January 2026

AI-RAN for Token Communication

Jihong Park

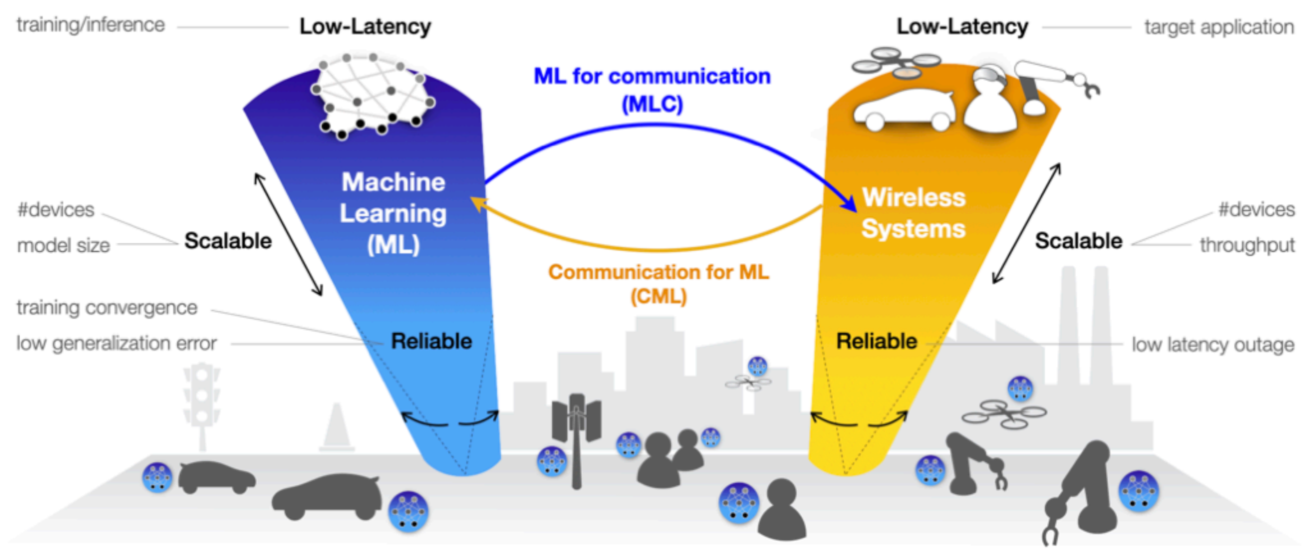
Associate Professor

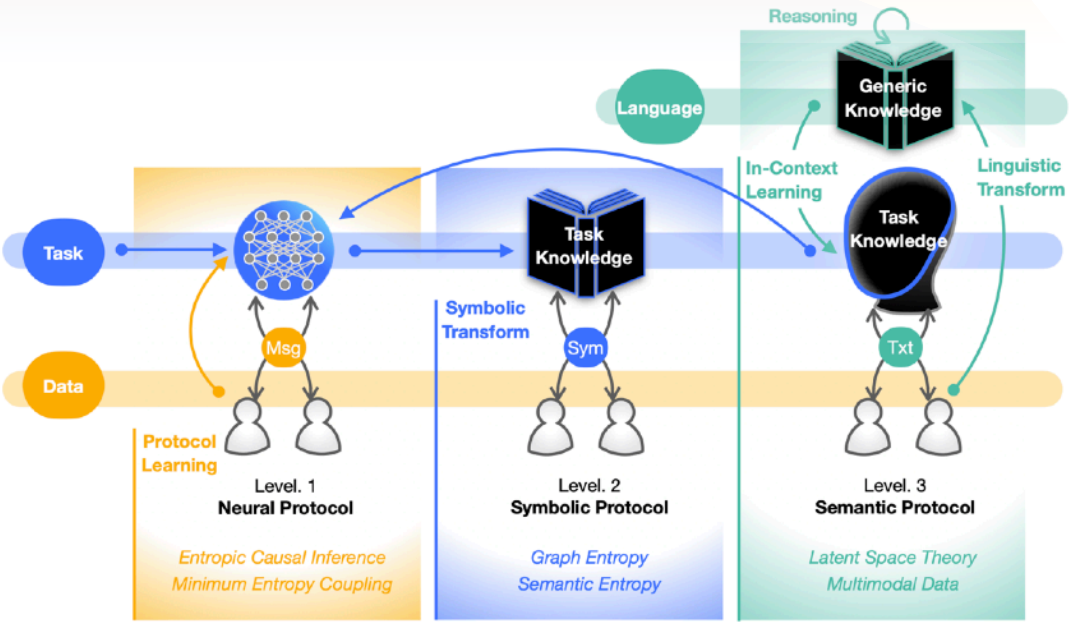
Deputy Director (Research) of FCP

AI-RAN Alliance WG3 Vice-Chair

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Jihong Park is an Associate Professor at the Singapore University of Technology and Design (SUTD) and an Honorary Associate Professor at Deakin University. He serves as the Deputy Director for the Future Communications Research and Development Programme (FCP) in **Singapore**. Prior to joining SUTD, he was a Lecturer at Deakin University, **Australia** (2020-2024). He obtained his B.S. and Ph.D. degrees from Yonsei University, Seoul, **Korea**, in 2009 and 2016, respectively. He was a Post-Doctoral Researcher at Aalborg University, **Denmark** (2016-2017), and at the University of Oulu, **Finland** (2018-2019). He was a Visiting Researcher at Aalborg University, KTH in **Sweden**, NJIT in **USA**, and Hong Kong Polytechnic University. His recent research focus includes **Distributed Machine Learning** and **AI-Native Semantic Communications** for their 6G and robotic system applications. Dr. Park has received several prestigious awards, including the 2023 IEEE Communication Society Heinrich Hertz Award and the 2022 FL-IJCAI Best Paper Award. He has served as the Symposium Chair and Track Chair for leading conferences, including IEEE GLOBECOM 2023, IEEE ICCC 2025, IJCNN 2025, and IEEE WCNC 2026. Currently, Dr. Park is an Editor of IEEE Transactions on Communications, a Member of IEEE Signal Processing Society's Machine Learning for Signal Processing Technical Committee, a Senior Member of IEEE, and Vice Chair of AI-RAN Alliance AI-on-RAN Working Group.





J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless Network Intelligence at the Edge," *Proceedings of the IEEE*, Nov. 2019 (Cover Article)
J. Park, S. Ko, J. Choi, S.-L. Kim, J. Choi, and M. Bennis, "Towards Semantic MAC Protocols for 6G: From Protocol Learning to Language-Oriented Approaches," *IEEE BITS Magazine* (Monthly Popular Article)

Future Comms R&D Programme (FCP)

FCP set up in May 2021 under RIE 2020 with S$68.7m initial investment

- We have under-invested in comms, which has 10 year cycles for development
- For 5G, now focusing on **regulations and facilitating commercialisation, not fundamental tech**
- Geo-political tech contestation could risk Singapore's access to key comms technology, we **need to build indigenous capabilities/ talent to afford resilience**
- Important to begin sustained investment to **develop niche leadership in beyond 5G and 6G**, as the first tranche

Evolution of mobile communications technologies – "10-year cycles"				
1990s	2000s	2010s	2020s	2030s onwards
2G	3G	4G	5G	6G & beyond

新科大举办5G无线接入网络峰会 促进本地通信科技发展

订户 赠阅文章

赵世慧

发布 /2023年8月23日 11:48 PM

SUTD to Launch South-East Asia's First O-RAN Open Testing and Integration Centre (OTIC)

28 Feb 2023

Information Systems Technology and Design

5G/6G/Network Communications

资媒局与新科大设实验室 推动新一代通信研究

胡洁梅

发布 /2022年9月19日 05:06 PM

通讯及新闻部长兼内政部第二部长杨莉明（左一）9月19日在新科大为未来通信互联实验室主持开幕。（张思庆摄）

CNBC

FUTURE TECH ASIA

Singapore is launching a $50 million program to advance research on AI and cybersecurity

PUBLISHED TUE, JUL 13 2021 1:22 AM EDT

UPDATED TUE, JUL 13 2021 10:57 PM EDT

Saheli Roy Choudhury

@SAHELIRC

FCP Leadership

Prof Tony Quek
FCP Director

SUTD Professor

Prof Quek is the Associate Provost (AI & Digital Innovation), Cheng Tsang Man Chair Professor and ST Engineering Distinguished Professor with the Singapore University of Technology and Design (SUTD), leading the Wireless Networks and Decision Systems (WNDS) Group. He is also the Chair of AI on RAN Working Group, Sector Lead of SUTD AI Program, and Deputy Director of the SUTD-ZJU IDEA, an IEEE Fellow, a WWRF Fellow, and a Fellow of Academy of Engineering Singapore. He has vast industry experience that includes collaborating with companies like Quanta Cloud Technology (QCT), VIAVI, ST Engineering.

Prof Chen Binbin
FCP Deputy Director (Industry)

SUTD Associate Professor

Prof Chen's research has received several awards, including the Best Paper Award in ACM SIGCOMM conference 2010 for the work on Error Estimating Coding. His research capabilities has been acknowledged and funded by large agencies and organisations such as, National Research Foundation (NRF), Infocomm Media Development Authority (IMDA), Cyber Security Agency (CSA), Energy Market Authority (EMA), Agency for Science, Technology and Research (A*STAR), Building & Construction Authority (BCA), National Instruments, Keysight, LITEON, StarHub.

Prof Park Jihong
FCP Deputy Director (Research)

SUTD Associate Professor

Prof Park has served as the Track Chair and Workshop Organizer for leading conferences in communications and AI, including IEEE GLOBECOM, WCNC, ICML, and AAAI. He has received several prestigious awards, including the 2023 IEEE Communication Society Heinrich Hertz Award and the 2022 FL-IJCAI Best Paper Award. Currently, Dr. Park is an Editor of IEEE Transactions on Communications, a Member of IEEE Signal Processing Society's Machine Learning for Signal Processing Technical Committee, and Vice Chair of the AI-RAN Alliance's AI-on-RAN Working Group.

THE TEAM AT SUTD – FCCLAB/OTIC

LEADERSHIP

Prof Tony Quek
FCP Director

Prof Chen Binbin
FCP Deputy Director
(Industry)

Prof Park Jihong
FCP Deputy Director
(Research)

ADMIN TEAM

Ms Dawn Chia
Admin Team Lead

Ms Li Jiayan
Deputy Manager

Ms Stacey Zhang
Deputy Manager

Ms Fazilah Kasim
Associate

TECHNICAL TEAM

Dr Ngo Van Mao
Technical Team Lead

Mr Yeo Siow Long
Communications Testbed Engineer

Mr Le Thanh Long
Communications Testbed Engineer

Mr Nguyen Thanh Tam
GPU Testbed Engineer

Mr Nguyen Nam Duong
Communications Testbed Engineer

Dr Wang Peng
Research Fellow

Ms Wang Zhuoran
Testing and Certification Engineer

Mr Liang Xian Loong
Communications Testbed Engineer

Mr Hariz Yet
Communications Testbed Engineer

Mr Ngo Van Tuan
Software Engineer

Mr Ye Xiaodong
Senior Research Assistant

Mr Vishal Choudhary
Research Associate

FCP Partners

(as at 17-Oct-2025)

FCP1.0 (2021-2025)

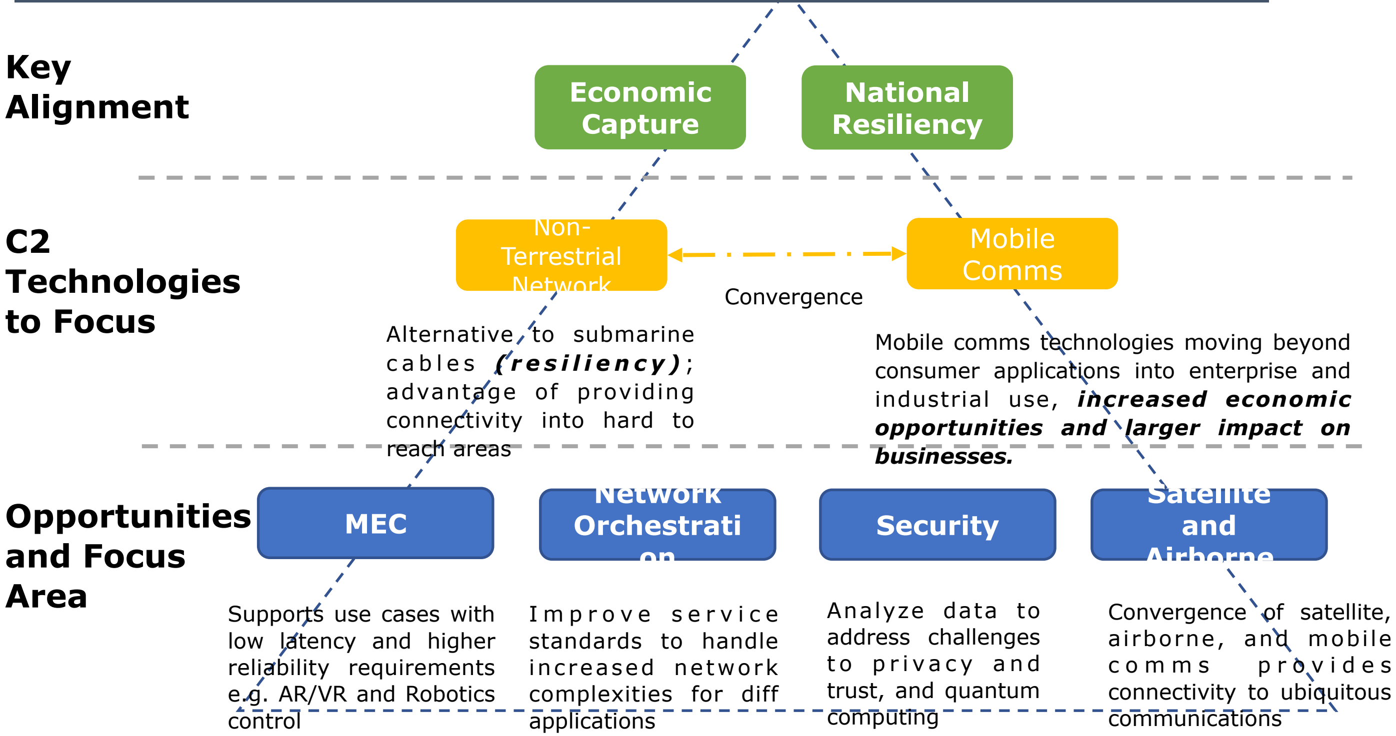
FCP FUTURE COMMUNICATIONS R&D PROGRAMME

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Select the Right C2 Technology to Focus

- C2 technologies are wide, therefore it is important to narrow down the scope to the key alignment of (i) **economic capture** and (ii) **national resiliency**;
- Focus on selected C2 technologies; and
- Focus in areas where we have higher chance/opportunities to succeed

Identified Opportunities and Focus Areas



How Do We Achieve This ?

- 22 research projects awarded
  - ➢ MEC & Network Orchestration (9)
  - ➢ URLLC (4)
  - ➢ Security (4)
  - ➢ Integrated Sensing & Comms (3)
  - ➢ NTN (2)
- Research projects with key partners (academia and companies)
- Explore collaboration with adjacent research programmes in SG (AI.SG, QEP, FME2.0)

FCP1.0 (2021-2025): **5G NTN Live Demo** (Osaka Expo'25)

- HD content can be delivered directly through satellites using OFDM signals in NR-NTN.
- E2E NTN testbed includes NTN UE-gNB emulators (VIAVI, R&S) and a **real GEO satellite** (JSAT, Japan)
- Live demonstration has been completed at the Singapore Pavilion in **Osaka 2025 Expo**, Japan.

* **gNB**: a base station using 3GPP New Radio (NR) technology
* **JSAT**: a largest satellite vendor in the Asia-Pacific region

FCP1.0 (2021-2025): O-RAN OTIC

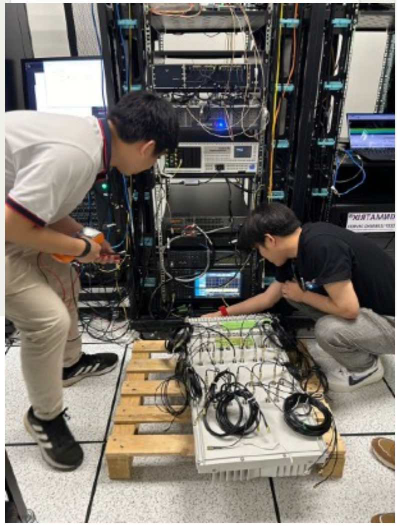
The first and only O-RAN Open Testing and Integration Centre (OTIC) in South-East Asia

- A member of the global OTIC network, approved by O-RAN Alliance in Feb. 2023
- To serve as the South-East Asia's hub & gateway to the global Open RAN ecosystem
- Co-located with FCCLab
- Focus areas: Security - Sustainability - AI/ML for strategic verticals (e.g., maritime)

Alex Jinsung Choi

Deutsche Telekom Senior Vice President, Chairman of O-RAN Alliance

View full profile



Alex Jinsung Choi · 1st

Deutsche Telekom Senior Vice President, Chairman of O-RAN Alliance

3w · Edited ·

O-RAN Alliance Global Open Testing and Integration (OTIC) Network

During this MWC Barcelona week, the 10th and 11th O-RAN Alliance OTICs were launched. One at SUTD in Singapore and the other at COSMOS Lab in NYC USA. With them, O-RAN Alliance's global OTIC network establishment is accelerating, and it is expected that 2-3 additional OTICs will be added within this year. In particular, SUTD in Singapore, the 10th OTIC, is the first in South-east Asia, and is supported by Singapore IMDA (Infocom Media Development Authority). They will provide rigorous cybersecurity testing and assurance for multi-vendor O-RAN solutions in addition to standard OTIC testing job. According to their PR, see the link, "The framework will facilitate security assessments modeled on diverse threats originated from multiple attack vectors, including compromised O-RAN RU and compromised network elements in fronthaul interfaces". So far there are 7 O-RAN Alliance security documents available on the website, and these documents will be the basis for the security tests conducted here. I hope that 'Test as a Service' framework for the regional specific use case (i.e., maritime industry in Singapore) will be successfully established here.

Achievement highlights:

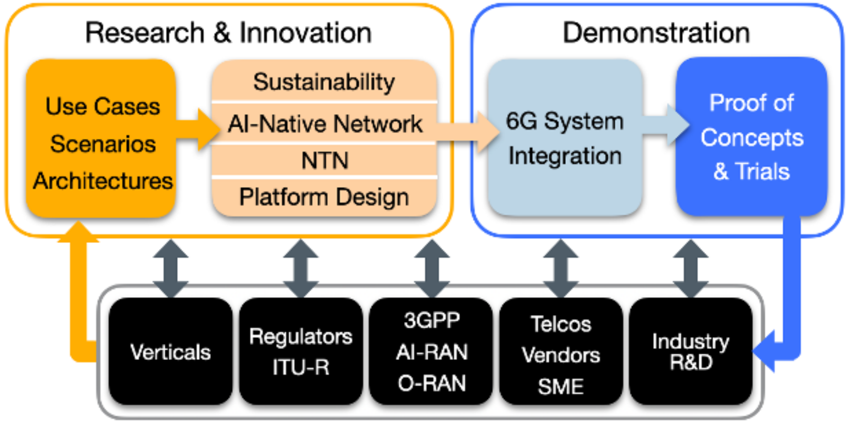
- Test-automation solutions to reduce massive MIMO O-RU testing time to 3-days (in collaboration with Japan OTIC)
- Integrating NVIDIA stack with multiple O-RU vendors
- Integrating O-RAN solution with a local telco's live core

FCP1.0 (2021-2025): MediaTek-SUTD Joint Lab

MEDIATEK

- **SUTD-MTK-R&S joint live demo** of FR1 NR-NTN Video Call (with commercial phones and an emulated LEO satellite) at the 2025 Osaka Expo

- **MediaTek-SUTD Joint Lab** to be launched this year, initially with 8 projects, including a pilot project on "High-Fidelity 3GPP NTN Testbed," which aims to provide the impact of key design parameters on KPIs, such as moving vs. fixed beams and other deployment plans



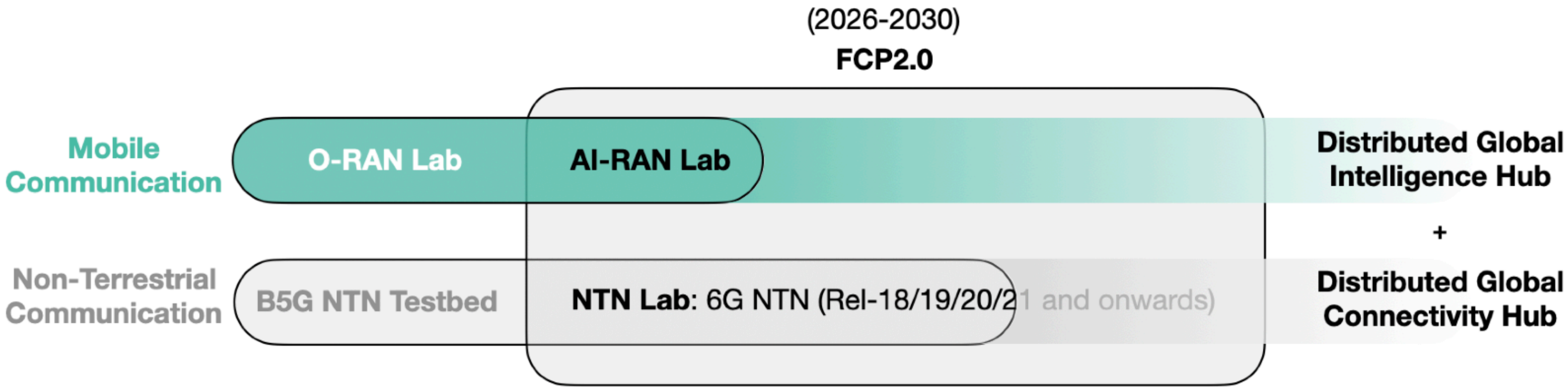
Agency for Science, Technology and Research SINGAPORE

EDB SINGAPORE

Enterprise Singapore

National Research Foundation
Prime Minister's Office
Singapore

FCP2.0 (2026-2030)



Background and Impetus to start FCP2.0

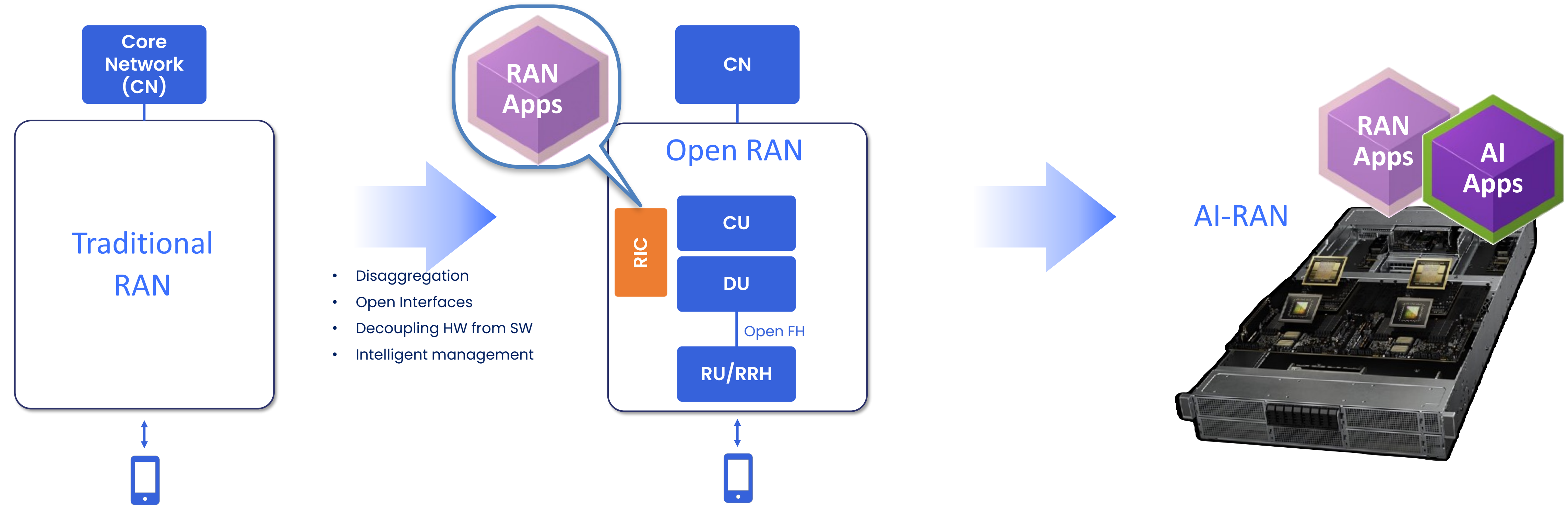
1. Wireless technologies like 5G and beyond is no longer just about mere connectivity, but is increasingly about providing integrated and comprehensive services (such as metaverse) This will further accelerate with the introduction of NTN.
2. FCP has made initial R&D investments in future communications since May 2021. Other countries are already moving to the next bound, 6G, an area where we need to make sustained and patient investments.
3. Singapore has made good progress in Open-RAN with setting up of O-RAN OTIC and initial investments by some O-RAN vendors. Therefore, we need to continue to establish our global presence and gain mind share.
4. Given limited budget and talent, it is not possible for us to compete in all fronts and need to leverage on FCP. Hence, we will need to focus our efforts for the investment in **national resiliency and economic capture**.

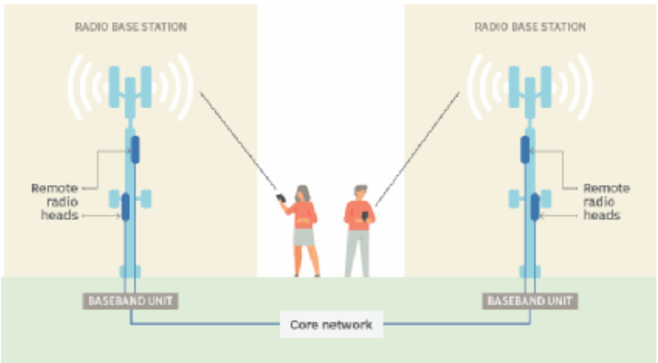
AI-RAN for Token Communication

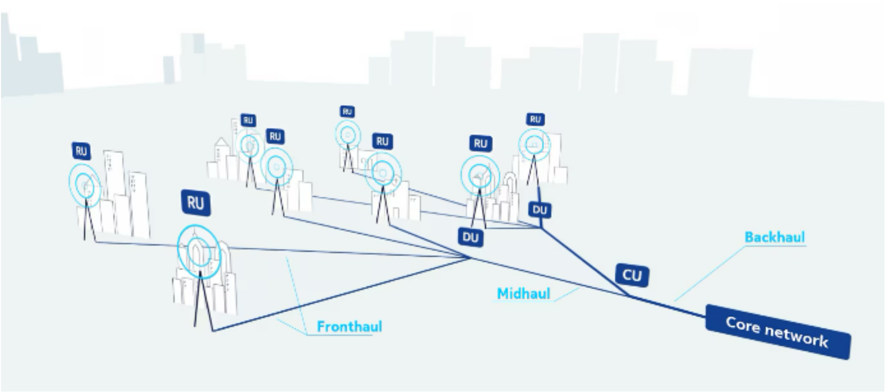
AI-RAN

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

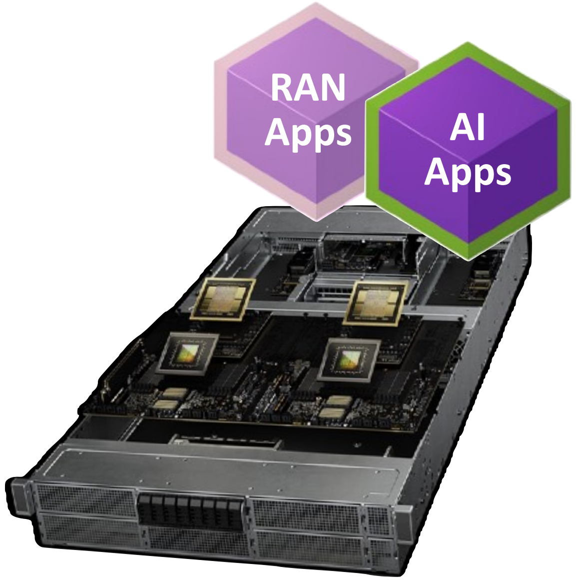
AI-RAN: RAN + Programmability + Multi-functionality



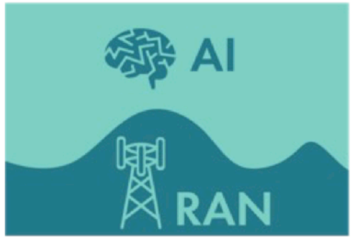




AI-RAN: Transforming RAN with AI



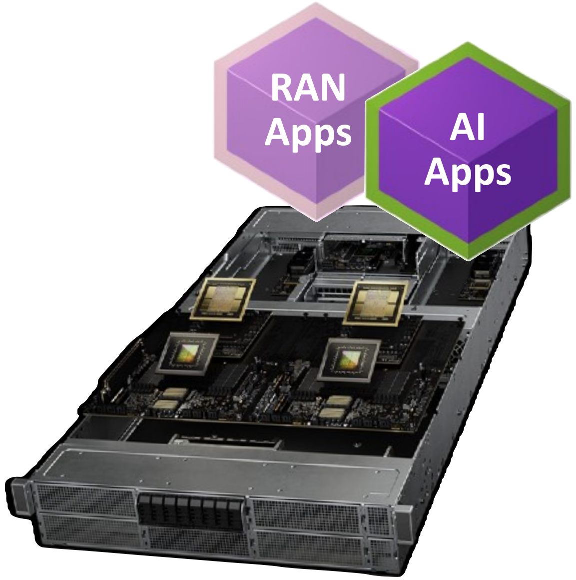
AI-for-RAN	AI-and-RAN	AI-on-RAN
AI for the enhancement of RAN	AI and RAN sharing the same infrastructure	AI applications enabled by RAN

