Qualcomm | @QCOMResearch | 6G@UT Forum 2024 | April 11th, 2024

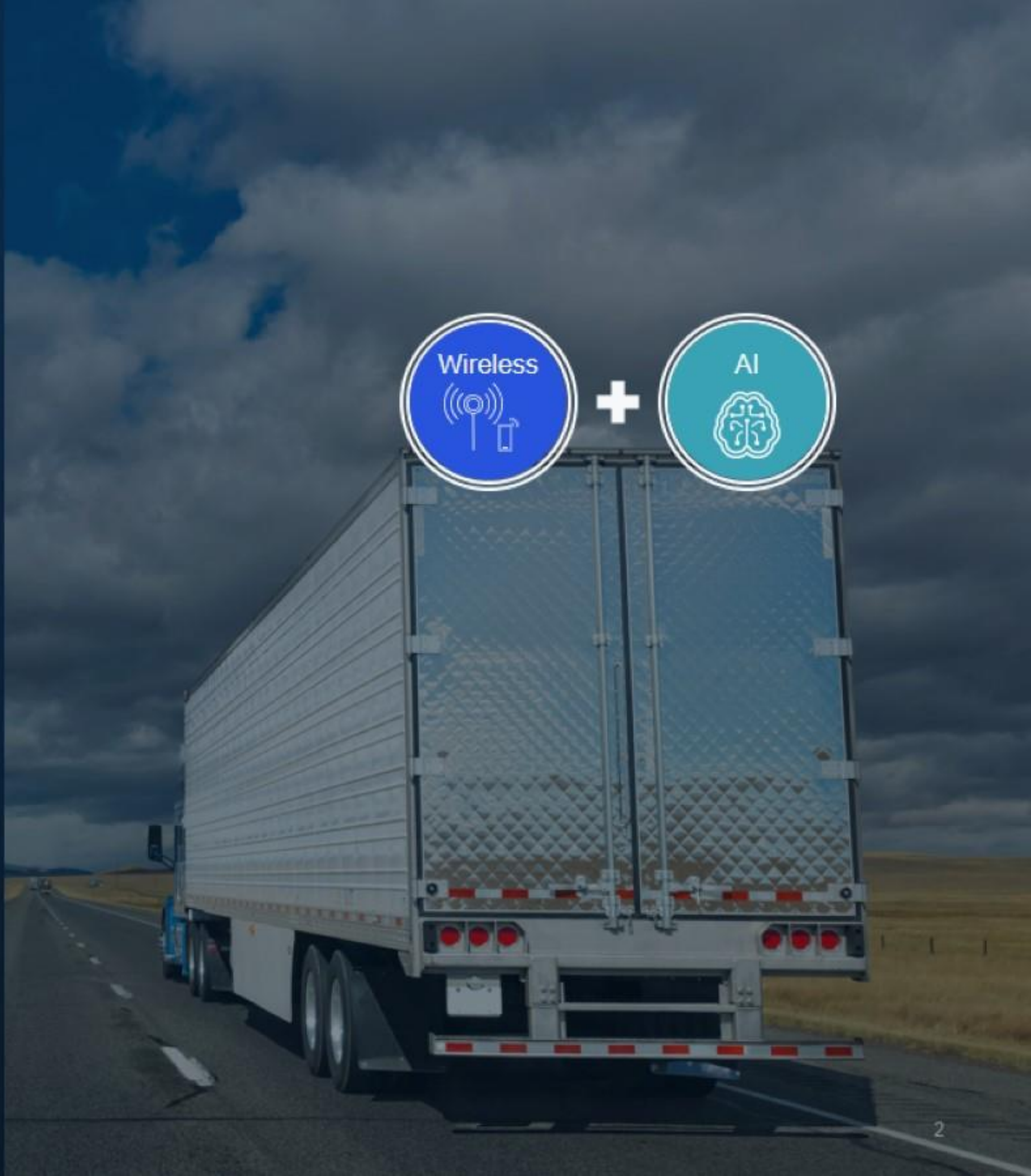
How machine learning will revolutionize 6G system designs

Devices, air interface and network operations

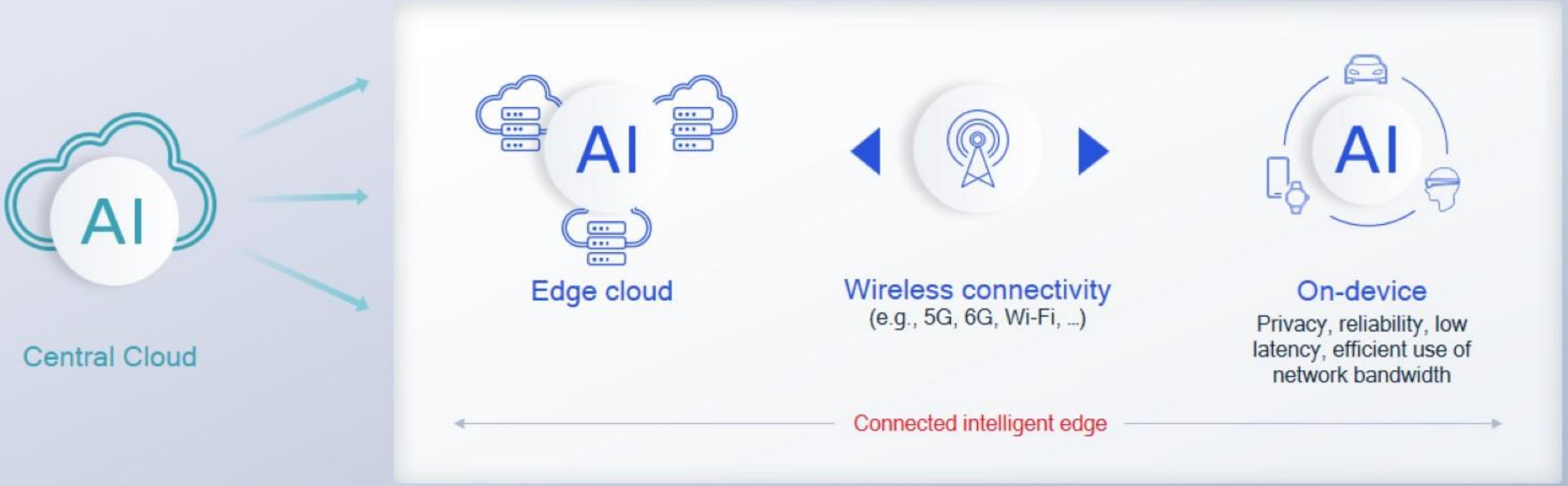
Dr. Tingfang Ji
Vice President, Engineering
Qualcomm Technologies, Inc.

Today’s agenda

1. AI enablement of 5G wireless today
2. AI-based air interface for 5G-Advanced
3. AI for 6G



To scale efficiently, AI processing is expanding towards the edge



Qualcomm is leading the realization of the connected intelligent edge

CONVERGENCE OF:

- Wireless connectivity
- Efficient computing
- Distributed AI
- Unleashing massive amount of data to fuel our digital future

Snapdragon
X80 5G Modem-RF

Qualcomm®
5G AI Suite
Gen 3

AI-enhanced 5G Advanced user experience

Multi-antenna management to improve user experience

Contextually-aware QoS and latency improvements

60%* faster CPE service acquisition (mmWave)

10%* lower power in connected mode (mmWave)

Location accuracy improvement by 30%*

Best-cell selection time reduced by 20%*

30%* faster link acquisition

* Up to indicated numbers

Snapdragon X Elite — PERFORMANCE REBORN

Best-in-class performance vs. x86

Up to **2X** faster CPU
- Qualcomm Oryon™ CPU
- 12 high-performance cores
- Dual-Core Boost

Up to **2X** faster GPU
- Qualcomm® Adreno™ GPU
- 4.6 teraflops
- Triple UHD monitor support

4nm process node

136GB/s memory bandwidth
LPDDR5x

Generative AI powerhouse

Capable to run **13B+** parameters on device

Generates **30** tokens/sec for 7B LLMs

Built for AI

75 TOPS Qualcomm® AI Engine

45 TOPS Qualcomm® Hexagon™ NPU

<1 seconds Stable Diffusion

4.5X faster AI NPU processing power than competitors

1st PC processor with integrated Always-Sensing ISP

Integrated Micro NPU on Qualcomm Sensing Hub

Smart user experiences

- Lightning-fast 5G | Wi-Fi 7
- Immersive lossless audio
- Advanced camera ISP
- Snapdragon Seamless
- Chip-to-cloud security

Leading PC performance per watt

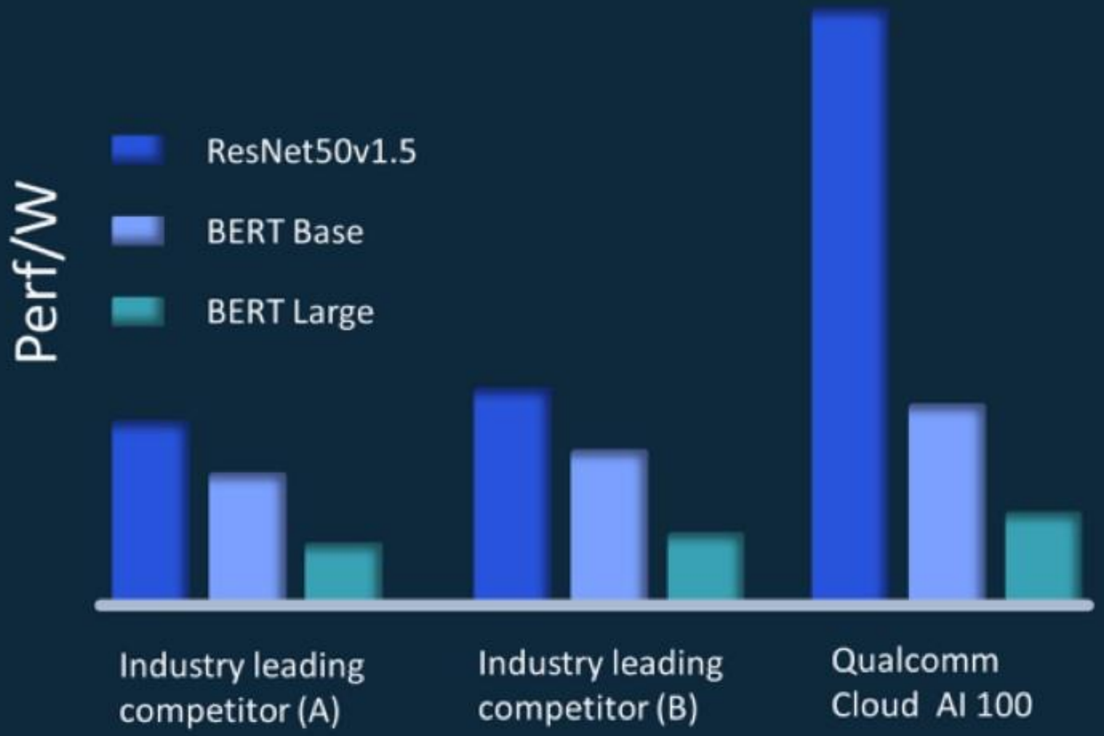
Matches peak PC performance at **68%** less power vs. competition

Scalable across a range of thermal designs and form factors

 | CPU Performance is based on Geekbench v6.2 Multi-Thread on Windows OS run in October 2023. Snapdragon X Elite was tested using a Qualcomm laptop reference design on Windows OS. The i7-1360P (12 core) and i7-1355U (10 core) were tested using a Samsung Galaxy Book3 360 13" 2023 (NP730QFG) laptop and Samsung Galaxy Book3 15.6" 2023 (NP750XFG) laptop, respectively on Windows 11. Maximum performance reflected by i7-1360P and i7-1355U represent maximum achievable results in given platforms under unconstrained PL1/PL2 settings and no thermal limitations. Power and performance comparison reflects results based on measurements and hardware instrumentation of given devices. GPU Performance is based on 3DMark WildLife Extreme run in October 2023. Snapdragon X Elite was tested using a Qualcomm laptop reference design on Windows PS and a pre-commercial build of Wildlife Extreme. The i7-13800H (14 core) was tested using a Razer Blade 15 2023 (RZ09-0485) laptop on Windows 11. Maximum performance reflected by i7-13800H represents maximum achievable results in given platform under unconstrained PL1/PL2 settings and no thermal limitations. Power and performance comparison reflects results based on measurements and hardware instrumentation of given devices.

Building high performance, power efficient, AI inference accelerator

Qualcomm® Cloud AI 100

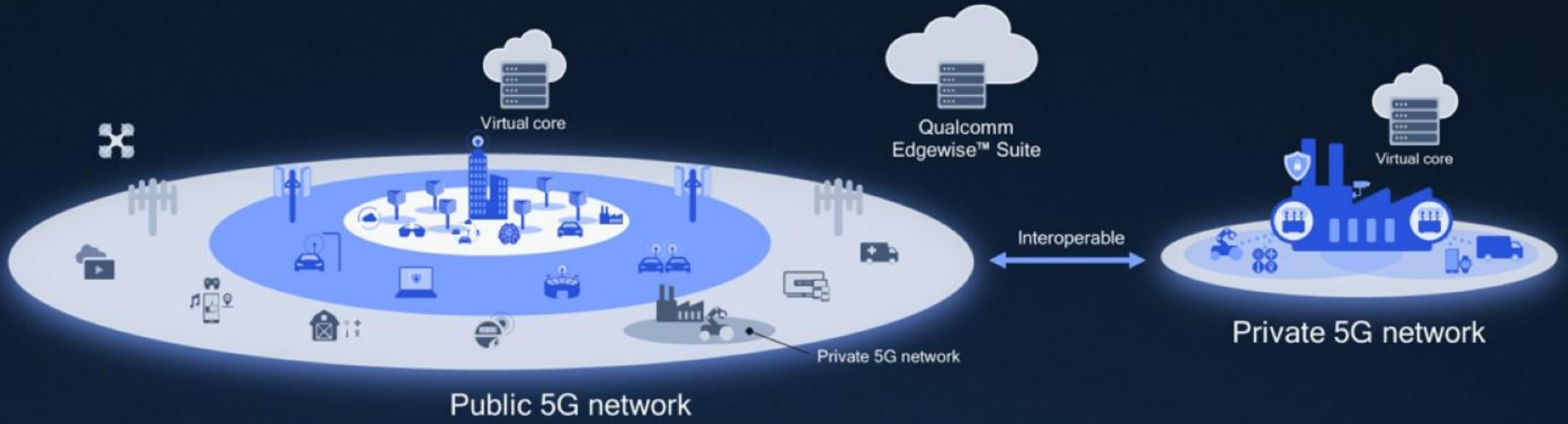


Perf/TCO$ = Inf/sec / (CapEx + OpEx)

e.g., 4x Perf at same OpEx delivers up to 5x Perf/TCO$ benefit

Qualcomm solution's 2-4x Performance advantage at similar power leads to 2-5x Perf/TCO$ advantage

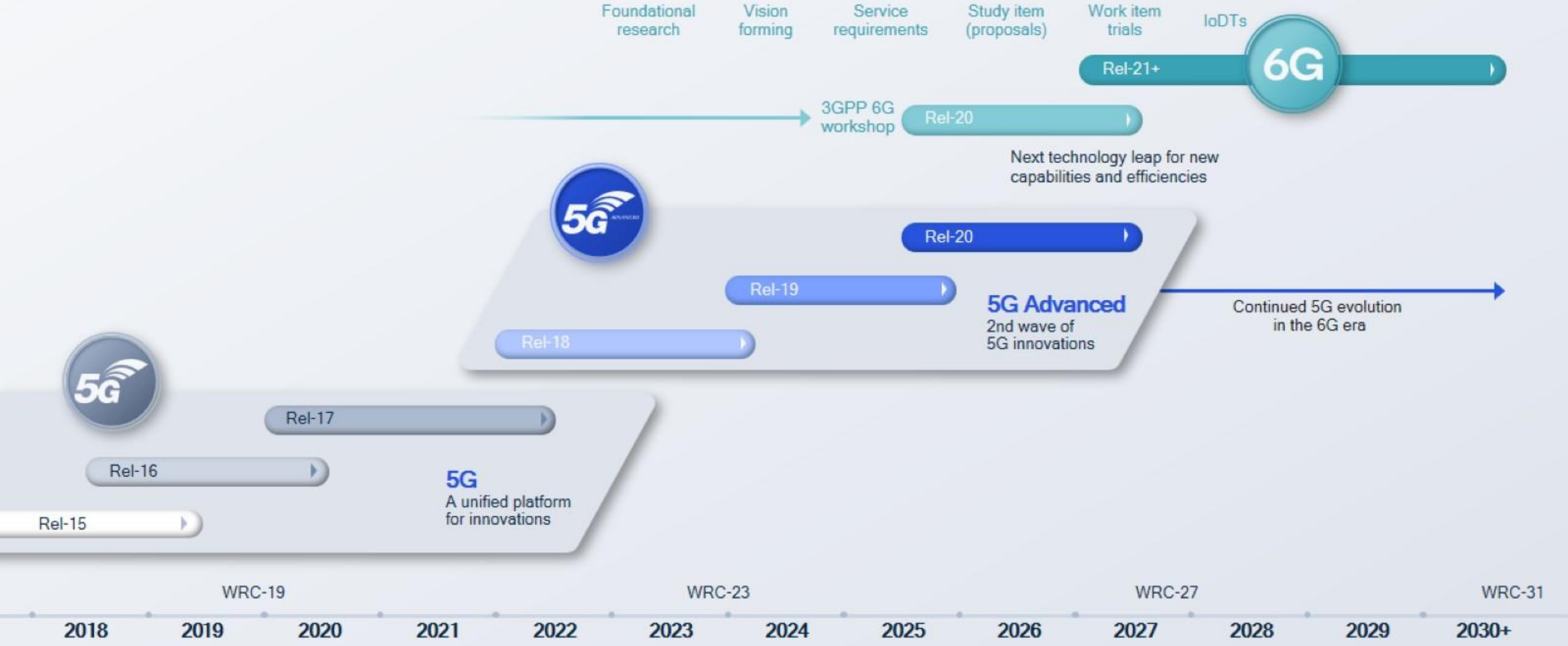
Qualcomm 5G RAN platforms



Qualcomm® 5G RAN Platforms

Qualcomm® X100 5G RAN Accelerator Card

5G NR AI-based air interface design in Rel-18+



A KEY PILLAR OF THE
5G ADVANCED ERA

Wireless AI

3 projects in Release 19

Study on AI/ML for Next-Gen Radio Access Network[1]

New use cases including network slicing and coverage and capacity optimization (CCO)

Continued studies on mobility optimization for NR-DC, split architecture support, enhanced energy saving, continuous MDT, and multi-hop device trajectory

Study on AI/ML to enhance 5G NR mobility[2]

Focusing on L3 device mobility, including RRM measurement & event prediction, device assistance information for network-side model, enhanced LCM, evaluation on testability, interoperability, impacts on RRM requirements and performance

1 RAN 3 led; 2 RAN 2 led; 3 RAN 1 led; 4 Continued study with corresponding checkpoints in RAN#105 (Sept '24)

Source: RP-234039 (AI/ML for NR air interface); RP-234054 (Study on AI/ML for NG-RAN); RP-234055 (Study on AI/ML for mobility in NR)

Work on AI/ML Air Interface[3]

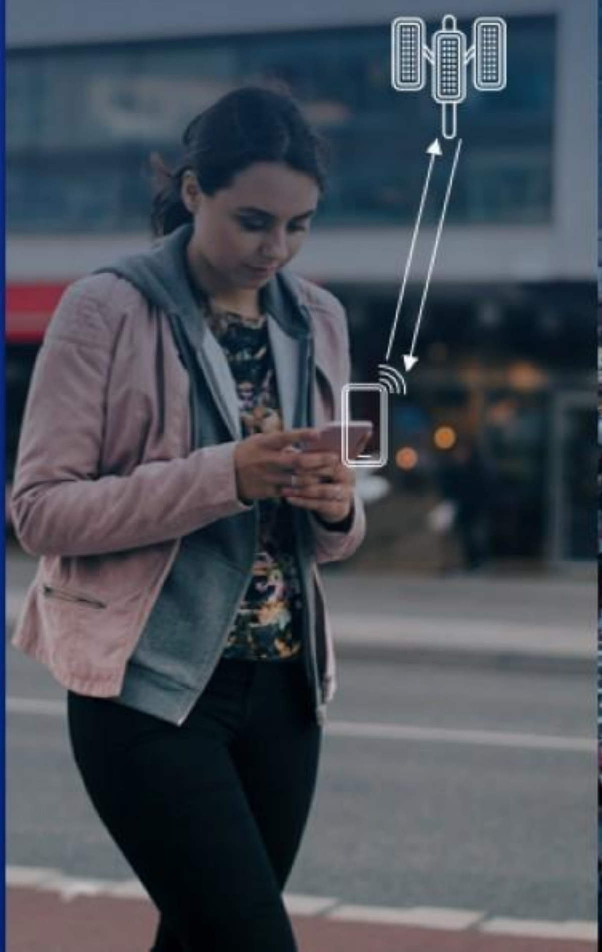
General Wireless AI Framework

Support Life Cycle Management (LCM) of one-sided (i.e., device or network) AI/ML models

Channel feedback[4]

Further study 2-sided CSI compression, 1-sided CSI prediction, model transfer/deliver, ...

Improve user downlink throughput and reduce uplink overhead



Beam management

Support device/network-sided beam prediction model in time/spatial domain

Reduce overhead, latency, and improve beam selection accuracy

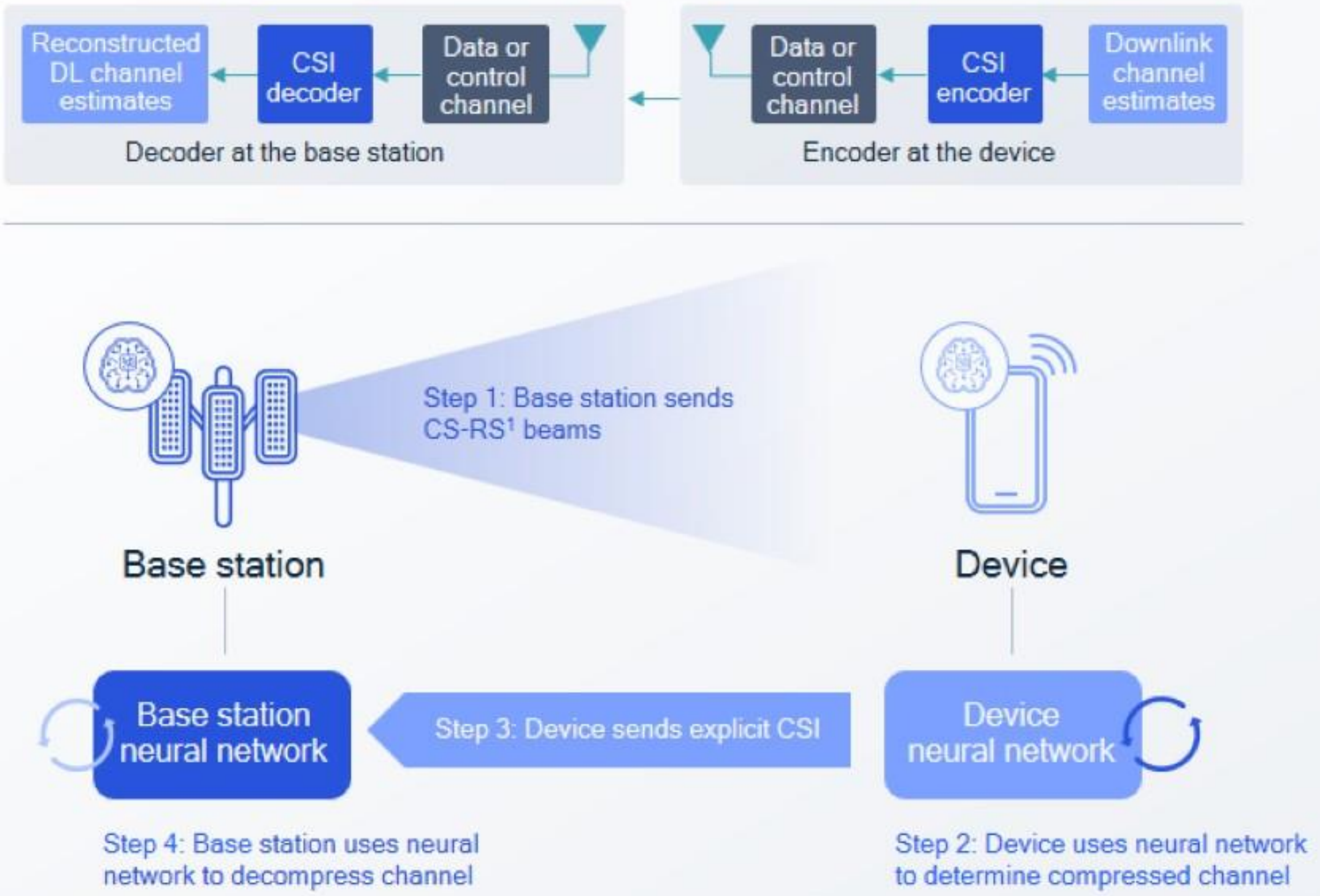
Precise positioning

Support single-sided model for both AI-direct and AI-assisted positioning

Improve positioning accuracy for different indoor/outdoor scenarios

Cross-node machine learning based channel state information

Explicit channel feedback framework for CSI compression and prediction utilizing domain knowledge and neural networks



Improve system efficiency with neural network framework for CSI on non-linear encoding and decoding

More effective multi-user multiplexing minimizing interference

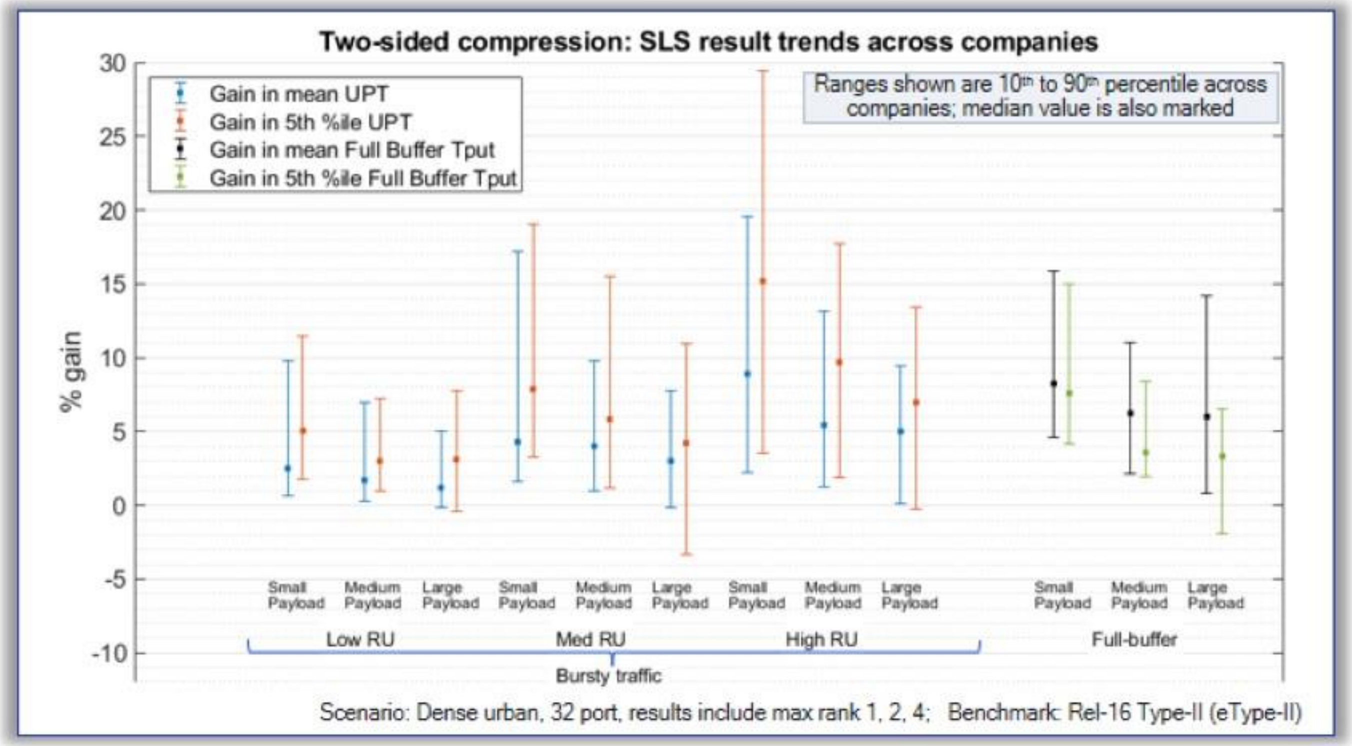
Customized, lower overhead feedback based on individual device

[1] Channel State Information Reference Signal

Performance gains from 3gpp spatio-frequency CSI compression study

System-level results: Mean and 5th %ile Throughput

Assumptions: 3gpp Dense Urban scenarios, 4GHz, 32 CSI-RS ports, 4 Rx, 20MHz comparing ML CSF and eType 2



- CSI feedback overhead reduction: 30-70% reduction in feedback overhead
- Throughput gains: up to 20-30% gain in median and tail user perceived throughput
  - Gains are more pronounced for smaller payload, higher resource utilization, and larger cells.

Ongoing and future directions

- CSI prediction
- Spatial-frequency-temporal compression
- Joint CSI prediction and compression
- Hyperlocal models
- Joint CSI-RS optimization, feedback, and precoder
- Joint source channel coding on CSI
- Utilizing DL/UL reciprocity for CSI

Qualcomm

MWCB 2024

Wireless AI Interoperability

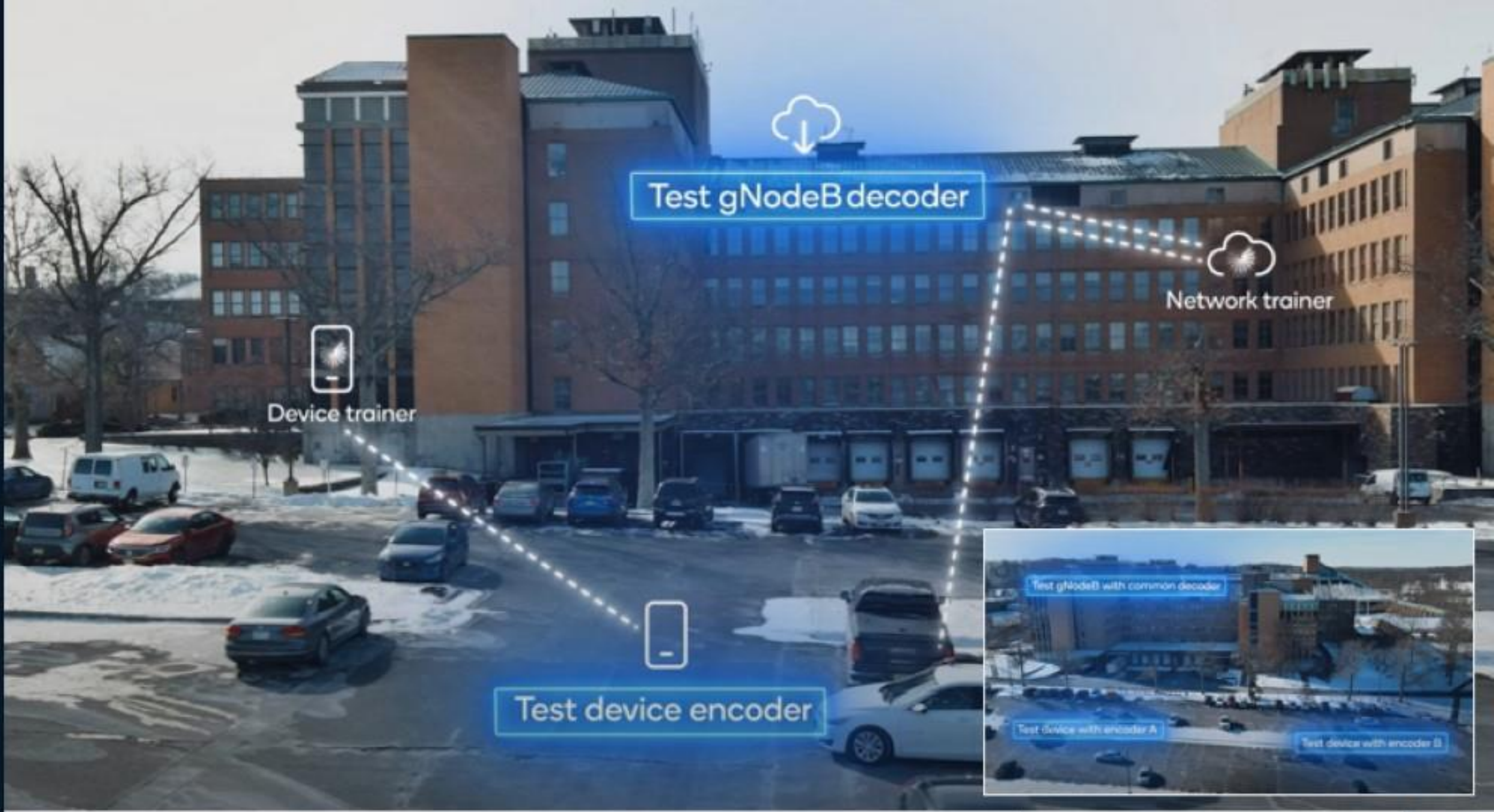
For multi-vendor system

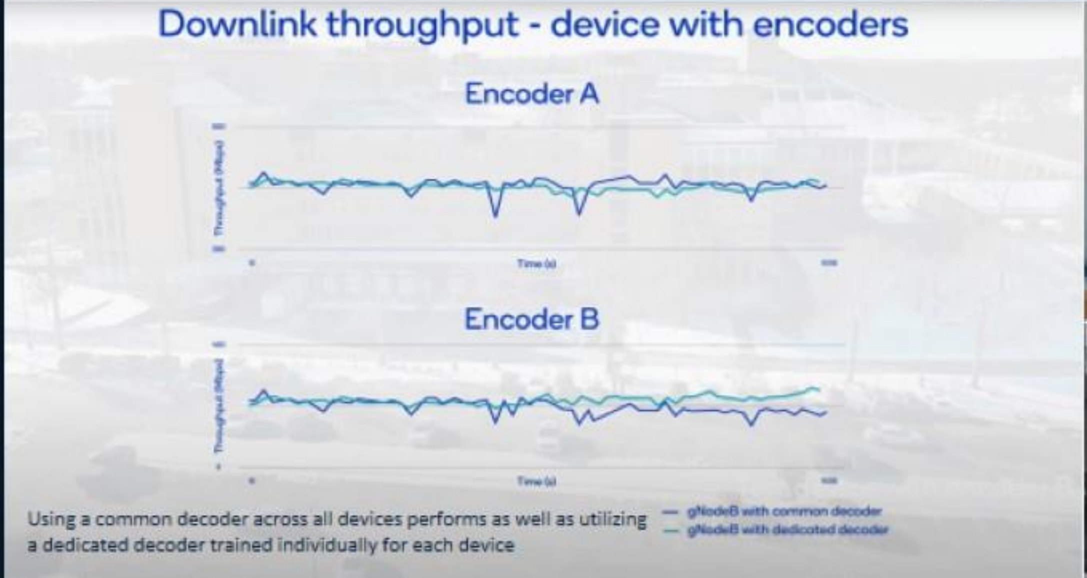
Close collaboration with Nokia Bell Labs on an over-the-air prototype of two-sided channel state information (CSI) feedback

Test network in Murray Hill, NJ, with Nokia infra and Snapdragon 5G Modem-RF

Sequential training enables data sharing but not neural network structures (models)

3GPP global standards compliant, potentially a part of 5G Advanced Rel-19+





Qualcomm

MWCB 2024

AI-enabled 5G mmWave Beam Management

Time-domain and spatial-domain device beam prediction

Reduced signaling overhead yields a more energy-efficient system design

Global standards-compliant design based on 3GPP Rel-18/19 projects

Over-the-air testbed operating in the 28 GHz band with up to 800 MHz bandwidth

Performance gains from 3gpp positioning study

AI/ML learns multipath and enhances positioning in challenging NLOS scenarios, reducing positioning error from >10 meters to submeter level

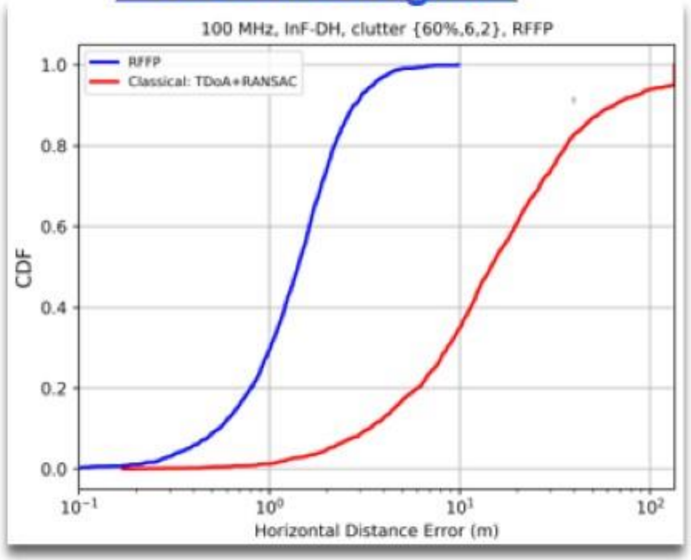
Direct AI/ML positioning

PRS/SRS measurement(s), e.g., CIR/PDP/DP[1] → **AI/ML model** → UE location

AI/ML assisted positioning

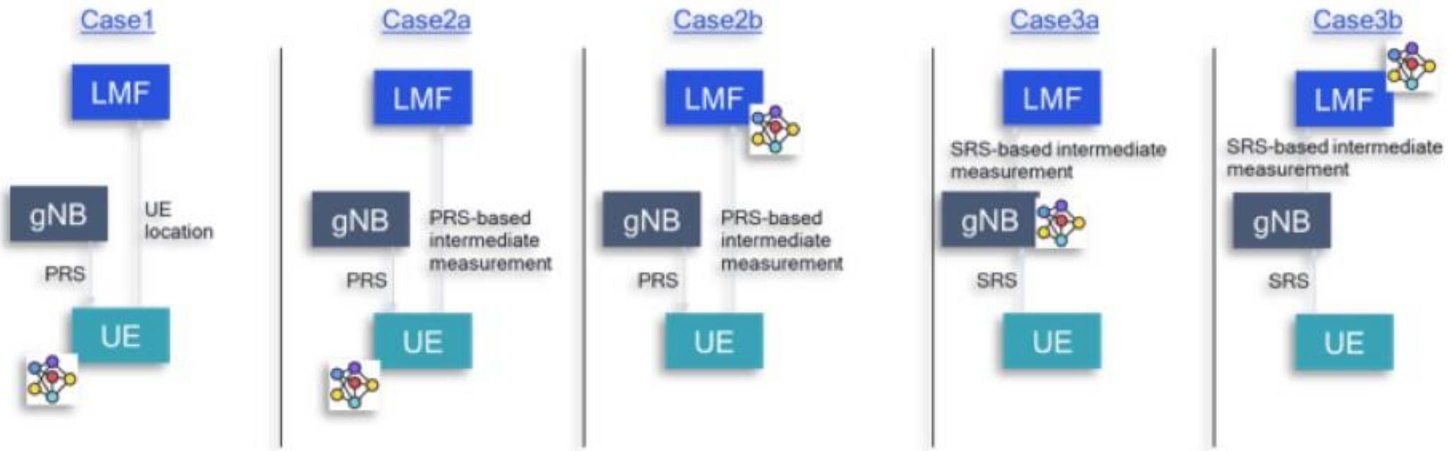
PRS/SRS measurement(s)[1], e.g., CIR/PDP/DP → **AI/ML model** → Positioning intermediate measurements (e.g., timing, LOS identification, etc.) → **Classical positioning model** → UE location

Performance gains



RFFP: direct AI/ML positioning | Extreme NLOS condition (LOS < 1%) | 3GPP InF-DH scenario| Classical: Time difference of arrival with outlier rejection using RANSAC algorithm

Deployment cases



Note[1]: CIR: Channel impulse response | PDP: Power delay profile | DP: Delay profile

Fundamental 6G motivation

Operators networks

- Increasing revenue with new use cases/services

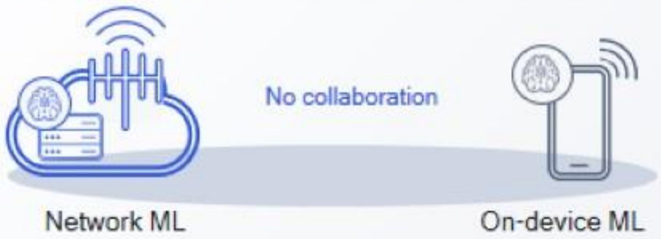
- Reducing network TCO (total cost of ownership)

Evolving towards an AI-native wireless system

Multiple wireless AI/ML training and inference scenarios

Overlay AI/ML

INDEPENDENTLY AT THE DEVICE OR NETWORK



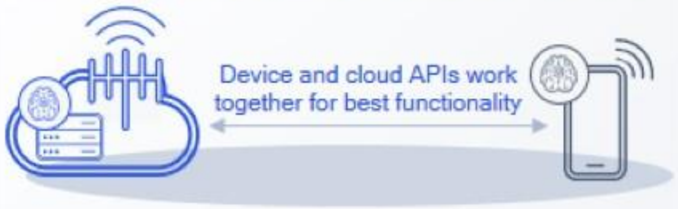
ML operates independently at the device and network as an optimization of existing functions

Proprietary ML procedures including model development and management

Proprietary and standardized data collection used as input to training

Cross-node AI/ML

COORDINATED BETWEEN DEVICE AND NETWORK



ML operates in a coordinated manner between the device and network

Proprietary and standardized ML procedures including model development and management

Further data collection used as input to training as well as monitoring

Native AI/ML

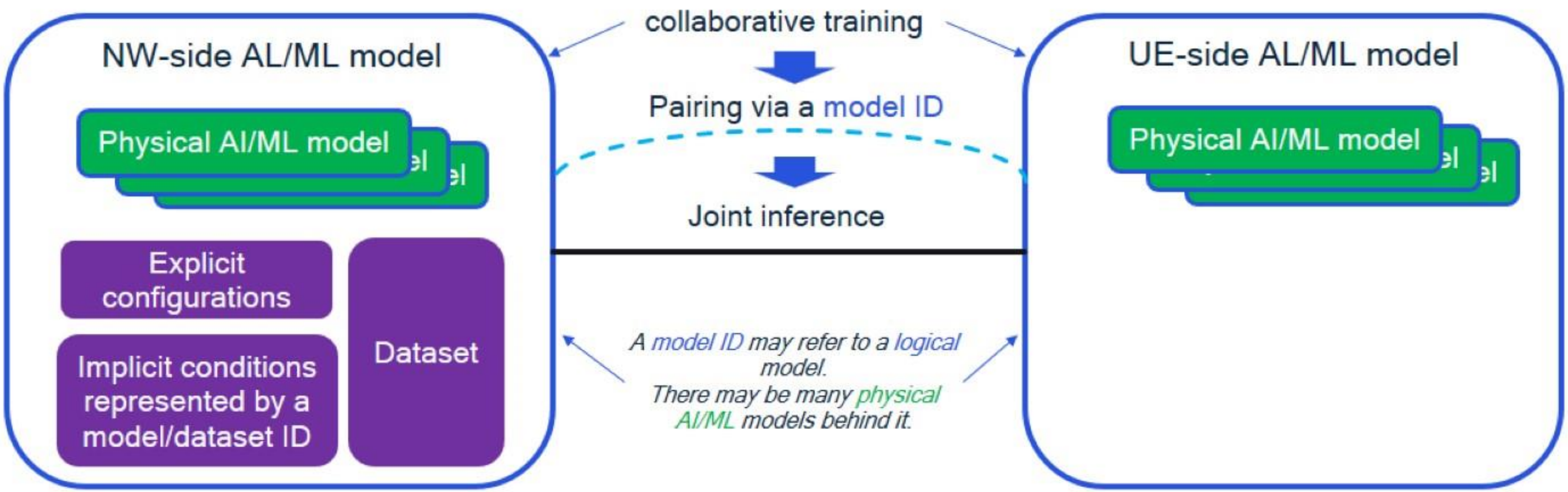
AT ALL DEVICE AND NETWORK LAYERS

ML operates autonomously between the device and network across all protocols and layers

Integrated ML procedures across to train performance and adapt to different environments

Data fusion for integrated dynamic ML lifecycle management

AI-native air interface open issues – Model pairing

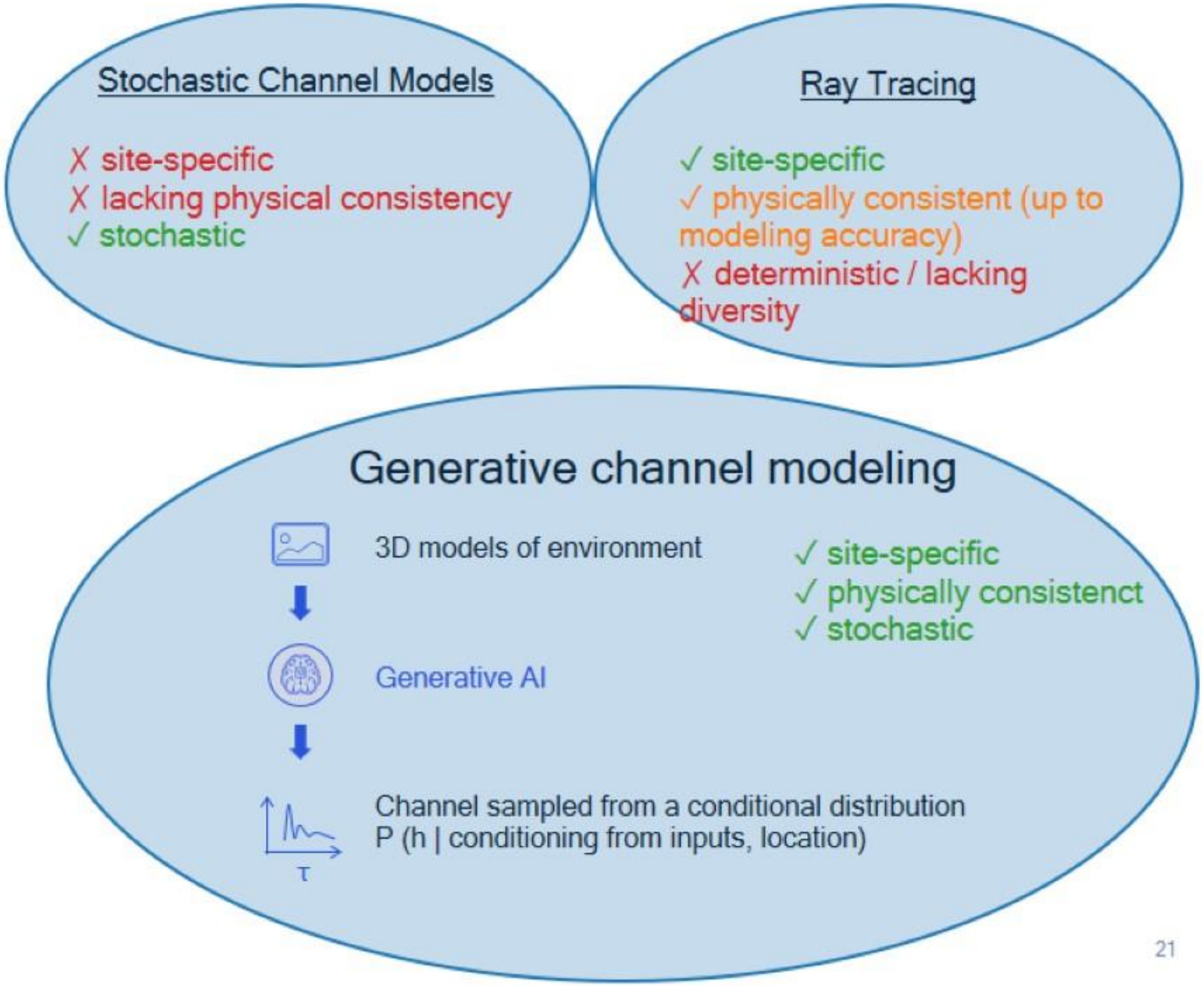


Or there may be no physical AI/ML model at all behind a model ID.

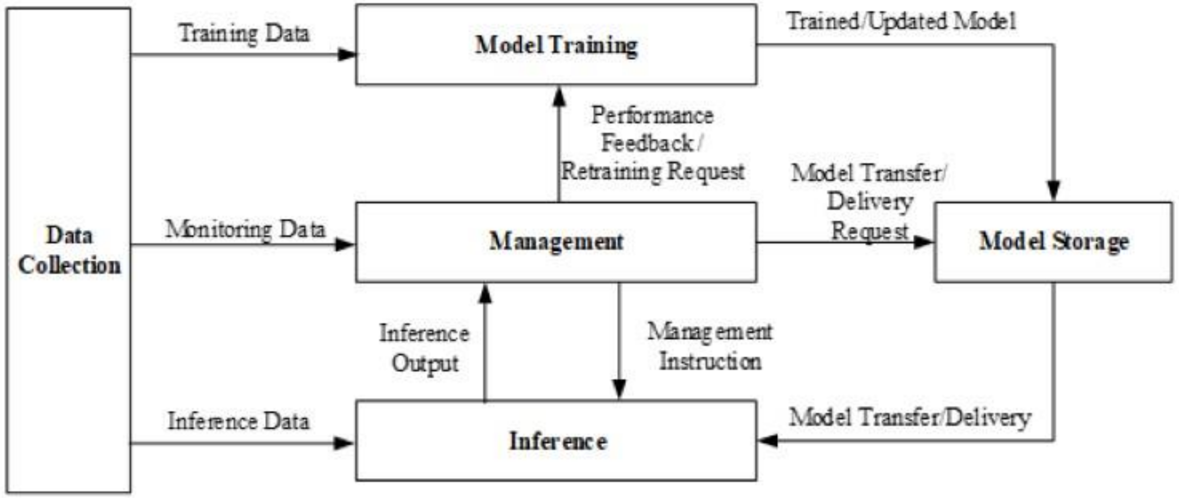
model ID == a set of conditions that determine data distribution, or a dataset itself

AI-native air interface open issues - channel models

- Conventional communication algorithms have been designed and evaluated based on synthetic channels following stochastic modeling assumptions.
- AI-native air interface design necessitates new channel modeling framework.
  - AI/ML models often achieve higher gain when optimized to a given site-specific propagation environment.
  - AI/ML models may not work well in real-world channels when trained on synthetic channels.
  - Channels from Ray Tracing is too deterministic and easily lead to AI/ML model overfitting.
- Data-driven approach may be used toward new channel modeling.

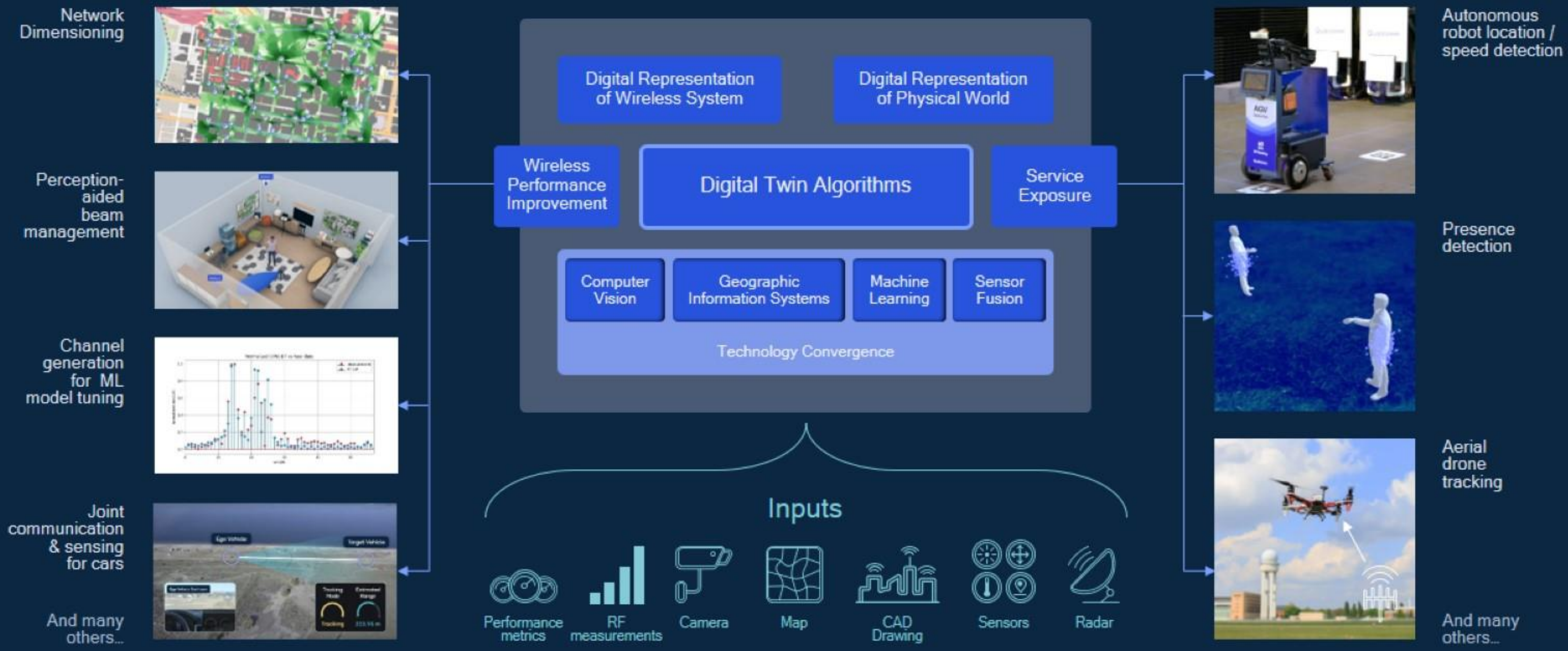


AI-native air interface open issues - Life cycle management



Source: TR 38.843 v18.0.0, "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface" (Release 18), December 2023

- Backend ML infra for data collection and model Life Cycle Management (LCM) are important.
- What and how much to specify?
  - Proprietary vs. standardized data collection
  - Proprietary vs. standardized training procedure
  - Proprietary vs. standardized model performance monitoring
- Proprietary innovation vs. standardized inter-operability



Digital Twin - virtual representation of physical system

Monitoring and performance optimization of its real-world counterpart

Vision for gen AI-augmented and autonomous networks

Intelligent monitoring and management

On-the-fly modeling

Proactive alerts

Programming AI-assistants

'Level-3' autonomous networks

6G

Qualcomm

MWCB 2024

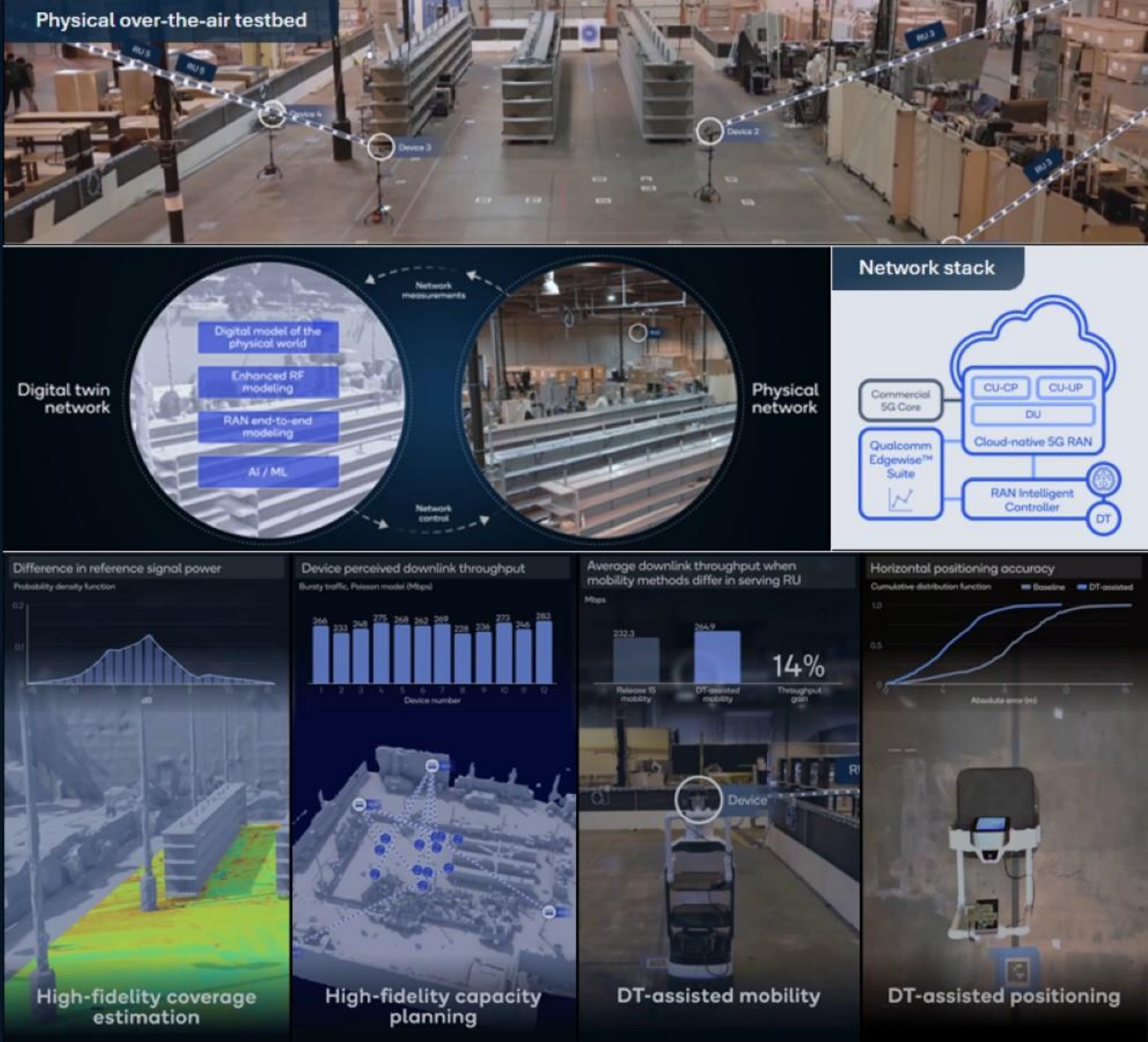
Digital Twin Network (DTN)

Converging expertise in wireless system modeling, computer vision and AI to create the high-fidelity DTN

Generating synthetic data to address the data collection challenge from real world deployments

Sophisticated dynamic modeling of 5G RAN infrastructure

Over-the-air testbed operating in the 3.35 GHz band with cloud-native 5G RAN, RAN Intelligent Controller (RIC) and the Qualcomm Edgewise™ Suite



Applying generative modeling to improve wireless communications system design

Wide applicability for Generative Modeling

Real-time use cases for air interface

- Propagation channel
- Beam management
- Interference prediction
- Scheduler optimization
- Traffic source
- Mobility enhancement

- Link / system simulation
- Deployment optimization
- Positioning and sensing
- Network and device optimization
- Others...

Application examples

Channel rendering

Text description of image or semantic map

⬇

Diffusion model
(To generate channel information)

⬇

Channel sampled from a conditional distribution
P (h | conditioning from inputs, location)

Network / device prediction

Context in text, e.g., history of device reports and base station responses

⬇

Large language model
To learn link, beam, protocol languages

⬇

Next action for base station and/or device, sampled from a conditional distribution
P (next action | conditioning from inputs)

Our on-going wireless generative AI research areas

- 3D mapping and material learning
- Foundation models (e.g., link and protocol level use cases, beam prediction, and others)
- Neural channel rendering (e.g., map-based, ray tracer augmented, site-specific, and others)
- Customized ML-based stochastic channel
- Neural surrogate for base station scheduler and applications traffic
- And others...

University Collaborations

Qualcomm Innovators Development Kit AI/ML and Compute

- Develop AI/ML and compute applications
- Snapdragon® 8 Gen 2 SoC with AI HW/SW

Snapdragon Spaces™ XR Developer Platform

- Develop immersive AR experiences
- Lenovo Think Reality A3 Glass and Motorola edge+ smartphone kit

Robotics

- Develop power-efficient robots
- Qualcomm® RB5 Development Kit with Qualcomm® QRB5165 processor (15 TOPS AI, 7 cameras, VSP)
- 5G mezzanine card accessory

Expanding to other platforms in the future

University Courses	Research projects	Platform Training
Hackathons	Private-Public projects	Onboarding Assistance

Thank you

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog