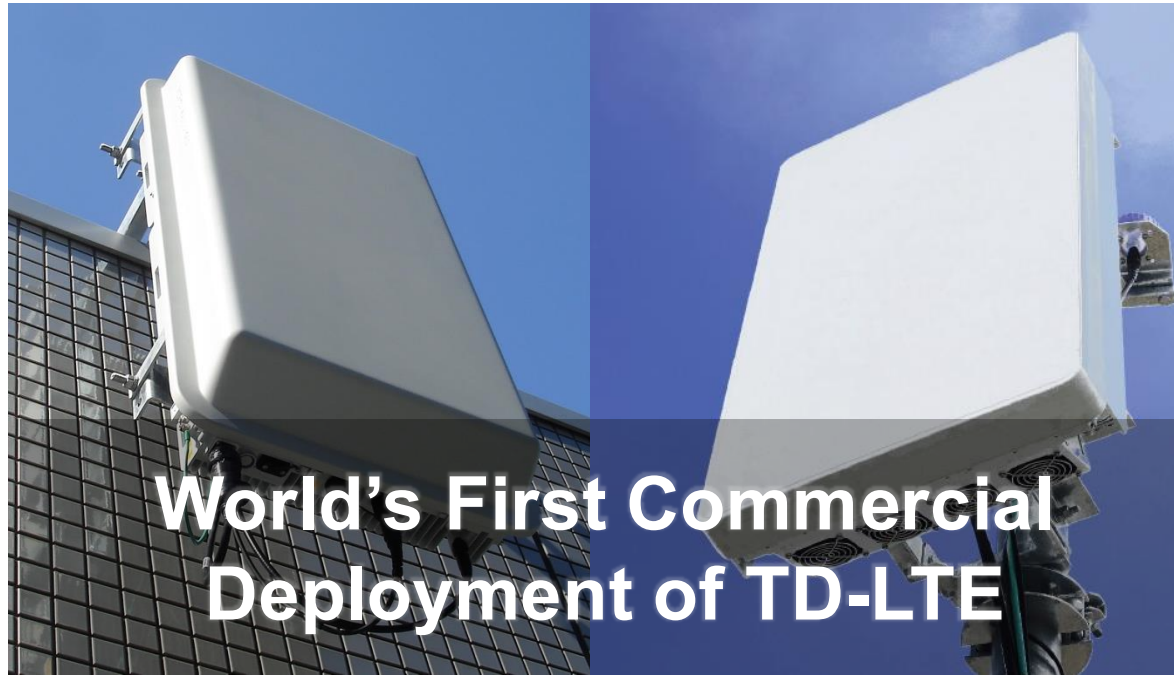


**C-RAN implemented  
to enable better coordination  
between base stations**



# Implementation of Massive MIMO in LTE

Developed pricing strategies to address high traffic needs



**First launch in  
2016**

Spectral efficiency:  
Achieving breakthrough improvement in  
wireless capacity

# OpenRAN



# Performance Limitations of oRAN (TDMA)

	Ant # (Element #)	MIMO stream #	Bandwidth (MHz)	
<div>O-RAN</div>	8T8R (8)	8	100	<div><div></div><div>x 2</div></div>
<div>SoftBank</div>	32T32R (128)	8	200	
	64T64R (192)	32	500	

- \* FD-LTE can be processed by software (bandwidth is ~20 MHz)

AI x Next Generation Mobile Network

AI - RAN



nvidia

arm

SoftBank



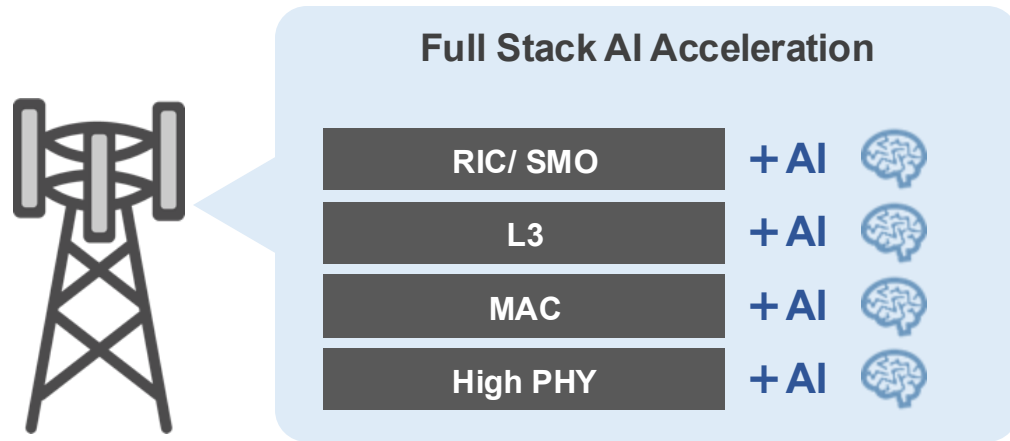
# AI-RAN

# oRAN + AI + Chip(GPU)

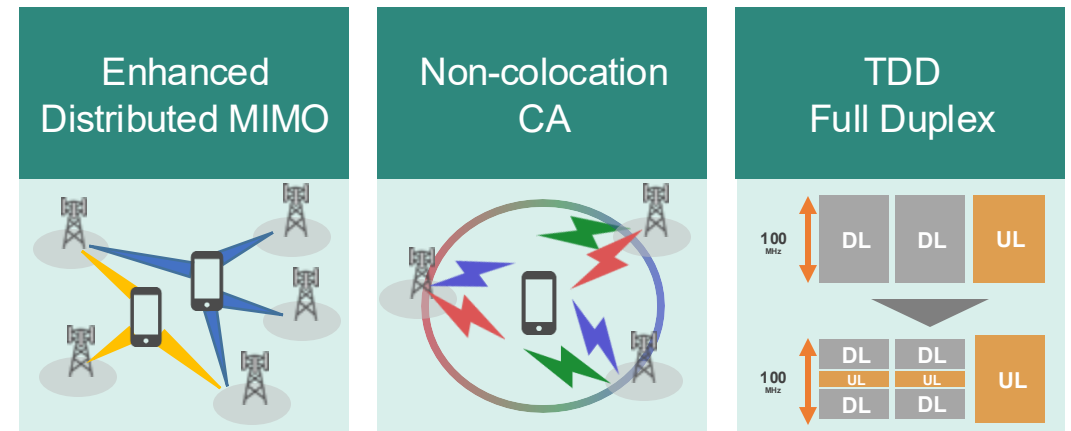
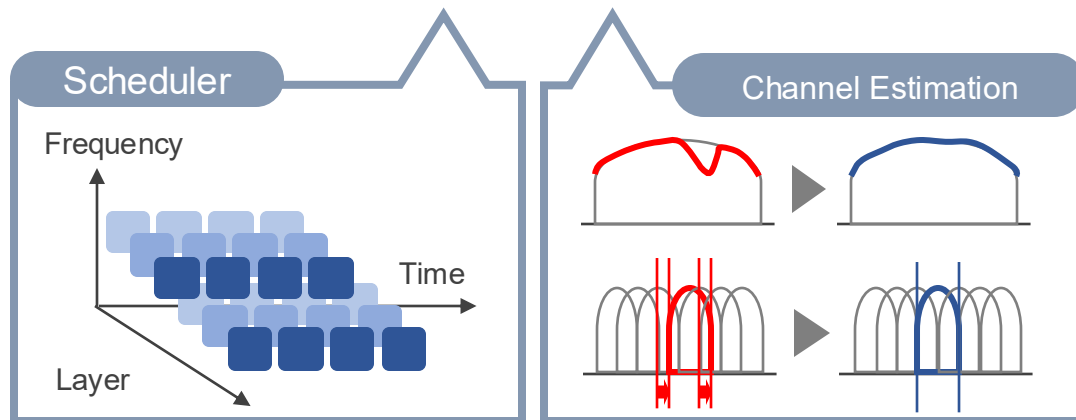
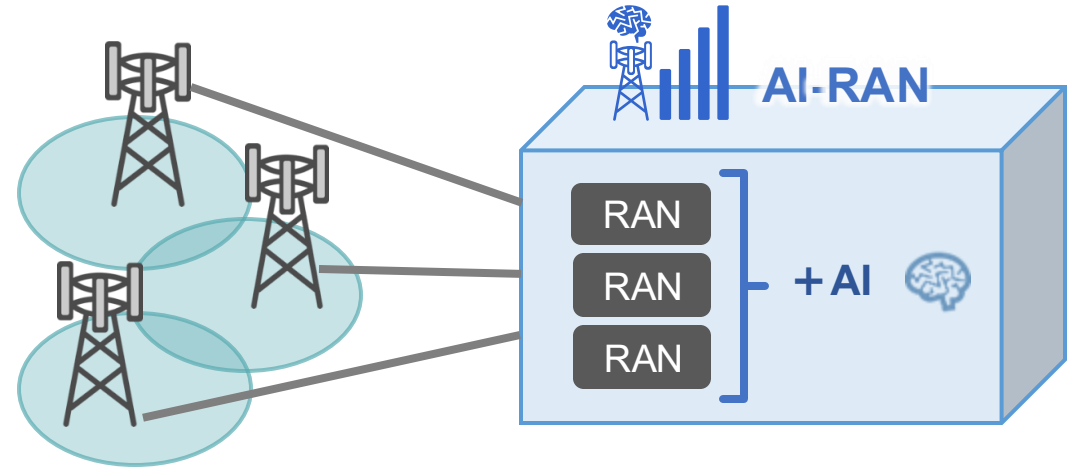
*This is where Innovation occurs*

# Radio Improvement (AI-for-RAN)

## Individual Cell



## Inter-Cell Coordination

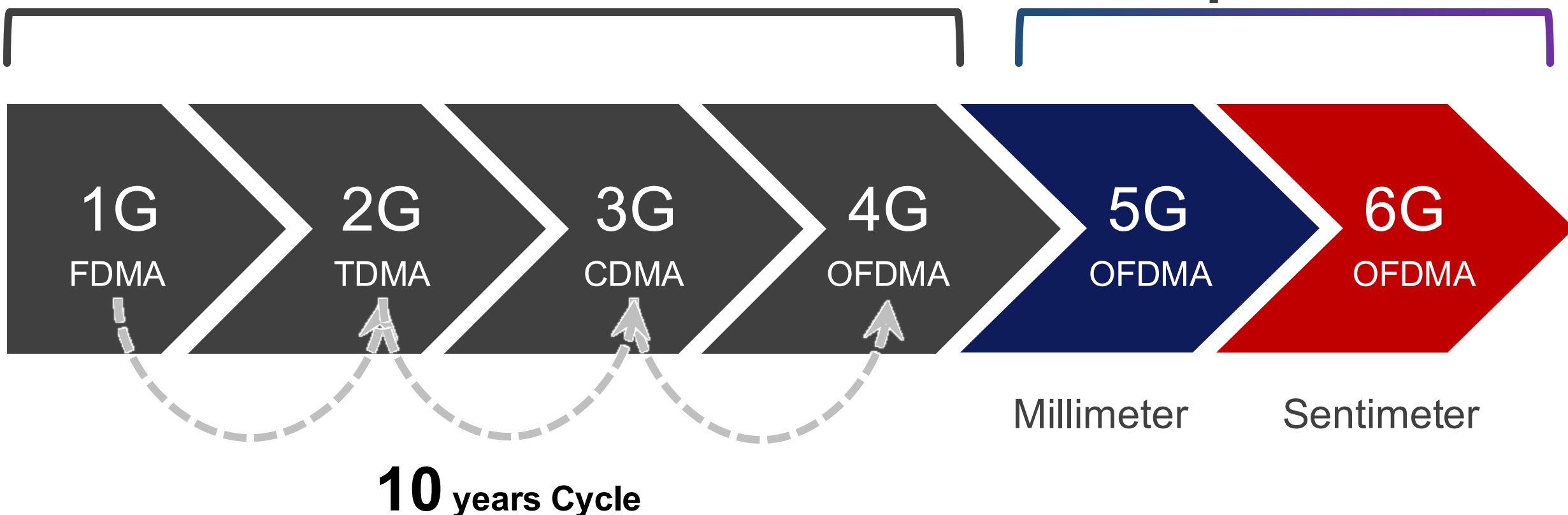




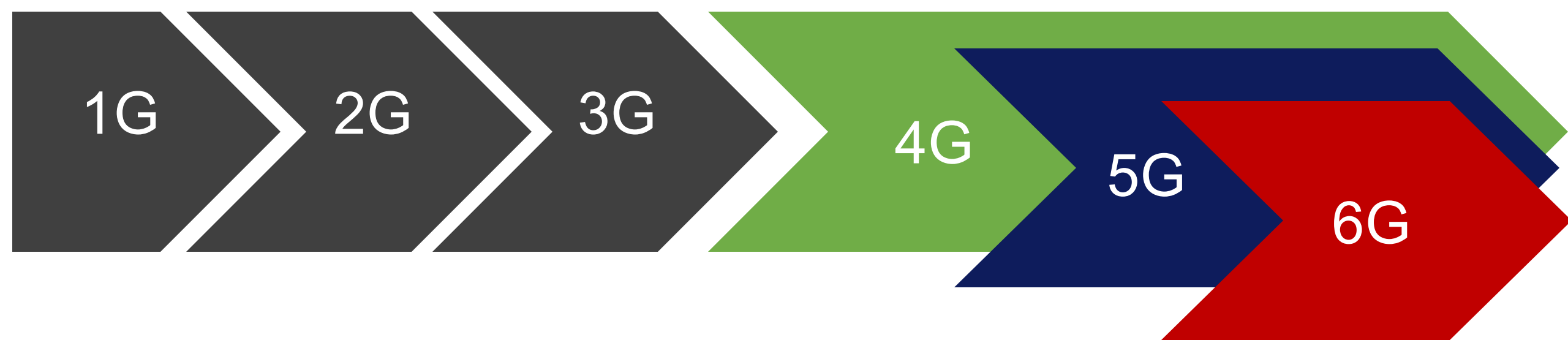
# Technological Advancements Limited with 5G

Technological Innovation

Frequency Expansion

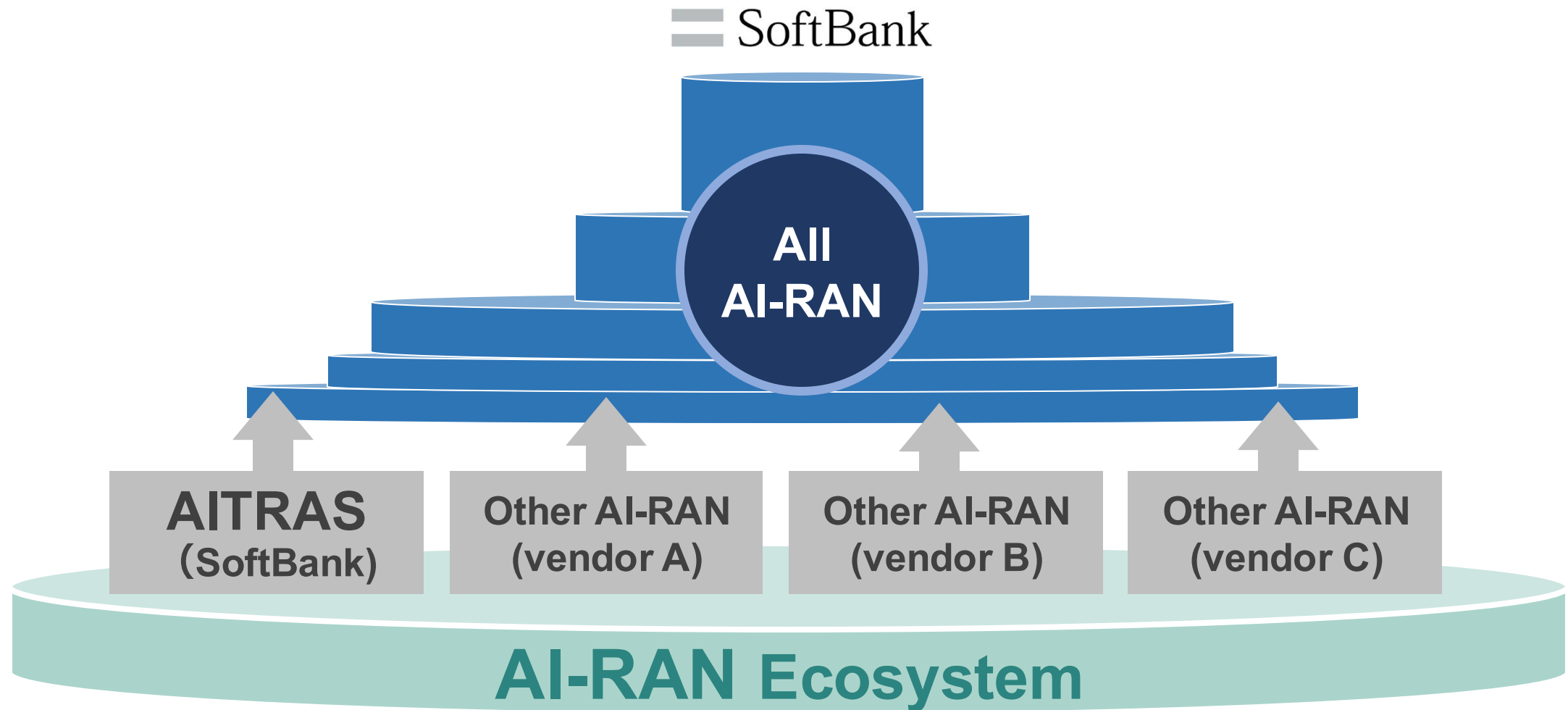


# AI-RAN : Towards continuous evolution





# Beyond Full AI-RAN Transformation



 SoftBank

# Speaker



## Alex Jinsung Choi

Principal Fellow  
SoftBank Corp.



The background features a complex geometric design with overlapping triangles in shades of light blue, white, and dark blue. At the bottom, a network diagram is visible, consisting of small white and blue dots connected by thin lines, suggesting a digital or technological theme.

# AITRAS

# AI-RAN Definitions

**AI-RAN (Artificial Intelligence Radio Access Network)** refers to the integration of AI technologies into Radio Access Networks (RAN) to optimize network performance, automate resource management, and enable new AI-driven services. It transforms traditional network infrastructure into intelligent, adaptive, and revenue-generating platforms by leveraging AI-powered data analysis, prediction models, and automated control mechanisms.

**AI-RAN** integrates AI technology in RAN to improve mobile network efficiency.



Three items established by **AI-RAN Alliance** include **AI-for-RAN, AI-and-RAN, AI-on-RAN.**

1. **AI-for-RAN:** Enhancing network operations such as load balancing, interference management, and network optimization using AI/ML models.
2. **AI-and-RAN:** Sharing infrastructure to run both RAN and AI workloads, enabling parallel processing and creating new revenue streams.
3. **AI-on-RAN:** Embedding AI capabilities directly into RAN components for real-time decision-making and operational automation.

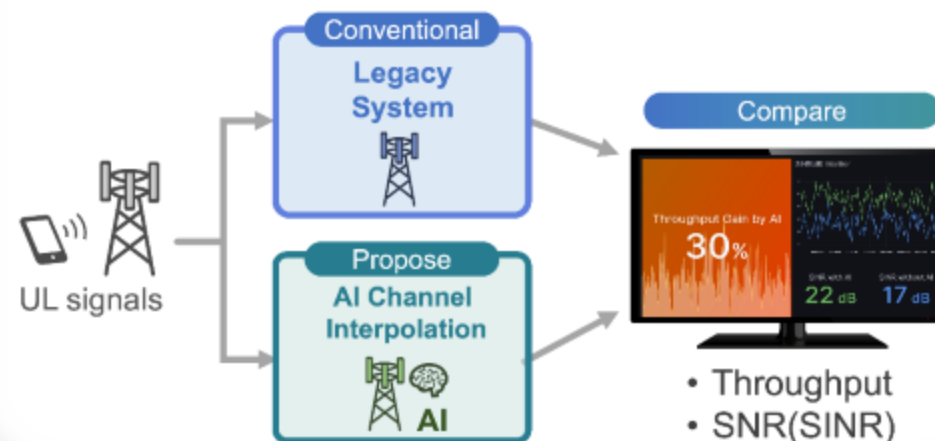
# Early Research and AI-RAN Development

## Initial Efforts

SoftBank's early AI-RAN research focused on integrating AI-driven signal processing into RAN to enhance 5G performance. The research explored channel estimation, AI-based beamforming, and AI-powered traffic management to overcome network congestion and improve service reliability. The development of an AI-RAN data center enabled parallel operation of RAN and AI workloads, unlocking new revenue streams through AI-powered services.

AI for channel interpolation in lower layers of wireless communication

30% Throughput Gain





# Introduction of GPU based vRAN - gRAN



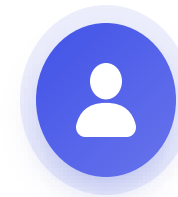
01

## GPU for vRAN Evolution

GPUs handle highly parallel processing workloads efficiently.



\*NVIDIA GH200 Grace Hopper Superchip



02

## Technological Leap with GPU-based vRAN

Enables modern AI services like LLM inferencing at the edge.

# Key Components of AITRAS

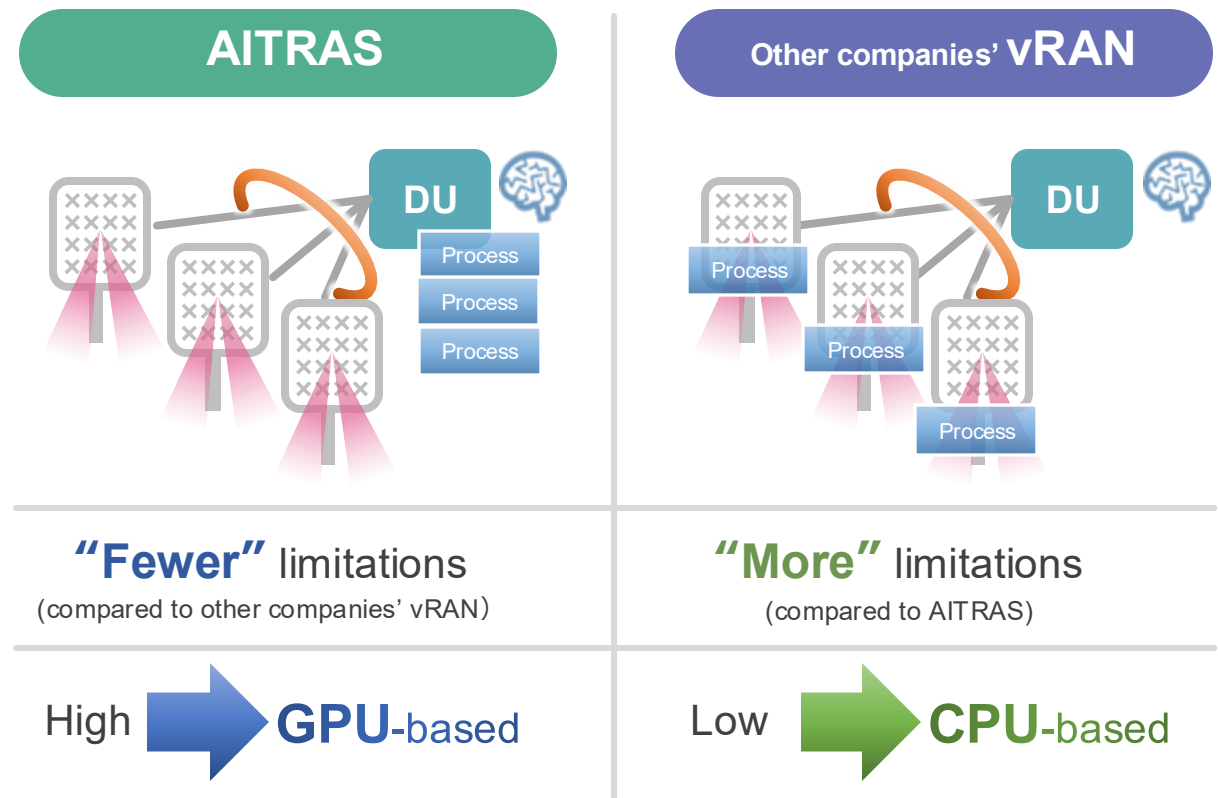
## System Components and Architecture

01

Virtualization platform with RAN functions structured with L1, L2, and L3 layers.

02

Edge AI and orchestrator for dynamic resource allocation.



# AI-Driven Orchestration

## AI Orchestration Layer

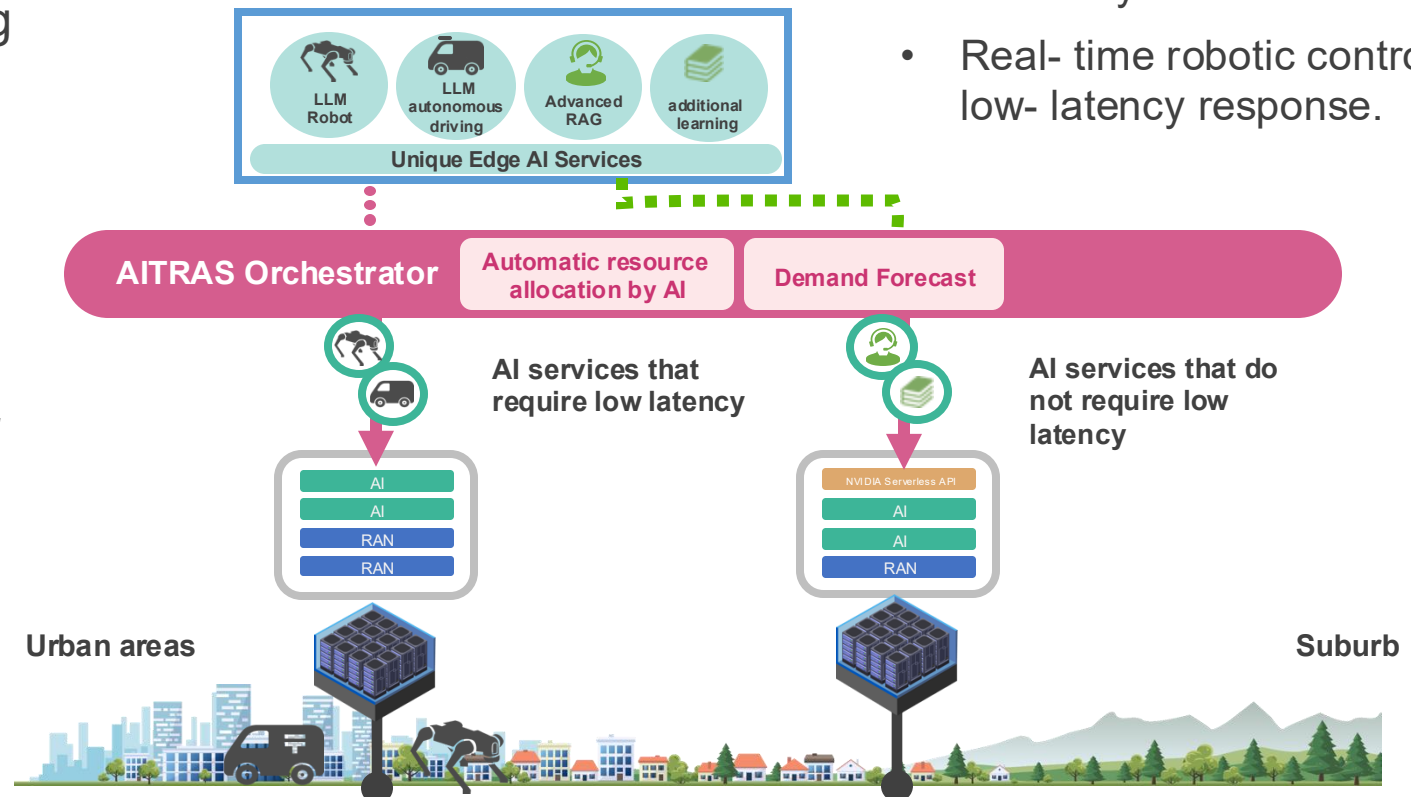
01

AI-driven orchestration for managing and optimizing workloads.

02

Includes automatic resource allocation by AI and dynamic changes of server roles.


- Multi- Modal AI for remote autonomous vehicle support.
- Edge RAG for operational efficiency.
- Real- time robotic control with low- latency response.






# Key Benefits of AITRAS

## Core Benefits



Cost reduction through  
consolidated AI and  
RAN workloads.



Optimized resource  
utilization and new  
revenue opportunities.

# Outdoor Testbed for AITRAS

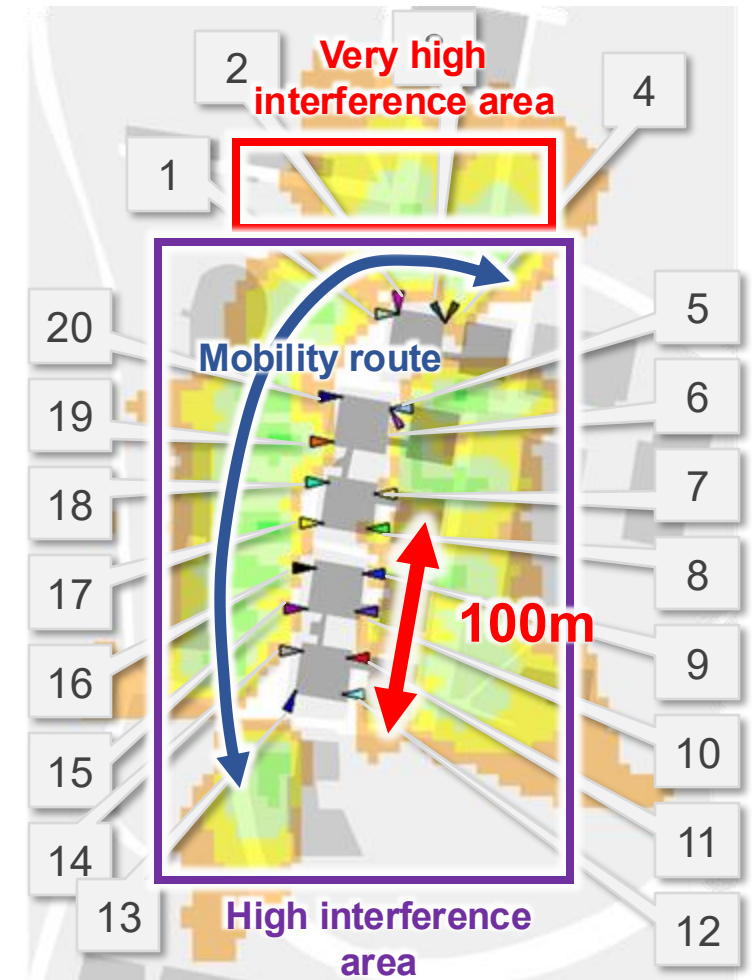
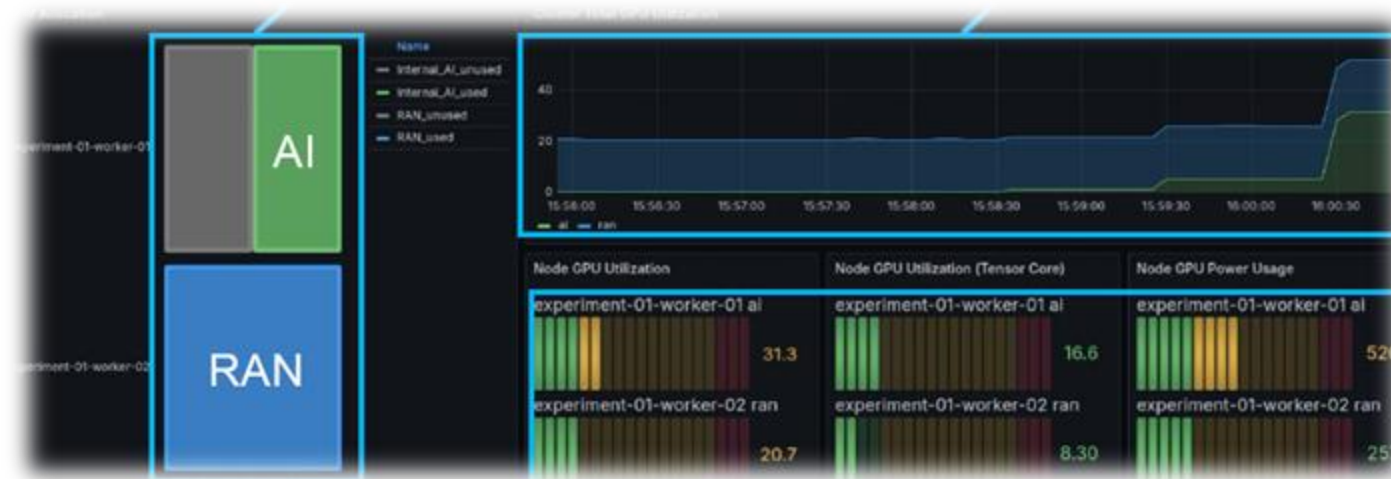
## Trial Results



Carrier-grade stability in the AITRAS trial in Kanagawa.



Tested with 100 User Equipment streaming videos simultaneously.



# AITRAS Performance Evaluation

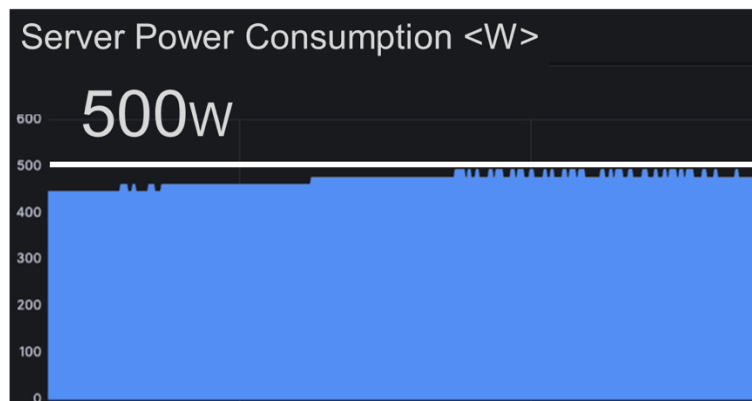
## Performance Metrics

Server's power consumption compared to current RANs.

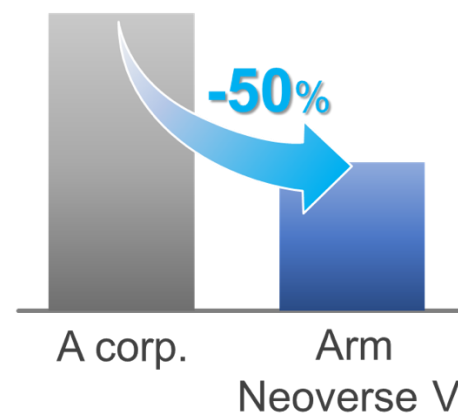
Evaluated UE accumulation and radio resource occupation.

Power rating  
1,500W  
(CPU+GPU)

**500W** (25W / cell)



**CPU power consumption※**

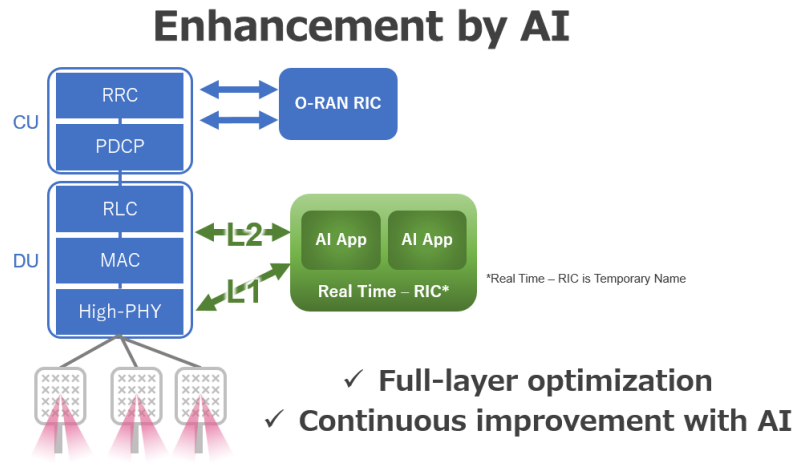


※Under the same processing conditions  
SoftBank lab measurement results



# SoftBank's L1 Enhancements in AITRAS

## RAN Performance Improvements



DL : **1.3Gbps** / Cell  
UL : **180Mbps** / Cell

GPU acceleration for higher processing speeds and throughput improvements.

AI- driven L1 optimizations for cell capacity and power consumption.

# SoftBank AI-and-RAN Approach

Management Cluster and Workload Clusters for optimized placement.

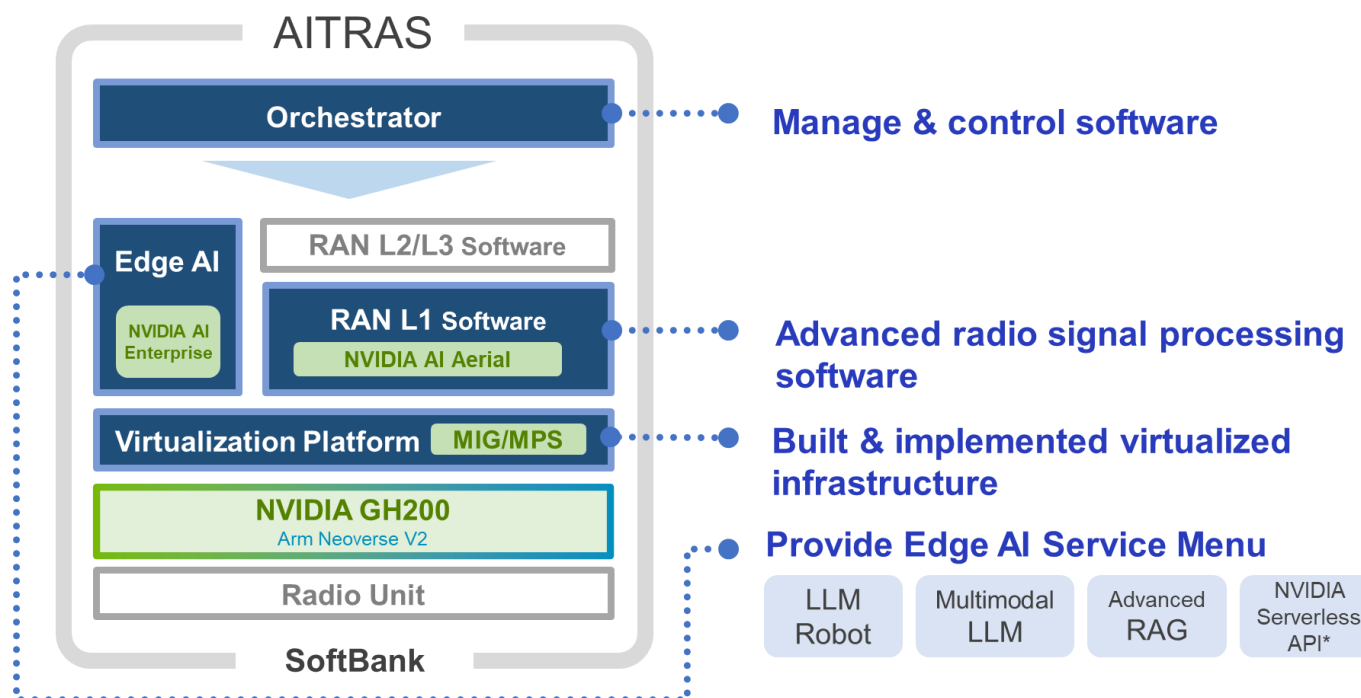


Cluster level, node level, and core level resource management.



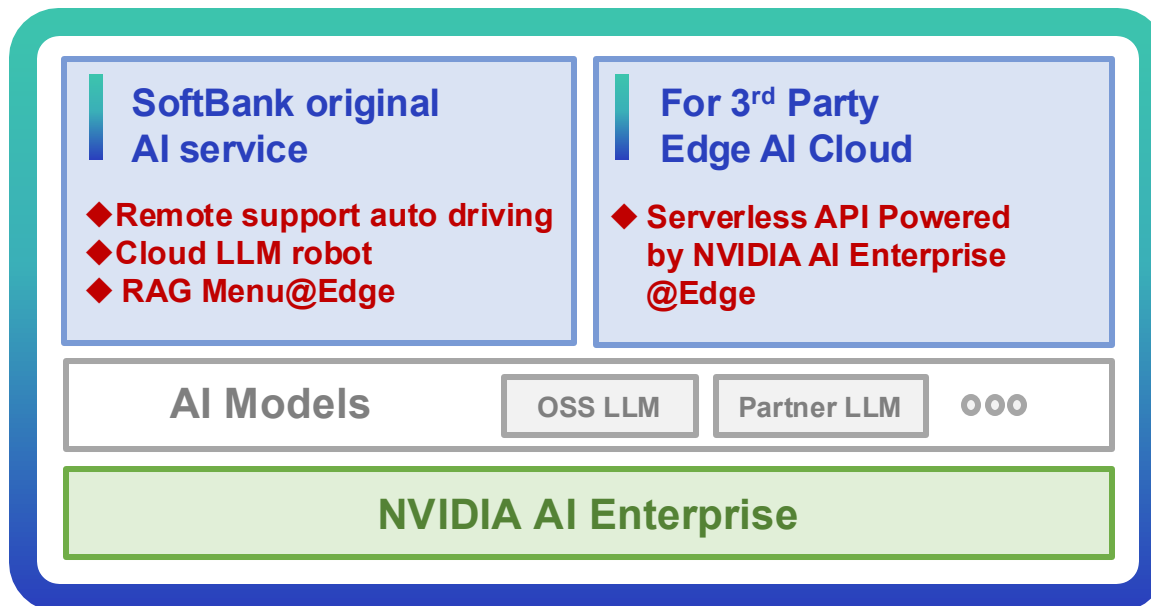
Use of CPU- GPU integrated servers for efficient handling of tasks.

## Virtualized Clusters



# Agentic AI – Serverless API powered by NVIDIA AI Enterprise

## Serverless API Benefits



Supports flexible  
and scalable AI  
deployments.

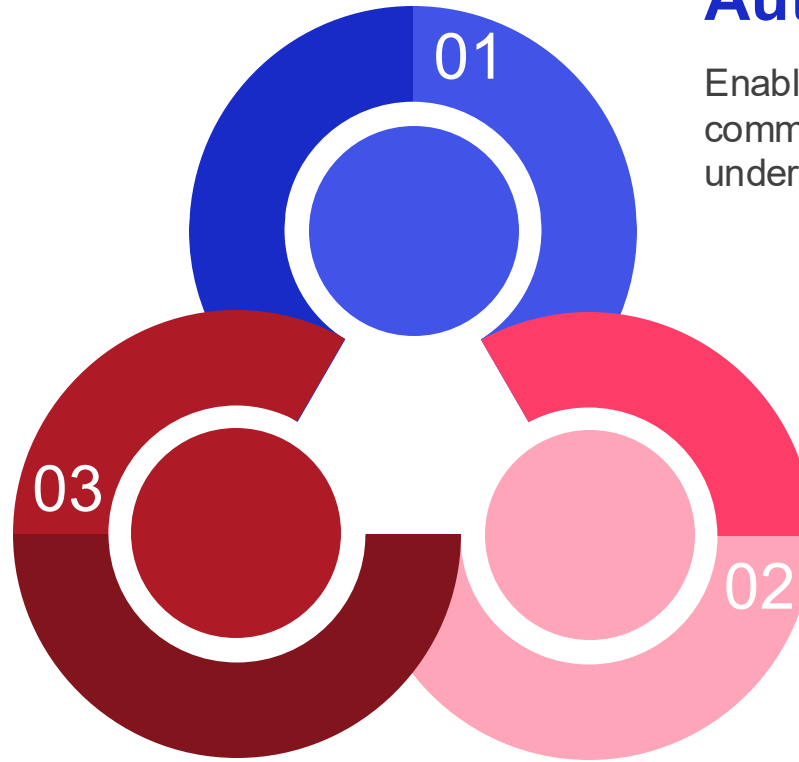


Interfacing with  
the AI-and-RAN  
infrastructure for  
optimized GPU  
usage.

# Use Cases for the AITRAS AI-on-RAN

## Retrieval-Augmented Generation (RAG) Chatbots

Use RAG chatbots for customer service and network support utilizing secure processing.



## Autonomous Driving

Enables real-time V2X communication and multimodal traffic understanding AI.

## LLM Robots

LLM robots perform with high-speed control LLM on AITRAS.

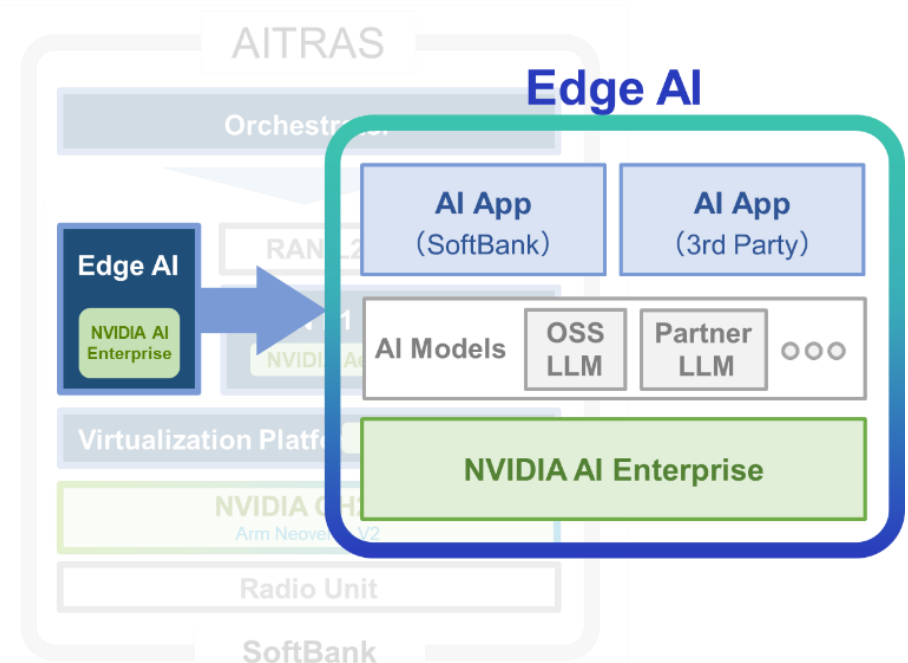


# AITRAS

## AI-and-RAN for New Revenue Generation

### Strategic AI Integration

Edge AI inferencing and containerized infrastructure supporting digital transformation.



 SoftBank

# Speaker



## Rajeev Koodli

Principal Fellow  
SoftBank Corp.



# AI for Telco Vision and Concept



# Outline

- Introduction: Telco and AI Inflection point
- SoftBank Vision
- Use Cases: Human AI, Machine AI
- AI Pillars: Data, Infrastructure, Models
- AI Tool Kit, Operationalizing AI
- Next Steps & Conclusion

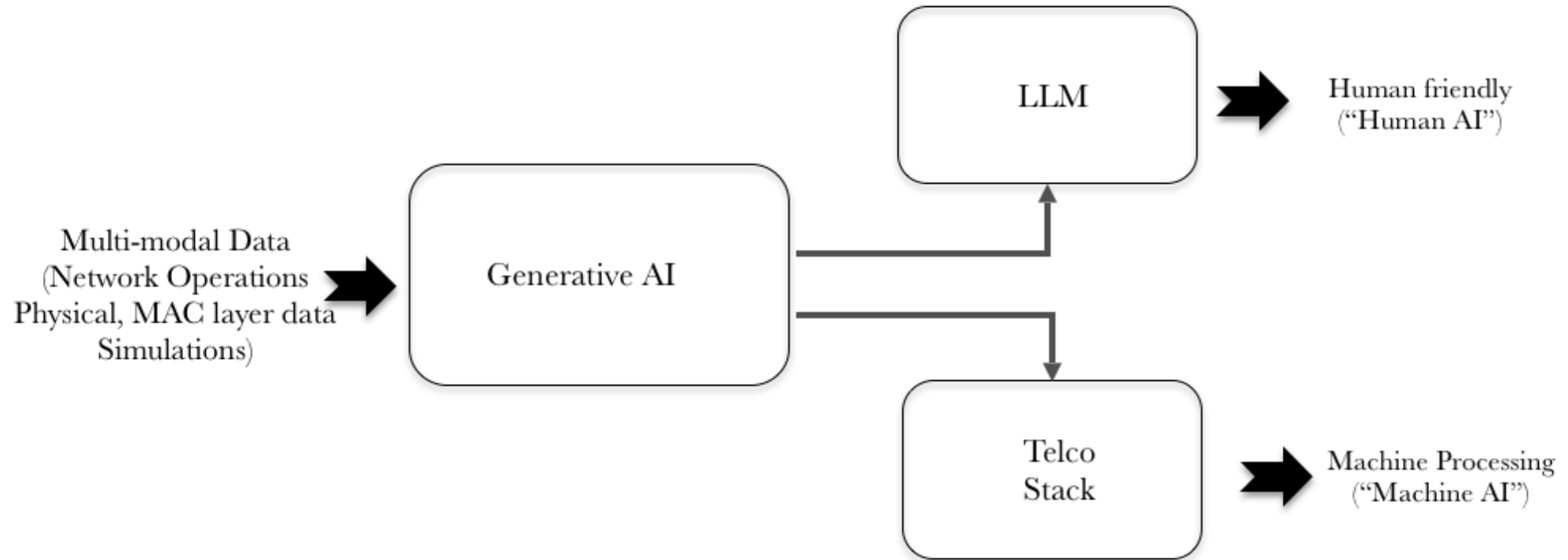
# Telco and AI: Inflection Point?

	AI-RAN Alliance	O-RAN Alliance	3GPP & ITU	Global Telco AI Alliance
Key Initiatives	<ul style="list-style-type: none"> <li>AI-for-RAN: AI to enhance capability of RANs</li> <li>AI-and-RAN: Integration of AI and RAN processing infrastructure</li> <li>AI-on-RAN: Deploying AI services at the network edge</li> </ul>	<ul style="list-style-type: none"> <li>RAN Intelligent Controller (RIC): AI/ML-based solutions for traffic steering, energy savings, anomaly detection</li> <li>Open Software Development: Collaboration with Linux Foundation</li> <li>Global PlugFests: Testing and integration of O-RAN solutions</li> </ul>	<p>3GPP</p> <ul style="list-style-type: none"> <li>AI/ML for 5G and Beyond: Enhancements for beam management, positioning, energy savings</li> <li>Lifecycle Management: Frameworks for managing AI/ML models</li> </ul> <p>ITU</p> <ul style="list-style-type: none"> <li>AI for Good: Promoting AI to address global challenges.</li> <li>AI Standards: Framework for AI governance</li> </ul>	<ul style="list-style-type: none"> <li>Telco LLM Development: AI-powered customer service and network management</li> <li>AI Governance: Focus on ethical and sustainable AI deployment</li> <li>Ecosystem Building: Integration of AI solutions into telco services and new business models</li> </ul>
Outcomes	<ul style="list-style-type: none"> <li>Demonstrated feasibility of AI-native RANs in trials and pilot projects</li> <li>Blueprints and benchmarks</li> </ul>	<ul style="list-style-type: none"> <li>Standardization of AI/ML frameworks for RANs</li> <li>Acceleration of open RAN adoption</li> </ul>	<ul style="list-style-type: none"> <li>Promoting AI standardization and governance in telecom networks</li> </ul>	<ul style="list-style-type: none"> <li>Initial applications of Telco LLM in customer care and network optimization</li> </ul>

Telcos and Industry Organizations are embracing AI

Developing and Deploying AI is necessary (prior to standardization)

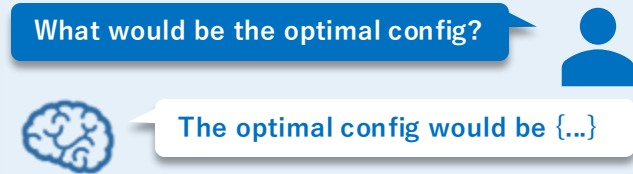

# SoftBank Vision



## Generative AI for Telco:

1. Network Operations Efficiency (Opex)
2. Network Stack Processing (Capex)

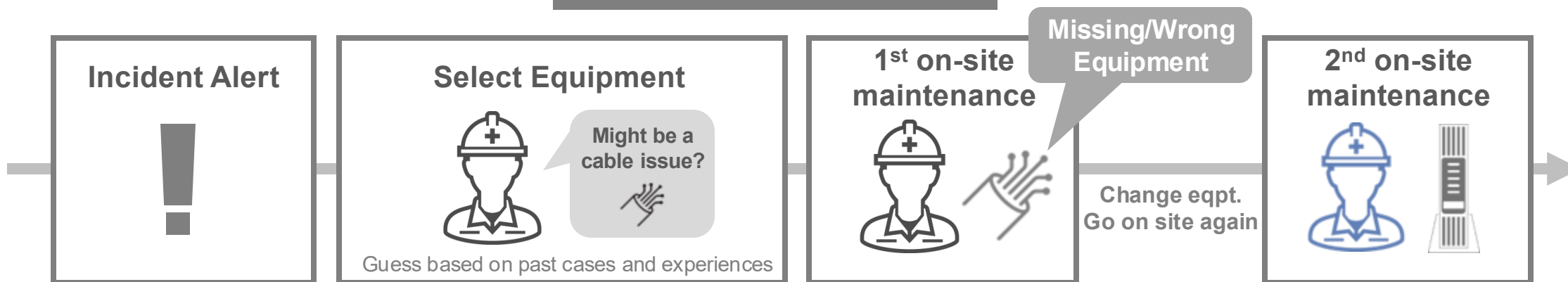
# Human AI vs Machine AI

	AI Model(s)	
	Human AI	Machine AI
	Use case A   Use case B   ...	Use case C   Use case D   ...
Interaction	Human <-> AI	Machine <-> AI
Target Layer	Higher (Operation, etc.)	Lower (Signal Processing, etc.)
Interaction Freq.	Daily~	~us, ~ms
Data Modality	including Natural Language	not including Natural Language
Architecture	DNNs+LLMs	DNNs
# of parameters	Billions~	Millions~
Example		
	An AI interacting with humans in daily operations	An AI integrated in real-time signal processing



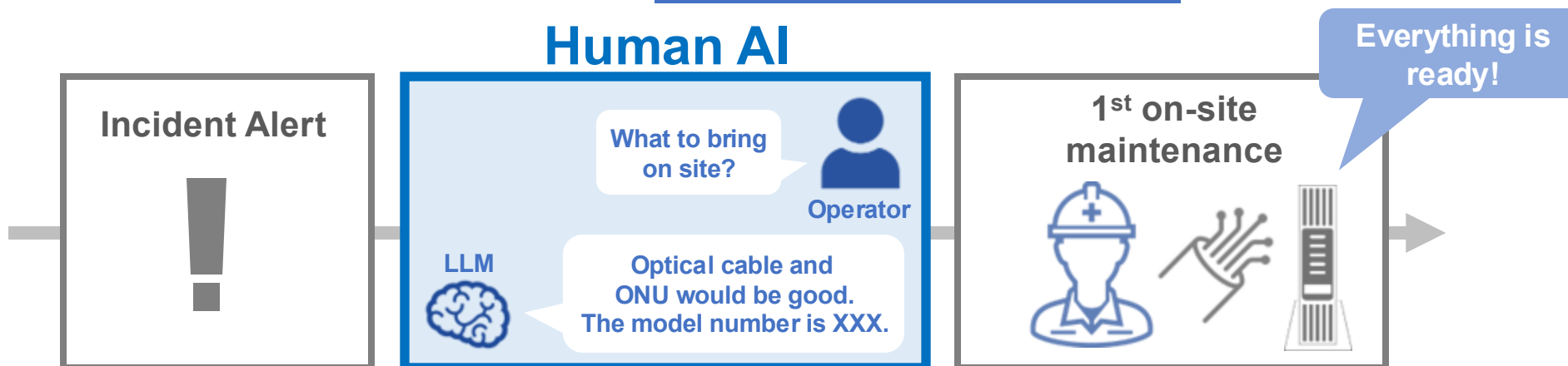
# Human AI Use Case

## Traditional Flow



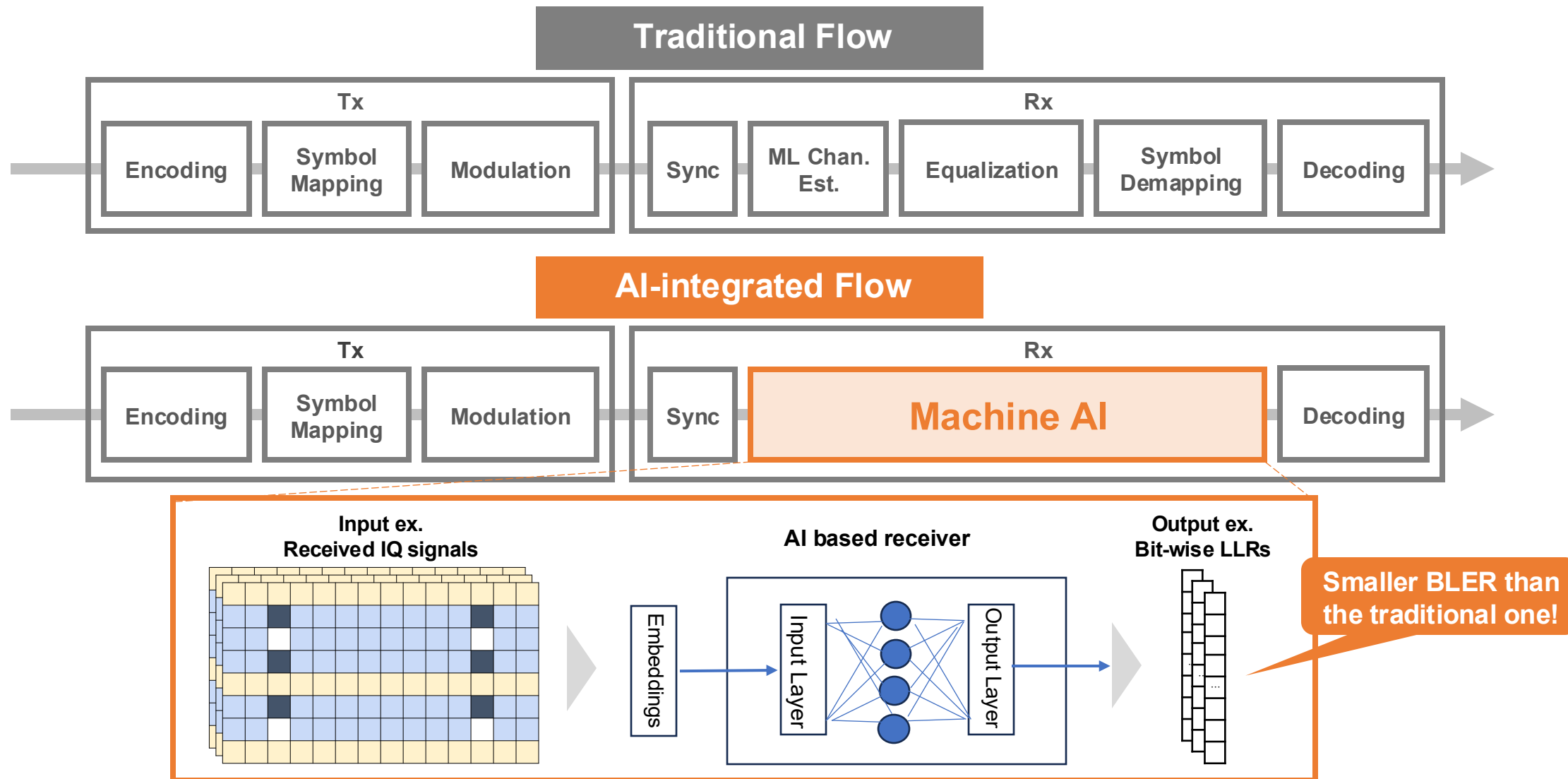
## AI-integrated Flow

### Human AI



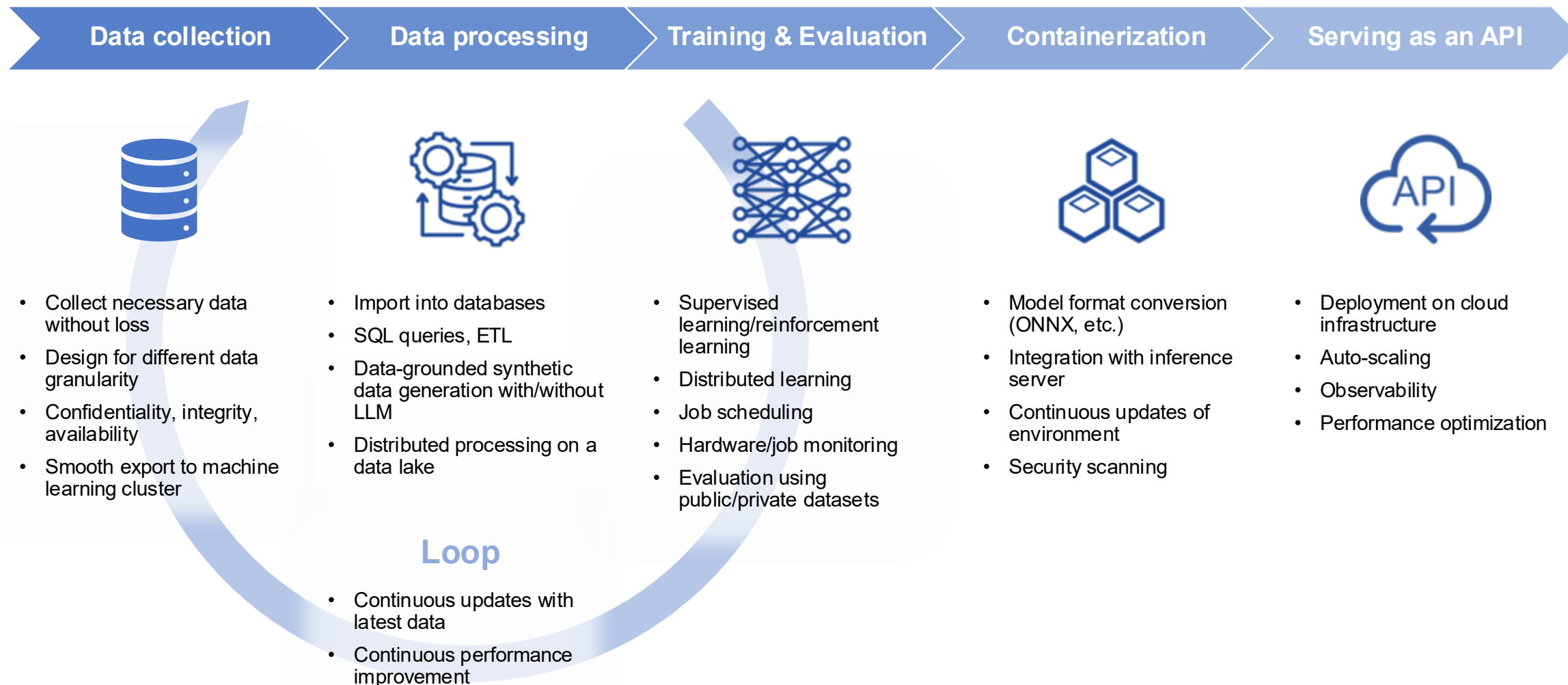
Inference based on the hardware alarm, transmission-side behavior, actual data logs from multiple nodes, and past cases.

# Machine AI Use Case



# AI Pillars

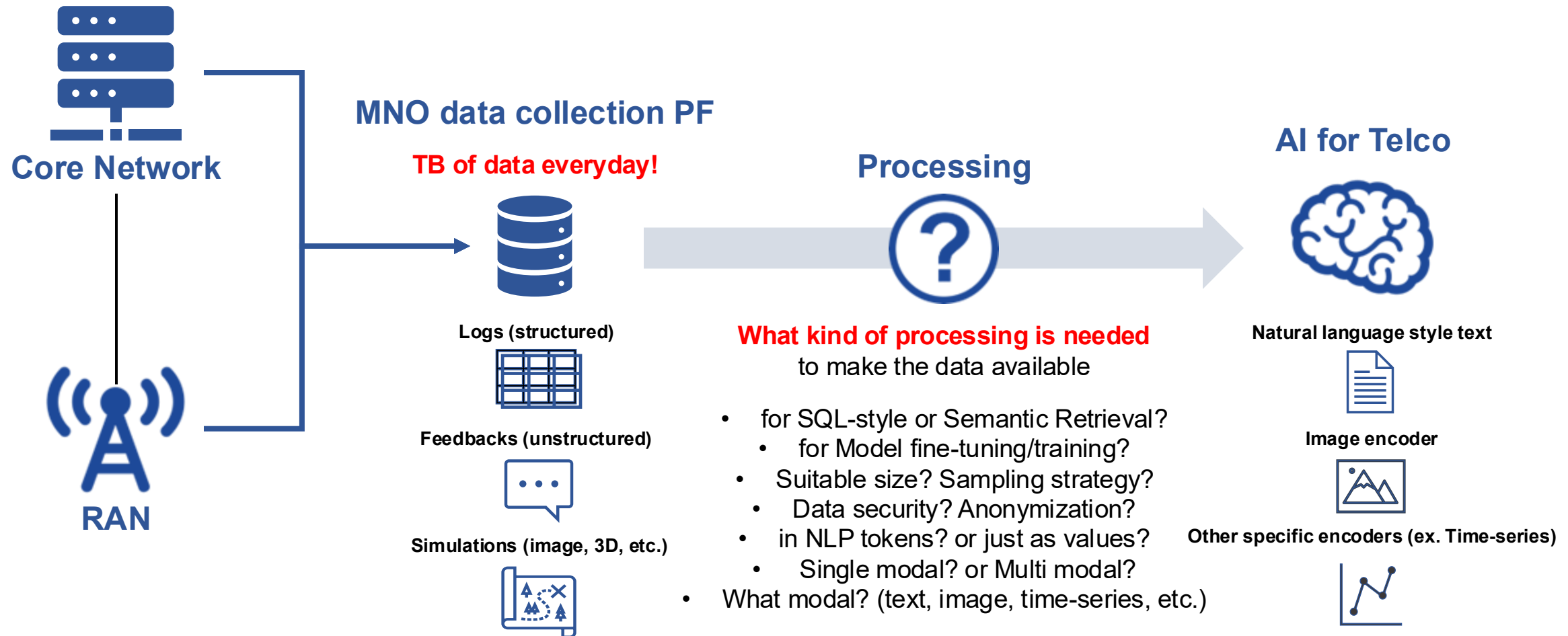
# AI Model End-End Pipeline





# Challenges with data

For all AI generation, we need to tailor the data (Data Processing)



# Example Data Size

## Total of ~4TB of Network Data

	15 minutes	hourly	daily
RAN	2024/05/07-2024/08/12 (1TB~)	2023/08/01-2024/08/12 (1TB~)	-
Core	2024/05/05-2024/08/12 (~50GB)	2023/08/01-2024/08/12 (~50GB)	-
Config	-	-	2024/07/16-2024/08/12 (500GB~ *zipped)

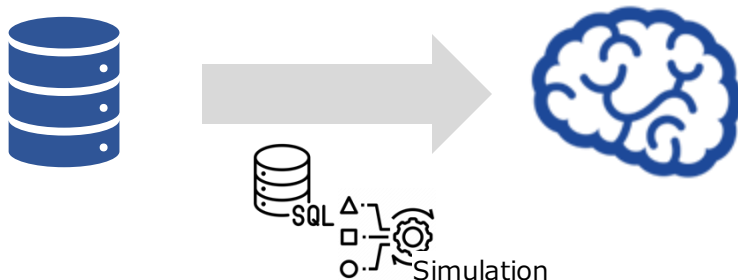
# Challenges with infrastructure

We need infrastructure to process the data and train the models

## Data Processing

Huge raw data

AI ready data



**Significant compute capacity needed to process the raw data**

## Training

Fine tuning



Full training



Computation resource required (Estimated, Unit: GPU Hour = 1 Hour spent on 1 DGX A100)  
Based on LoRA Parameter Efficient Fine Tuning (PEFT)

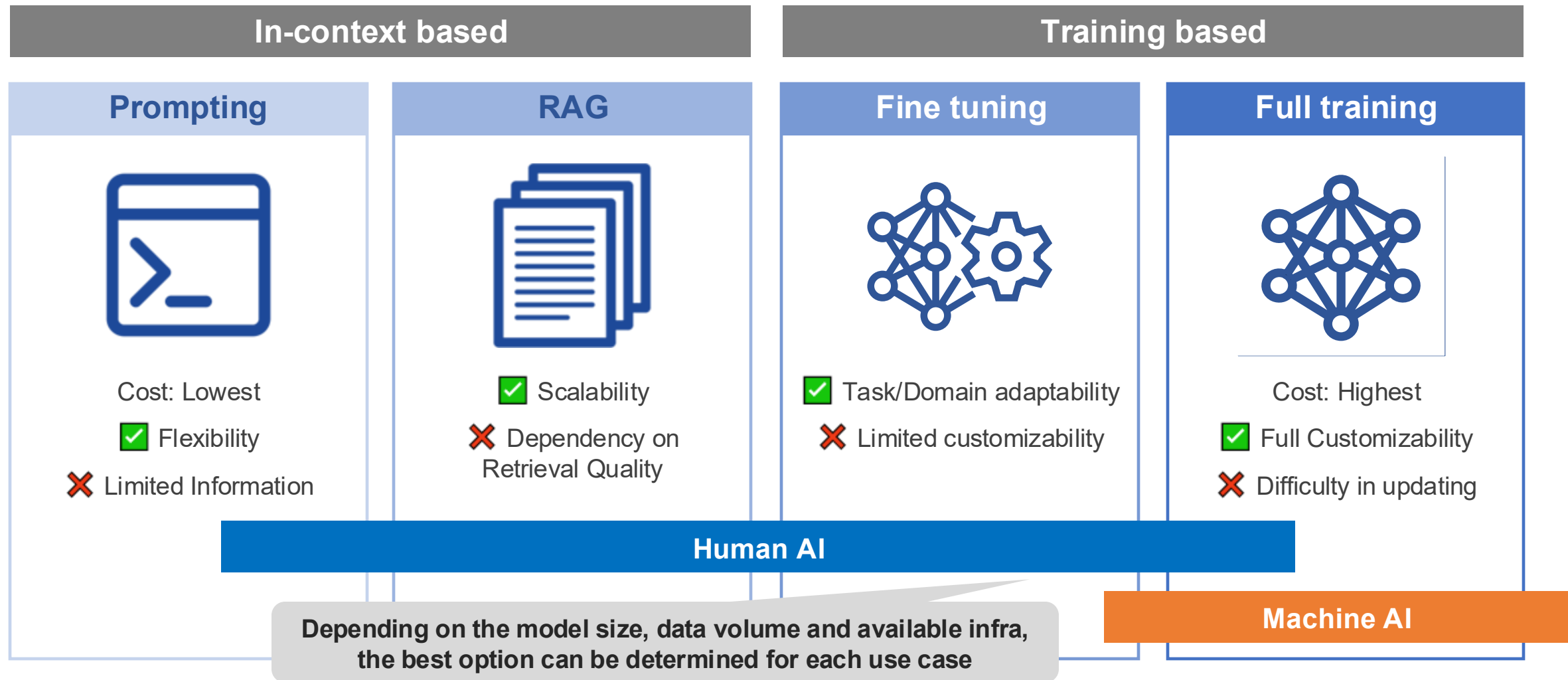
Data size	1k cells	5k cells	10k cells	200k cells
7 days of data	42 GPU Hours	9x24 GPU Hours	18x24 GPU Hours	24x365 GPU Hours
30 days	8x24 GPU Hours	38x24 GPU Hours	75x24 GPU Hours	4x24x365 GPU Hours
365 days	91x24 GPU Hours	24x365 GPU Hours	3x24x365 GPU Hours	50x24x365 GPU Hours

**GPU infrastructure is assumed for any kind of AI training**

**In LLM full training, Data Center level GPU infrastructure is needed**

# AI Tool Kit, Operationalizing AI

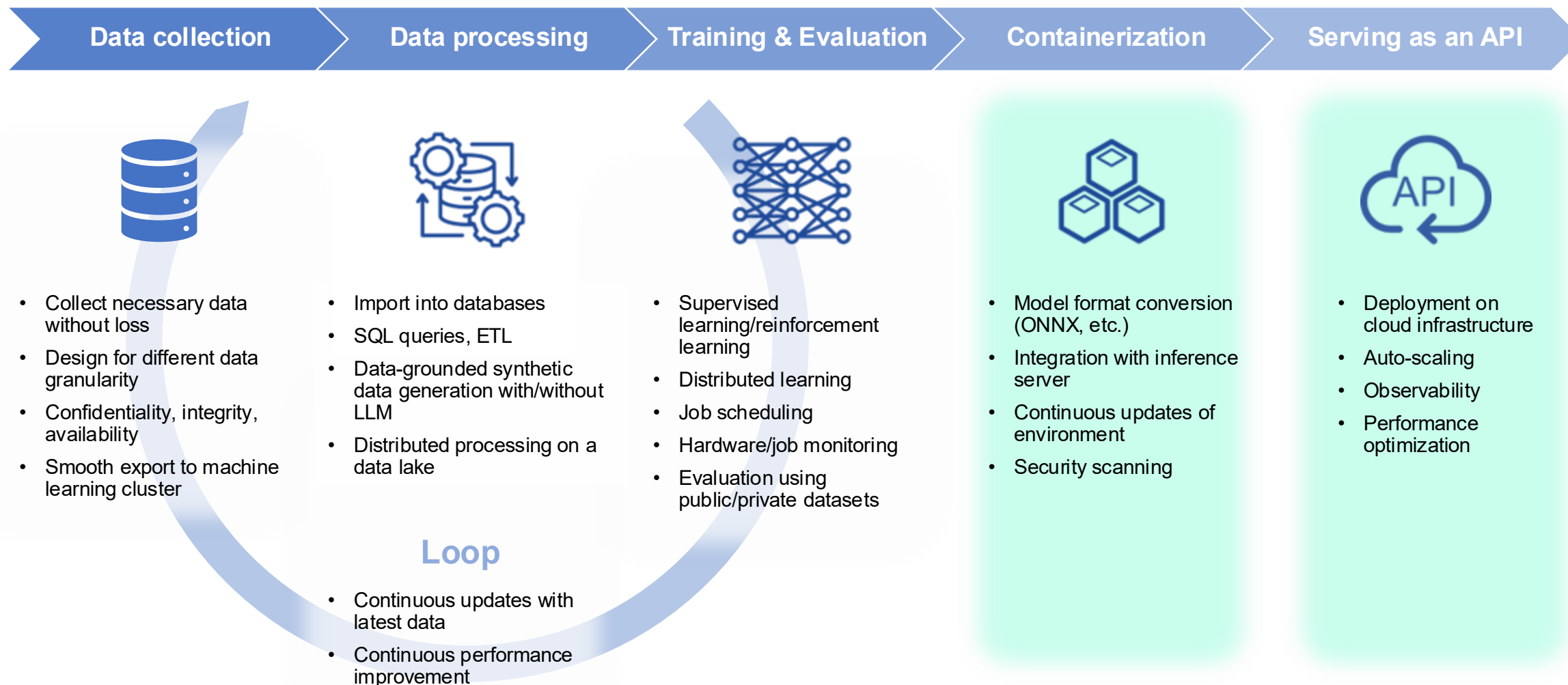
# AI for Telco: Tool Kits



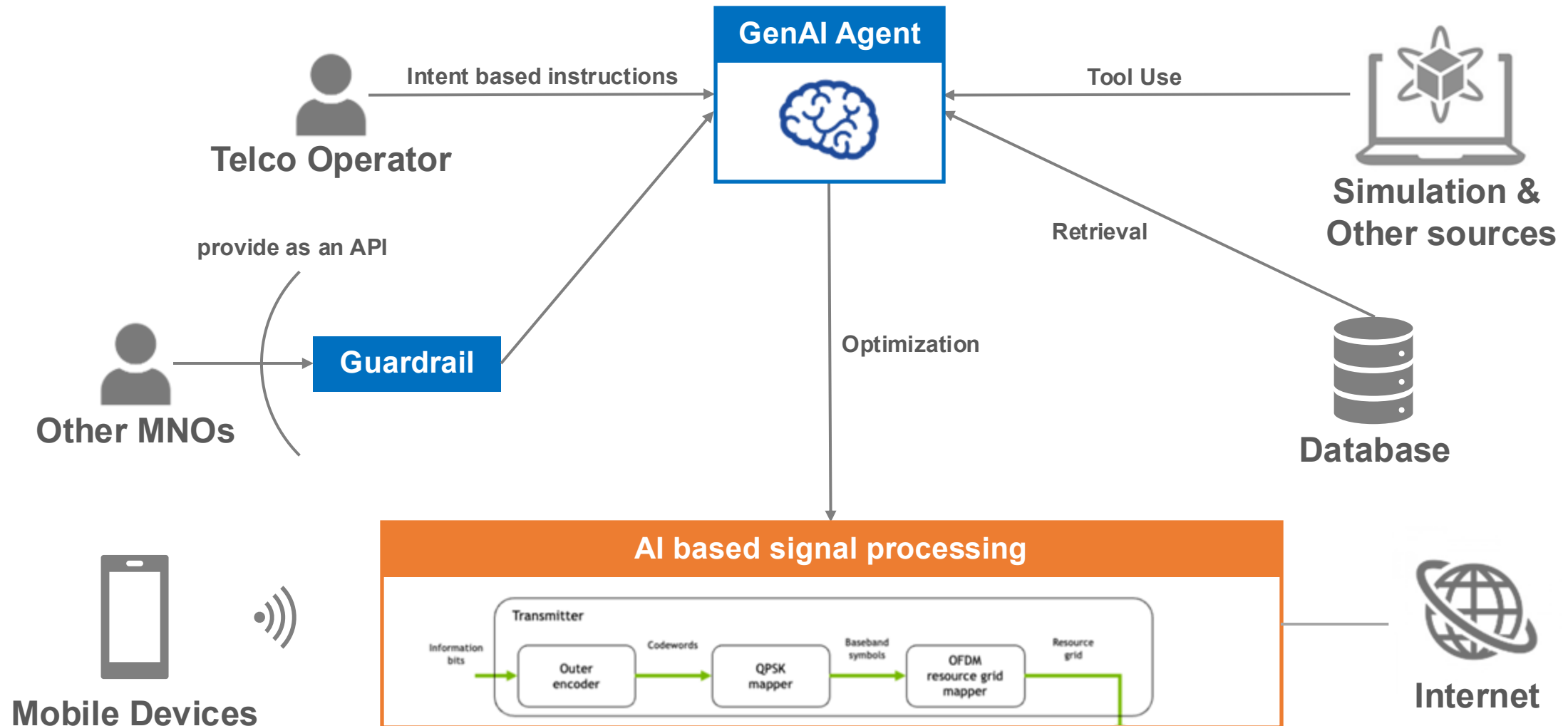
**In-context approaches augment the base or finetuned models (without updating model weights)**  
**Training Paradigm can start with Finetuning and move onto Continual Pretraining (and possibly Full Training)**



# Operationalizing AI Models



# Towards Productionization



# Conclusions & Next Steps

- Several bodies actively pursuing Telco AI initiatives. 6G is seen as a pivotal moment for AI-native air interfaces (around 2030). AI-RAN Alliance is spearheading the industry activities.
- Human AI and Machine AI have significant potential in improving the operational expenses (opex) and fundamentally reshaping the capital expenses (capex) for Telcos.
- Telcos have enormous amounts of invaluable multi-modal data. However, creating programmatic and flexible data sets out of the voluminous data is daunting yet crucial.
- Creating Telco AI models with general expertise or finetuned for specific use cases requires significant amount of GPU computing infrastructure.
- Operationalizing Telco AI models follows the same trajectory as other AI models but the scale and new skill set needed for Telcos can be vastly different compared to the current practices.

**SoftBank has embraced this journey of adopting AI for both its operational and infrastructure redesign purposes. We plan to further share our experiences in the near future.**

 SoftBank



# Speaker



## Koichiro Furueda

Deputy Director, Wireless System  
Development Department, Wireless System  
Innovation Division  
SoftBank Corp.



# AI-for-RAN R&D Update

# - Agenda -

- **What's AI-for-RAN?**
  - The Significance and Expectations of AI-for-RAN
  - Target Goal
  - SoftBank's Approach
- **Under Researching Themes of AI-for-RAN**
  - Reason why We Choose These Themes:
  - Details of Each Theme status:
    - *MU-MIMO Advanced scheduler*
    - *SRS Prediction*
    - *UL CH Interpolation*
- **SoftBank Developed System Level Simulator**
  - Development and Purpose of SoftBank-SLS

# What's AI-for-RAN?

---

- The Significance and Expectations of AI-for-RAN
- Target Goal
- SoftBank's Approach

# AI-RAN Alliance Key Focus Areas

## AI-for-RAN

AI for the enhancement of RAN



**Improving TCO and User Experience**

## AI-and-RAN

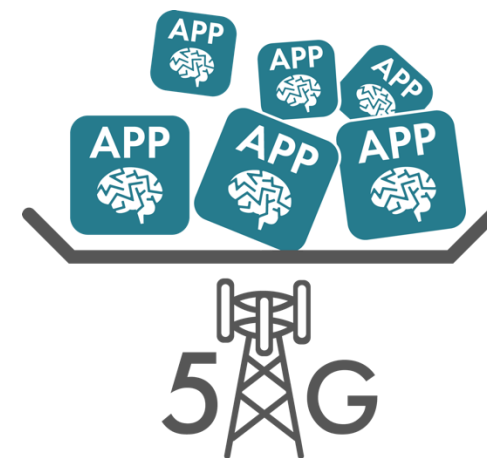
AI and RAN sharing the same infrastructure



**Biz Model Transformation**

## AI-on-RAN

AI applications enabled by RAN



**New Service Creation**

# Improving TCO and User Experience

- Usually, the improvement of user throughput has been achieved primarily through the expansion of deployed cell sites. This increase in the number of cell sites has led to an optimized expansion of coverage along user behavior and an increase in cell density, thereby enhancing the average user throughput.

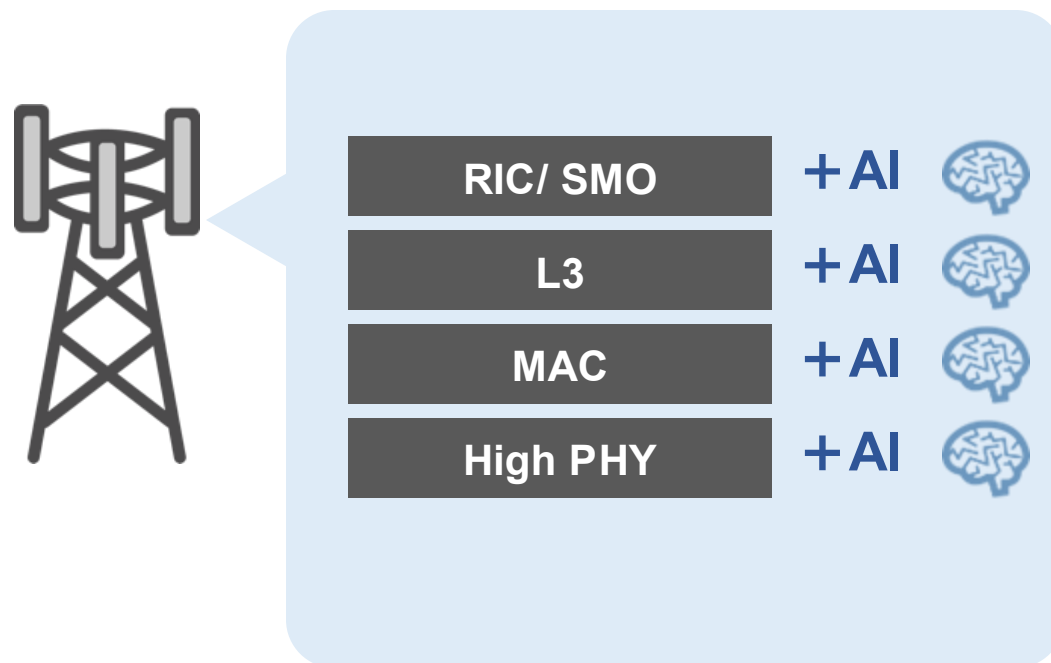


➡ If SINR will be up (=Throughput up) without increasing number of Cell by AI-for-RAN, it is equal improving TCO.



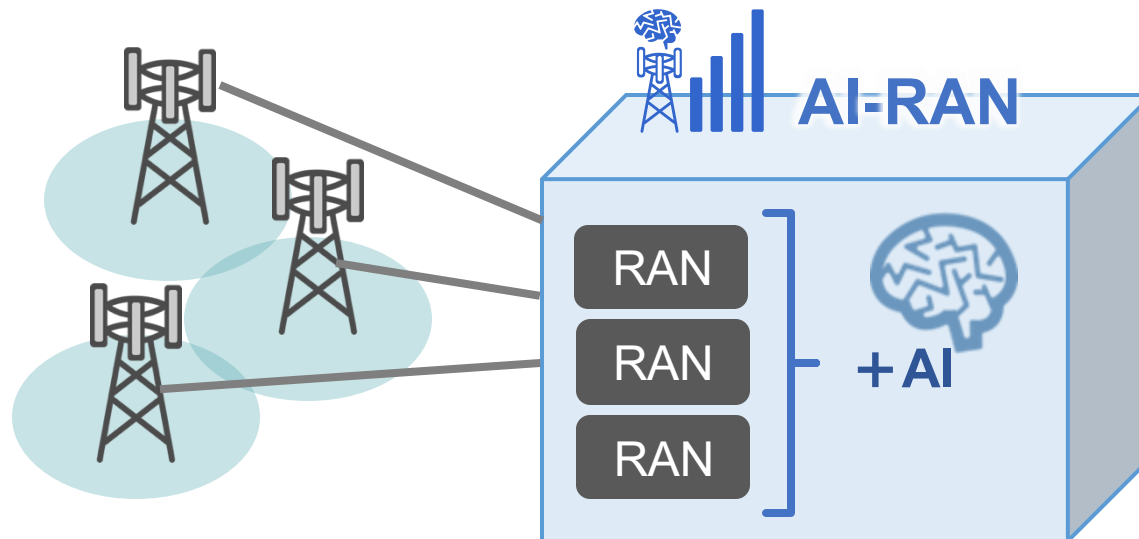
# Target of AI-for-RAN

## Individual Cell

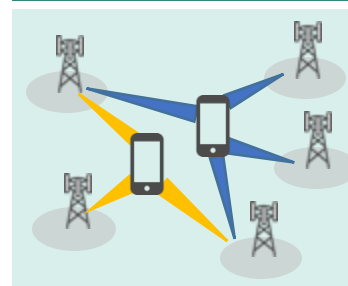


Full Stack AI Acceleration

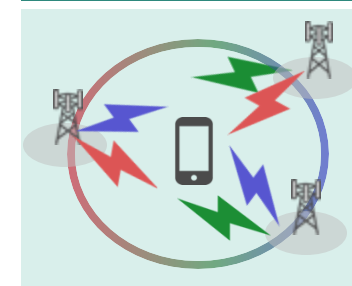
## Inter-Cell Coordination



Enhanced MIMO



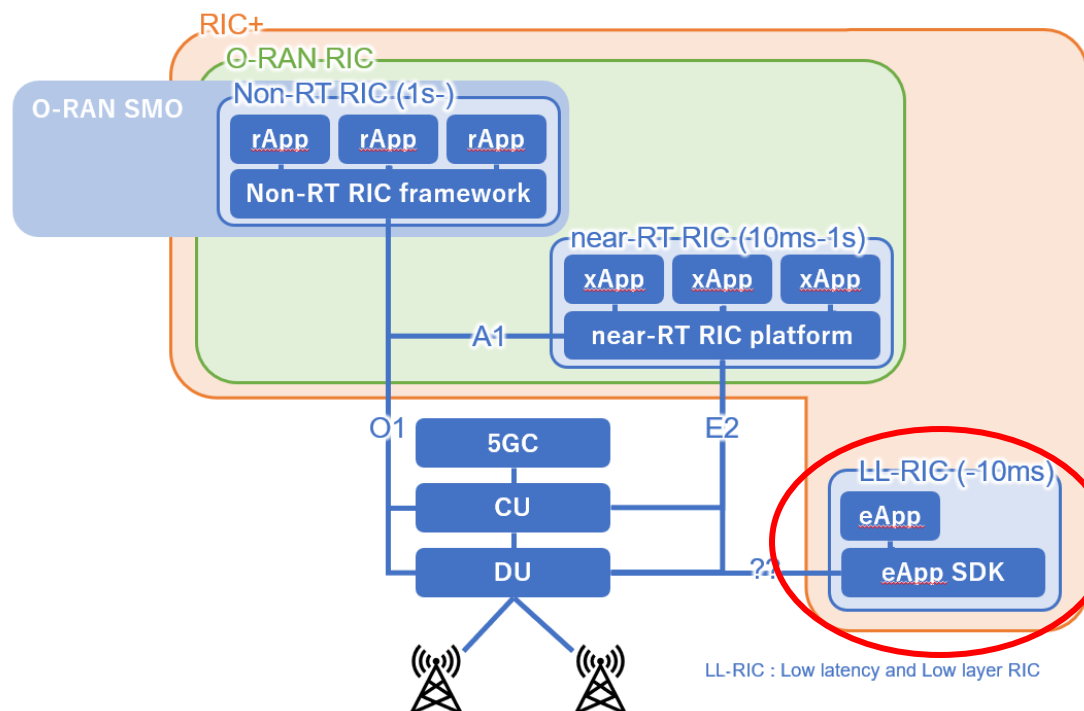
User Centric



Multi Cell Drive

# Approach of AI-for-RAN

- As a part of our initiative, we are starting with AI implementation at the lower layers.
- While the O-RAN Alliance is advancing the standardization of AI implementation at the upper layers by using Non-real time RIC and Near-real time RIC, there remains extensive research needed for the lower layers.
- However, this approach is extremely challenging and has a high failure rate.
- Through AI-RAN Alliance, we intend to explore diverse use cases and continuously expand them, promoting active utilization among our members.



This low layer part of RAN Intelligence Controller is ongoing research activity now. It has an opportunity to realize and extend new technologies of AI.

※LL-RIC and eAPP are defined by SoftBank, it might be different naming discussed within O-RAN Alliance.

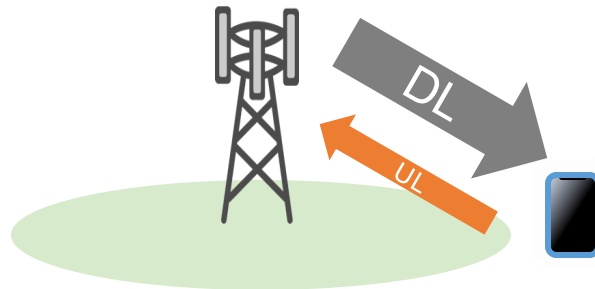
# Under Researching Themes of AI-for-RAN

---

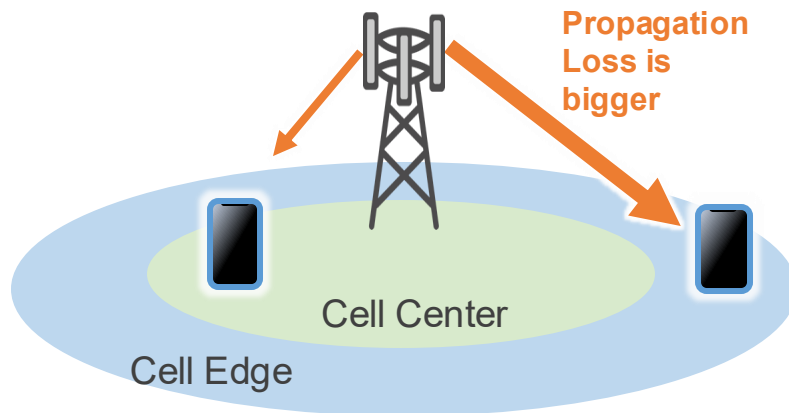
- Reason why We Choose These Themes:
- Details of Each Theme status:
  - MU-MIMO Advanced scheduler
  - SRS Prediction
  - UL CH Interpolation
- *Evaluation results: MWC'25*

# Reason why We Choose These Themes

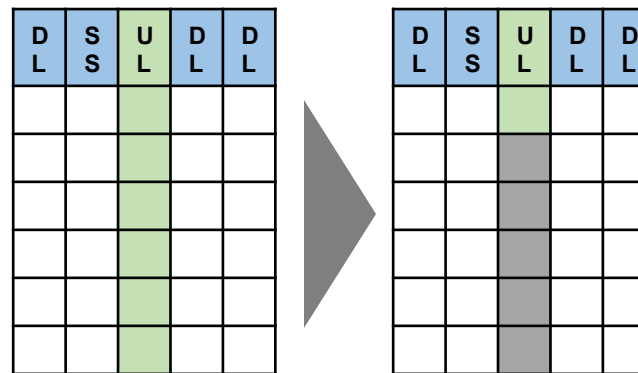
UL Data included ACK/NACK  
transmission opportunity has limited



TDD frame configuration in Japan  
DL:UL=7:2

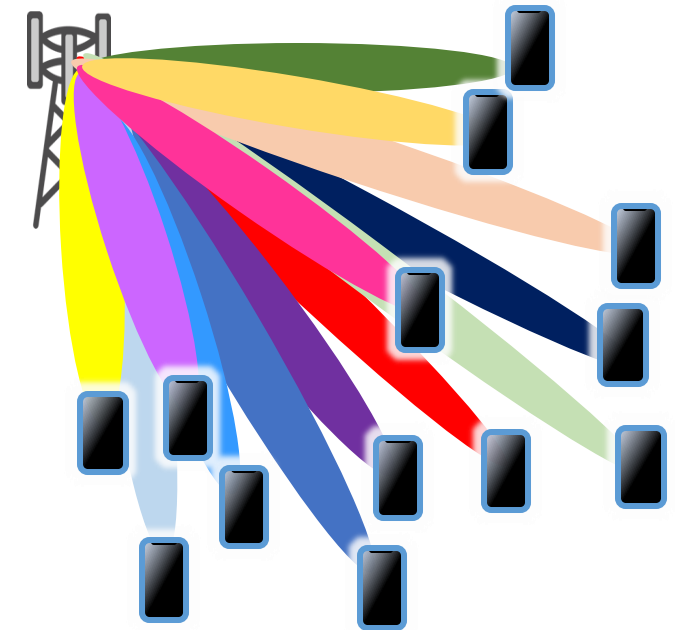


**For Improving TCO  
and  
User Experience**



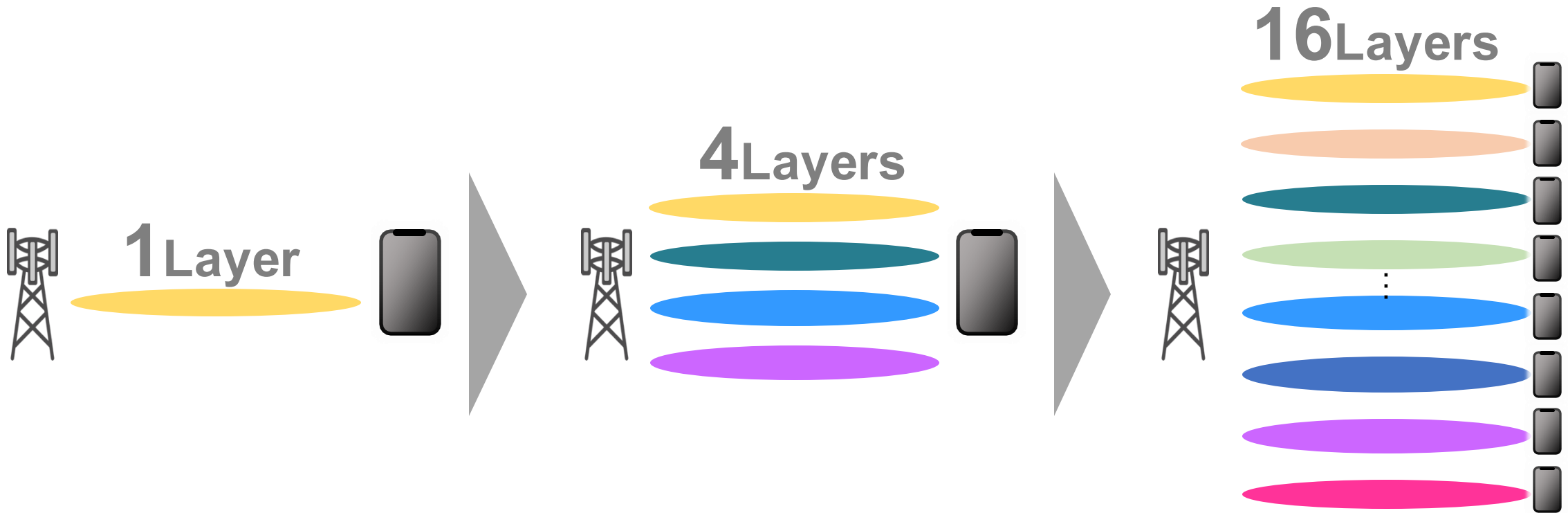
Reduced PRB # for  
power concentration

mMIMO performance equal NW  
Capacity!



MU-MIMO is more important feature  
for higher density population area

# Theme1:MU-MIMO Advanced Scheduler Necessity



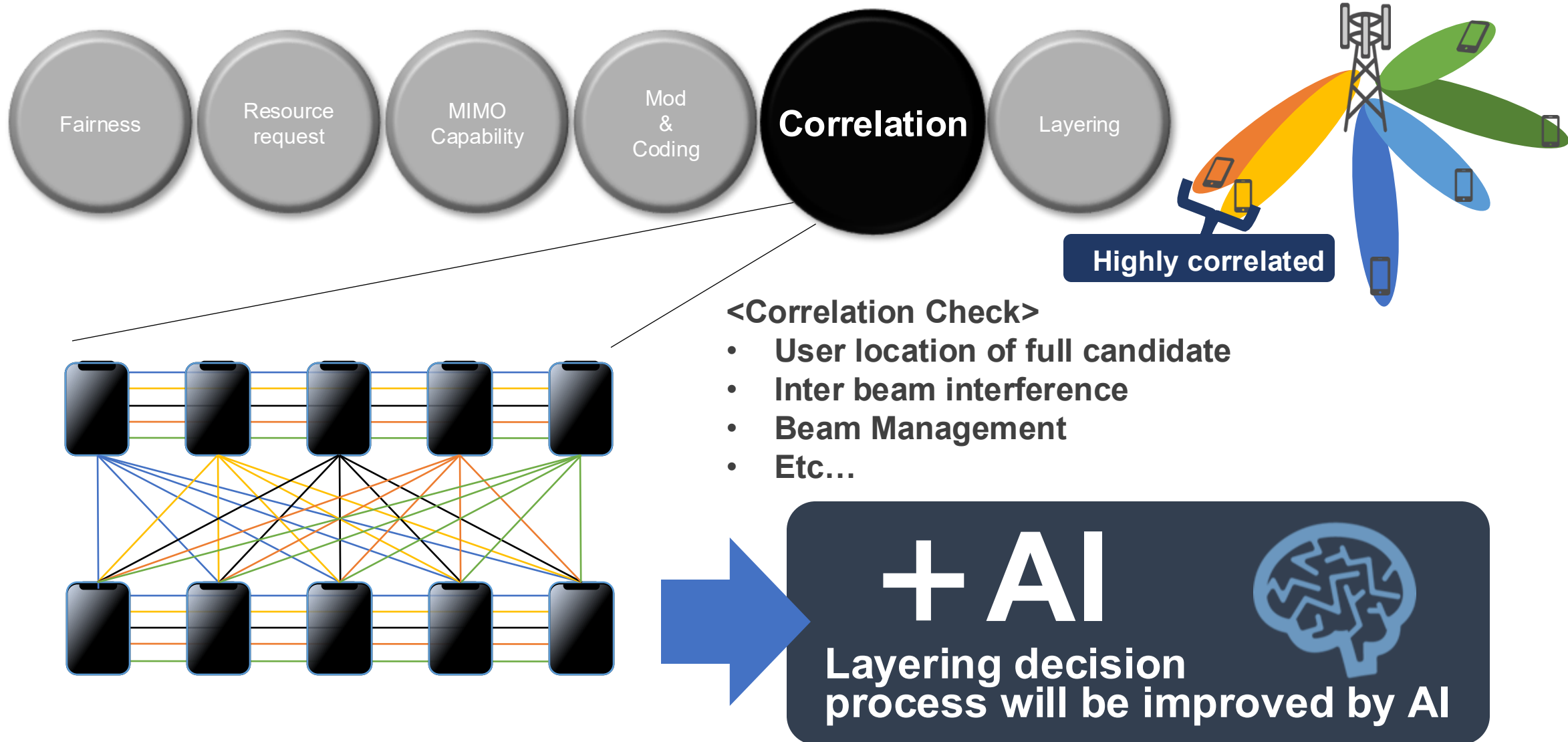
■ without MIMO case,  
resources will be shared.  
1data stream only

■ Single User-MIMO is effective for  
**The Peak User Throughput**  
4data stream at DL by smartphone

■ Multi Users-MIMO can achieve  
**Peak Cell Throughput**  
16data stream at DL  
**NW Capacity improved!**

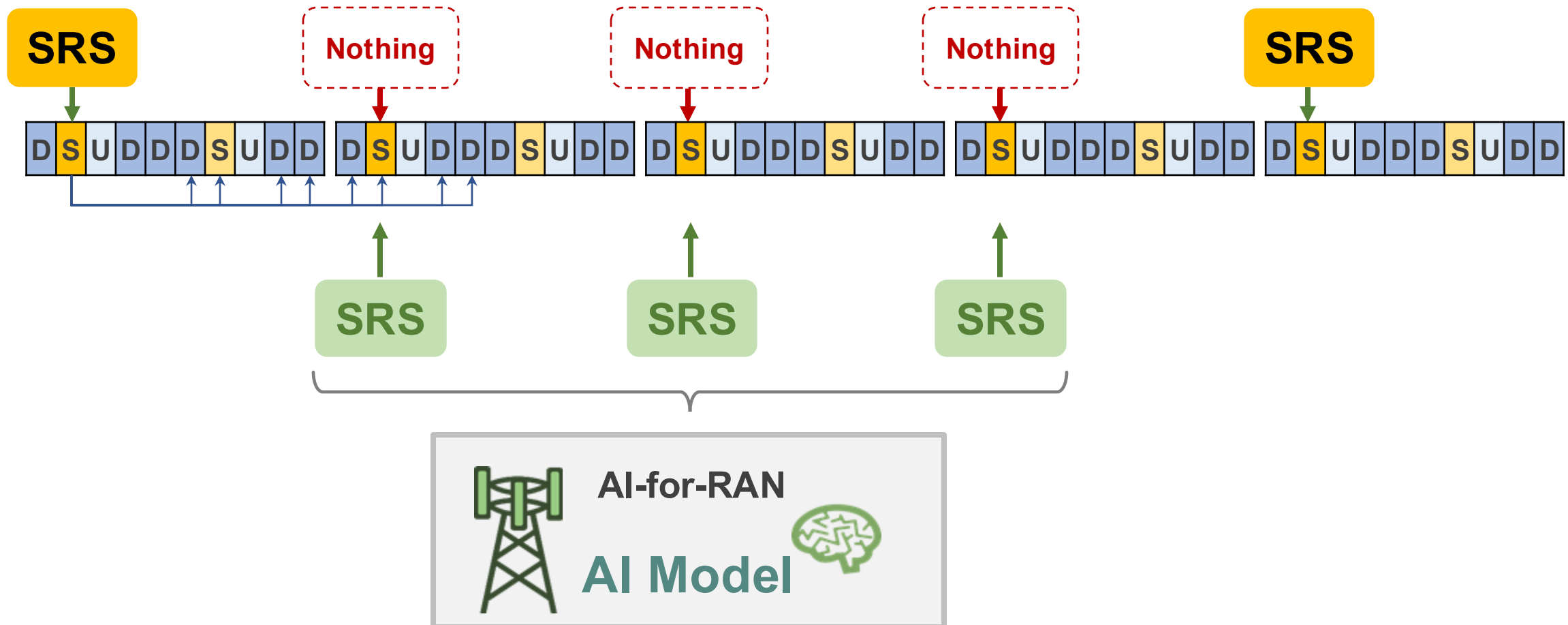


# MU-MIMO Difficulty

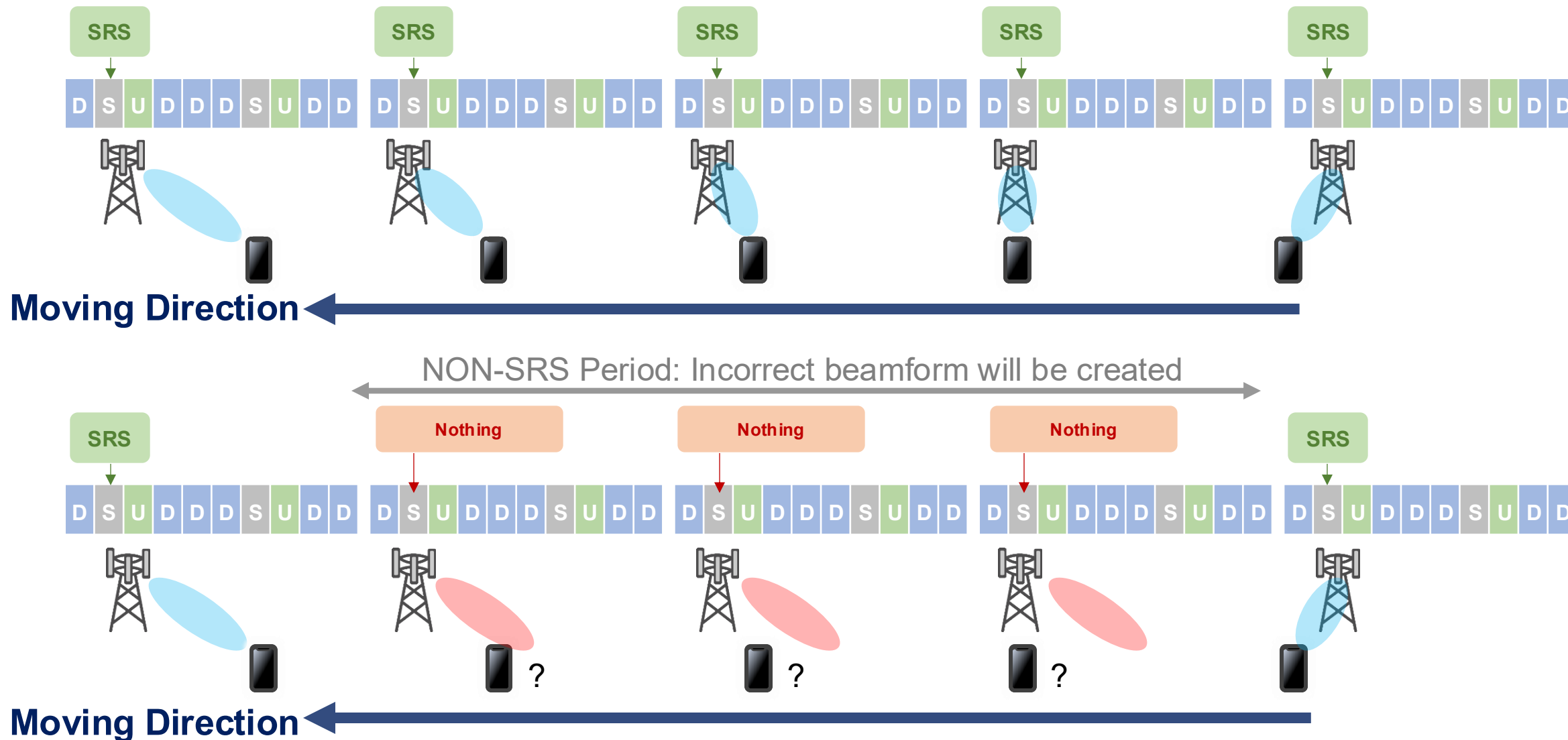


# Theme2: SRS Prediction Necessity

**Non-SRS Period : Beam Forming Accuracy will decrease**



# Beamforming Track Accuracy



# Theme3: UL Performance Improvement Necessity

$$\text{SINR} = \frac{\text{Desired Signal}}{\text{Interference} + \text{AWGN}}$$

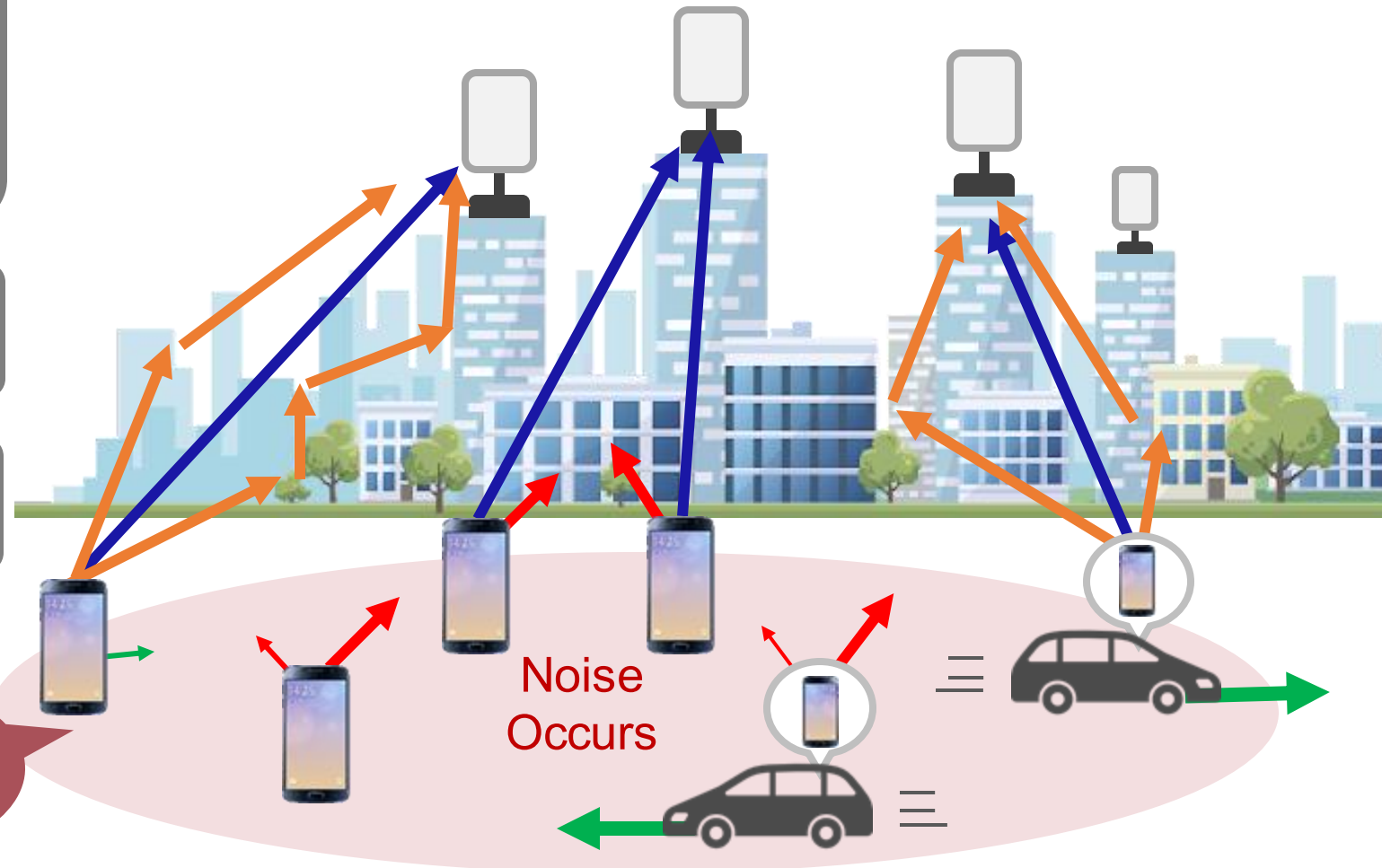
Doppler frequency shift occurs

Delayed signal due to multipath

Many UE (User Equipment) complex at the urban environment

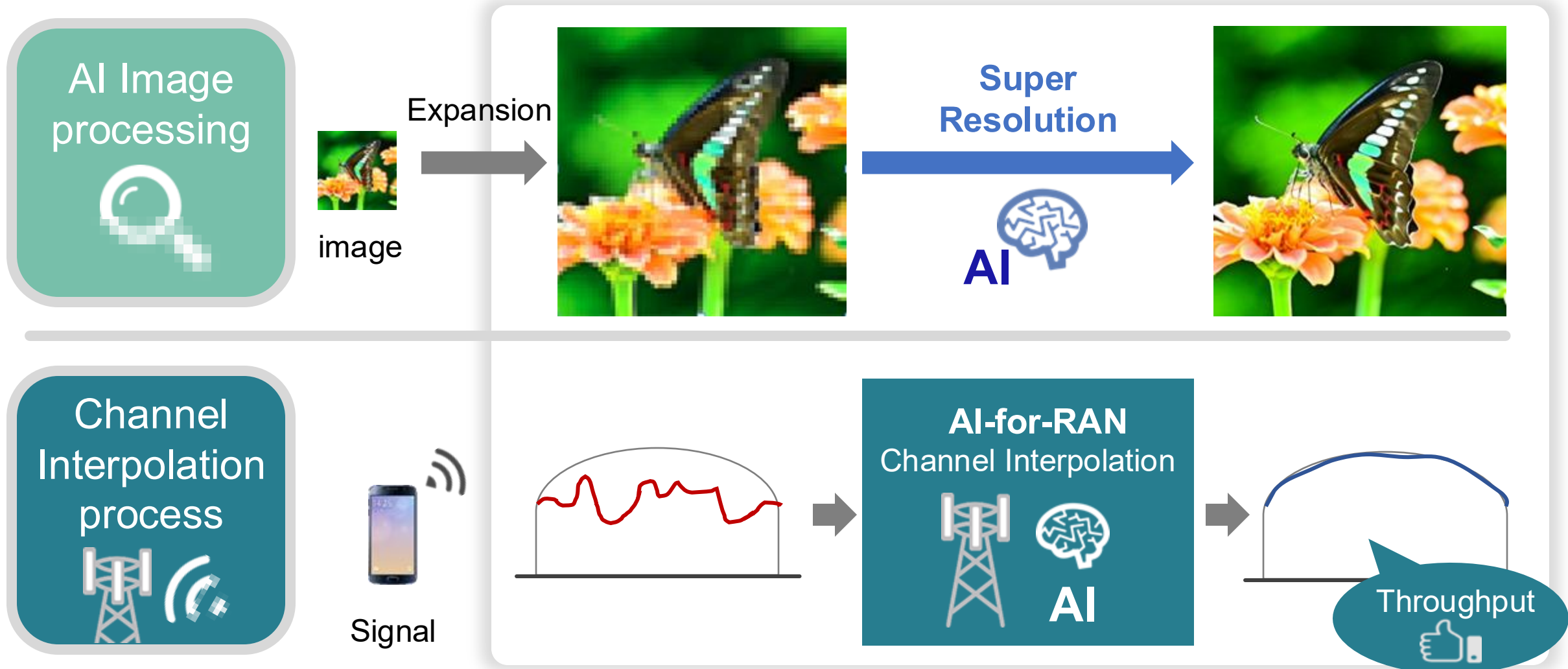
- Desired Signal
- Other Signal
- Moving direction
- Delayed signal

SINR



# Applied CNN Algorithm

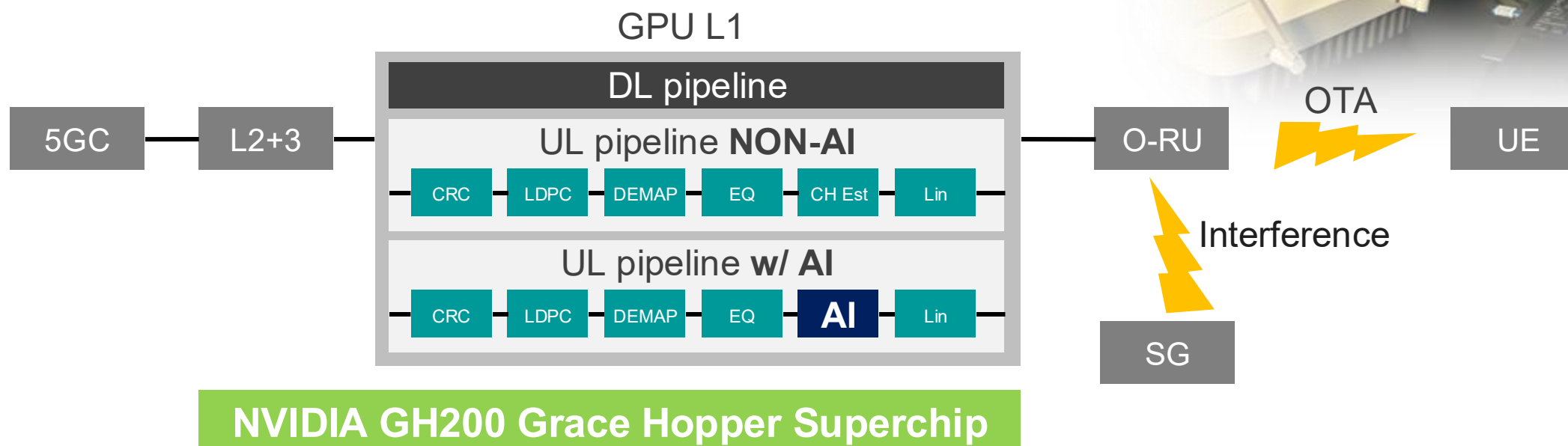
Super resolution in AI image processing and signal interpolation are similar





# Lab Test Environment

- Built a lab environment by implementing CPU Layer 2+3 and GPU Layer 1 on NVIDIA GH200.
- Modified the GPU Layer1 stack to embedded AI capabilities.
- Evaluated Uplink Channel Interpolation with AI on this circumstance.



# SoftBank developed System Level Simulator

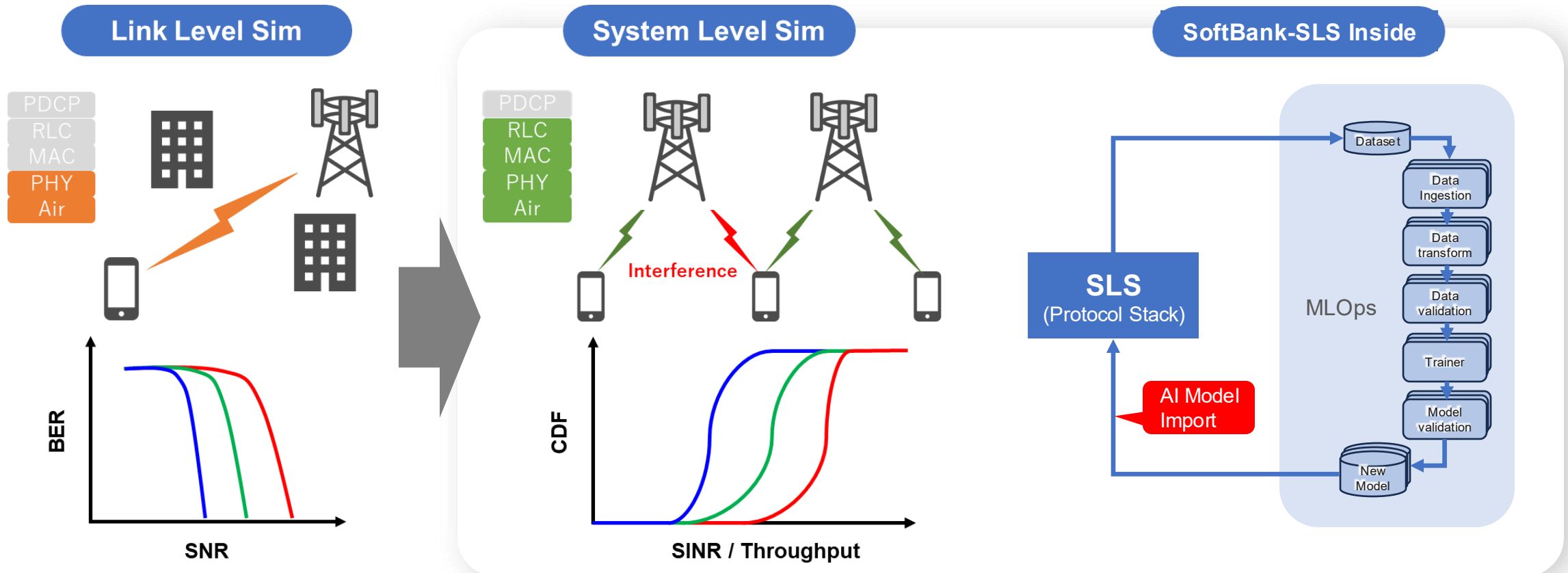
---

- Development and Purpose of SoftBank-SLS
- *Features and Implementation Details: MWC'25*

# System Level Simulator Necessity

## ■ Why need to develop the SLS (System-Level Simulation)?

- Link Level Simulation is involved 1 on 1 connection by Layer 1 only. It can NOT estimate neighbor cell interference, Multiple UE conditions and etc with PHY, MAC, RLC, and application layers protocol stack.
- And supported AI model porting with MLOps inside of SoftBank-SLS.



# MWC'25

Please keep an eye out and  
look forward to more  
exciting announcements  
on AI-for-RAN at MWC!

 SoftBank



# Q&A



Visit our website for  
the latest updates on  
SoftBank's AI-RAN initiatives!

[https://x.gd/SoftBank\\_RIAT](https://x.gd/SoftBank_RIAT)

**Thank you!**

 SoftBank