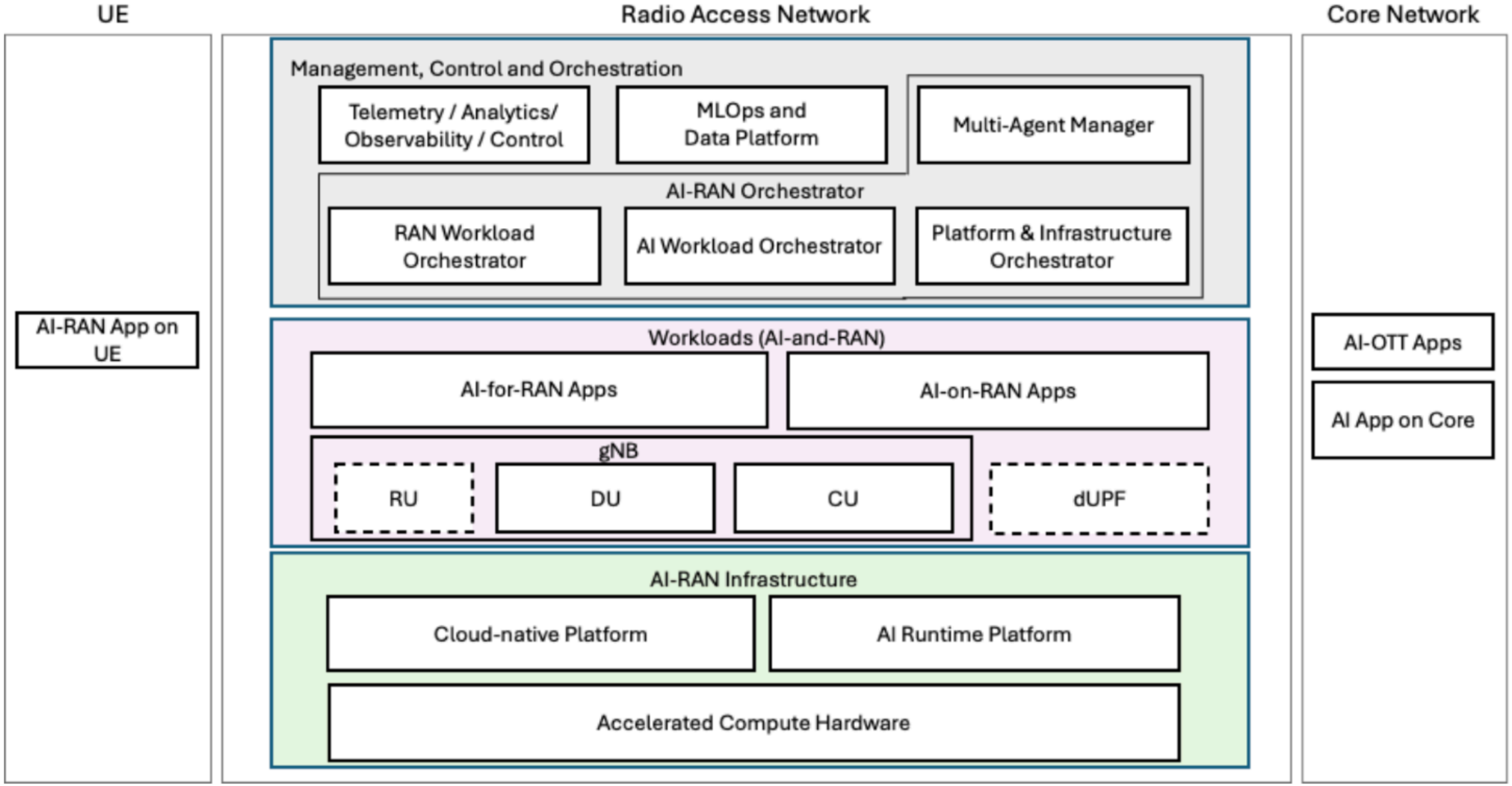


Spectral Efficiency	Asset Utilization	New Applications

- Transforming mobile network into distributed global inference engine
- By integrating AI and mobile network
- Towards accelerating innovation in AI and mobile

AI-RAN: Transforming RAN with AI





Candidate AI-RAN Reference Architecture

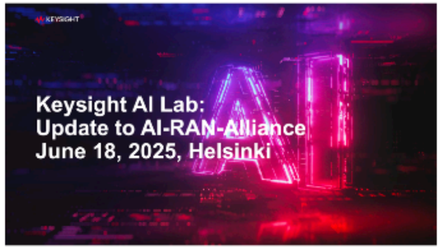
AI-RAN: AI-RAN Lab @ SUTD

SUTD has established the Asia & Pacific Open Testing and Integration Centre (OTIC) in Singapore (APOS). **The first and only O-RAN Open Testing and Integration Centre (OTIC) in South-East Asia**

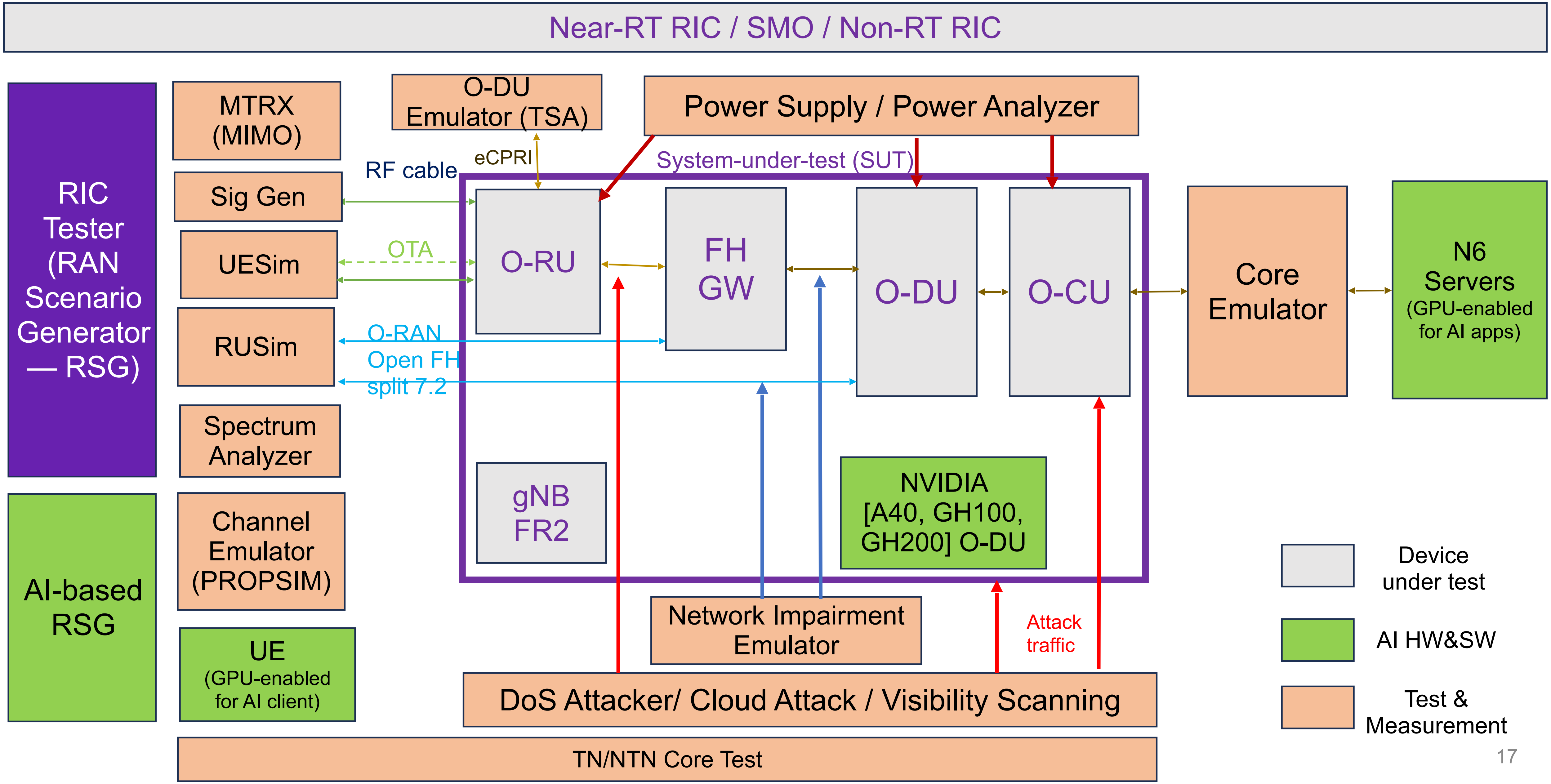
- To serve as the South-East Asia's hub & gateway to the global Open RAN ecosystem
- Focus areas: Security - Sustainability - AI/ML for strategic verticals (e.g., maritime)

SUTD has established **one of the world's first AI-RAN Alliance-Endorsed labs**, providing a platform for:

- Developing and testing of AI-RAN solutions that enhance network efficiency, adaptability, and performance
- Bridging academia and industry, fostering collaboration between researchers, students, and AI-RAN Alliance member organisations
- Exploring next-generation AI algorithms that optimise network orchestration, energy efficiency, and spectrum utilisation



AI-RAN: AI-RAN Lab @ SUTD



AI-RAN: AI-RAN Lab @ SUTD

FCP FUTURE COMMUNICATIONS R&D PROGRAMME

SUTD SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

5G Core

OPEN AIR INTERFACE
OAI 5G Core Network

Open5GS
Core Network

Open 5G Core
Fraunhofer open5gcore

KEYSIGHT TECHNOLOGIES
Keysight Core Sim

QCT
QCT 5G Core

ATAYA HARMONY
Ataya 5G Core

CU, DU

OPEN AIR INTERFACE, NVIDIA
OAI with GPU Accelerator, Grace Hopper MGX Systems

SRSRAN PROJECT, OPEN AIR INTERFACE
OAI, srsRAN with USRP on Bare Metal Server

SYNDESIGNTEK
Synergy O-CU & O-DU

SynaXG BBU

G REIGNS
HTC BBU

QCT
QCT BBU

All-in-One

SYNDESIGNTEK
mmWave gNB

LITEON
ST Engineering SC-250

Fronthaul

Keysight TSA

PTP and SyncE Provider and Tester

Paragon Neo

Falcon Switch PTP Grandmaster

Dell PTP Switch

NE3 Network Emulator

Packet Capture Appliance

RU

LITEON
LiteOn O-RU

DELTA
Delta O-RU

FOXCONN 鴻海科技集團

COMPAL 5G
Compal O-RU

SERA
SERA O-RU

METANOIA

WNC
WNC O-RU

X410, X310, B210

QCT
Indoor O-RU

QCT
Outdoor O-RU

KEYSIGHT TECHNOLOGIES RuSIM

Channel

Adjustable RF Attenuator J720x

MTRX

PropSIM

KEYSIGHT TECHNOLOGIES

RF shield box

OTA
FR1: 3400 – 3450 MHz
FR2: 25900 – 26300 MHz

SUTD Campus

Aerial Arena

UE

5G-native Drone

Humanoid robot

Quadruped robot dog

AR/VR Goggle

5G 4K Camera (Pegatron Nura4K)

5G Cell Phone

TEMS Pocket

Wheeled Robots (AMR, Limo ROS)

Pegatron FR2 CPE
PEGATRON 5G

5G Dongle (Pegatron, APAL)

UeSIM

VIAVI TM500

Server for RIC x/rAPP

RIC Test

Vector signal generator

Spectrum Analyzer

Power Analyzer

ALifecom NE7500/ NE6000

CMX500

CMW500

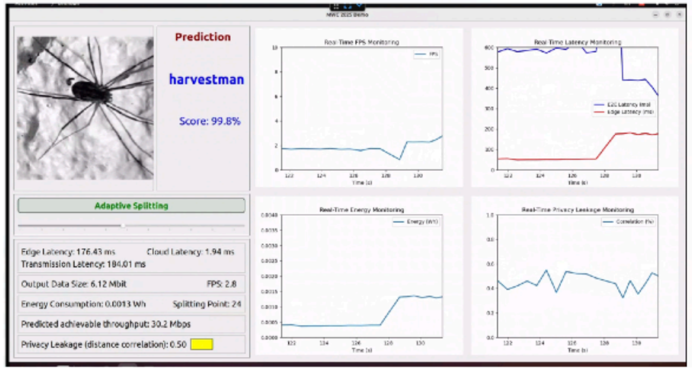
AI-RAN: Demo

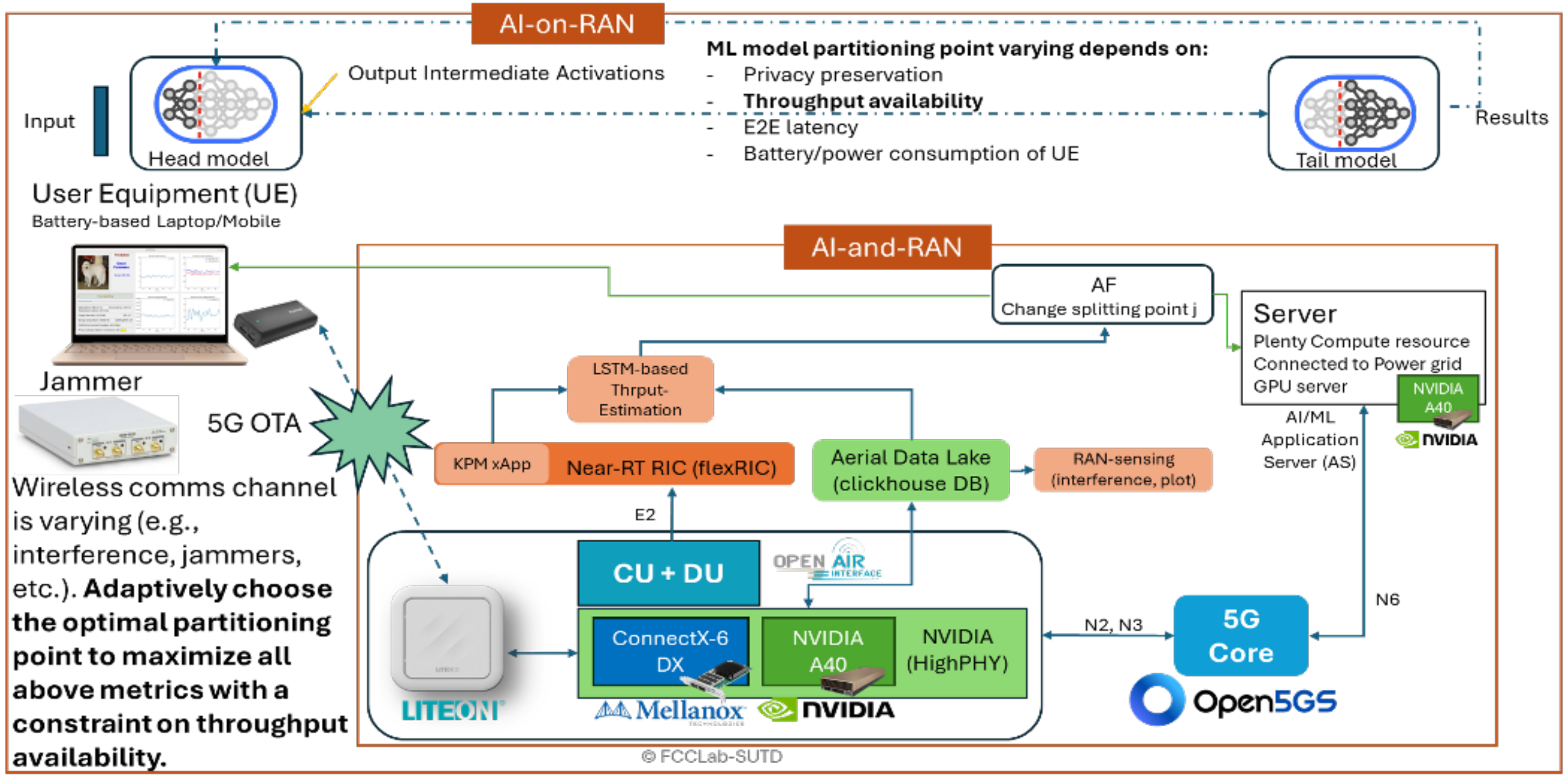
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN — NEURORAN — KEYSIGHT TECHNOLOGIES — LITEON

Channel-Adaptive Split Inference via Spectrum Sensing

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**

Plan: Collect IQ data and RAN's KPM, E2E performance for the use case under different wireless environment in Lab setup, share the results and dataset.





AI-RAN: Industry Trials & Research Tools

Samsung, KT confirm viability of 6G AI-RAN tech on commercial network

Kim Boram

All News · 10:07 December 11, 2025

NVIDIA and Nokia to Pioneer the AI Platform for 6G — Powering America's Return to Telecommunications Leadership

NVIDIA to Invest $1 Billion in Nokia to Accelerate AI-RAN Innovation and Lead Transition from 5G to 6G

October 28, 2025

AI-RAN Goes Live and Unlocks a New AI Opportunity for Telcos

Nov 12, 2024 +26 Like Discuss (2)

Figure 4. AI-RAN economics for covering one Tokyo district with 600 cells

	33% AI and 67% RAN	67% AI and 33% RAN
$ of revenue per $ of CapEx	2x	5x
ROI %	33%	219%

Powering AI-Native 6G Research with the NVIDIA Sionna Research Kit

Oct 28, 2025 +15 Like Discuss (0)

By Sebastian Cammerer and Alexander Keller

THE LINUX FOUNDATION PROJECTS

OCUDU

Driving the Vision of Open Source RAN

AI-RAN for Token Communication

Token-based Communication

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Motivation. **AI-RAN Key Applications**

Q. What are the key emerging applications of AI-RAN, in the era of GenAI and multimodal large language models?

AI-on-RAN

AI applications enabled by RAN

New Applications

On-Device

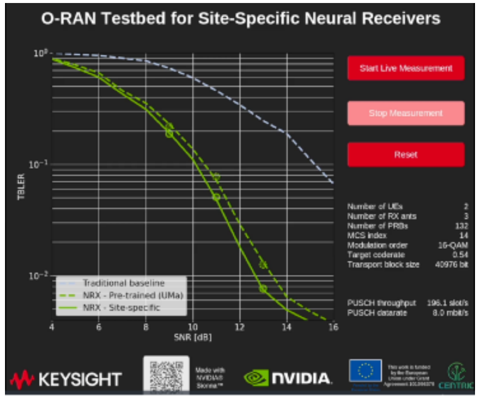
Generative

Multimodal

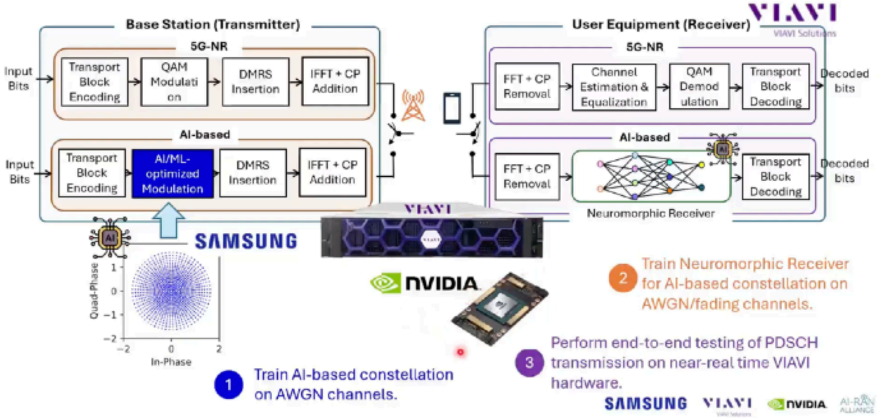
AI-for-RAN

AI for the enhancement of RAN

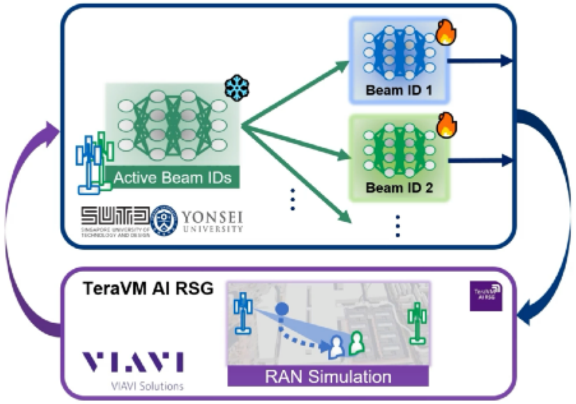
Spectral Efficiency



Site-specific



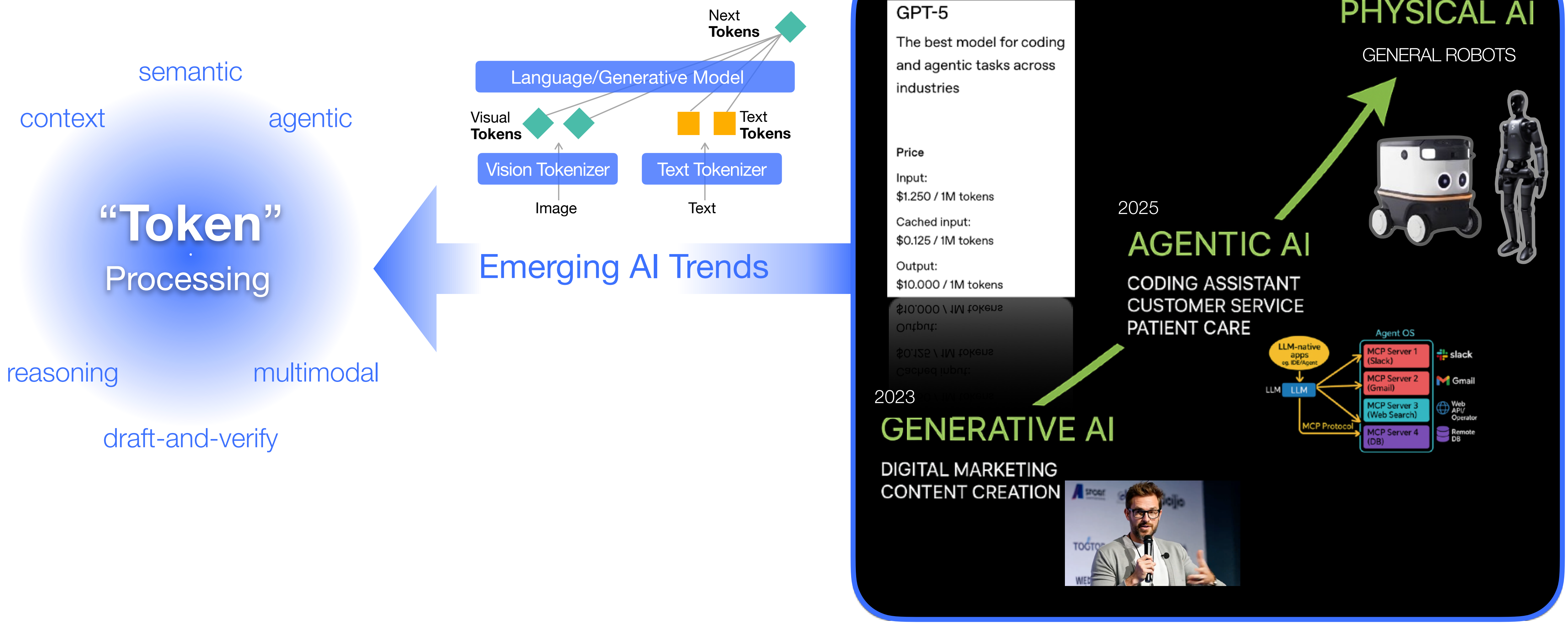
Joint Optimization



Predictive

Emerging AI Trends: **Token-based Processing**

Tokens are the fundamental units of processing in **generative** & **large language model-based applications**.



Source: NVIDIA

Emerging Communication Trends: **Token-based Communication**

To address the shift from bits to tokens, **token-aware** communication & **token-level** multiple access have emerged.

semantic

context

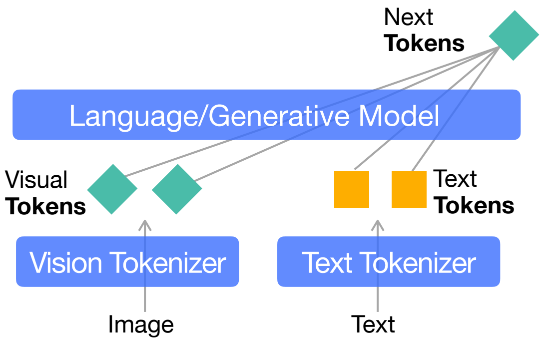
agentic

"Token"
Communication

reasoning

multimodal

draft-and-verify

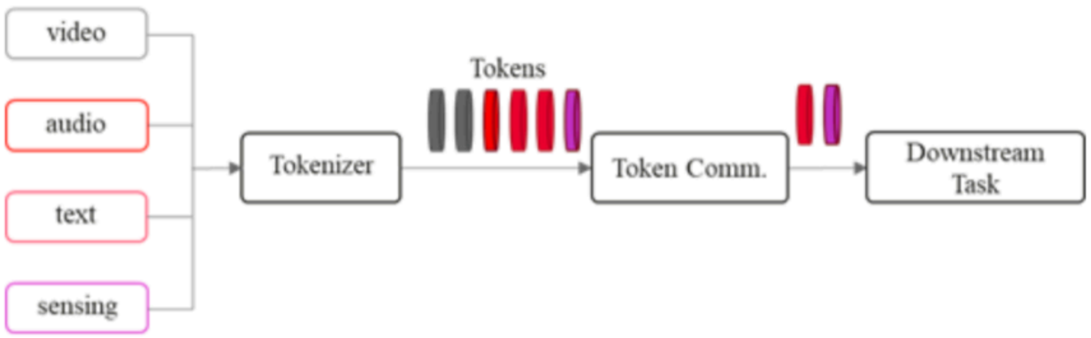


Emerging Communication

Source: Huawei, MediaTek

Token-based Communication: **Multimedia Communication**

Multimodal data transmission: **Token traffic model** for generative multimodal data transmission



New model 3:

Motivated by new services with AI related, e.g., immersive communication, token communication, etc.

- ***Mentioned by:*** *MediaTek, AT&T, Google, NVIDIA, Sharp, Huawei,*

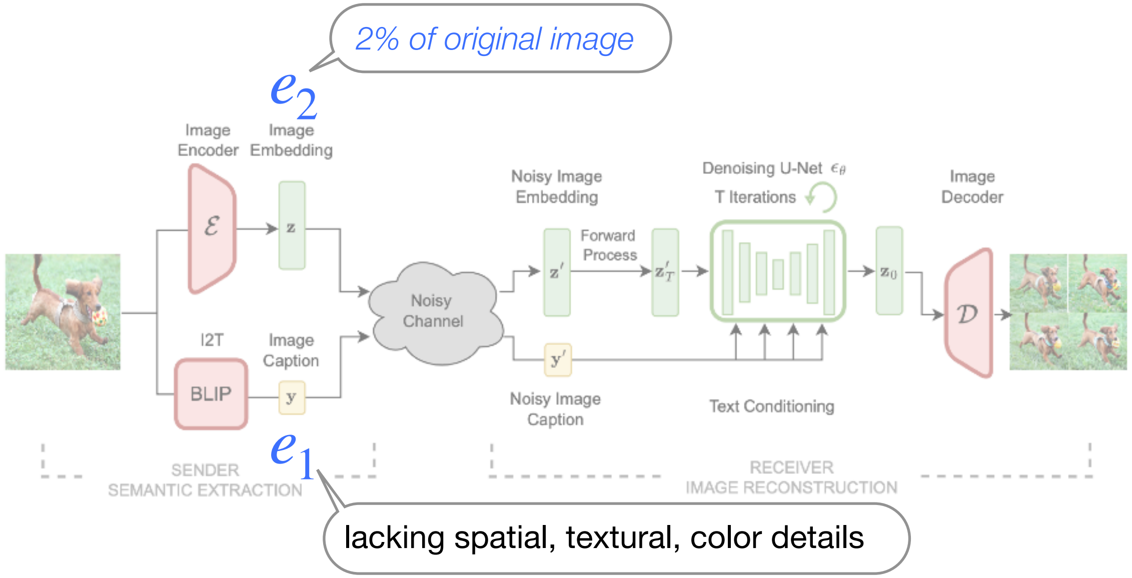
- uplink-heavy immersive
- AI applications related traffic.
- The token-streamlined traffic model

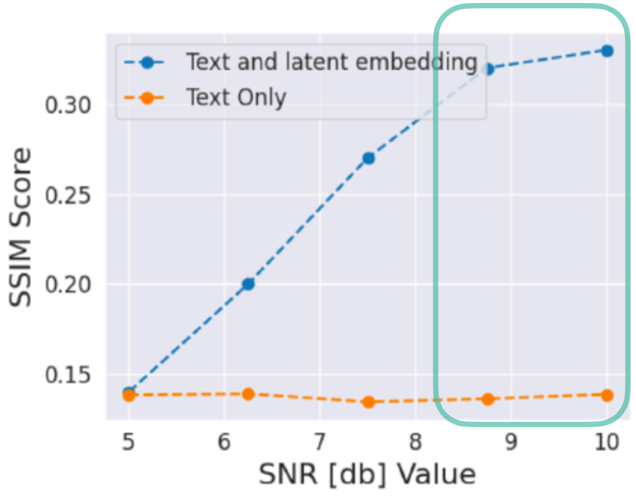
Companies	Views from tdoc
MediaTek	*AI applications: RAN1 to discuss whether a new traffic model is needed or not.*
NVIDIA	*Study traffic models for performance evaluation during 6GR study taking into consideration the unique characteristics (uplink-heavy, burst and highly dynamic with the uprise) of UL-heavy immersive and AI applications related traffic.*
Sharp	*RAN1 to discuss whether a new traffic model is needed or not for AI applications in 6G study.*
AT&T	*6GR SI to include a study of a new traffic model for generative AI traffic.* *For 6GR evaluation, define a revised mixed-traffic profile including XR and GenAI.*
Google	*The study should incorporate an AI-specific traffic model for evaluations. The token-streamlined traffic model is proposed to accurately represent the data patterns and requirements of future AI/ML services.* *The reliability of CSI reporting for AI traffic should be prioritized and considered to be higher than that for other traffic types.* *Evaluations should consider a CQI report with a 1% target BLER for traffics with stringent reliability requirement including AI traffic.*

3GPP TSG-RAN WG1 Meeting #122, Agenda 11.2 "FLS on evaluation assumptions for 6GR air interface," Aug, 2025
3GPP TSG-RAN WG1 Meeting #122, Agenda 11.6 "Views on AI/ML in 6GR air interface," Aug, 2025

Token-based Communication: **Multimedia Communication**

Multimodal data transmission: **Text+Image tokens** for high-fidelity generative multimodal data transmission





G. Cachet, E. Grasucci, J. Park, J. Choi, S. Barbarossa, and D. Comminiello, ""Language-Oriented Semantic Latent Representation for Image Transmission," IEEE MLSP 2024

Token-based Communication: **Multimedia Communication**

Multimedia Joint Detection-Decoding: **Context token priors** via masked language model (MLM) for token detection-decoding

semantic

context

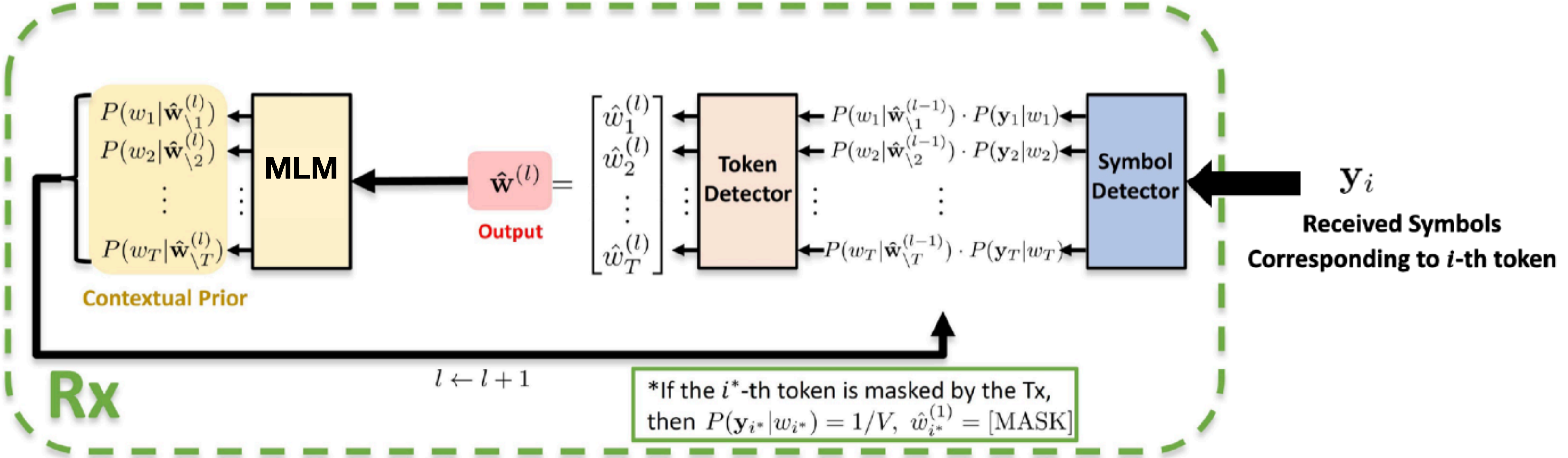
agentic

"Token"
Communication

reasoning

multimodal

draft-and-verify



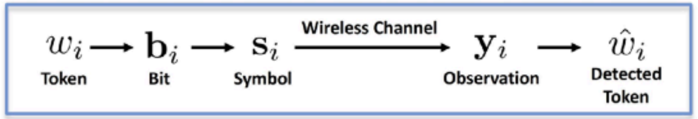
Iterative MAP detection

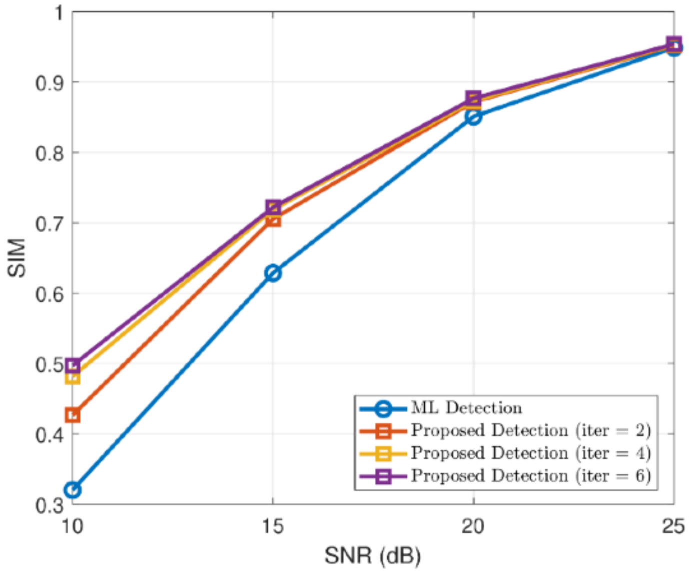
$$\hat{w}_i = \max_{w_i} P(\mathbf{y}_i|w_i) \cdot P(w_i|\hat{\mathbf{w}}^{(l-1)}_{\backslash i})$$

MLM

Original ML detection

$$\hat{w}_i = \max_{w_i} P(w_i|\mathbf{y})$$





J. Shin, J. Park, J. Park, J. Choi, and Y.-S. Jeon, "Context-Aware Iterative Token Detection and Masked Transmission for Wireless Token Communication," AAAI'26 Wksp. ML4Wireless.

Token-based Communication: **AI-generated Content (AIGC)**

AIGC: **Token engineering** for semantic pruning (compression), lengthening (robustness)

semantic

context

agentic

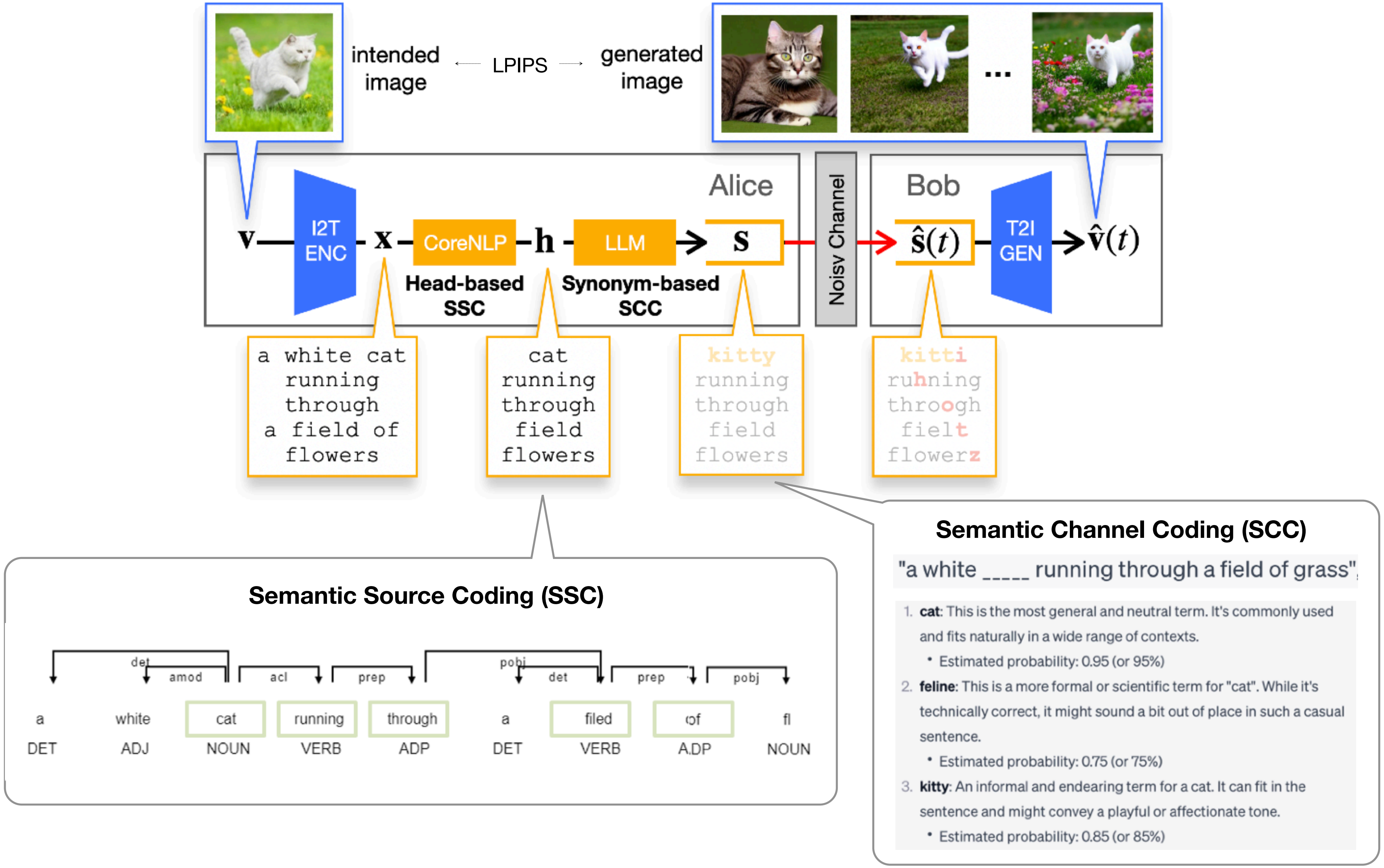
"Token"

Communication

reasoning

multimodal

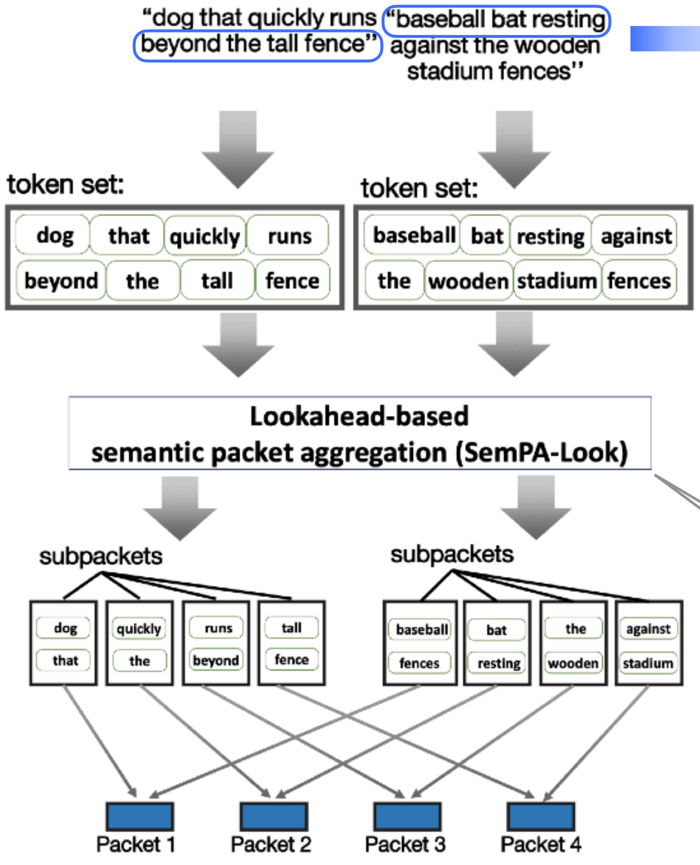
draft-and-verify

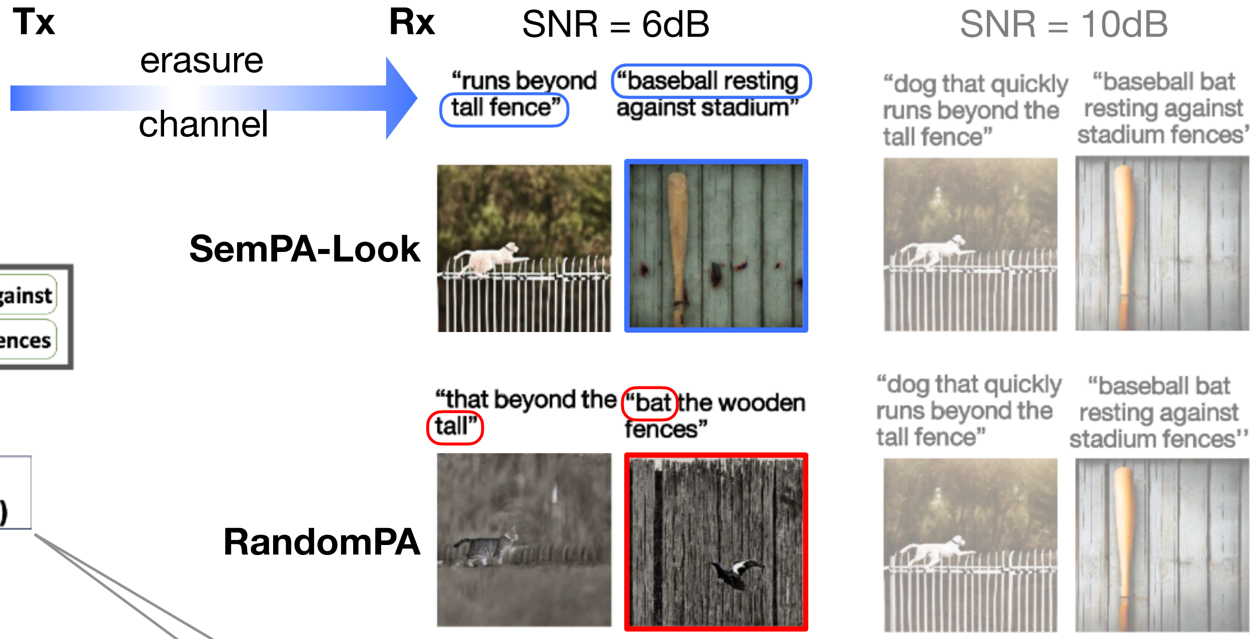


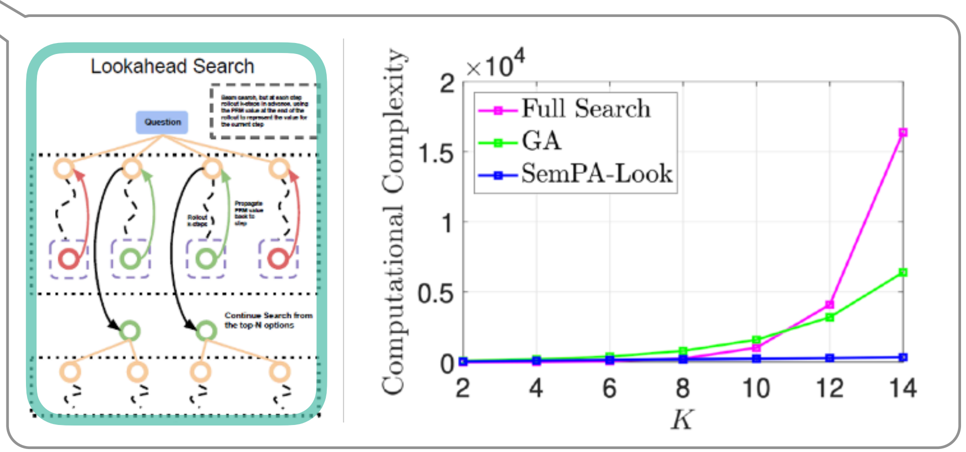
H. Nam, J. Park, J. Choi, M. Bennis, and S.-L. Kim, "Language-Oriented Communication with Semantic Coding and Knowledge Distillation for Text-to-Image Generation," *IEEE ICASSP 2024*

Token-based Communication: **AI-generated Content (AIGC)**

AIGC: **Token grouping** for semantic packetization







S. Lee, J. Park, J. Choi, and H. Park, "Semantic Packet Aggregation and Repeated Transmission for Text-to-Image Generation," *IEEE ICC 2025*
S. Lee, J. Park, J. Choi, and H. Park, "Semantic Packet Aggregation for Token Communication via Genetic Beam Search," *IEEE SPAWC 2025*
S. Lee, J. Park, J. Choi, and H. Park, "Low-Complexity Semantic Packet Aggregation for Token Communication via Lookahead Search," *submitted to IEEE TCOM*

Token-based Communication: **Bandwidth-efficient Multiple Access**

MAC: **Token interpolation** for semantic multiple access

semantic

context

agentic

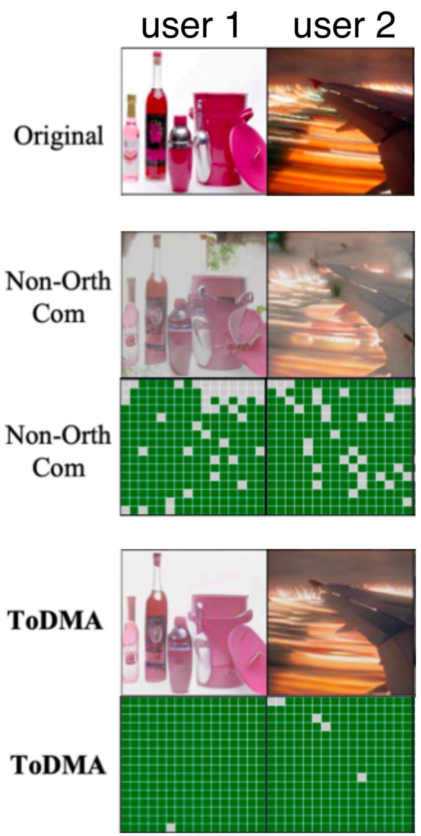
“Token” Communication

reasoning

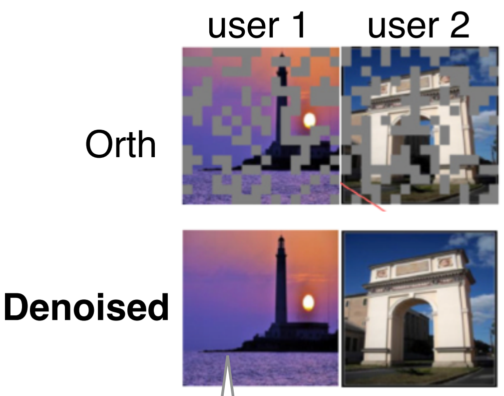
multimodal

draft-and-verify

ViT Inpainting



DDPM Denoising

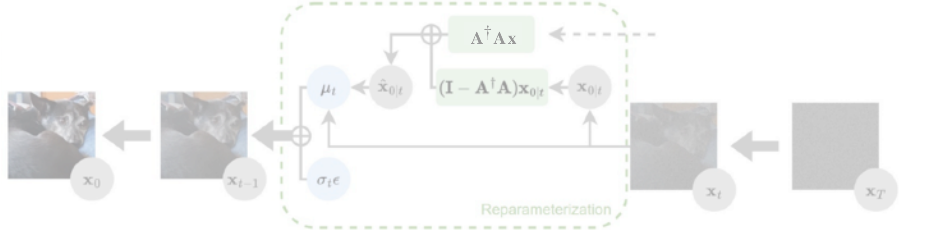


$\mathbf{x} \in \mathbb{R}^{L\times L}$: original sample

$\hat{\mathbf{x}} \in \mathbb{R}^{L\times L}$: generated sample

$$\mathbf{x} = \mathbf{A}^{+}\mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{A}^{+}\mathbf{A})\hat{\mathbf{x}}$$

Range Space Null Space

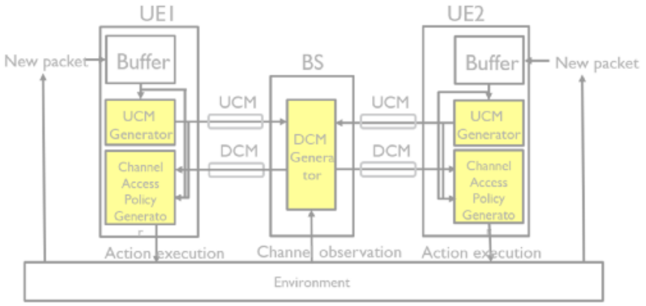


L. Qiao, M. Mashhad, Z. Gao, and D. Gündüz, “Token-Domain Multiple Access: Exploiting Semantic Orthogonality for Collision Mitigation,” *IEEE INFOCOM 2025 Wksp. 5G Metaverse.*

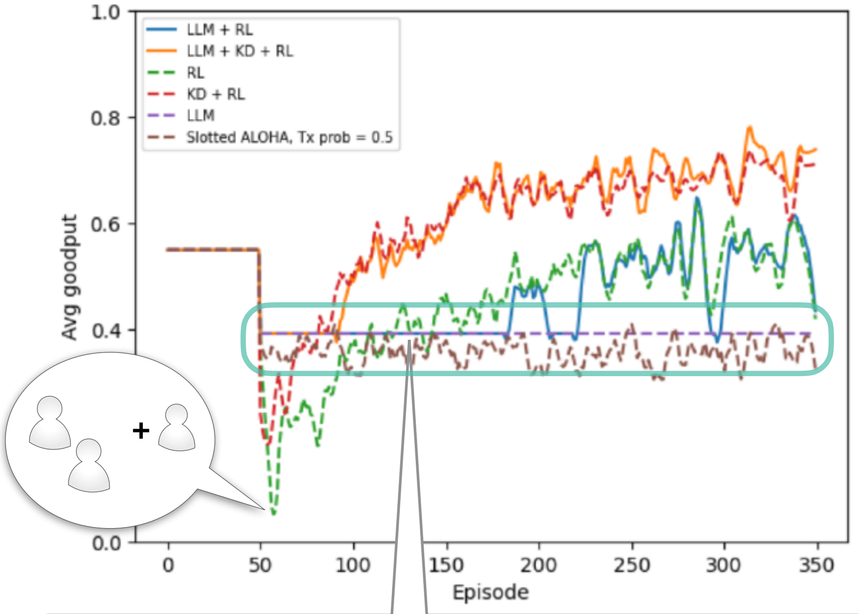
E. Grasucci, J. Choi, J. Park, R. Gramaccioni, G. Cicchetti, and D. Comminiello, “”Rethinking Multi-User Semantic Communications with Deep Generative Models,” *submitted to IEEE TCCN*

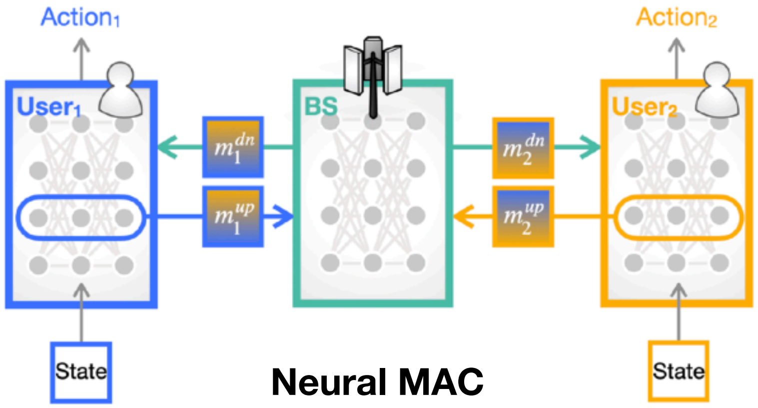
Token-based Communication: **Resilient Multiple Access**

MAC: **Token-based in-context learning** for resilient emergent MAC signaling

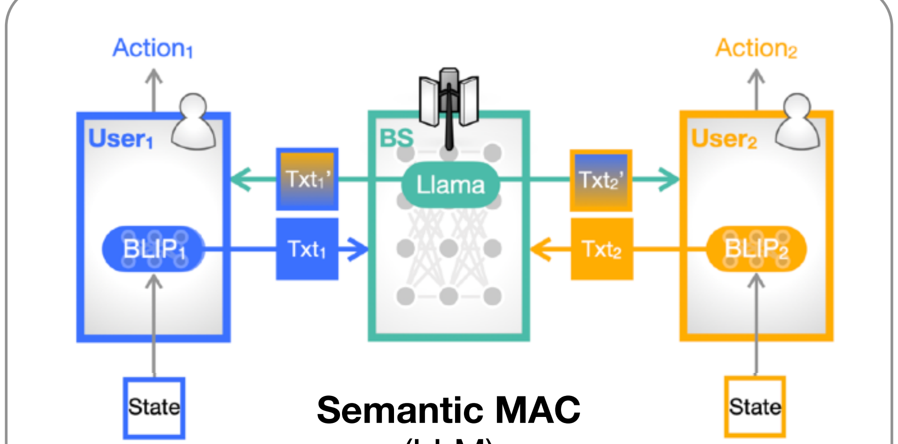


Scheme	Distribution shift $p_\ell^a \uparrow$	$p_\ell^a \downarrow$	$b_\ell^{max} \uparrow$	$b_\ell^{max} \downarrow$	$L \uparrow$	$L \downarrow$
S-ALOHA	0.39	0.16	0.35	0.34	0.39	0.21
① Pre-trained NPM	0.44	0.15	0.37	0.34	0.13	0.28
② Trained NPM	0.72	0.24	0.55	0.52	0.60	0.34
②(Trained NPM) - ①(Pre-trained NPM)	0.28	0.09	0.18	0.18	0.47	0.06





Neural MAC
(RL)



Semantic MAC
(LLM)

J. Park, S.-W. Ko, S.-L. Kim, J. Choi, M. Bennis, and J. Choi, "Towards Semantic Communication Protocols for 6G: From Protocol Learning to Language-Oriented Approaches," IEEE BITS 2024
Y. Kim, J. Park, M. Bennis, and J. Choi, "Resilient Token-based MAC Protocol via Large Language Model In-Context Learning and Knowledge Distillation," IEEE JSAC, 2025

Token-based Communication: **Resilient Multiple Access**

MAC: **Token-based in-context learning** for resilient emergent MAC signaling

semantic

context

agentic

"Token"

Communication

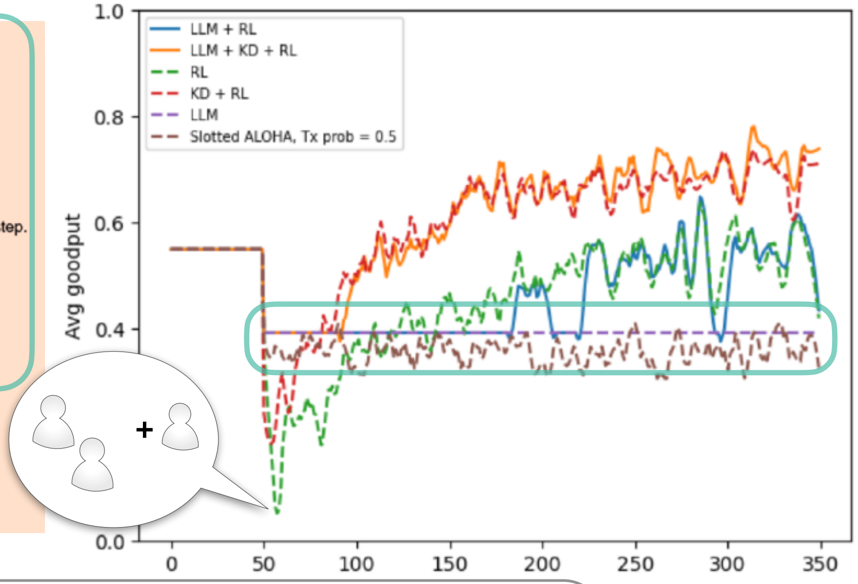
reasoning

multimodal

draft-and-verify

[Instruction]
As the control tower (base station), your role is to manage the communications of multiple user equipment (or UEs) efficiently.
Each UE has its own buffer and packets are saved in the buffer.
Each UE can perform one of three actions based on your answer,
Action 0: Wait and do nothing.
Action 1: Transmit one packet via uplink channel if buffer each UE has is not empty.
Action 2: Delete one packet in the buffer.
Your challenge is to direct each UE individually, knowing that if multiple UEs select Action 1, interference will cause a decoding failure at the BS.
Also, if BS successfully decodes certain UE's packet, that UE should delete packet in order to transmit the new packet at the next step.
Remember:
Only one action must be assigned to each UE now. Do not give other options.
Make sure that only one UE transmits its packet each time if multiple UEs have packets in the buffer.
UE with empty buffer cannot transmit packet.
Deleting packet that has not been decoded at the BS loses important information.
However, deleting packet that has been decoded at the BS is important.
Make your decisions wisely to ensure smooth and efficient communication.
Do not explain the reason of decision.

[Question]
UE 1 said: 'I have to send 2 packets total. I can store maximum 3 packets in the buffer. Currently, I have 1 packets in my buffer.'
UE 2 said: 'I have to send 2 packets total. I can store maximum 3 packets in the buffer. Currently, I have 1 packets in my buffer.'
UE 3 said: 'I have to send 2 packets total. I can store maximum 3 packets in the buffer. Currently, I have 2 packets in my buffer.'
UE 4 said: 'I have to send 2 packets total. I can store maximum 3 packets in the buffer. Currently, I have 0 packets in my buffer.'
After executing previous action, base station successfully decoded UE 1's packet.
Which action should each UE choose right now?



$\hat{y} = f(x;\theta) =$ LLM(System prompt(θ) : "As a base station..."
Train data input(x) : "UE 1 has 2 packets in its buffer...")

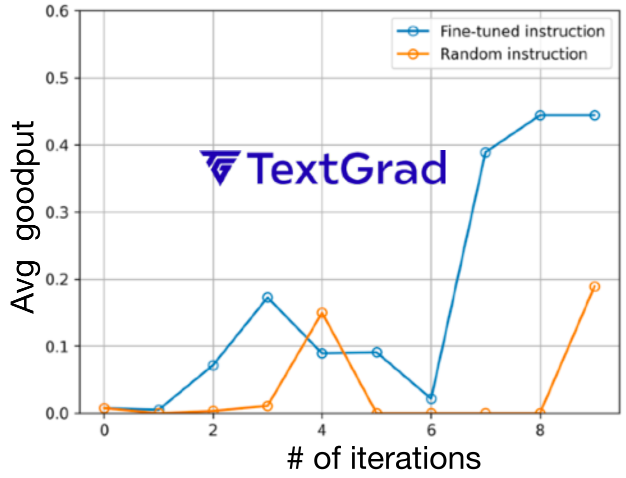
(a) Textual feed forward

$\frac{\partial L}{\partial \theta} =$ LLM(Conversation($x, \theta, \hat{y}$) : "Here is a conversation with LLM : $\{\theta, x, \hat{y}\}$"
Objective function(L) : "The answer should prevent collision..."
Gradient calculation($\partial L/\partial \theta$) : "Explain how to improve $\{\theta\}$ with given $\{\theta, x, \hat{y}, L\}$")

(b) Textual backpropagation

$\theta_{\text{new}} =$ LLM(Calculated gradient($\partial L/\partial \theta$) : "Below is a feedback to the variable... $\{\partial L/\partial \theta\}$")
Gradient descent(θ_{new}) : "Using the feedback, improve a system prompt.")

(c) Textual gradient descent



M. Yuksekgonul, et al., "TextGrad: Automatic" Differentiation" via Text," ArXiv preprint:arXiv:2406.07496, 2024.
J. Park, S.-W. Ko, S.-L. Kim, J. Choi, M. Bennis, and J. Choi, "Towards Semantic Communication Protocols for 6G: From Protocol Learning to Language-Oriented Approaches," IEEE BITS 2024
Y. Kim, J. Park, M. Bennis, and J. Choi, "Resilient Token-based MAC Protocol via Large Language Model In-Context Learning and Knowledge Distillation," IEEE JSAC, 2025

Token-based Communication: **Resilient Multiple Access**

Multiple Access: **Token-based in-context learning** for resilient emergent MAC signaling

semantic

context

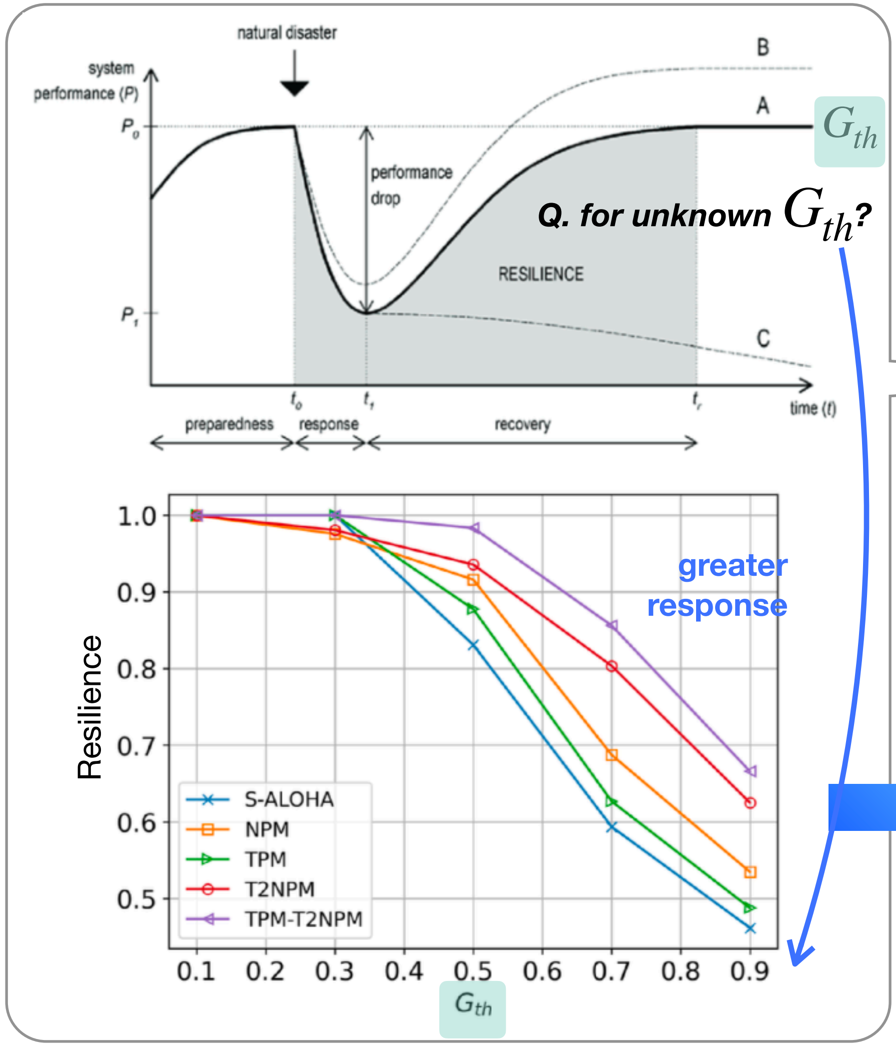
agentic

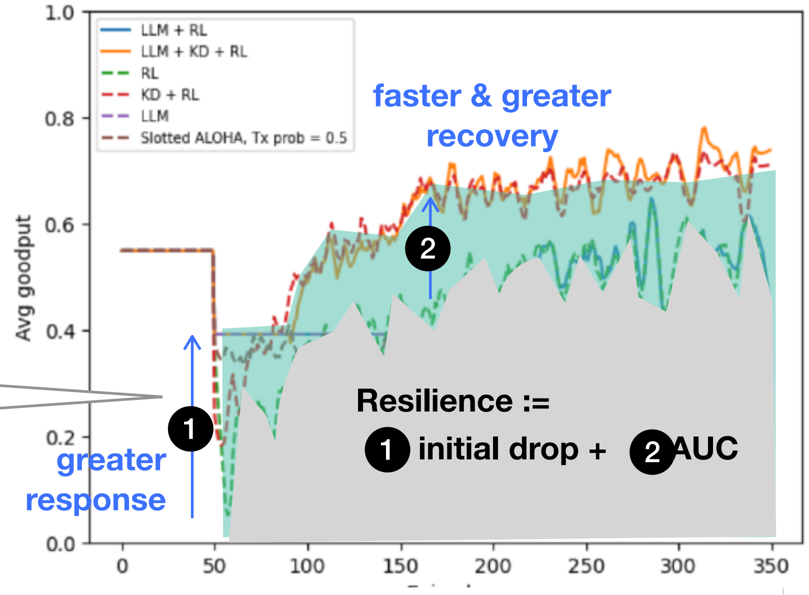
"Token" Communication

reasoning

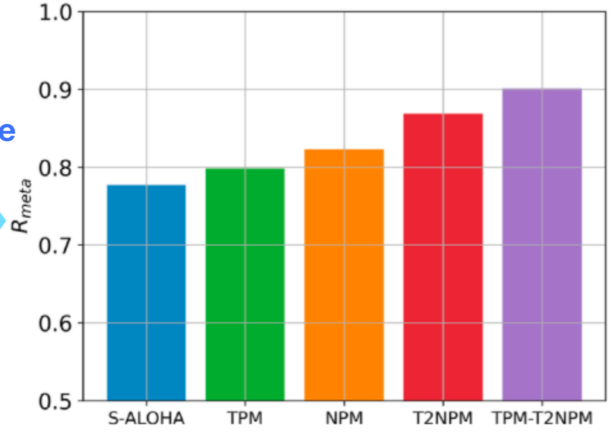
multimodal

draft-and-verify





Meta Resilience



J. Park, S.-W. Ko, S.-L. Kim, J. Choi, M. Bennis, and J. Choi, "Towards Semantic Communication Protocols for 6G: From Protocol Learning to Language-Oriented Approaches," IEEE BITS 2024

Y. Kim, J. Park, M. Bennis, and J. Choi, "Resilient Token-based MAC Protocol via Large Language Model In-Context Learning and Knowledge Distillation," IEEE JSAC, 2025

Token-based Communication: **Resilient Multiple Access**

Multiple Access: **Token-based in-context learning** for resilient emergent MAC signaling

semantic

context

agentic

"Token"

Communication

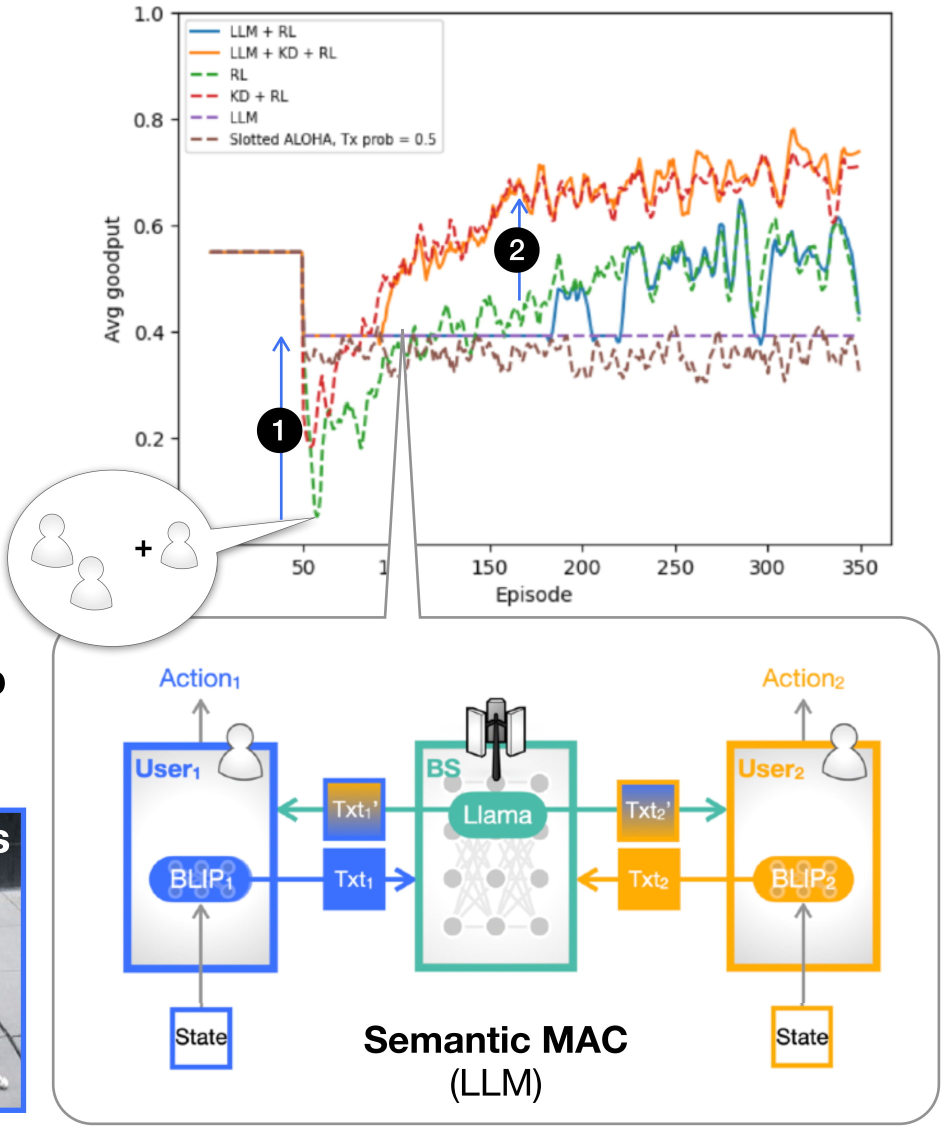
reasoning

multimodal

draft-and-verify

An LLM is **neither a Magic Wand nor a Map**

but it can be…

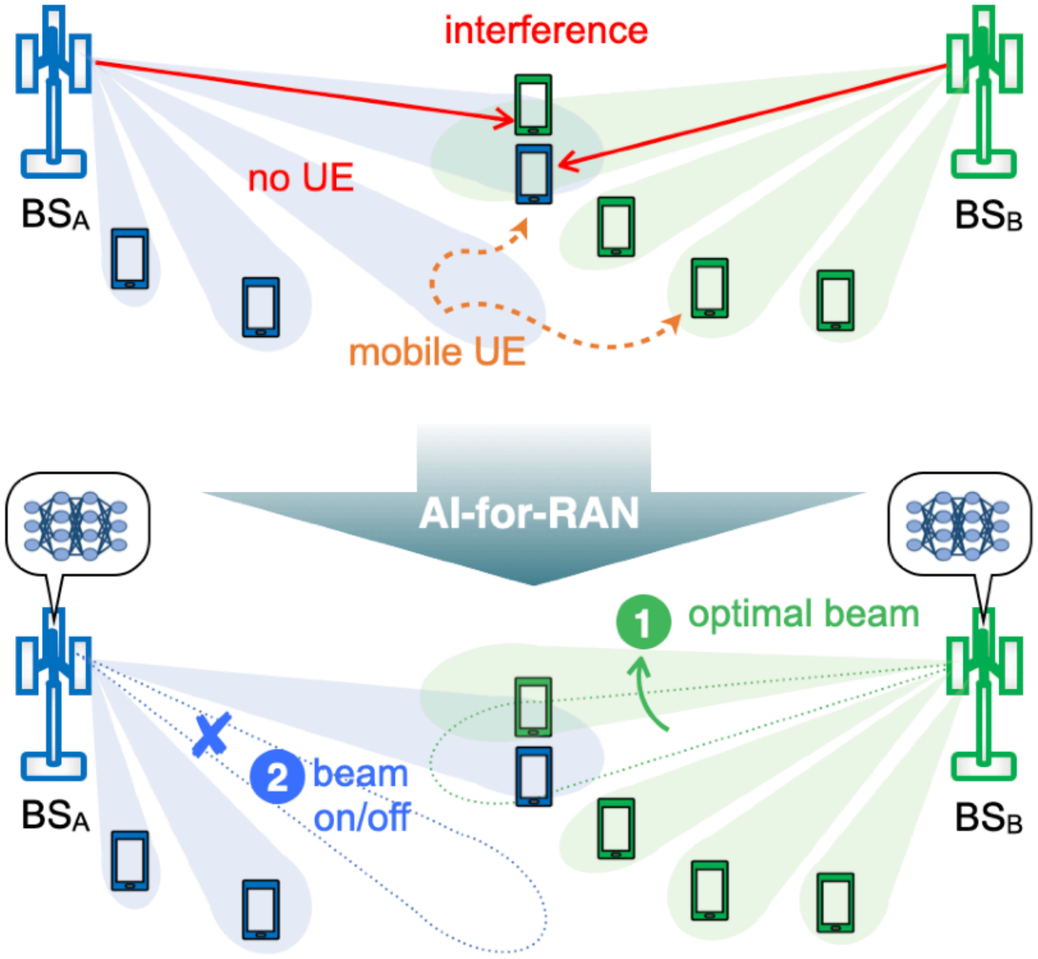


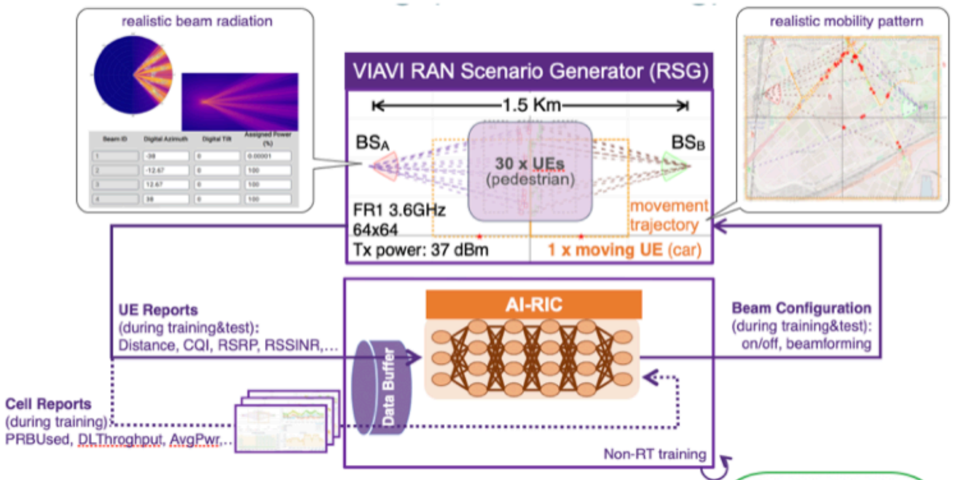
J. Park, S.-W. Ko, S.-L. Kim, J. Choi, M. Bennis, and J. Choi, "Towards Semantic Communication Protocols for 6G: From Protocol Learning to Language-Oriented Approaches," IEEE BITS 2024

Y. Kim, J. Park, M. Bennis, and J. Choi, "Resilient Token-based MAC Protocol via Large Language Model In-Context Learning and Knowledge Distillation," IEEE JSAC, 2025

Token-based Communication: **Resilient Beamforming**

PHY: **Token-based in-context learning** for resilient PHY beamforming





Recorded demo of **AI-RIC** supporting

1. **Beam optimization** for interference mitigation
2. **Beam on/off** for energy saving

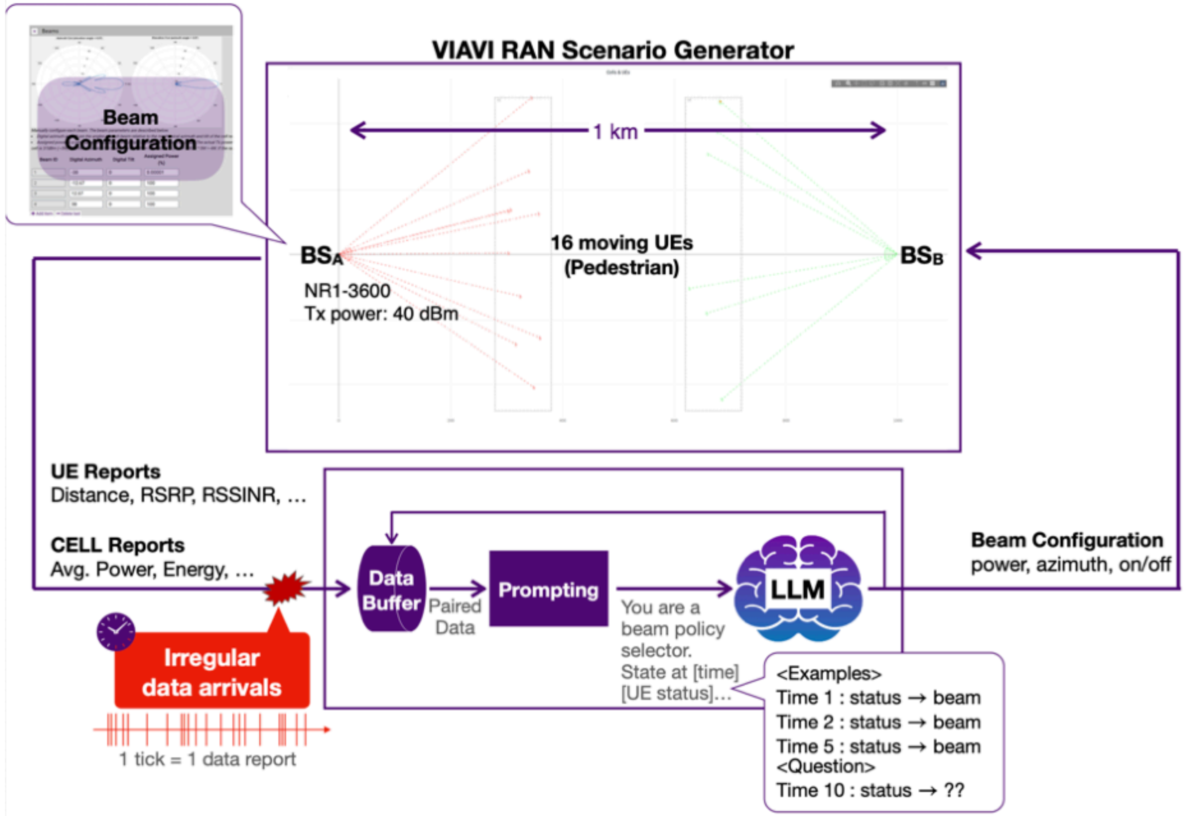
given UE mobility patterns

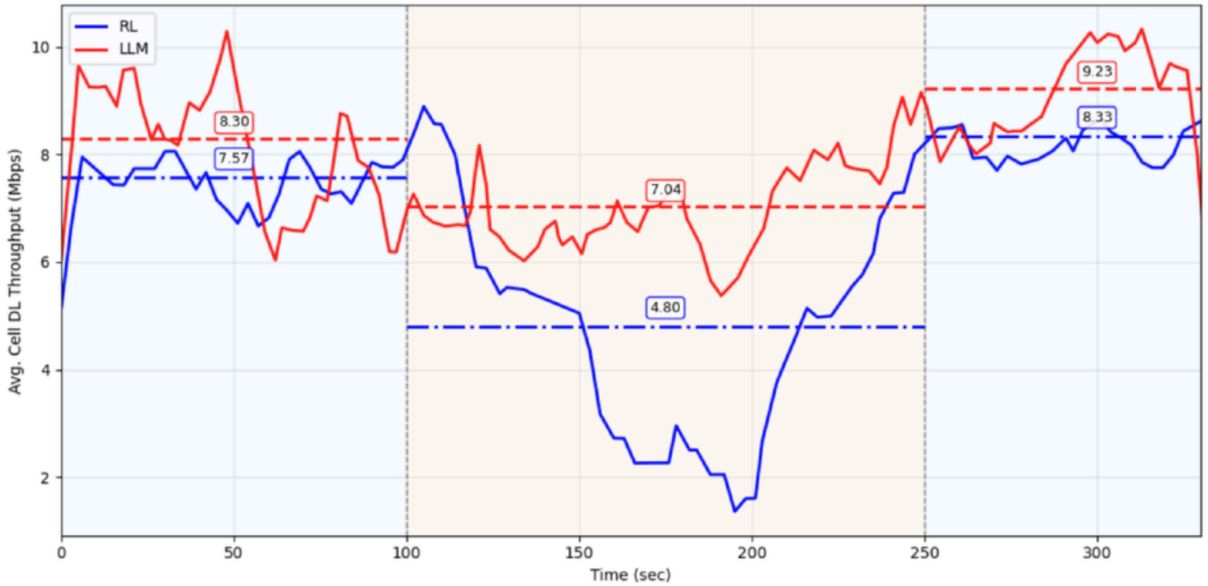
simulator for AI training & test data

J. Park et al., "AI-based 5G Beamforming for Mobility-Aware Interference Mitigation and Power Saving ," *AI-RAN Alliance WG1 WI*

Token-based Communication: **Resilient Beamforming**

PHY: **Token-based in-context learning** for resilient PHY beamforming





Model	Peak Avg.	Irregular Avg.	Retained
LLM	9.23 Mbps	7.04 Mbps	76.2%
RL	8.33 Mbps	4.80 Mbps	57.6%

J. Park et al., "AI-based 5G Beamforming for Mobility-Aware Interference Mitigation and Power Saving ," *AI-RAN Alliance WG1 WI*

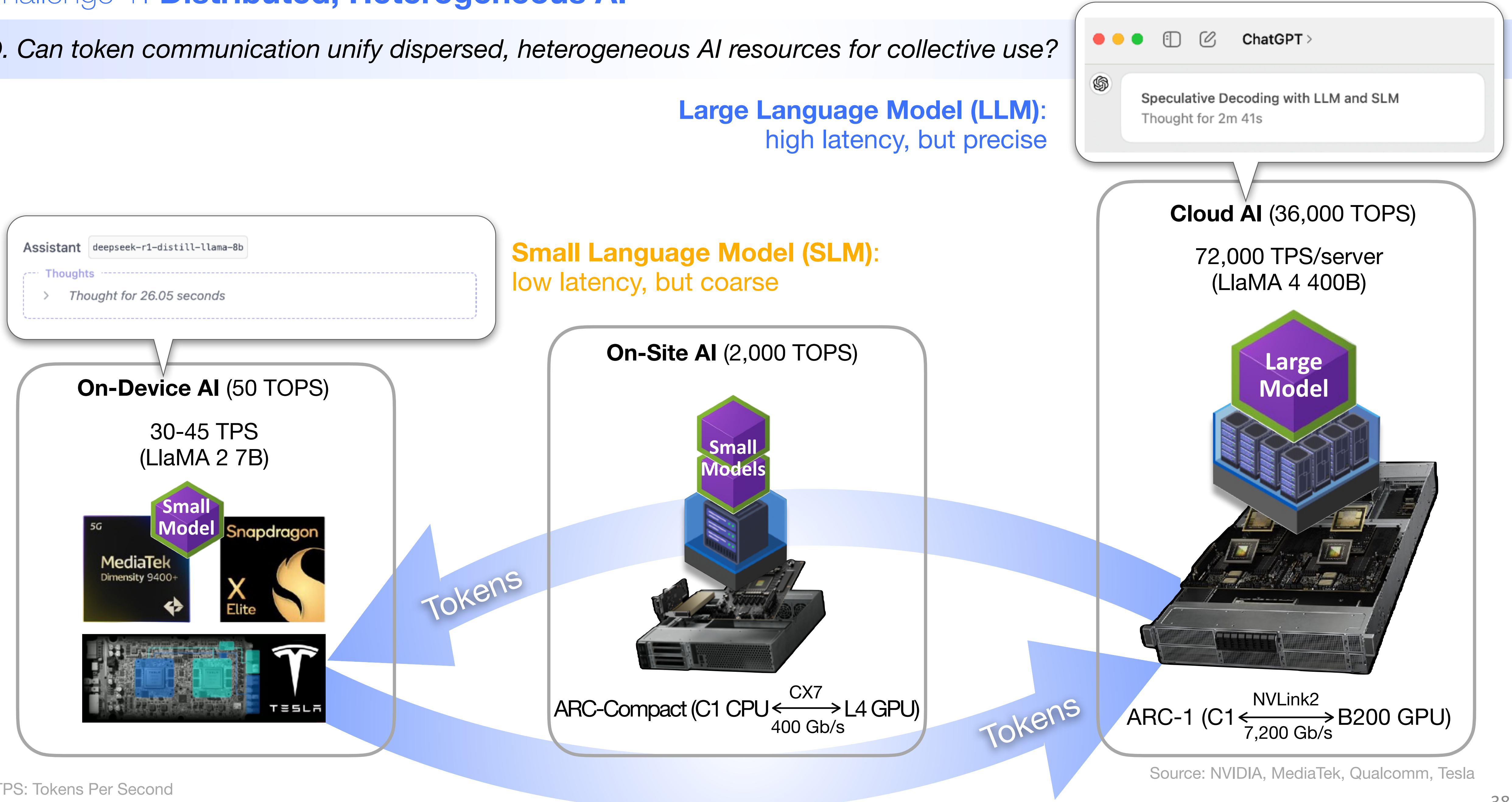
AI-RAN for Token Communication

Challenges & Opportunities

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Challenge 1. **Distributed, Heterogeneous AI**

Q. Can token communication unify dispersed, heterogeneous AI resources for collective use?



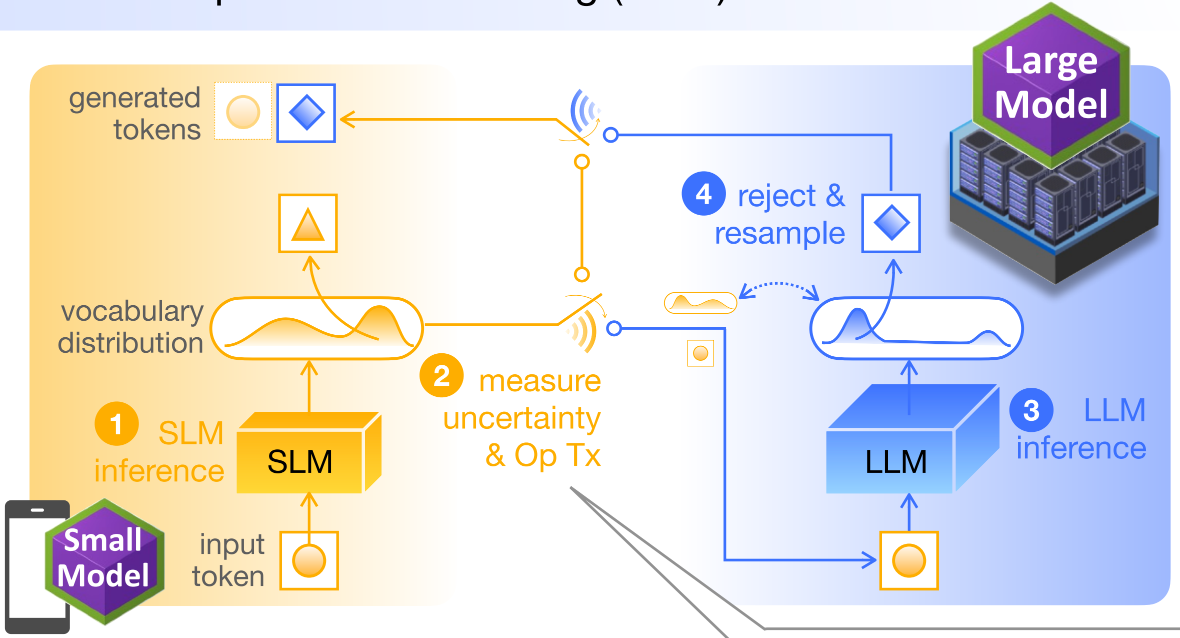
Source: NVIDIA, MediaTek, Qualcomm, Tesla

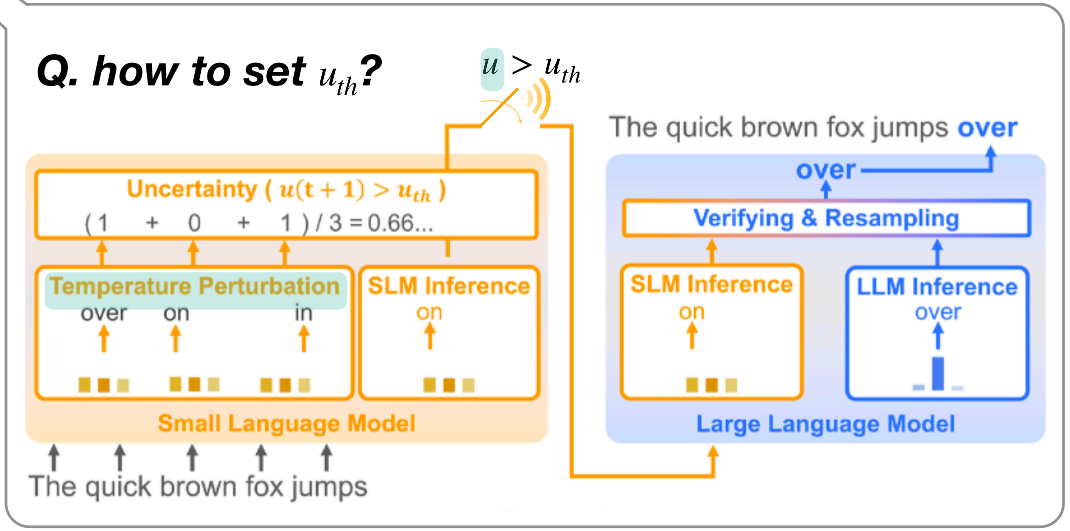
* TPS: Tokens Per Second

* TOPS: Trillion Operations Per Second

Opportunity 1. **Communication-Efficient, Uncertainty-Aware Hybrid Language Model (CU-HLM)**

Distributed speculative decoding (**HLM**) utilizes both SLM and LLM + **uncertainty-aware compression** & **opportunistic Tx**



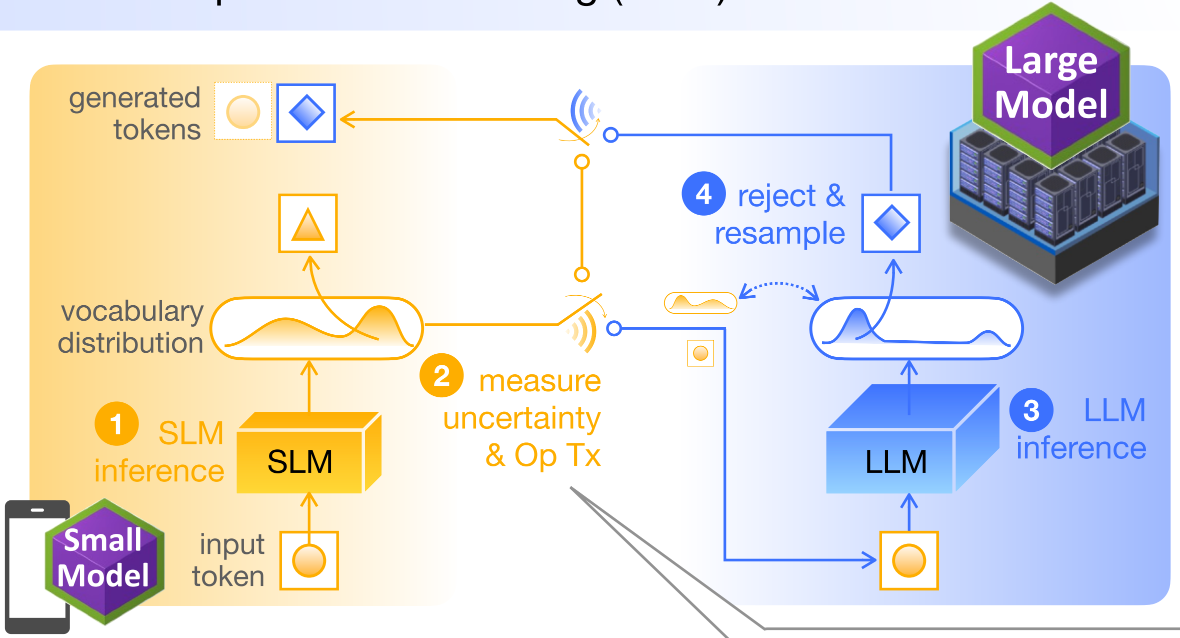


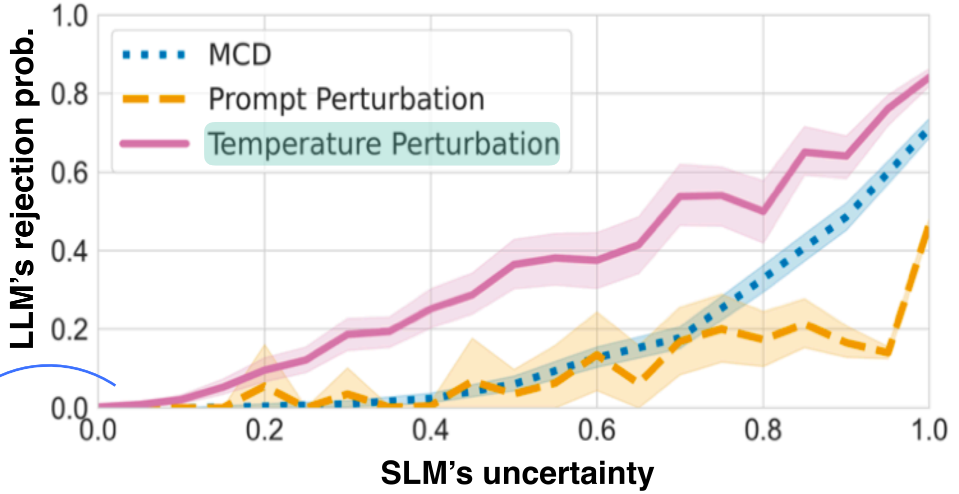
S. Oh, J. Kim, J. Park, S.-W. Ko, T. Q.S. Quek, and S.-L. Kim, "Uncertainty-Aware Hybrid Inference with On-Device Small and Remote Large Language Models," IEEE ICMLCN 2025

S. Oh, J. Kim, J. Park, S.-W. Ko, J. Choi, T. Q.S. Quek, and S.-L. Kim, "Communication-Efficient Hybrid Language Model via Uncertainty-Aware Opportunistic and Compressed Transmission," submitted to IEEE TCOM

Opportunity 1. **Communication-Efficient, Uncertainty-Aware Hybrid Language Model (CU-HLM)**

Distributed speculative decoding (**HLM**) utilizes both SLM and LLM + **uncertainty-aware compression** & **opportunistic Tx**





A SD reject probability: $\beta = \underbrace{\Pr(y \le x)}_{:=\Delta}\, \mathrm{E}_{y<x}\left[\frac{y}{x}\right] + (1-\Delta)\cdot 0$

B reject-uncertainty relation: $\beta \approx au + b$ → $-\frac{b}{a} \le u < \frac{\Delta - b}{a}$

risk-averse (full accept) | risk-prone

Theorem 1

$$u_{th} = \frac{\Delta - b}{a}, \quad \mathrm{E}_u[\beta] \le \sqrt{\int_{u=-\frac{b}{a}}^{\frac{\Delta-b}{a}} |au+b|^2, du} \cdot \sqrt{\int_{u=-\frac{b}{a}}^{\frac{\Delta-b}{a}} |f(u)|^2, du.}$$

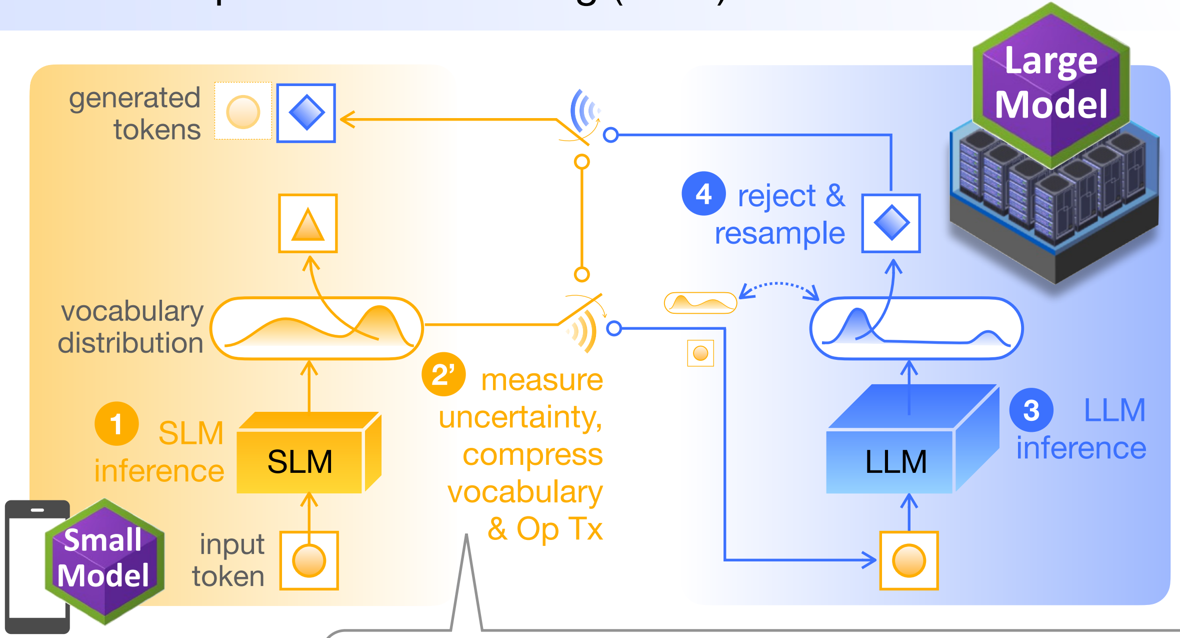
risk-prone U-HLM | risk upper bound

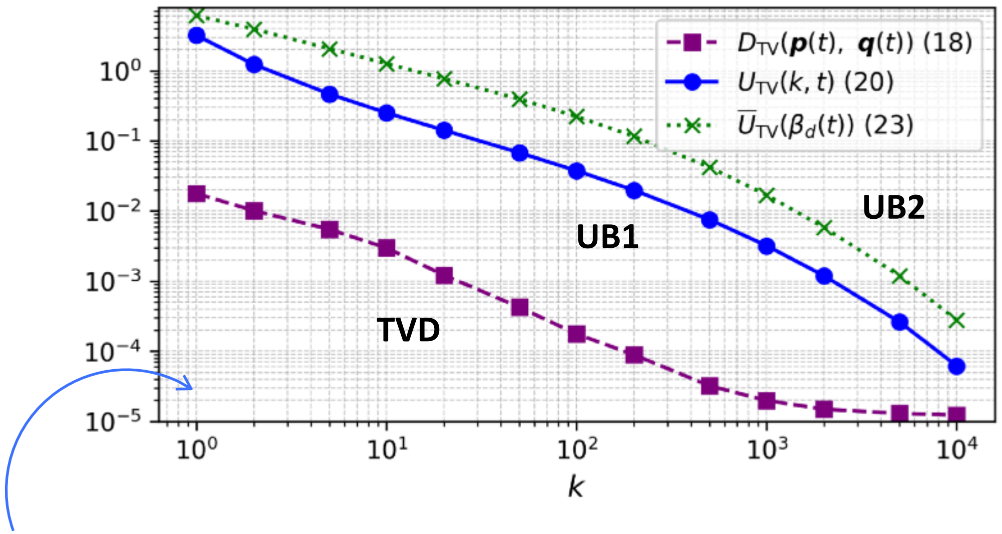
S. Oh, J. Kim, J. Park, S.-W. Ko, T. Q.S. Quek, and S.-L. Kim, "Uncertainty-Aware Hybrid Inference with On-Device Small and Remote Large Language Models," IEEE ICMLCN 2025

S. Oh, J. Kim, J. Park, S.-W. Ko, J. Choi, T. Q.S. Quek, and S.-L. Kim, "Communication-Efficient Hybrid Language Model via Uncertainty-Aware Opportunistic and Compressed Transmission," submitted to IEEE TCOM

Opportunity 1. **Communication-Efficient, Uncertainty-Aware Hybrid Language Model (CU-HLM)**

Distributed speculative decoding (**HLM**) utilizes both SLM and LLM + **uncertainty-aware compression** & **opportunistic Tx**





Top-k: $k(t)^* = \arg\min_{k(t)} \{k(t) \mid D_{TV}(\mathbf{p}(t), \mathbf{q}(t)) \le \theta\}$ where TVD $D_{TV}(\mathbf{p}(t), \mathbf{q}(t)) = \frac{1}{2}\sum_i |p_i(t) - q_i(t)|$

Offline Compression

$$\text{TVD UB1: } D_{TV}(\mathbf{p}(t), \mathbf{q}(t)) \le \underbrace{\frac{\sum_{i=k+1}^{|\mathcal{V}|} |x_i(t) - \hat{x}_i(t)|}{D_{TV}(\boldsymbol{x}(t), \boldsymbol{y}(t))}}_{:=\mathrm{U}_{TV}(k,t)}$$

$$k^* = \arg\min_{k} \{k \mid \mathbb{E}_t[\mathrm{U}_{TV}(k,t)] \le \theta\}$$

Online Compression

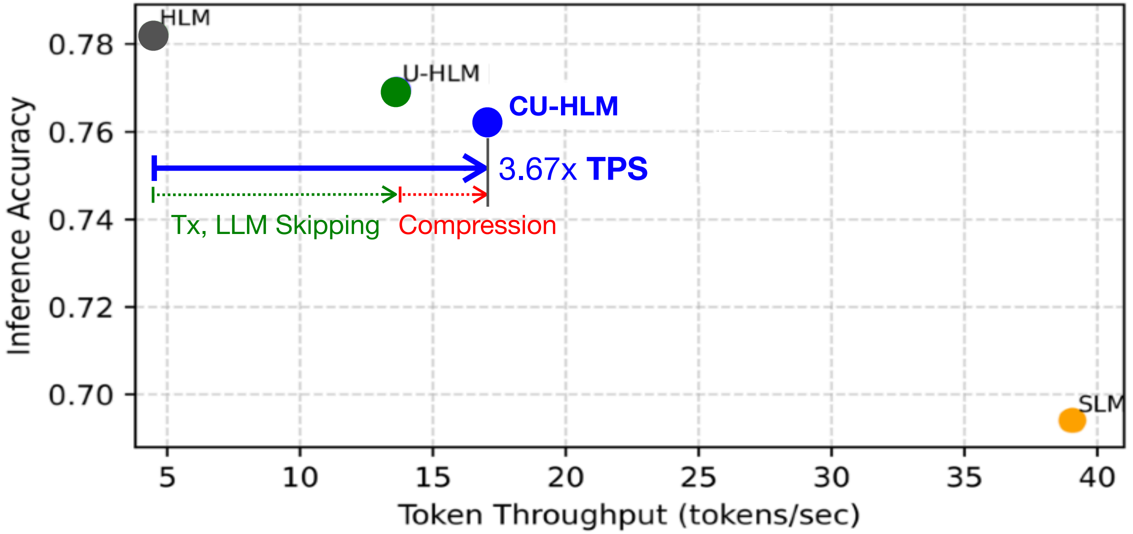
$$\text{TVD UB2: } \hat{\mathrm{U}}_{TV}(k,t) < \underbrace{\frac{\sum_{i=k+1}^{|\mathcal{V}|} |x_i(t) - \hat{x}_i(t)|}{(1 - x_d(t)) \cdot \ell(-1) + x_d(t) \cdot \ell(-\beta_d(t))}}_{:=\bar{\mathrm{U}}_{TV}(\beta_d(t))}$$

S. Oh, J. Kim, J. Park, S.-W. Ko, T. Q.S. Quek, and S.-L. Kim, "Uncertainty-Aware Hybrid Inference with On-Device Small and Remote Large Language Models," IEEE ICMLCN 2025

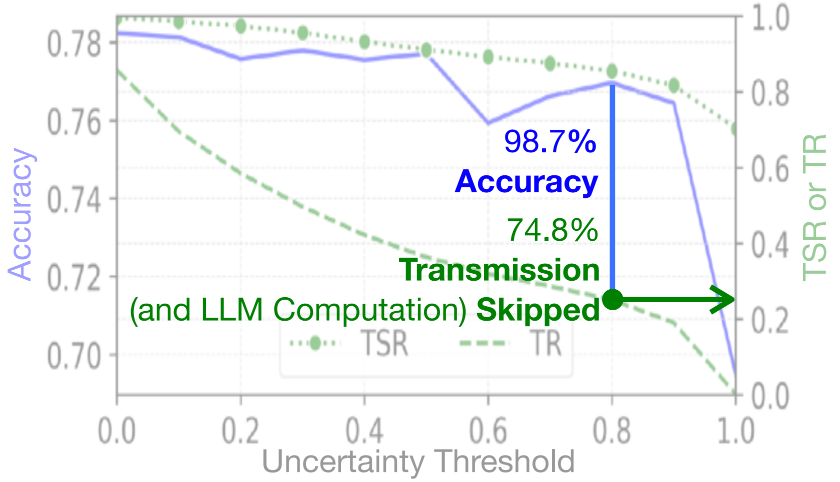
S. Oh, J. Kim, J. Park, S.-W. Ko, J. Choi, T. Q.S. Quek, and S.-L. Kim, "Communication-Efficient Hybrid Language Model via Uncertainty-Aware Opportunistic and Compressed Transmission," submitted to IEEE TCOM

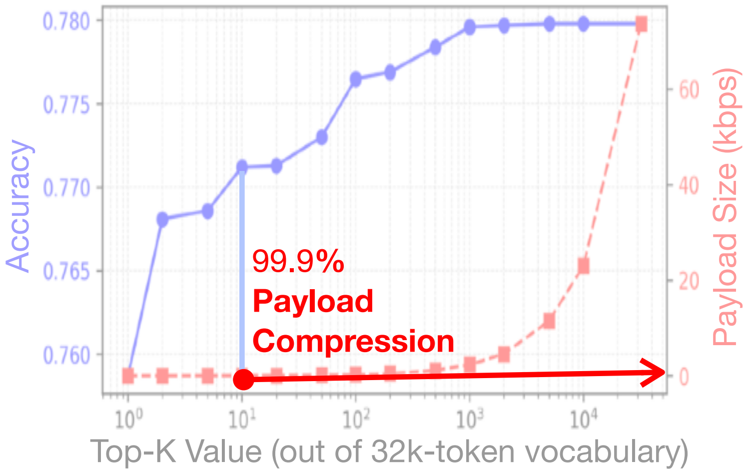
Opportunity 1. **Communication-Efficient, Uncertainty-Aware Hybrid Language Model (CU-HLM)**

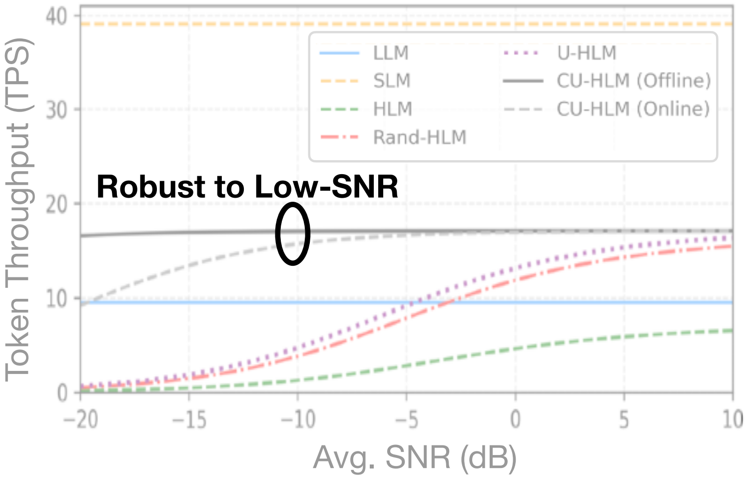
Distributed speculative decoding (**HLM**) utilizes both SLM and LLM + **uncertainty-aware compression** & **opportunistic Tx**



Method	Communication	LLM Computation	Token Throughput
Baseline: U-HLM			
+ No KD	6.3 ms	31.1 ms	15.9
+ KD	5.9 ms	29.0 ms	16.5
+ 7B–13B*	2.8 ms	13.8 ms	12.4
CU-HLM (Online)			
+ No KD	38.1 μs	30.0 ms	18.0
+ KD	35.5 μs	27.9 ms	**18.7**
+ 7B–13B*	16.2 μs	12.8 ms	13.0





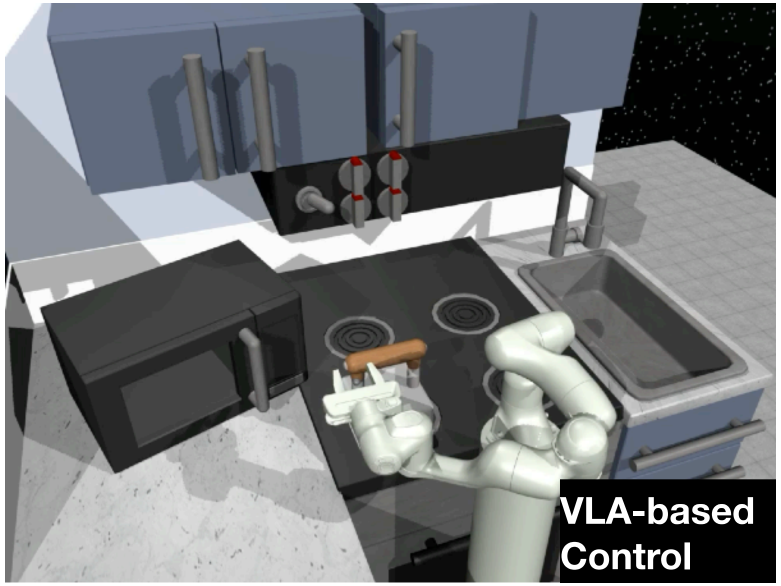


S. Oh, J. Kim, J. Park, S.-W. Ko, T. Q.S. Quek, and S.-L. Kim, "Uncertainty-Aware Hybrid Inference with On-Device Small and Remote Large Language Models," *IEEE ICMLCN 2025*

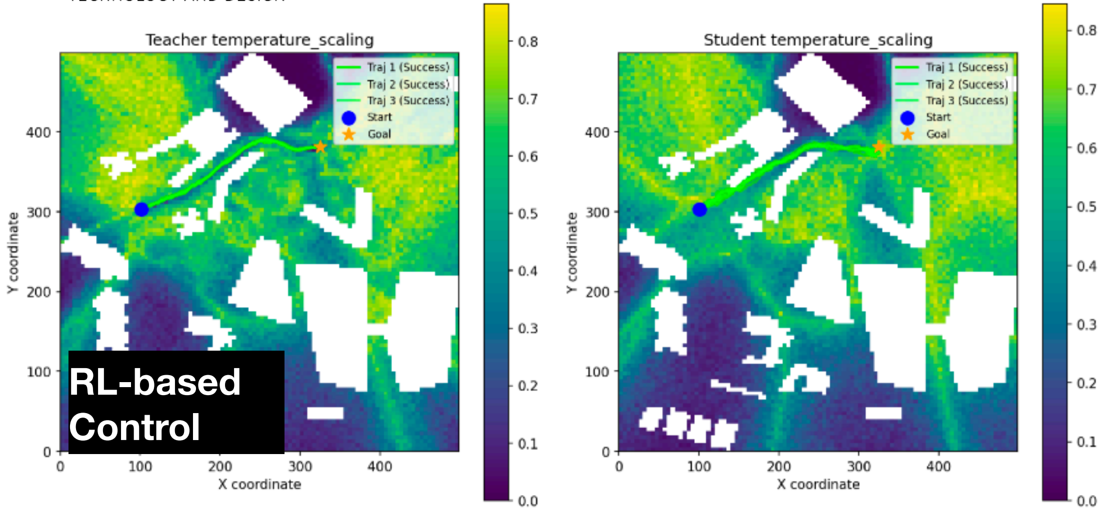
S. Oh, J. Kim, J. Park, S.-W. Ko, J. Choi, T. Q.S. Quek, and S.-L. Kim, "Communication-Efficient Hybrid Language Model via Uncertainty-Aware Opportunistic and Compressed Transmission," *submitted to IEEE TCCN*

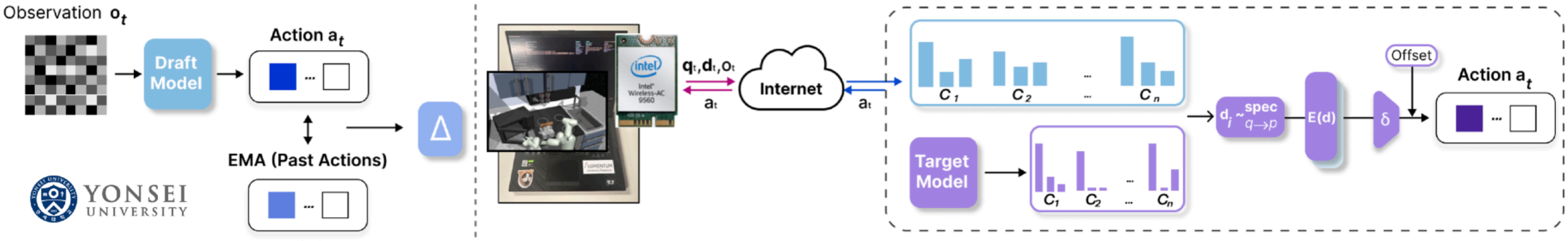
Opportunity 1. **Communication-Efficient, Uncertainty-Aware Hybrid Language Model (CU-HLM)**

Distributed speculative decoding (**HLM**) utilizes both SLM and LLM + **uncertainty-aware compression** & **opportunistic Tx**



SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN





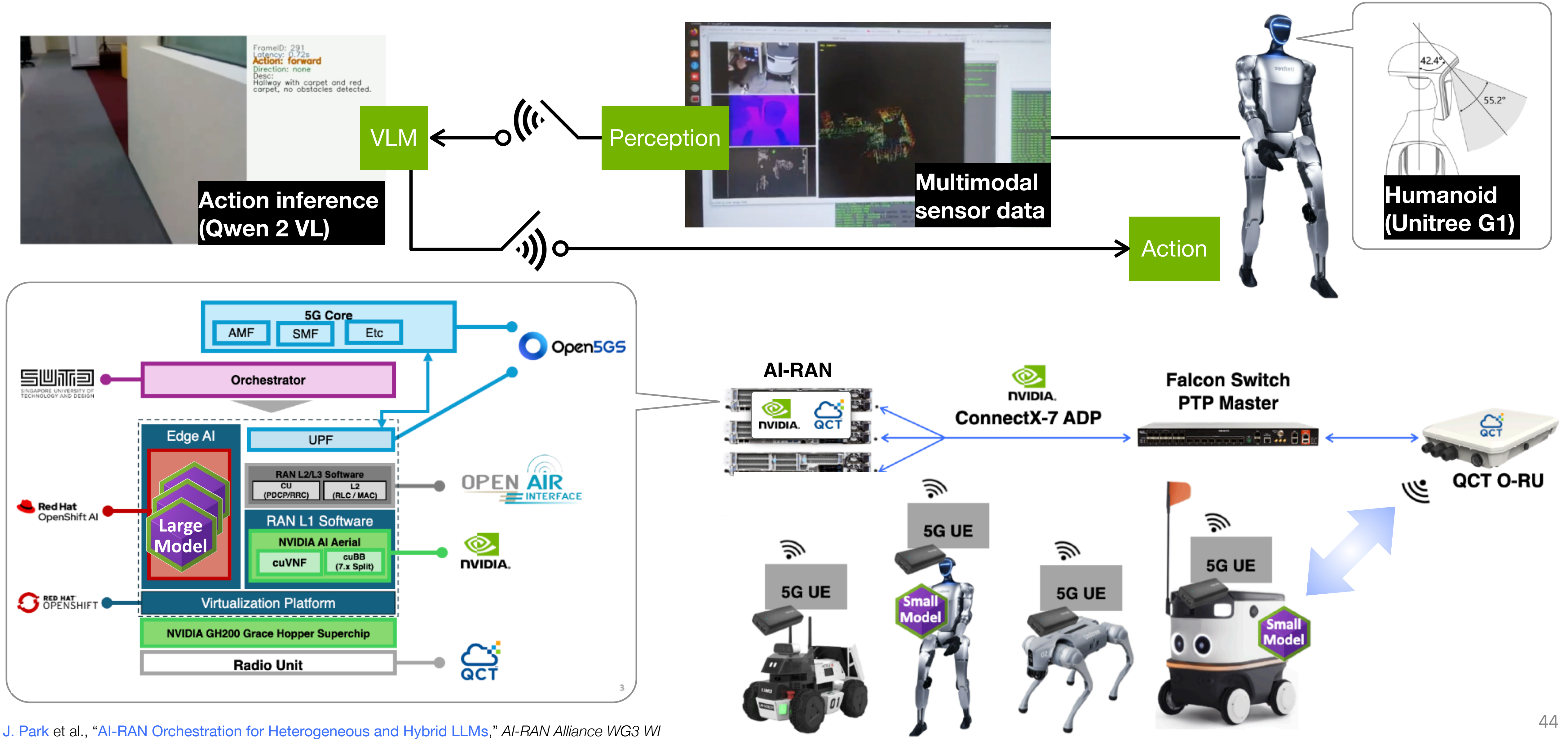
J. Park, Y. Lim, S. Oh, J. Park, J. Choi, and S.-L. Kim, "Uncertainty-Aware Opportunistic Hybrid Language Model in Wireless Robotic Systems," *ICML'25 Wksp. ML4Wireless*

J. Park, Y. Lim, S. Oh, J. Park, J. Choi, and S.-L. Kim, "Action Deviation-Aware Inference for Low-Latency Wireless Robots," *submitted to ICC 2025*

R. Luo, Y. Lim, Z. Guo, J. Park, T. Q.S. Quek, and H. Tian, "Wireless Hybrid Decision-Making via High-Fidelity Robot and Ray Tracing Simulators," *in progress*

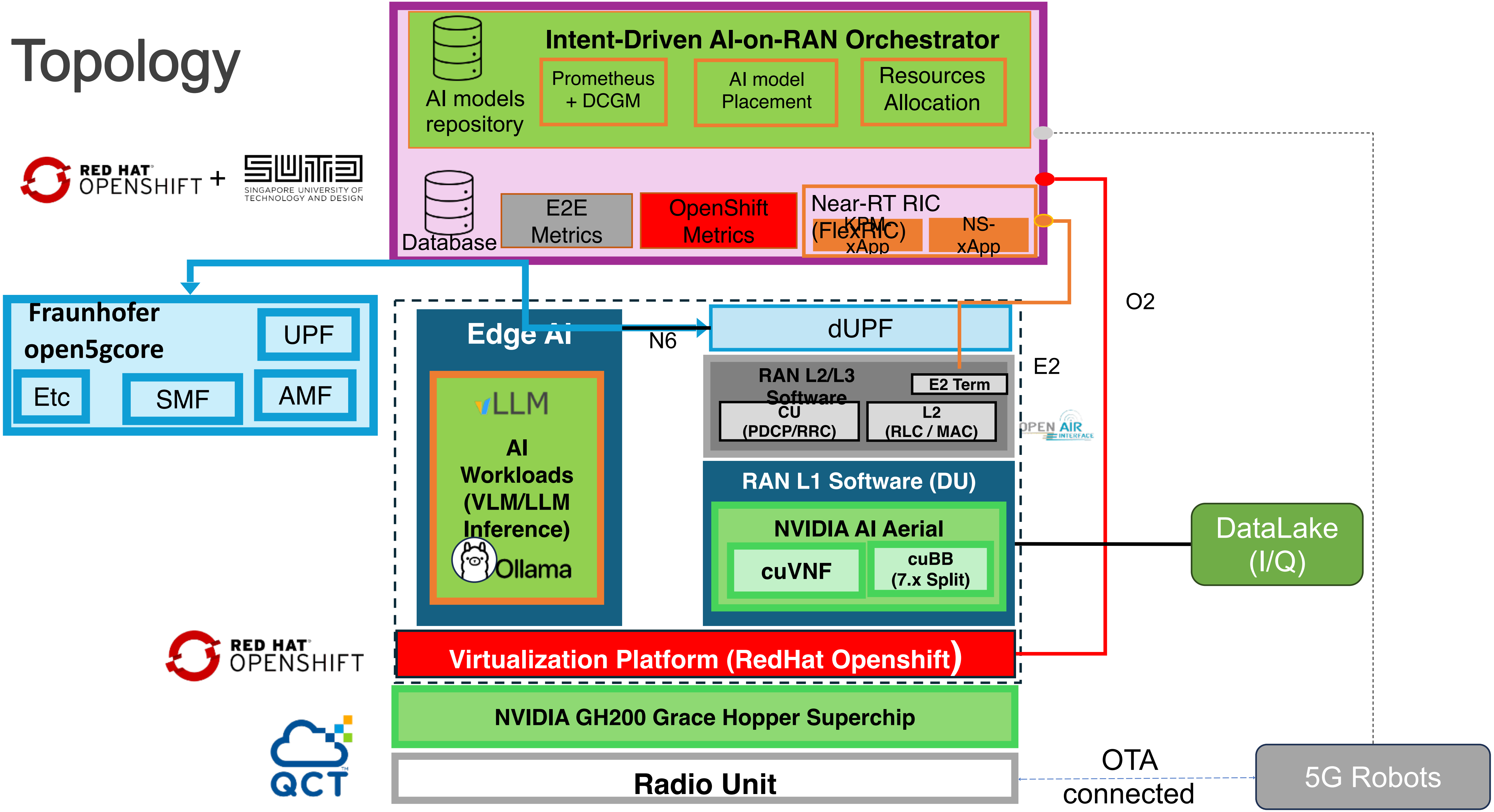
Opportunity 1. Communication-Efficient, Uncertainty-Aware Hybrid Language Model (CU-HLM)

Distributed speculative decoding (**HLM**) utilizes both SLM and LLM + **uncertainty-aware compression** & **opportunistic Tx**



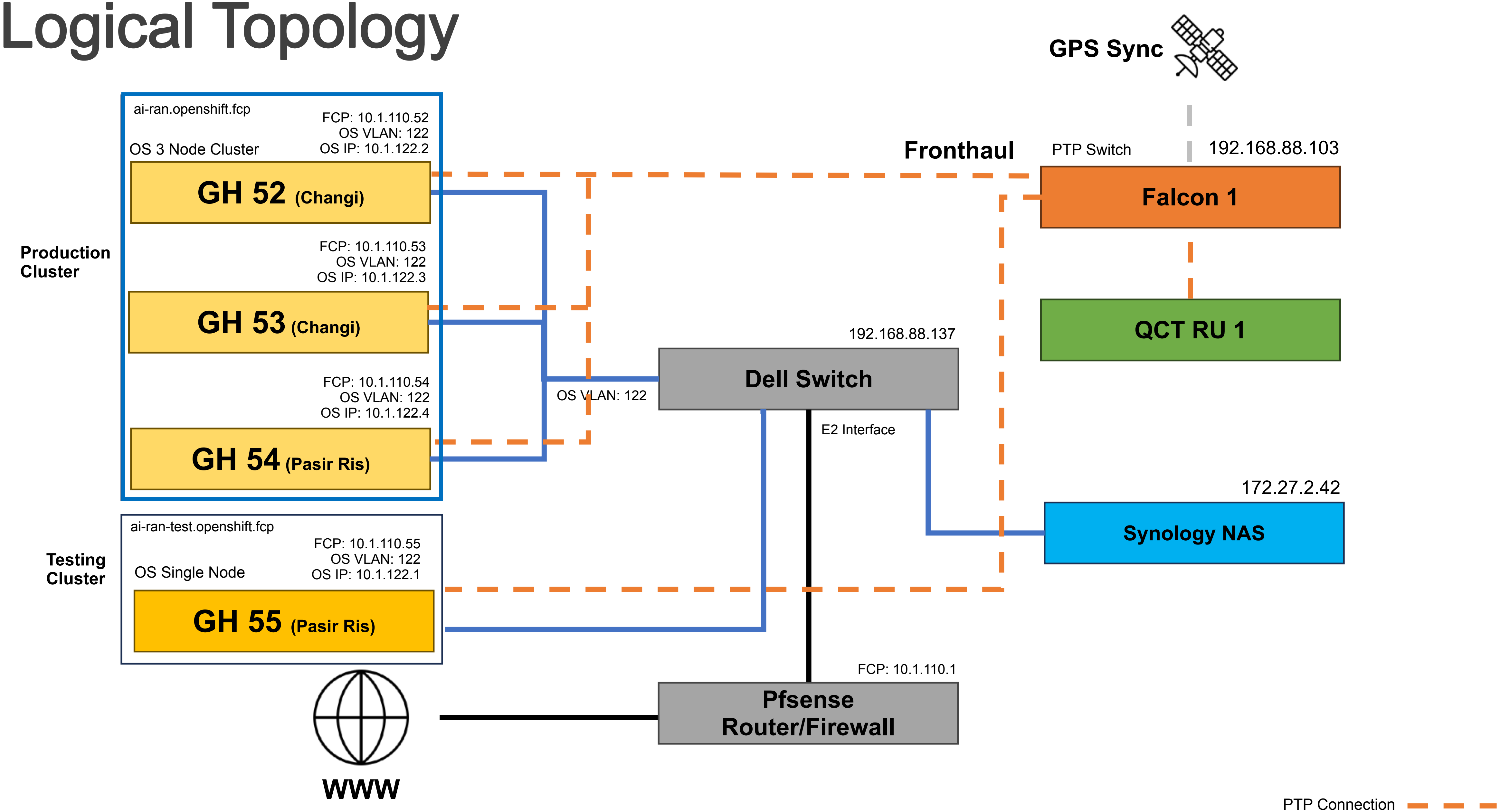
J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

Topology



J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

Logical Topology



J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

Multi-Instance GPU (MIG) Availability

Node 1
Balanced workloads (gh200-3node-eq-balanced)

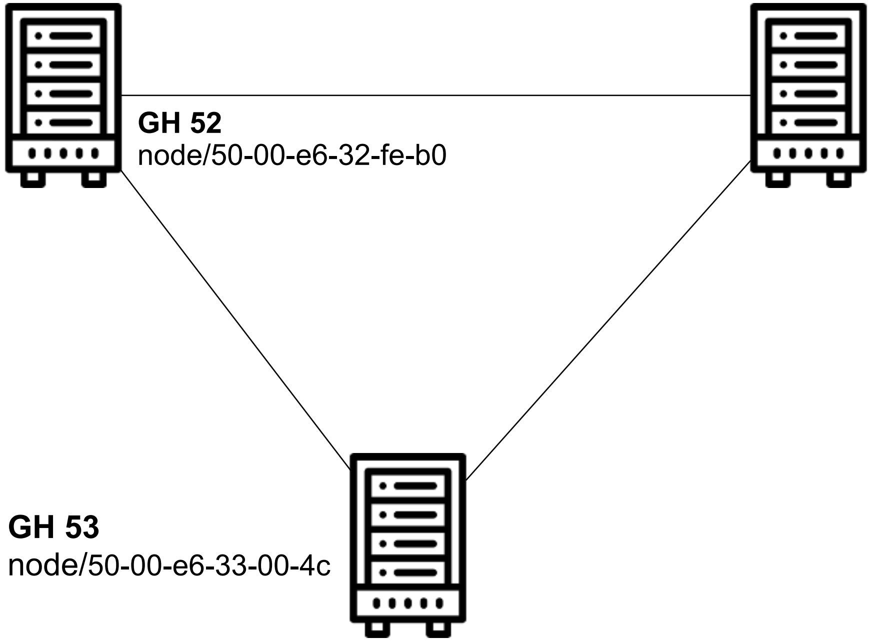
- 2 × 1g.12GB (small models)
- 1 × 2g.24GB (medium models)
- 1 × 3g.48GB (critical app eligible/ medium models)

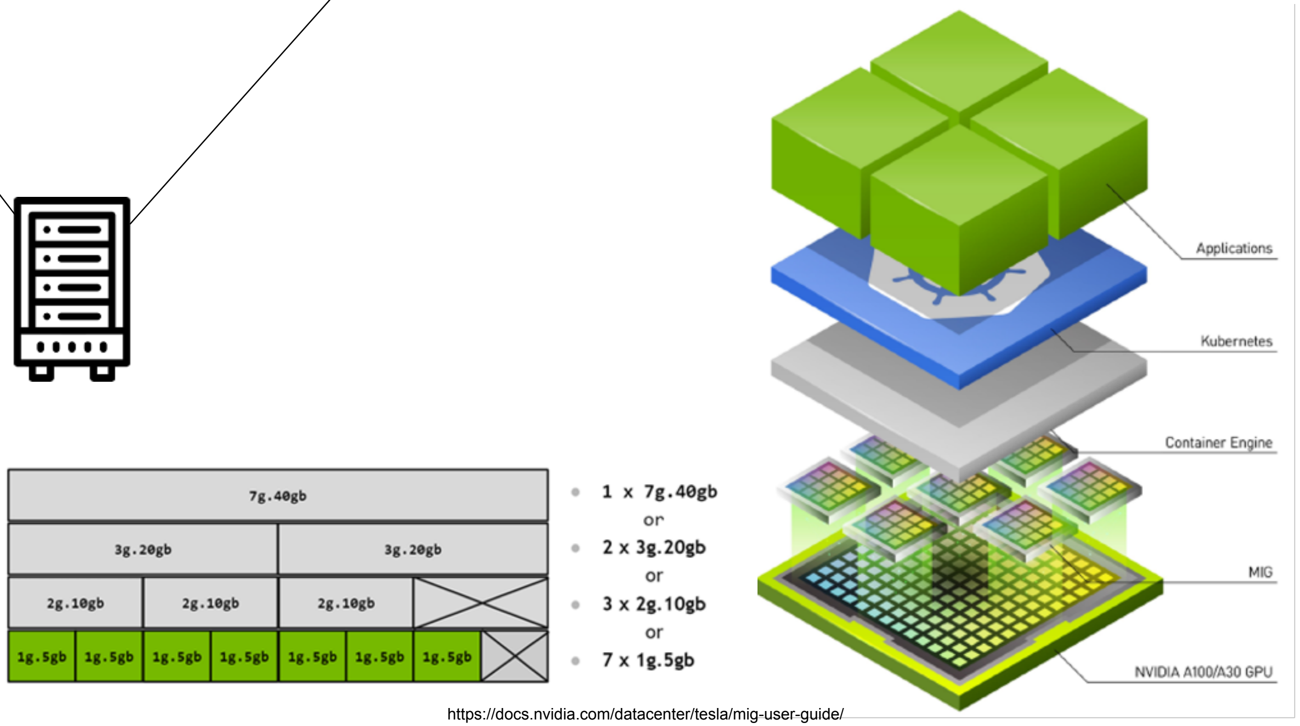
Node 2
Balanced workloads 2 (gh200-3node-eq-balanced)

- 2 × 1g.12GB (small models)
- 1 × 2g.24GB (medium models)
- 1 × 3g.48GB (critical app eligible/ medium models)

Node 3
Large workloads (all-disabled)

- Full GPU (critical app eligible/ large models)





https://docs.nvidia.com/datacenter/tesla/mig-user-guide/

Local/Edge AI Model Test (without Robots)

Inference Accuracy (Multimodal Benchmarks)

Benchmark	qwen2.5vl:3b (Jetson)	qwen2.5vl:7b (Jetson)	llama3.2-vision-11b (Jetson)	Qwen2.5-VL-7B-Instruct (GH200)
DocVQA	93.0%	94.5%	92.0%	95.7%
ChartQA	83.0%	85.0%	82.0%	87.3%
TextVQA	81.0%	83.0%	80.0%	84.9%
MathVista	63.0%	65.0%	60.0%	70.5%

On-Device AI

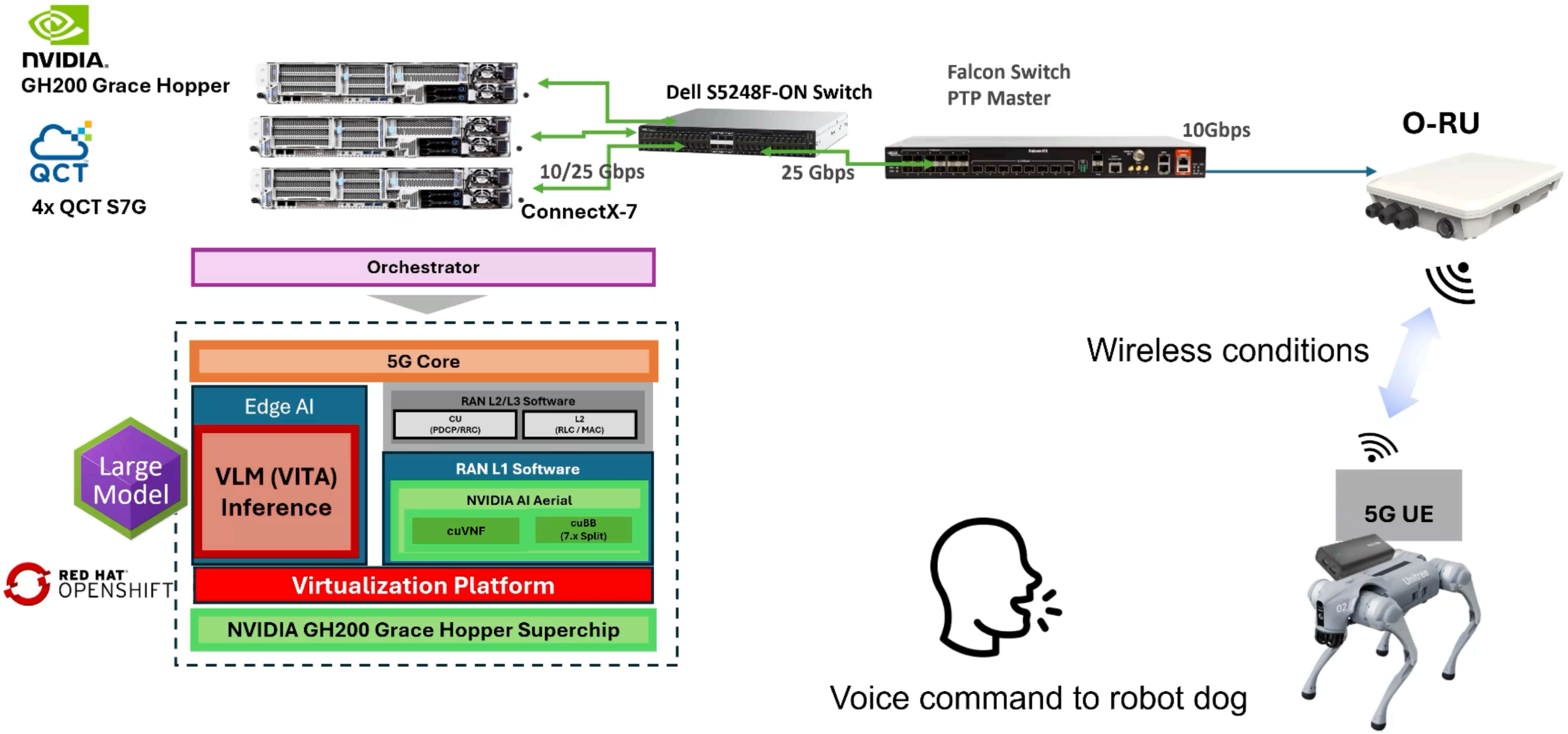
Edge AI

Inference Latency

Metric	qwen2.5vl:3b (Jetson Orin)	qwen2.5vl:7b (Jetson Orin)	llama3.2-vision-11b (Jetson Orin)	Qwen2.5-VL-7B-Instruct (GH200)
TTFT (avg, s)	0.340	0.440	0.520	0.024
TPT (avg, s)	0.156	0.272	0.298	0.0067
TPS (tokens/s)	6	4	3	1,017

J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

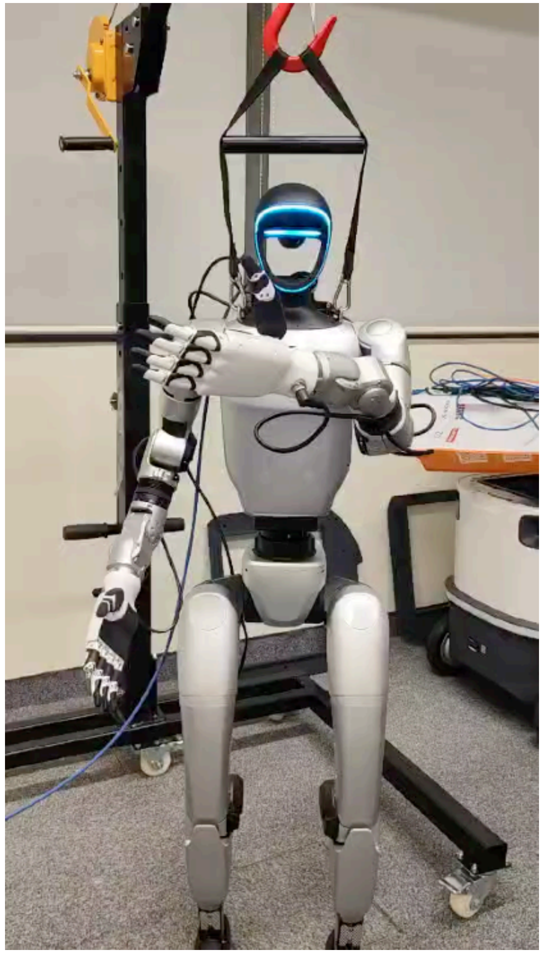
Robot Test 1: Voice-Controlled Intelligent Robotic Dog

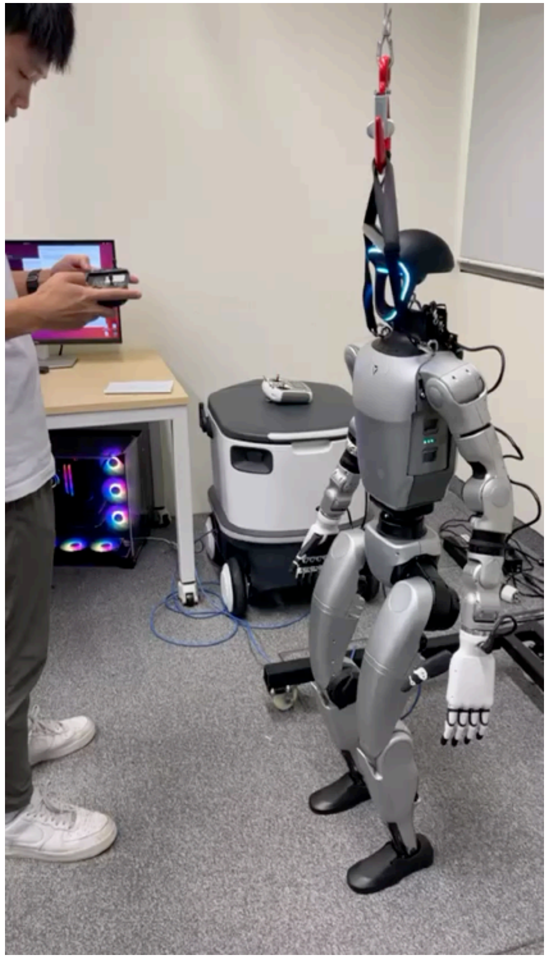


STT/TTS on Jetson Orin, Qwen2.5 7B LLM on GH200, connected over 5G-NR (FR1)

J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

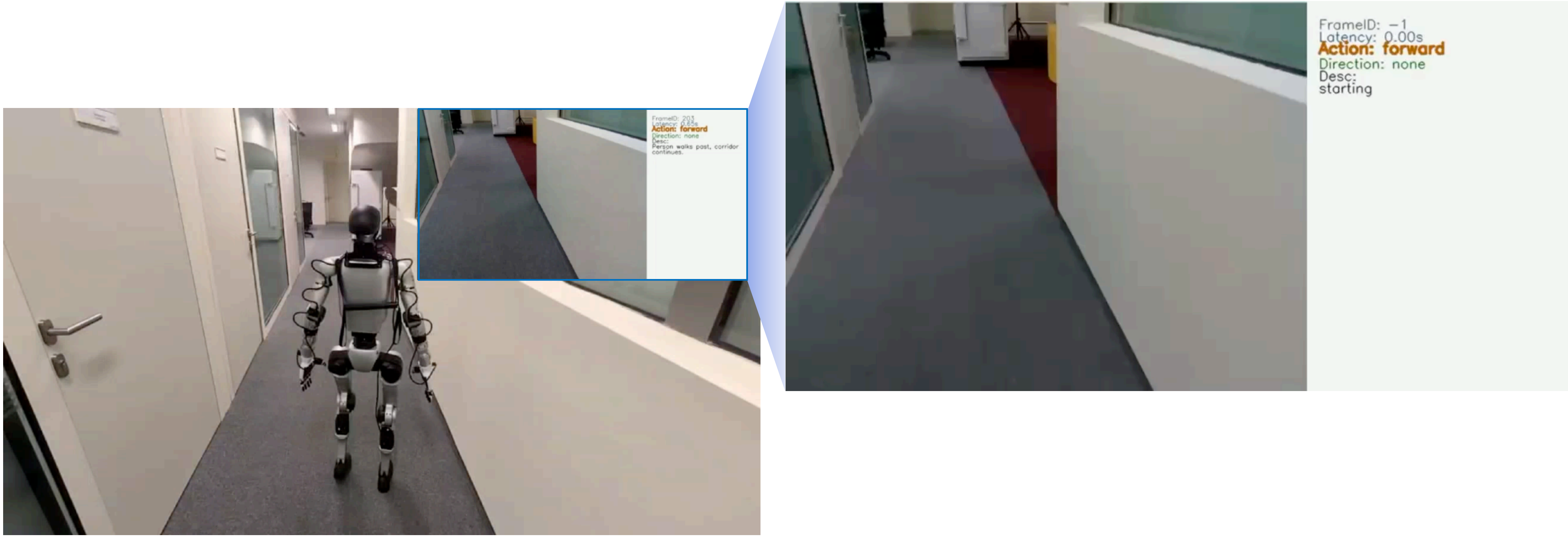
Robot Test 2: Voice-Controlled Humanoid





J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

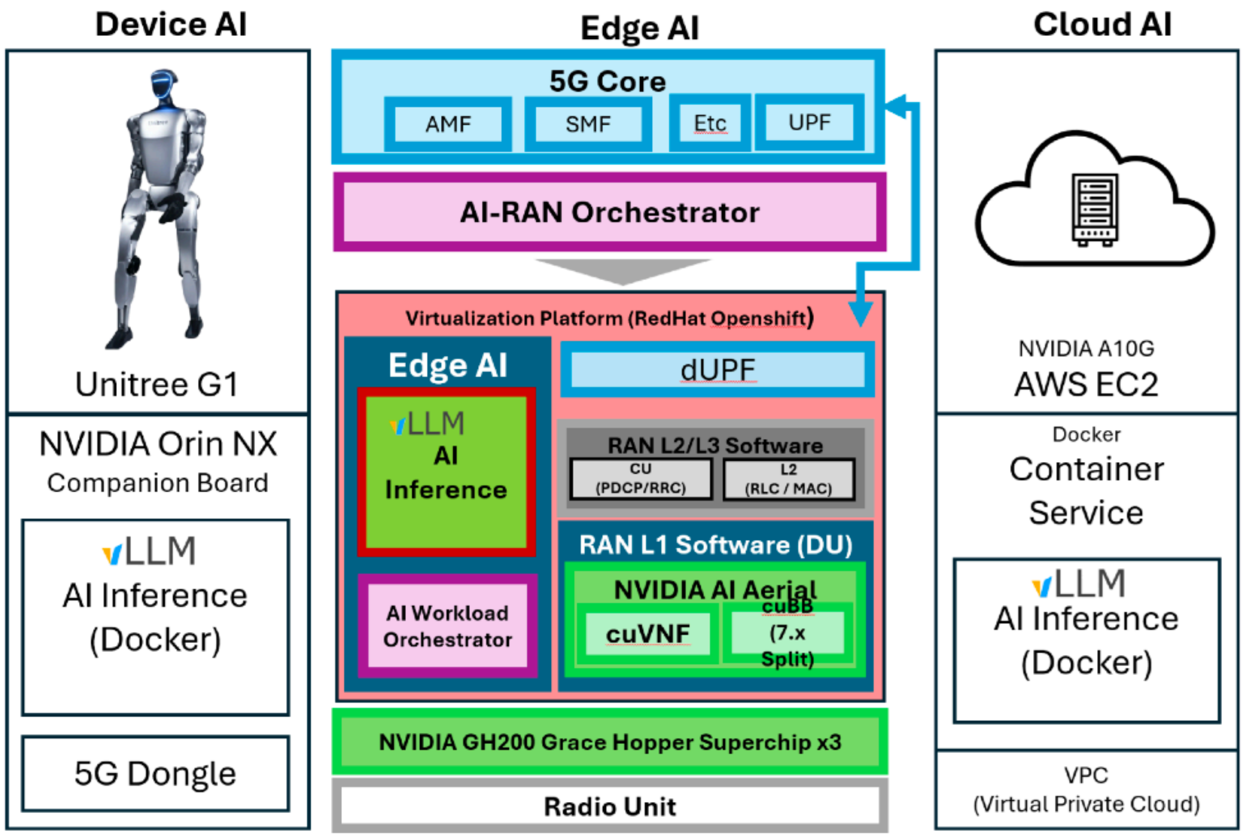
Robot Test 3: Vision-Controlled Humanoid



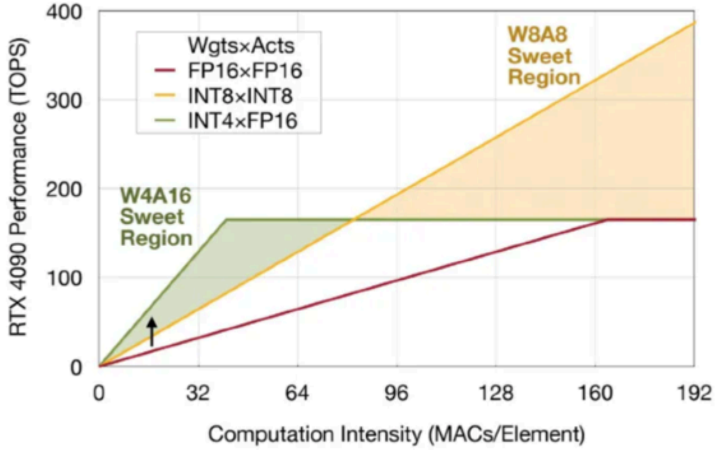
GPU	VLM Model	Connectivity	Throughput (tokens/s)	TTFT (ms)	Compute Delay (ms)	Comm Delay (ms)	E2E Delay (ms)
GH200	Gemma3-4B	Ethernet	125.10	599.497	599.497	6.800	606.297
GH200	Gemma3-4B	5G	120.54	597.295	597.295	11.664	608.959
Jetson Orin	Moondream2-1.8B	Local	42.13	1.741	4,000.005	0.000	4,000.005

J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

Robot Test 3: Vision-Controlled Humanoid



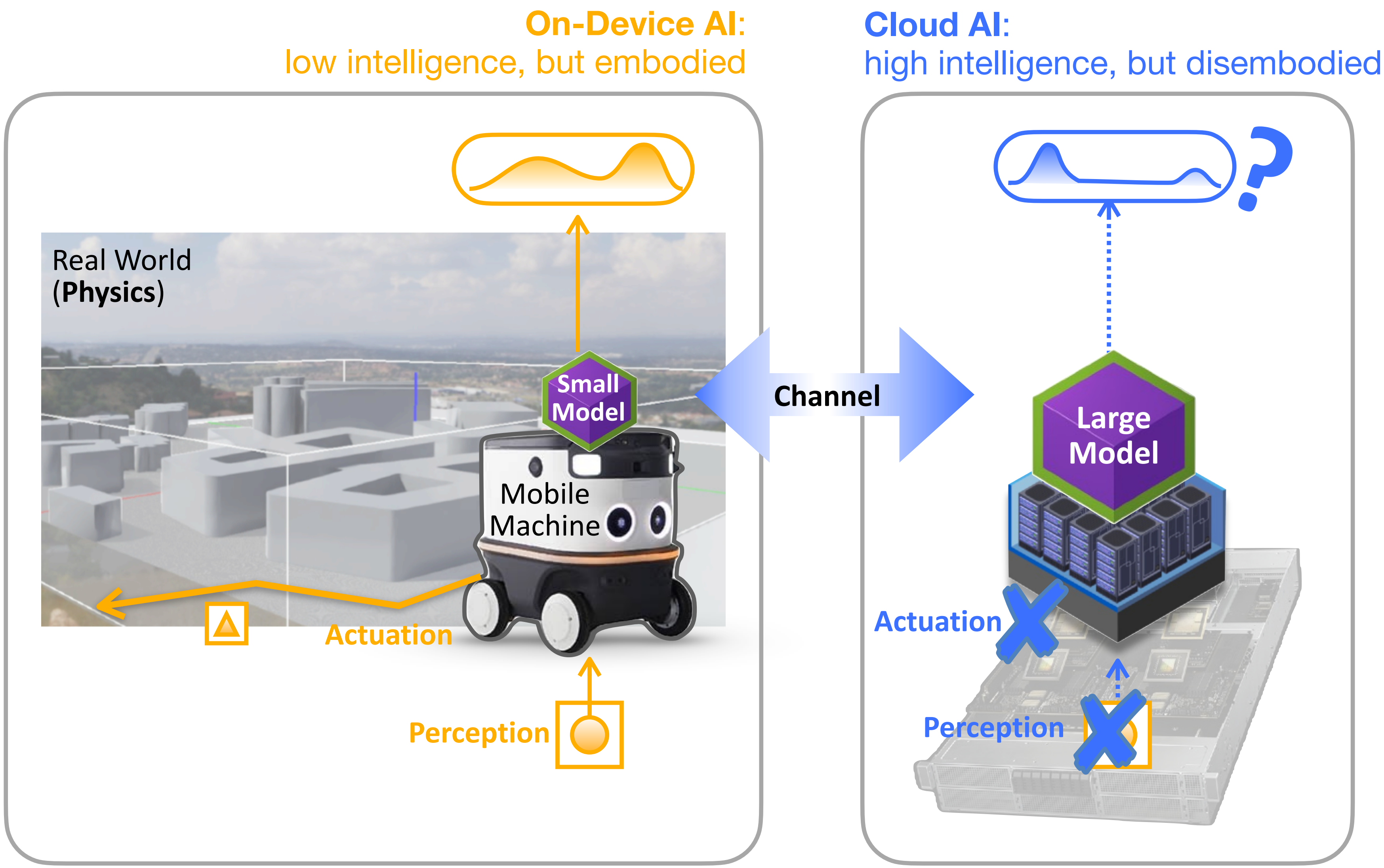
Variant	Platform	E2E (ms)	σ_{RTT} (ms)	TTFT (ms)
3B-FP16	Local	4710	0.0	4702
	Edge	551	1.1	537
	Cloud	1232	1.3	1218
3B-AWQ	Local	5230	0.0	5217
	Edge	407	1.1	393
	Cloud	713	1.2	700
3B-W4A16	Local	5390	0.0	5373
	Edge	515	1.2	501
	Cloud	733	1.2	720
3B-W8A8	Edge	419	1.2	405
	Cloud	1023	1.1	1010



J. Park et al., "AI-RAN Orchestration for Heterogeneous and Hybrid LLMs," *AI-RAN Alliance WG3 WI*

Challenge 2. **Disembodied AI** — Lack of Perception and Actuation in the Real World

Q. Can embodied on-device AI communicate tokens with disembodied cloud AI?



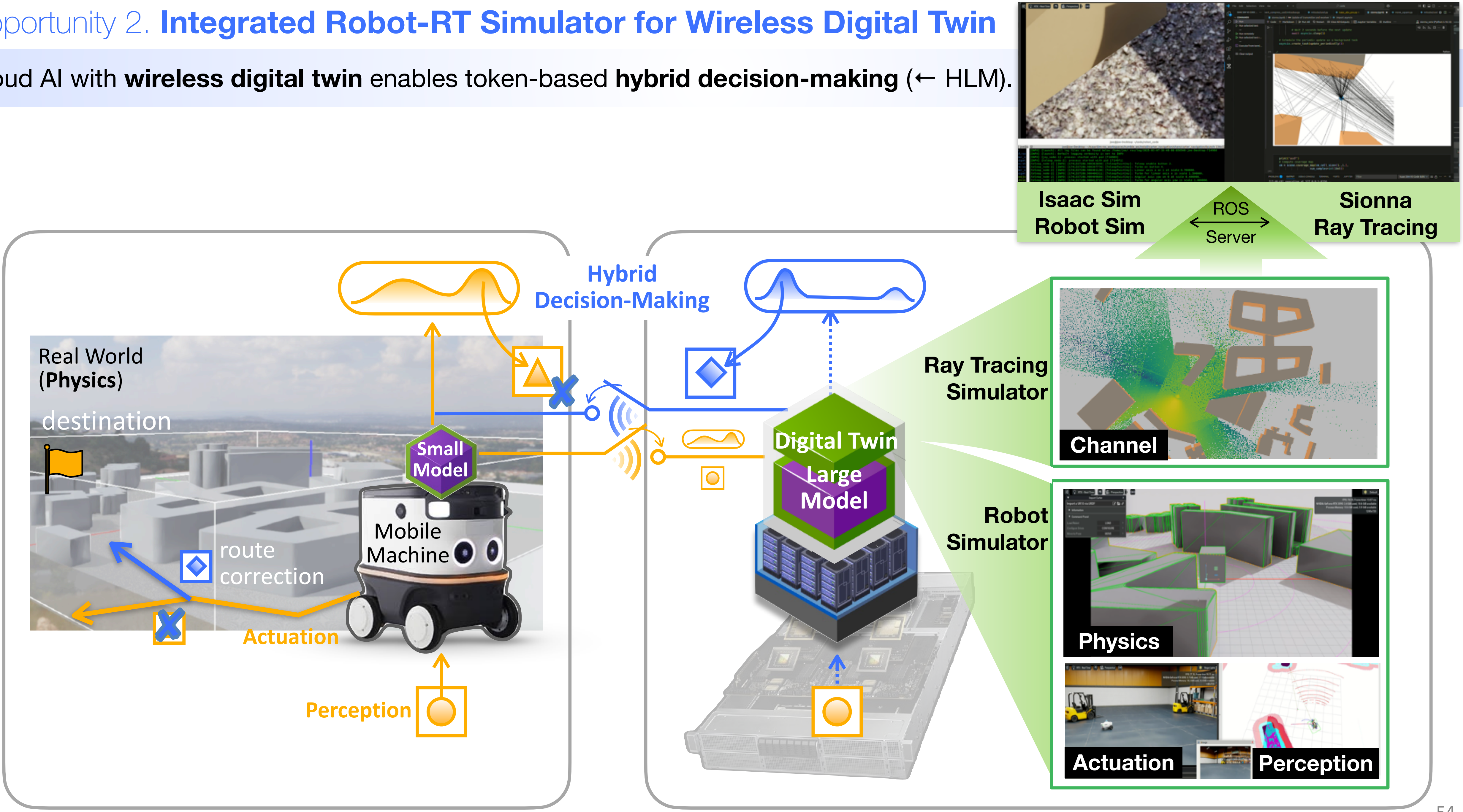
Enabling communication with embodied Cloud AI requires:

- **Perception**
- **Actuation**
- **Physics**
- **Channel**

Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, "Robot Communication Digital Twin for Wireless Hybrid Inference," *in progress*

Opportunity 2. **Integrated Robot-RT Simulator for Wireless Digital Twin**

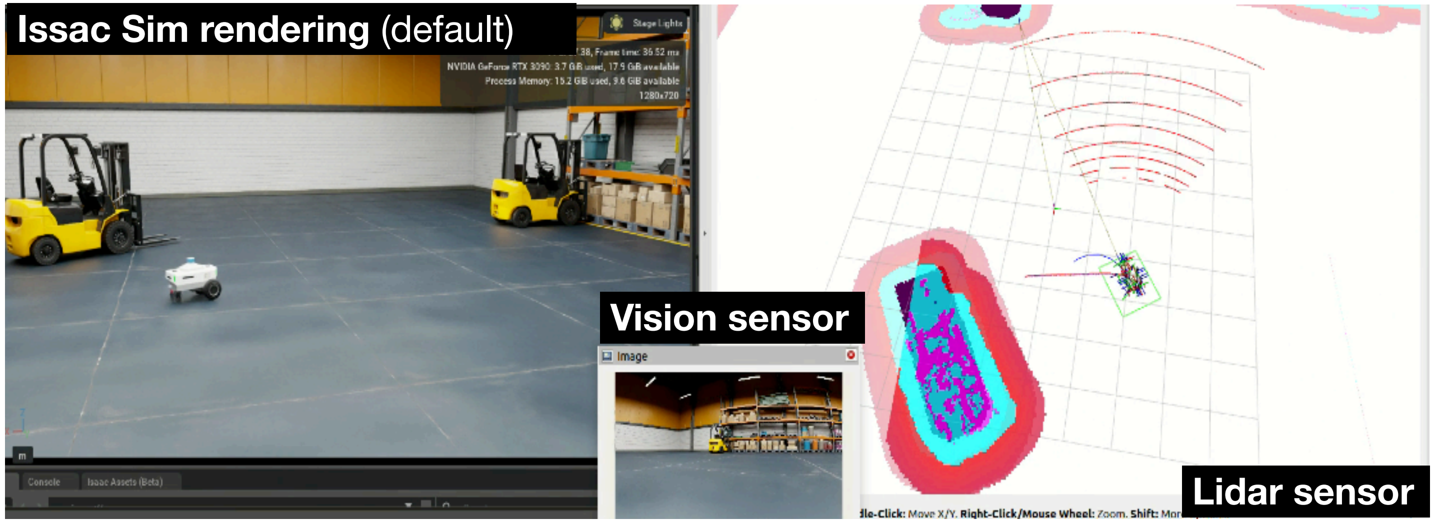
Cloud AI with **wireless digital twin** enables token-based **hybrid decision-making** (← HLM).

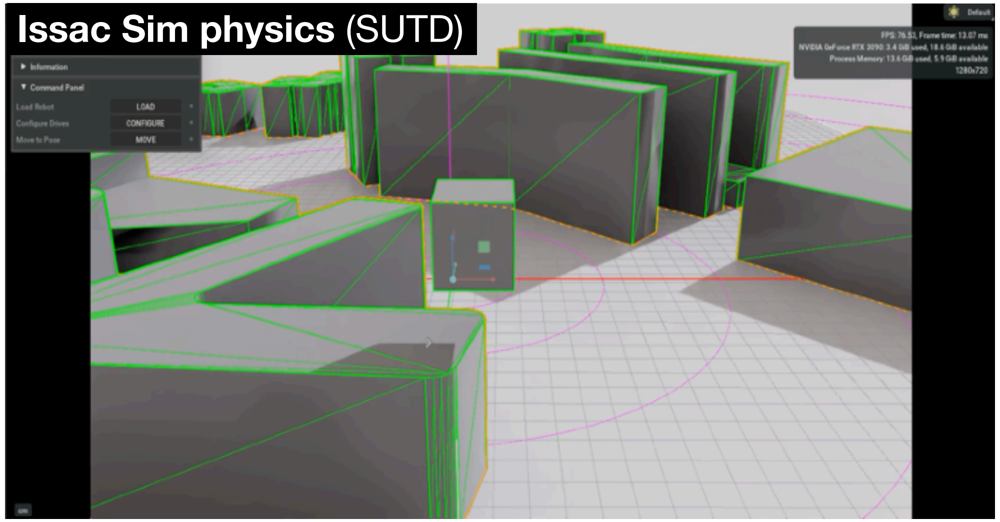


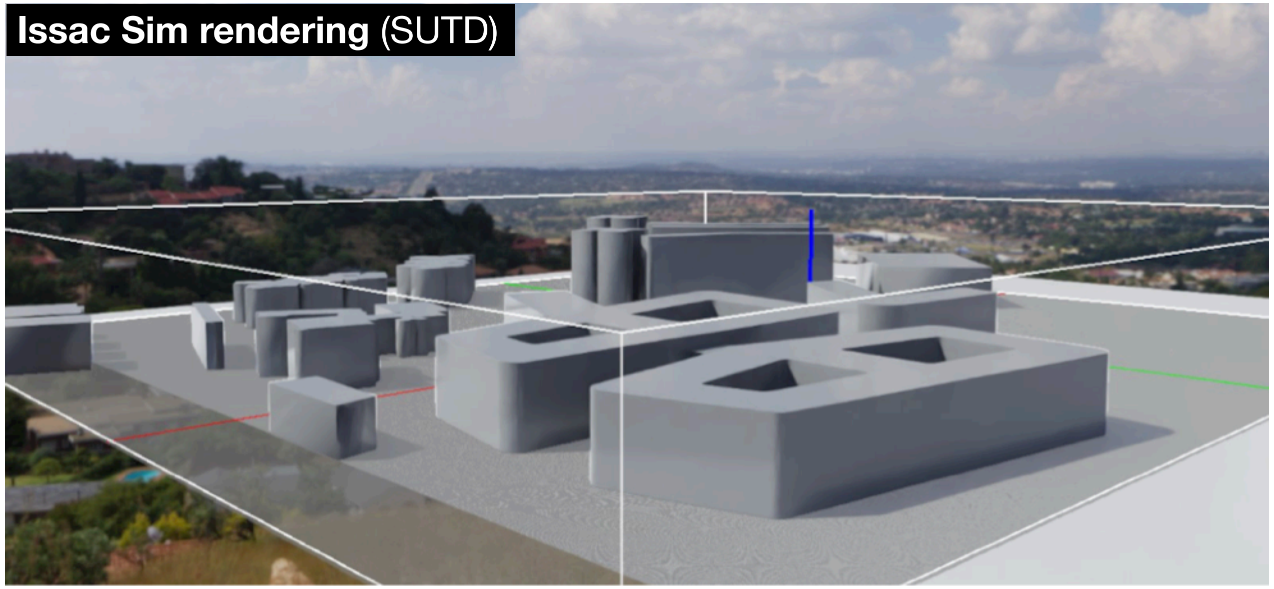
Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, "Robot Communication Digital Twin for Wireless Hybrid Inference," *in progress*

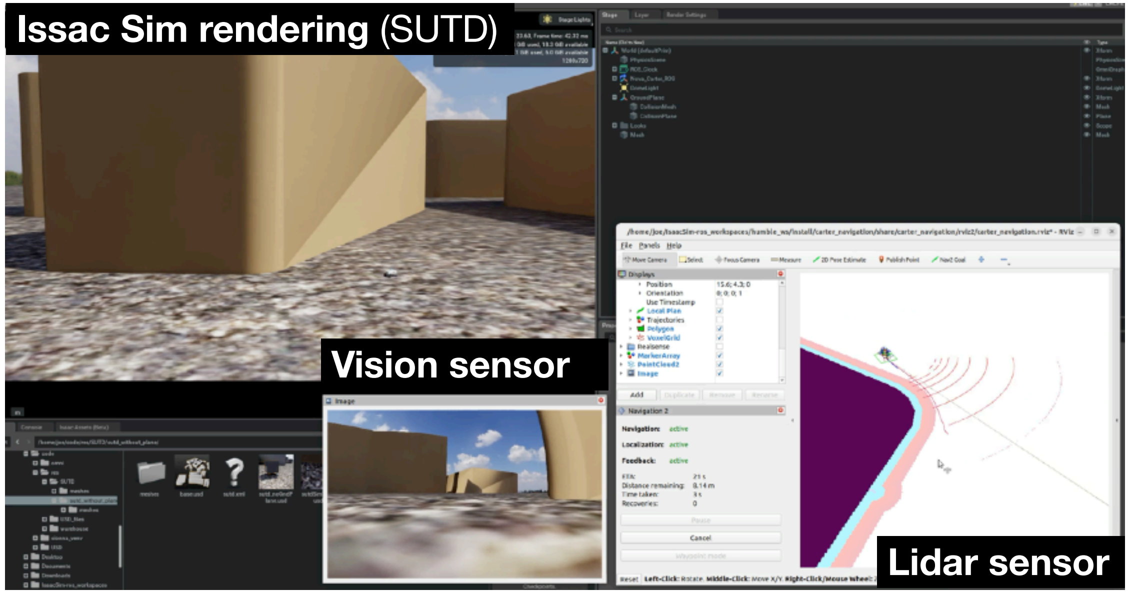
Opportunity 2. **Integrated Robot-RT Simulator for Wireless Digital Twin**

(Robot) **NVIDIA Isaac Sim**: GPU-accelerated robot simulator with real-time rendering and high fidelity physics modeling





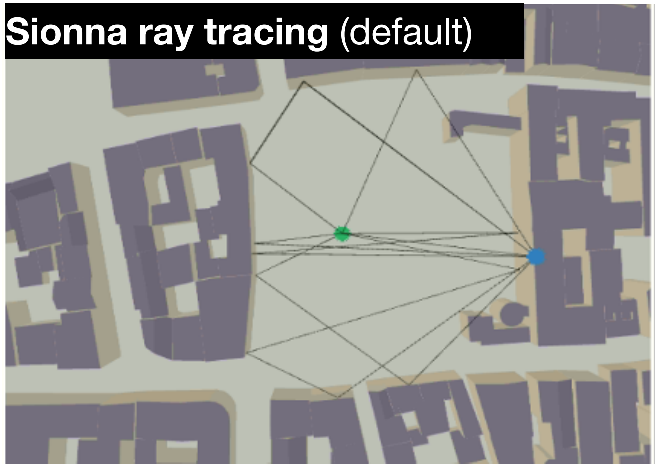


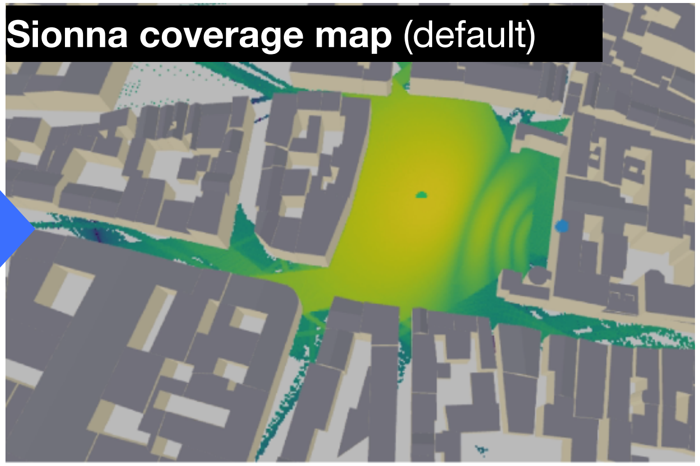


Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, "Robot Communication Digital Twin for Wireless Hybrid Inference," *in progress*

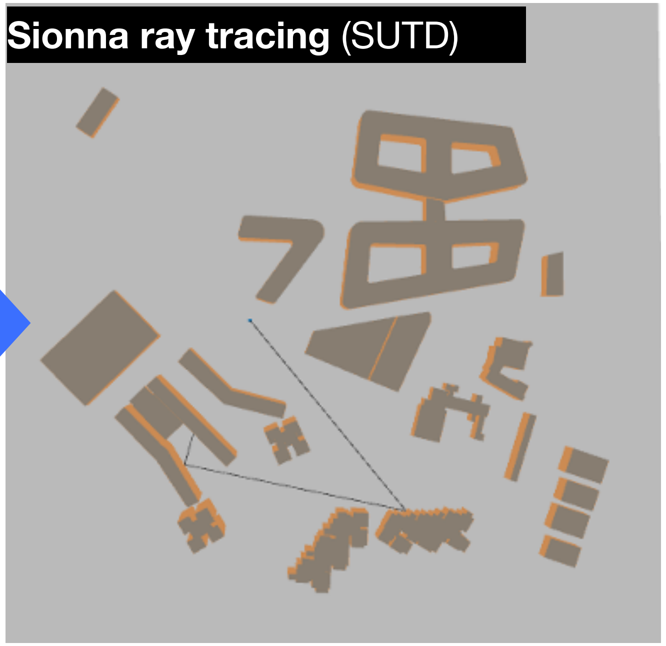
Opportunity 2. **Integrated Robot-RT Simulator for Wireless Digital Twin**

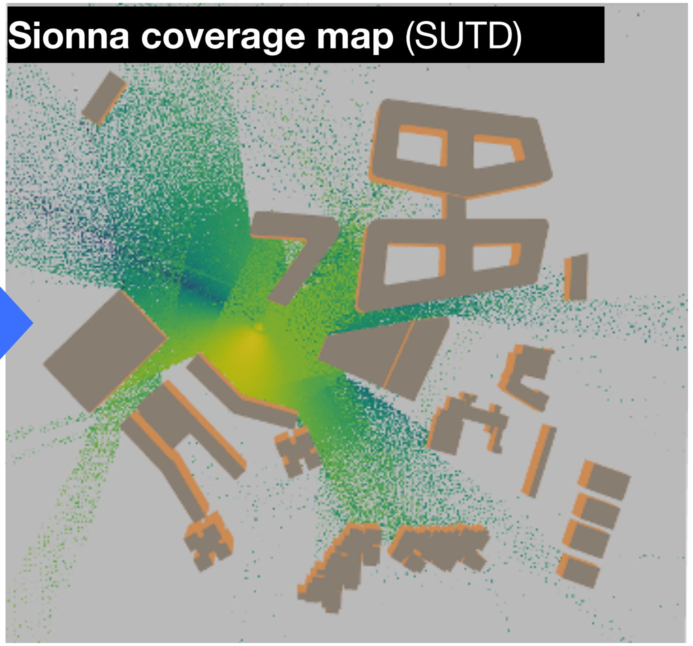
(Communication) **NVIDIA Sionna RT**: GPU-accelerated electromagnetic (EM) ray tracing simulator





SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

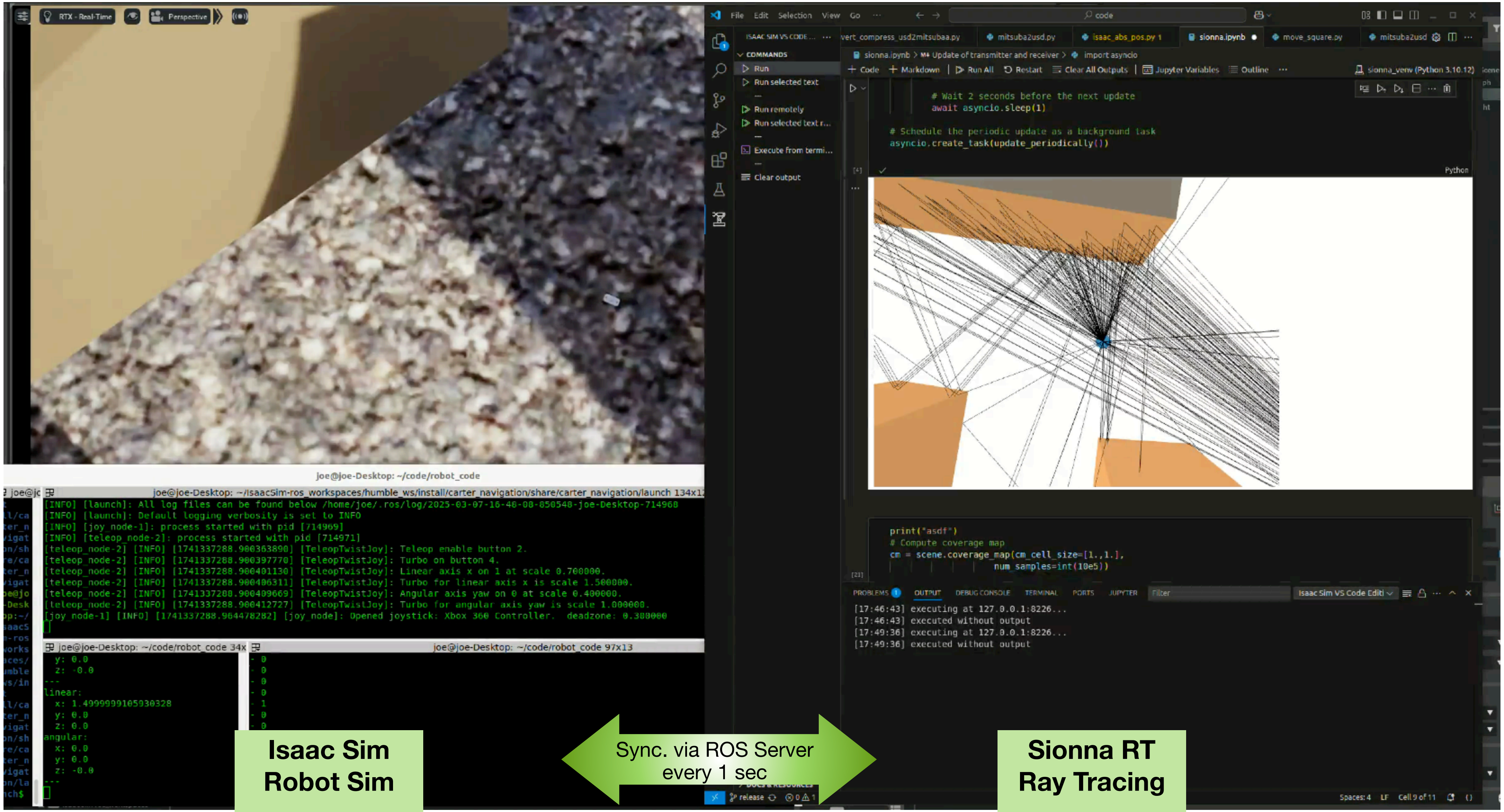




Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, “Robot Communication Digital Twin for Wireless Hybrid Inference,” *in progress*

Opportunity 2. **Integrated Robot-RT Simulator for Wireless Digital Twin**

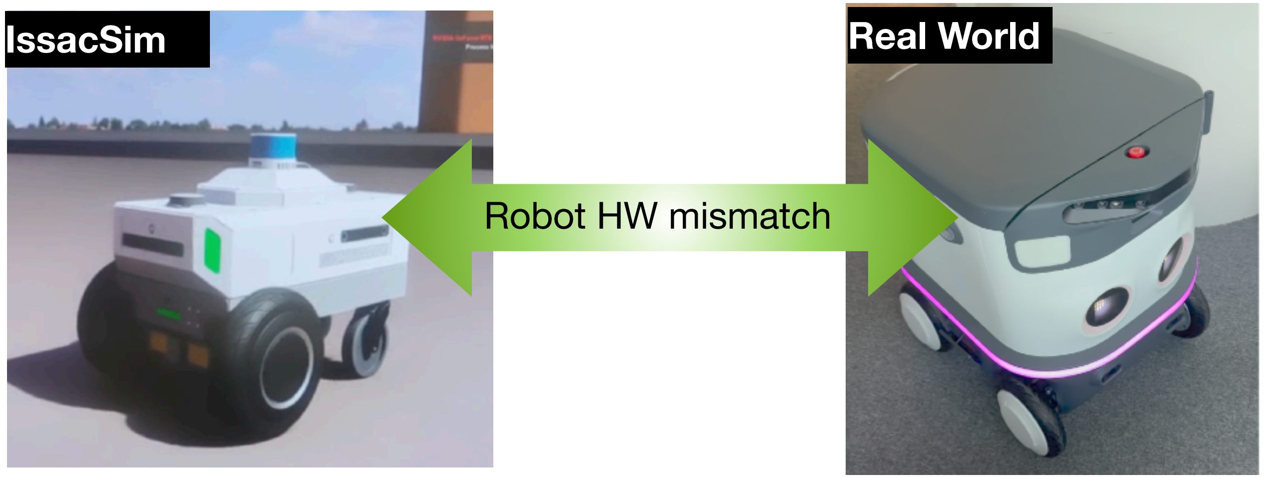
Integrated Issac Sim + Sionna RT via ROS server (in progress)

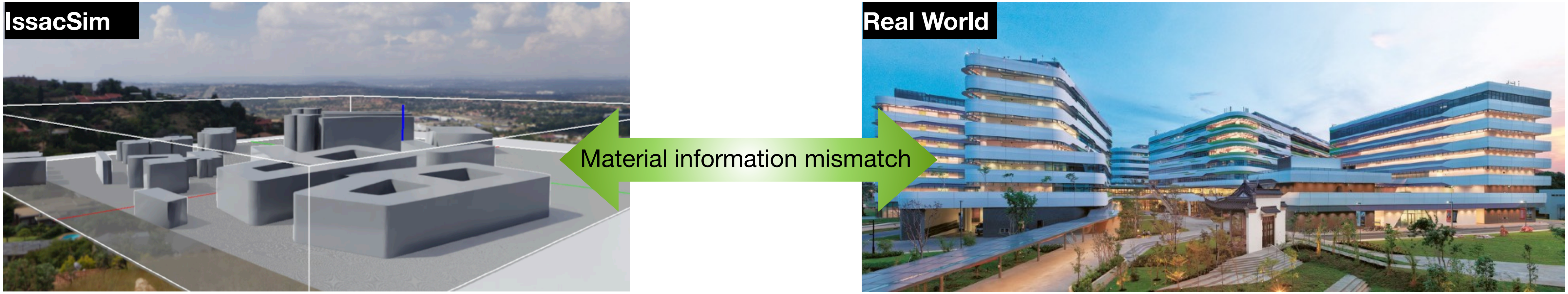


Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, "Robot Communication Digital Twin for Wireless Hybrid Inference," *in progress*

Challenge 3. **Wireless Digital Twin Construction** — Lack of High-Fidelity Environmental Information

Q. How can we synchronize digital-physical worlds, despite limited access to high-fidelity environmental information?



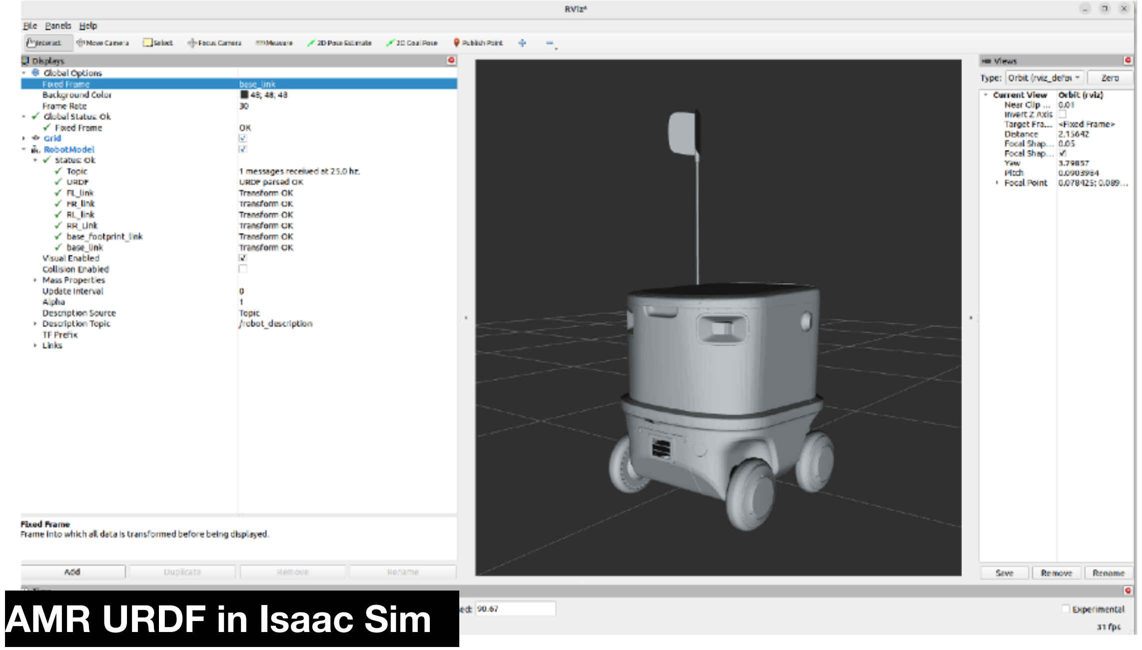


Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, "Robot Communication Digital Twin for Wireless Hybrid Inference," *in progress*

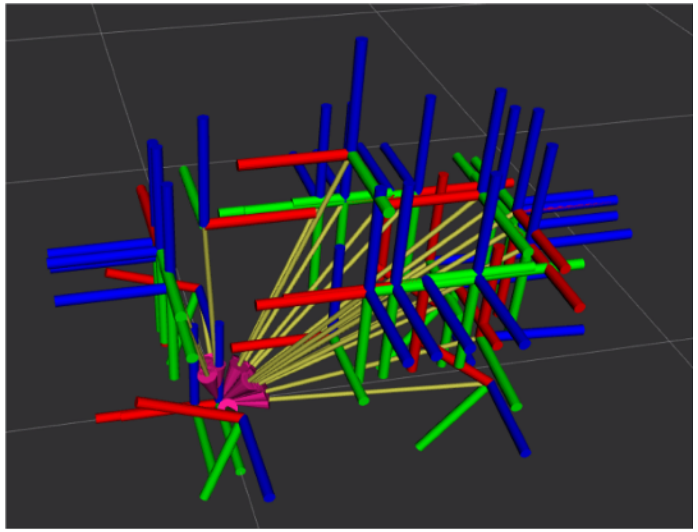
Opportunity 3. **Integrated Robot-RT Simulator for Wireless Digital Twin**

Differentiable RT enables to learn material parameters via GD, with Isaac Sim robot asset (**URDF**)

Autonomous mobile robot (AMR) in real world



AMR URDF in Isaac Sim



Y. Lim, Z. Kang, Z. Guo, J. Park, and T. Q.S. Quek, "Robot Communication Digital Twin for Wireless Hybrid Inference," *in progress*

Opportunity 3. **Differentiable RT with Warm-Start Initialization**

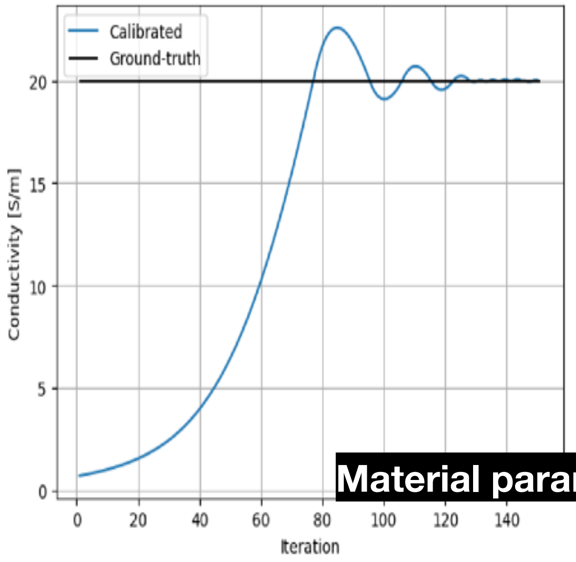
Differentiable RT enables to learn material parameters via GD, with initial parameters obtained using vision-language model (**VLM**)

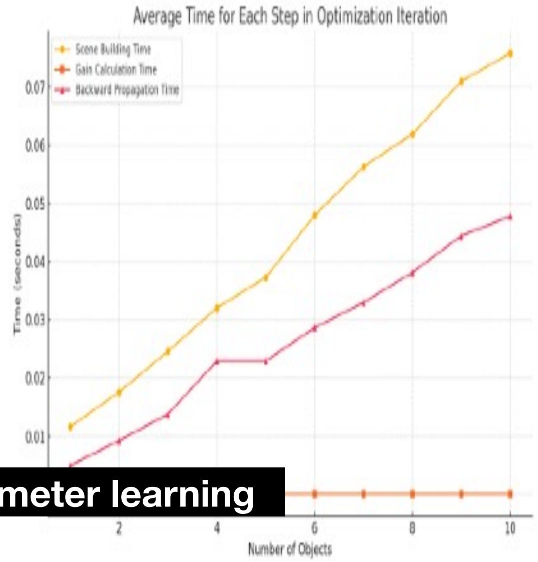
Rec. ITU-R P.2040-3

TABLE 3

Material properties

Material class	Real part of relative permittivity		Conductivity S/m		Frequency range
	a	*b*	*c*	*d*	GHz
Vacuum (≈ air)	1	0	0	0	0.001-100
Concrete	5.24	0	0.0462	0.7822	1-100
Brick	3.91	0	0.0238	0.16	1-40
Plasterboard	2.73	0	0.0085	0.9395	1-100
Wood	1.99	0	0.0047	1.0718	0.001-100
Glass	6.31	0	0.0036	1.3394	0.1-100





Material parameter learning

Issac Sim

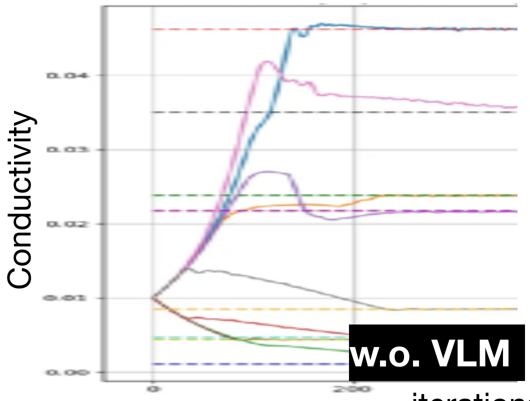
Visual scene from Issac Sim

VLM

yaml
Material: Wood-Plastic Flooring
Frequency: 3 GHz
Relative Permittivity (εr): 3.66
Conductivity (σ): 0.0096 S/m
Validity: Below recommended frequency range (50-100 GHz)
Confidence: Medium

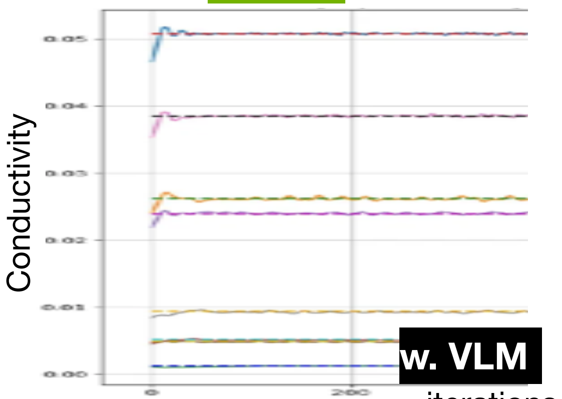
VLM-derived material parameters

Sionna RT



w.o. VLM

iterations



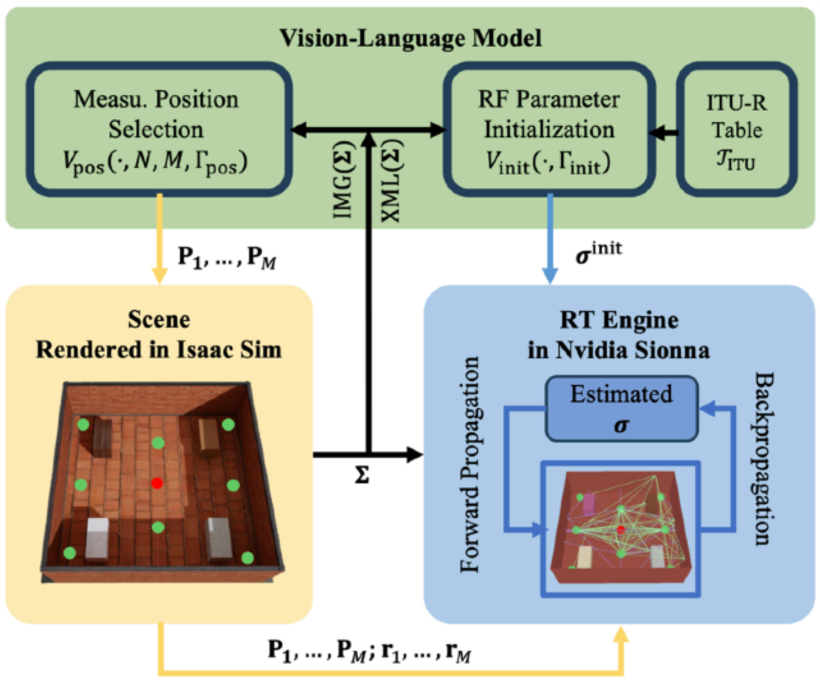
w. VLM

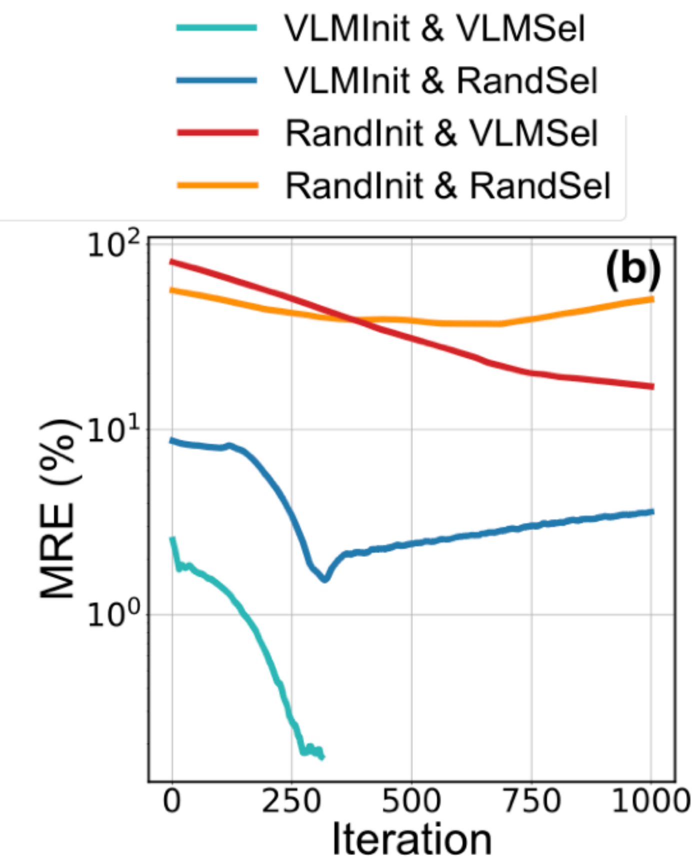
iterations

Z. Kang, Y. Lim, Z. Gu, S..-W. Ko, Tony Q.S. Quek, and J. Park, "Accelerated Multi-Material Parameter Active Learning via Vision-Language Model Aided Differentiable Ray Tracing," AAAI'26 Wksp. ML4Wireless.

Opportunity 3. **Differentiable RT with Warm-Start Initialization**

Differentiable RT enables to learn material parameters via GD, with initial parameters obtained using vision-language model (**VLM**)





Detailed Prompt 2: VLM-Aided Measurement Position Selection

Inputs:
1. Scene Image: An overhead view or a key 3D perspective of the environment.
2. Geometry XML: The structured XML description of the scene.
3. Measurement Counts: N (Number of Receivers) = N_VALUE], M (Number of Tx-Rx measurement configurations) = [M_VALUE].

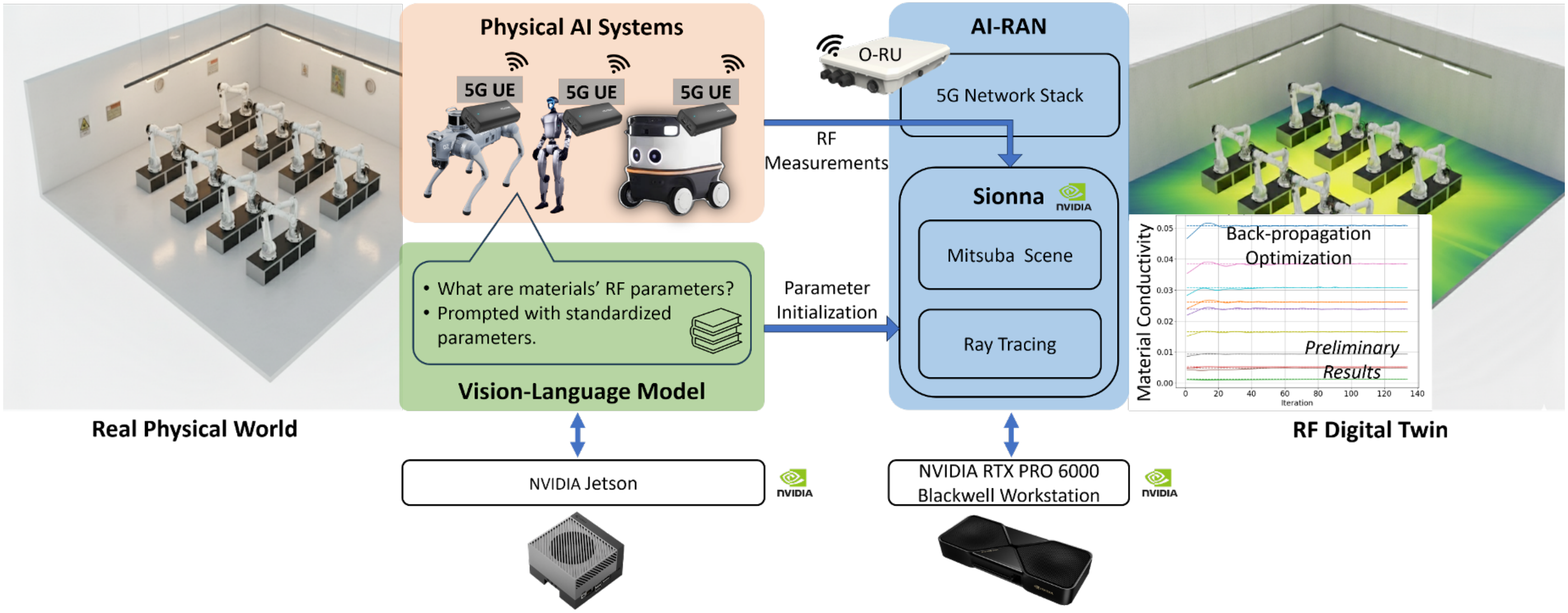
Prompt: You are an intelligent spatial planner optimizing RF measurement campaigns. Your goal is to select Transmitter (Tx) and Receiver (Rx) positions that maximize the diversity and information gain regarding the conductivity parameters of the materials in the scene.

Task:
1. Analyze the XML and Image: Identify all distinct material interfaces and complex geometrical features.
2. Strategy: Select N distinct Tx/Rx positions and form M pairs. Prioritize positions that force the RF path to interact with the different types of materials. For instance, select pairs to sample Non-Line-of-Sight (NLoS) paths or Reflection-dominant paths.
3. Provide the selected positions as a JSON list. The coordinates must be valid and directly correspond to the XML's coordinate system.

Output Schema: ["id": "P_1", "type": "Tx", "x": 2.5, "y": 1.0, "z": 1.5, "reasoning": "Placed to maximize transmission through the central concrete wall to the back area (NLoS).", "id": "P_2", "type": "Rx", "x": -3.0, "y": 5.0, "z": 1.5, "reasoning": "Paired with P_1 to measure loss through multiple materials (wood table and brick wall).",
// ... continue until R positions are defined and S pairs are formed implicitly by the list]

Opportunity 3. **Differentiable RT with Warm-Start Initialization**

Differentiable RT enables to learn material parameters via GD, with initial parameters obtained using vision-language model (**VLM**)



Z. Kang, Y. Lim, Z. Gu, S..-W. Ko, Tony Q.S. Quek, and J. Park, "Accelerated Multi-Material Parameter Active Learning via Vision-Language Model Aided Differentiable Ray Tracing," AAAI'26 Wksp. ML4Wireless.

Future Challenges. **Coarse Tokens, Agentic AI**

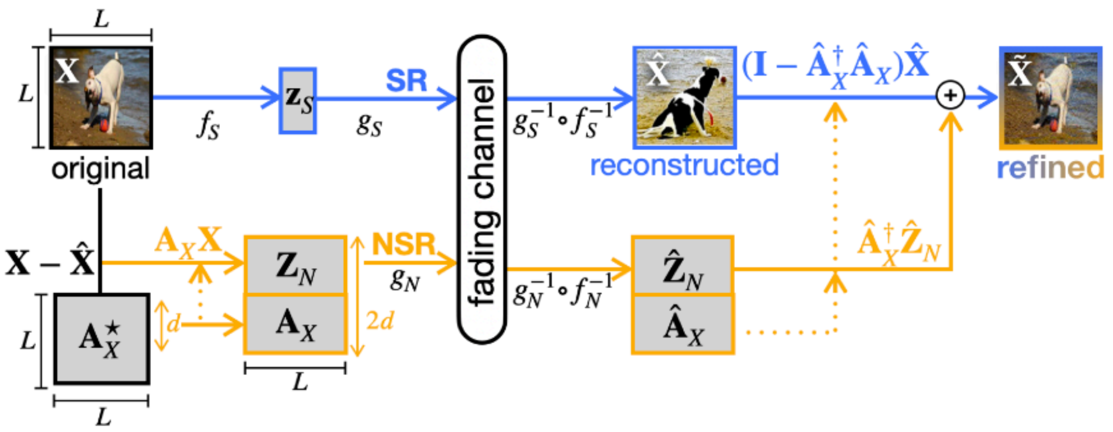
Source: Anthropic, Google

- **Tokens are more coarse** than bits:
  - Language tokens: ~ 16 bits (50K vocabulary size)
  - Vision tokens ~ 0.75 KB (16 x 16 patches, 8 bits/channel)

Q. For high-fidelity applications, should we rely solely on classical communication, or adopt a hybrid approach?

- **Model Context Protocol (MCP)** and **Agent-to-Agent protocol (A2A)** have emerged to support AI-to-App and inter-AI communication.
- MCP resembles URLLC+eMBB, A2A aligns with uRLLC+mMTC slices

Q. Should we ***optimize concurrent slices, or create a new slice for token communication****?*



Feature	HTTP	MCP (Streamable HTTP)	A2A (Agent-to-Agent)
Duplex Type	Half-duplex	Full-duplex	Full-duplex
Transport	TCP	TCP/QUIC	HTTP / SSE / JSON-RPC
Persistent Conn.	No (by default)	Yes	Yes
HTTP Semantics	Native	Native	Native (built on HTTP + JSON-RPC)
Best For	Static/REST	LLMs, real-time APIs	Cross-platform agent collaboration
Setup Complexity	Low	Low–Moderate	Moderate

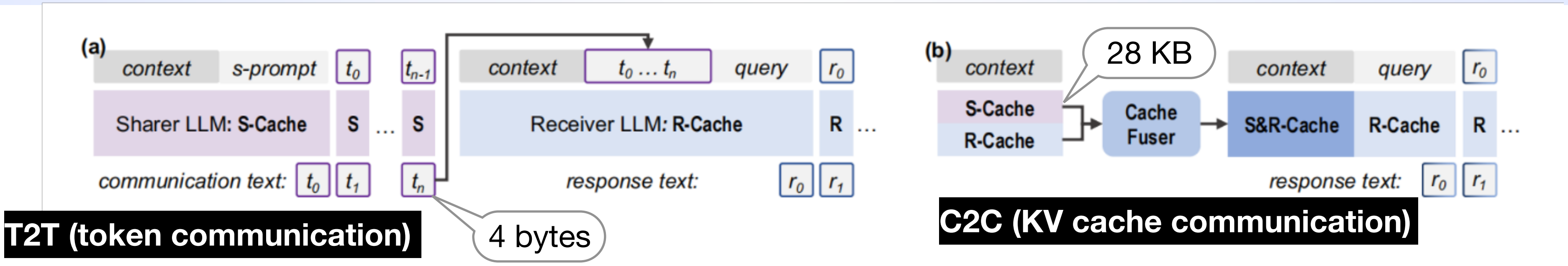
H. Nam, J. Park, J. Choi, S.-L. Kim, "Hybrid Semantic Communication Integrating Complementary Shannon Communication," *submitted to IEEE ICASSP'26*

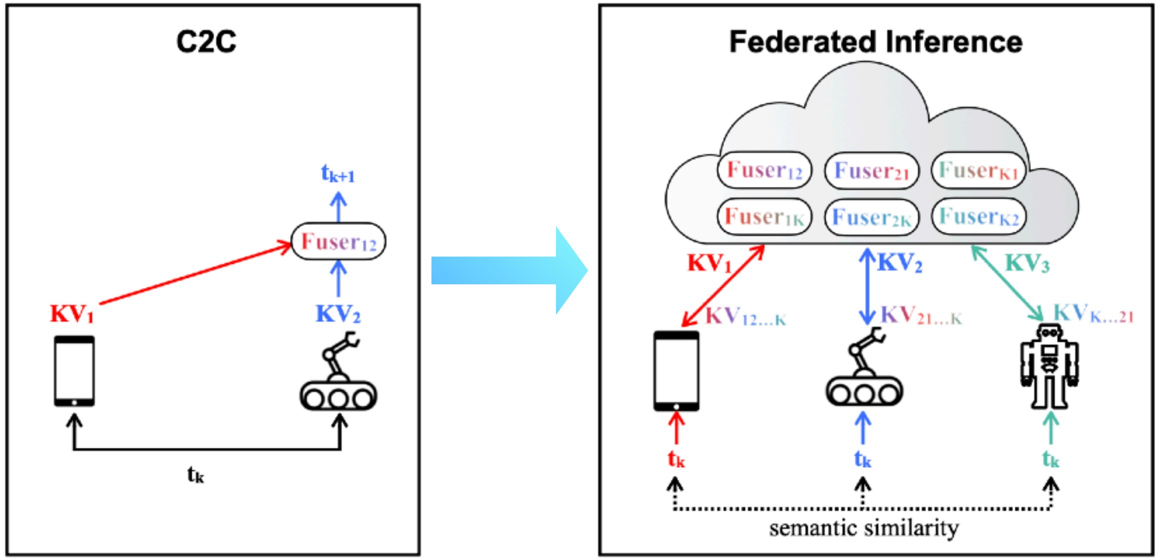
H. Nam, J. Park, J. Choi, and S.-L. Kim, "Hybrid Semantic Communication for Controllable Fidelity in Wireless Image Transmission," *submitted to IEEE TCCN*

Future Challenges. **Token vs. Cache Communication**

- Token communication: **$Computation ↑ ➨ $Communication ↓**

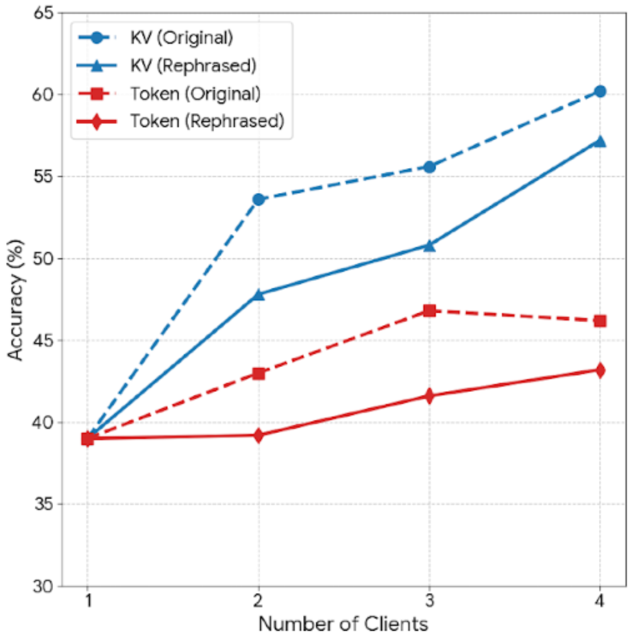
*Q. If computation dominates, can **$Communication** ↑ ➨ **$Computation** ↓ ? ➨ **KV cache communication***

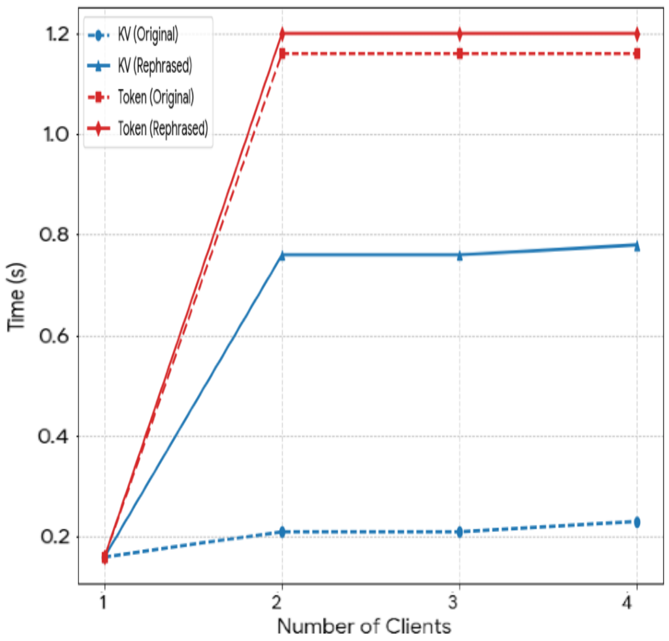




- Uni-directional
- Two clients
- Input token sharing

- **Multi-directional**
- **Multi-Clients**
- **Private input tokens** (w. similarity guarantee)





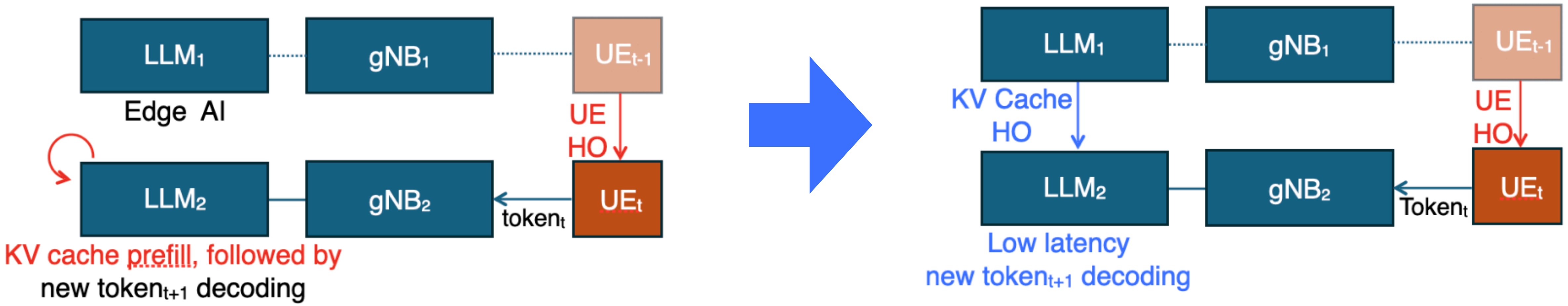
T. Fu et al., “Cache-to-Cache: Direct Semantic Communication Between Large Language Models,” arXiv:2510.315, 2025

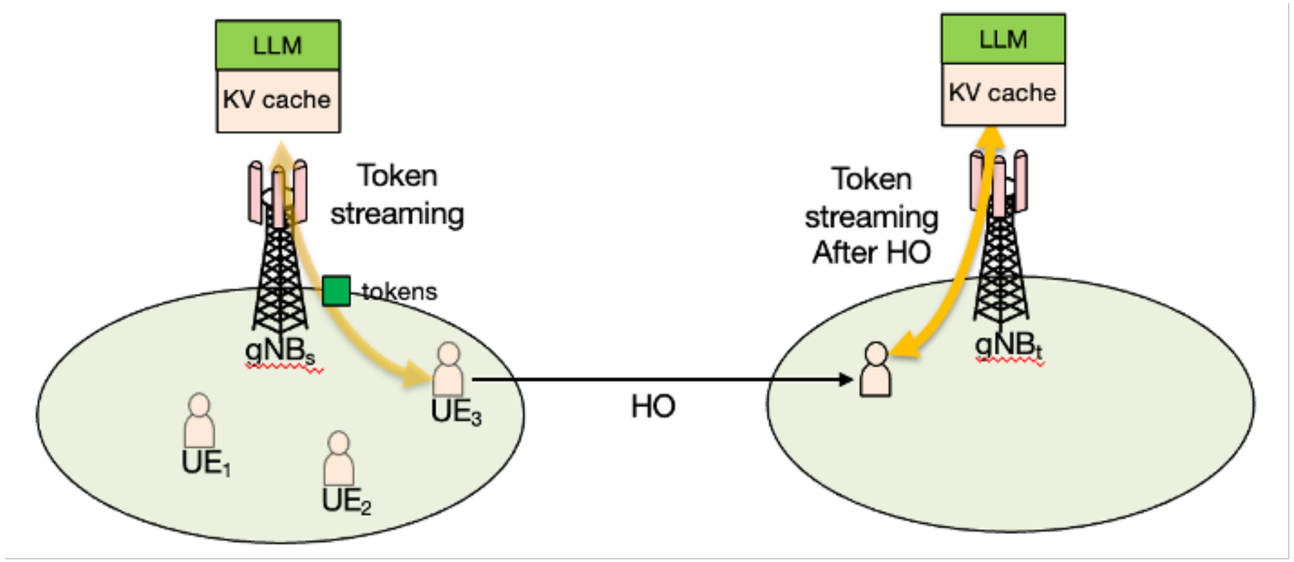
Z. Chen, Z. Li, H. H. Yang, S.-W. Ko, T. Q.S. Quek, and J. Park, "Federated Inference for Heterogeneous LLM Communication and Collaboration," AAAI'26 Wksp. ML4Wireless.

Future Challenges. **Token vs. Cache Communication**

- Token communication: **\$Computation ↑ ➨ \$Communication ↓**

*Q. If computation dominates, can **\$Communication ↑ ➨ \$Computation ↓ ?** ➨ **KV cache communication***





S. Lee, C. Zheng, H. Park, and J. Park, "KV Cache Handover for Mobile Wireless LLM Streaming," *work in progress*.


Thank You
FCP
FUTURE COMMUNICATIONS
R&D PROGRAMME
SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN