September 2021 San Diego, CA @QCOMResearch Qualcomm

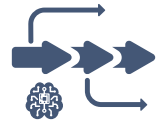
The essential role of AI in the 5G future

How machine learning is accelerating wireless innovations in the new decade and beyond

The essential role of AI in the 5G future

5G and AI are two synergistic, essential ingredients that are fueling future innovations

Applying AI to solve difficult wireless challenges and deliver new values



AI plays an expanding role in the evolution of 5G towards 6G

5G + AI

Unifying connectivity fabric that can efficiently connect virtually everything around us

5G

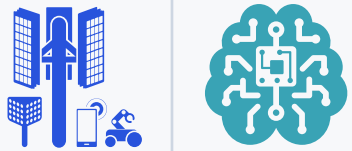
Extreme capacity

Multi-Gbps speeds

Ultra-high reliability and low latency

Robust end-to-end security and privacy

New and diverse services, spectrum, deployments



Learning platform that can make virtually everything around us intelligent

AI

Contextual awareness

Personalization at scale

Intelligent, intuitive, and automated actions

Continual improvement through self-learning

Solving seemingly impossible-to-model problems

Two synergistic and essential ingredients fueling future innovations

Advancement in AI is making

5G better

Elevated level of performance

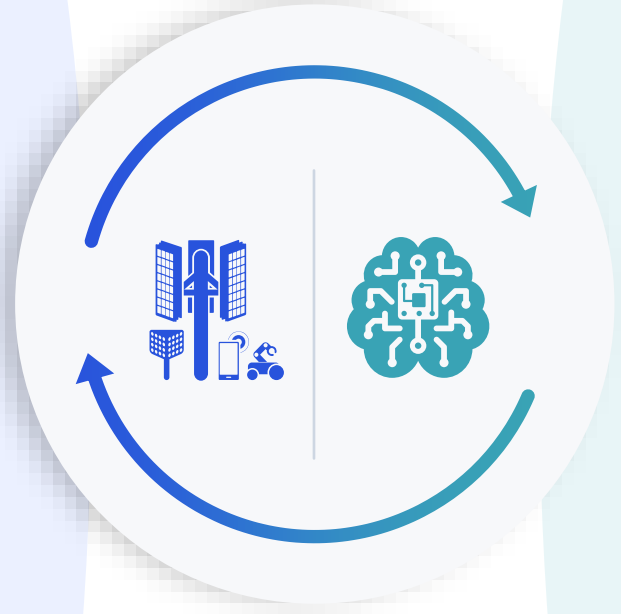
More efficient resource utilization

Energy reduction for longer battery life

Personalized security and privacy

Continuous enhancements over time

New and enhanced system capabilities



Proliferation of 5G is making

AI better

Responsive user experiences and services

Lifelong learning

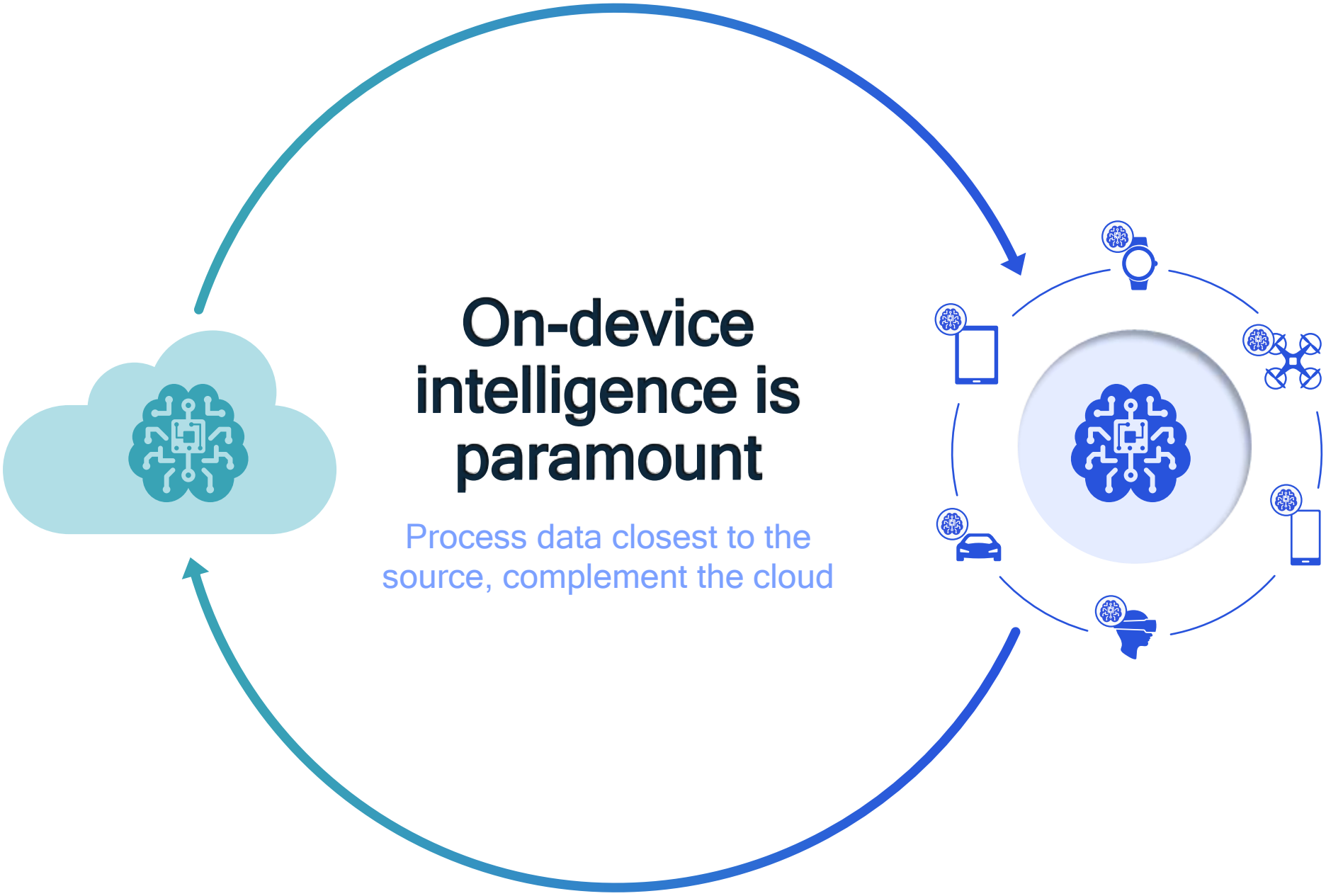
Flexibility for distributed functionality across devices

On-device intelligence assisted by cloud

Scale intelligence through distributed learning

Massive data aggregation for improved AI models

5G and AI are working together to accelerate innovations



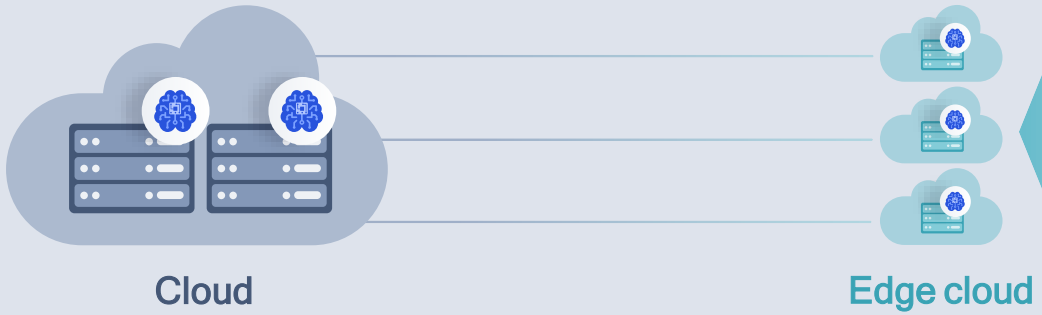
On-device intelligence is paramount

Process data closest to the source, complement the cloud

- Privacy
- Reliability
- Low latency
- Efficient use of network bandwidth

Transformation of the Connected Intelligent Edge has begun at scale

Processing data closer to devices at the edge derives new system values (e.g., lower latency, enhanced privacy)



Past
Cloud-centric AI
AI training and inference in the central cloud

Today
Partially-distributed AI
Power-efficient on-device AI inference

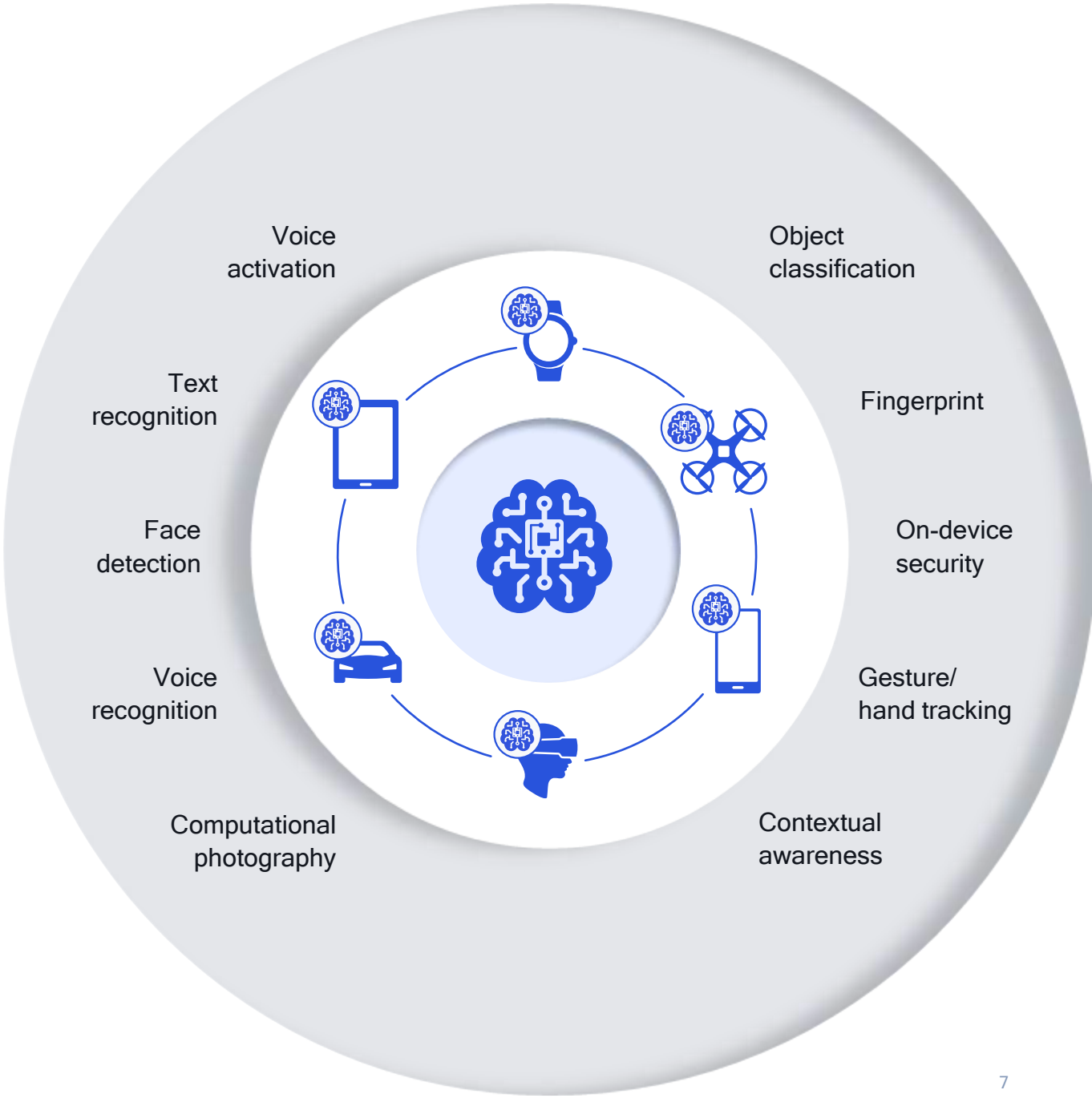
Future
Fully-distributed AI
With lifelong on-device learning

- Local network analytics
- Low-latency interactive content
- Boundless XR
- On-demand computing
- Industrial automation and control
- Enterprise data

Connected Intelligent Edge

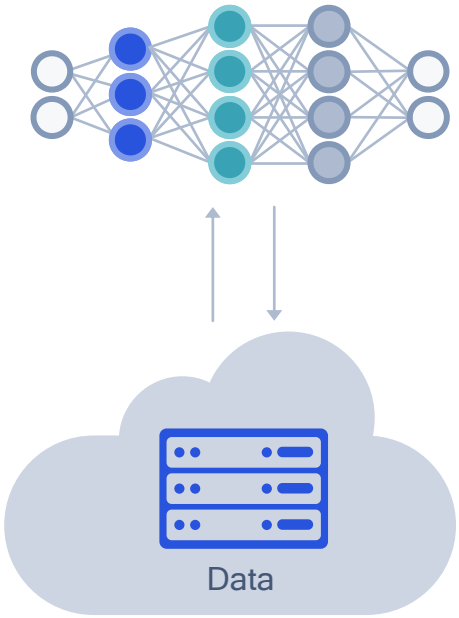
brings new and enhanced services

Edge Cloud AI + On-device AI



Federated learning brings on-device learning to new level

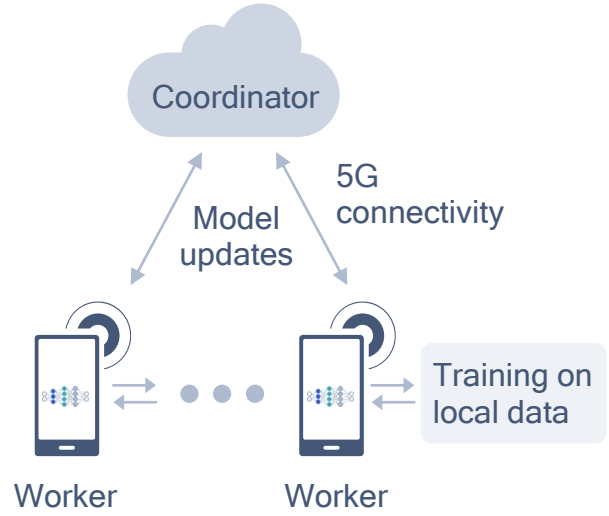
Offline learning



On-device learning

Locally adapt once to a few samples (e.g., few shot learning) or continuously (e.g., unsupervised learning)

Federated learning



Aggregate model updates across multiple users to globally improve model from more diverse data

Offline training prior to deployment | Local adaptation | Global adaptation

Our research focus to improve 5G with AI

With a cloud-to-network-to-device approach for data collection and learning

Distributed Cloud

Central and edge clouds

For end-to-end service optimization

Disaggregated Network

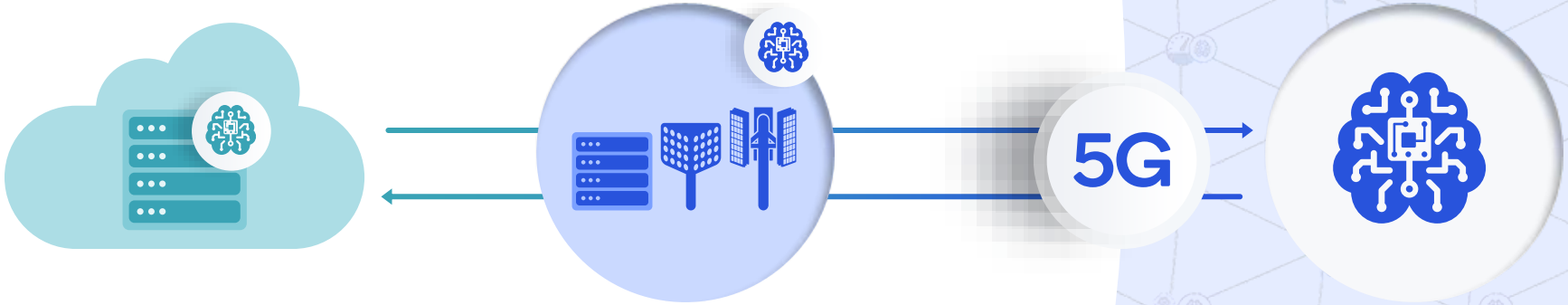
Virtualized and disaggregated RAN

For network and device optimization

Edge Devices

On-device intelligence

For local device optimization



Steps towards enabling AI/ML cloud and device platforms

Shorter Term

Continue to define data collection for new use cases hosted at the RAN

Medium Term

Enabling jointly optimized AI/ML use cases between RAN functions and device

Longer Term

Joint cloud, core, RAN, and device AI/ML functions

Applying AI

to solve difficult wireless challenges
and deliver new values

Wireless challenges

- Hard-to-model problems
- Computational infeasibility of optimal solution
- Efficient modem parameter optimization
- Dealing with non-linearity

AI-enhanced wireless communications

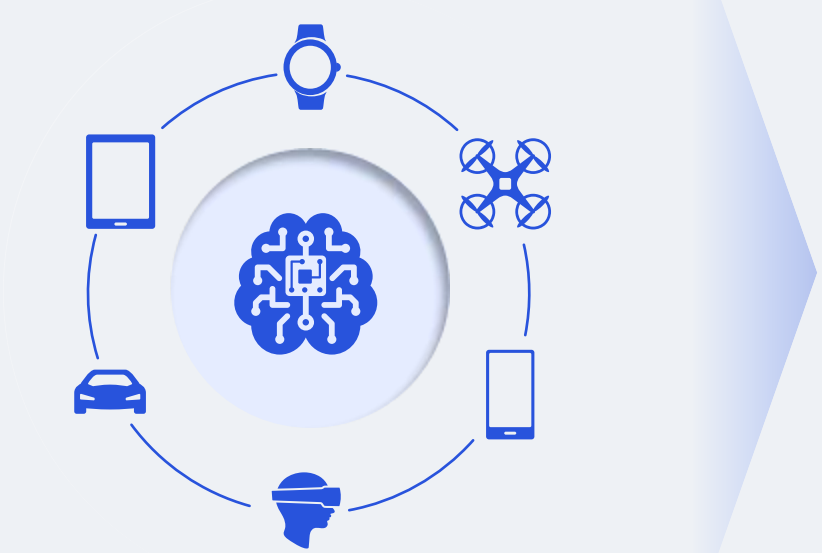
AI strengths

- Determining appropriate representations for hard-to-model problems
- Finding near-ideal and computationally realizable solutions
- Modeling non-linear functions

Applying AI to solve difficult wireless challenges

Deep wireless domain knowledge is required to optimally use AI capabilities

On-device AI improves the 5G end-to-end system



Radio awareness

Environmental and contextual sensing that reduces access overhead and latency

Enhanced device experience

More intelligent beamforming and power management improve throughput, robustness, and battery life

Improved system performance

On-device inference reduces network data traffic for more efficient mobility and spectrum utilization

Better radio security

Detecting and defending against malicious base station spoofing and jamming with fingerprinting

Radio awareness

achieved by advanced on-device AI algorithms

Spectrum sensing and access

Predict activities of other devices for more efficient access and better scheduling to improve 5G system performance

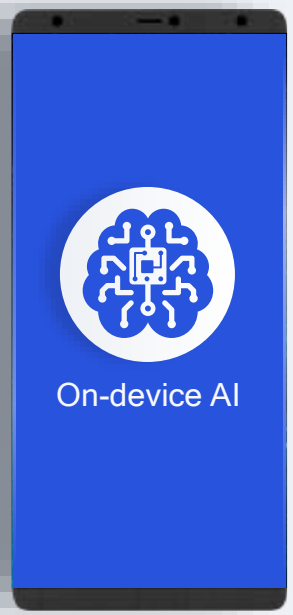
Contextual awareness

Use device context (e.g., position, velocity, or in-car) derived from RF, sensors, traffic activities to improve device experience

Environment (RF) sensing

Detect gestures, movements, and objects by monitoring signal reflection patterns to enable new use cases

On-device AI enhances 5G device experience

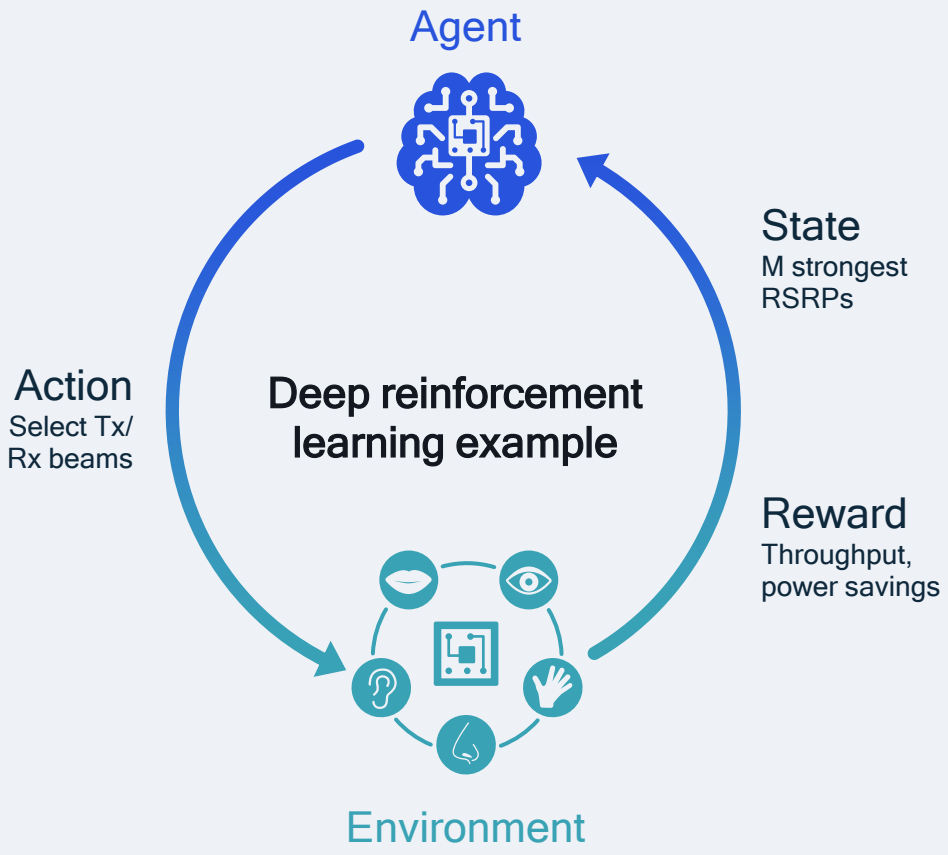


Better beam management

Incorporate location, velocity, other aspects of environmental and application awareness to improve robustness and throughput

More power saving

Optimize performance/power consumption tradeoffs by taking advantage of better contextual awareness on device

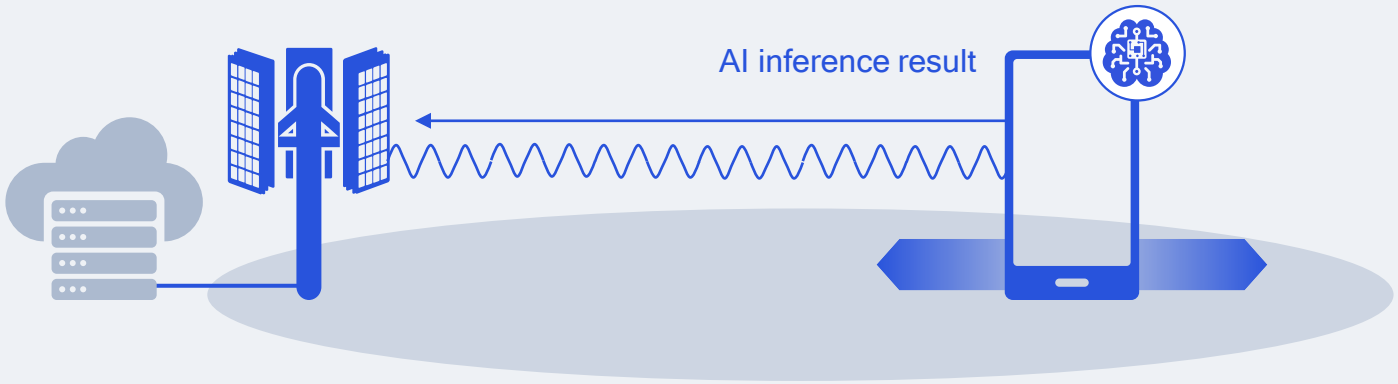


On-device AI

improves 5G system performance

Better link adaptation

Position-aware interference prediction can improve overall system throughput and spectral efficiency



Reduced network loading

On-device AI inference reduces the amount of raw data needed be sent across the network

More seamless mobility

Device-centric mobility utilizes on-device AI and sensors to predict handovers

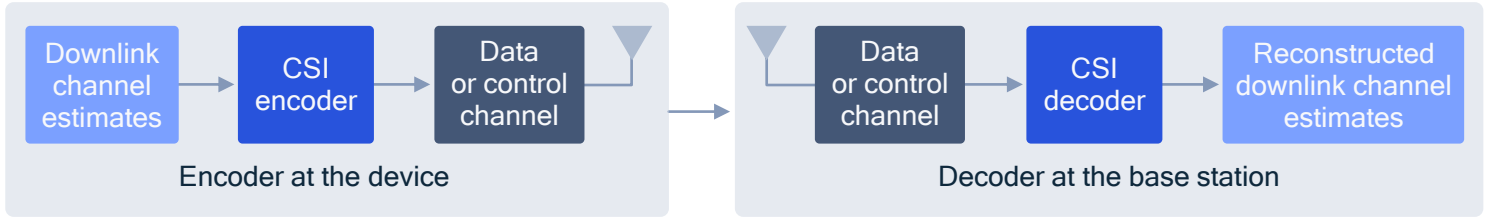
Demo: Mobile 5G mmWave beam prediction

Applying AI for enhanced 5G air interface efficiency

Example: for uplink transmissions

Improving system spectral efficiency

Implementing a neural network framework for CSI[1] on non-linear temporal encoding and decoding



Improving device power efficiency

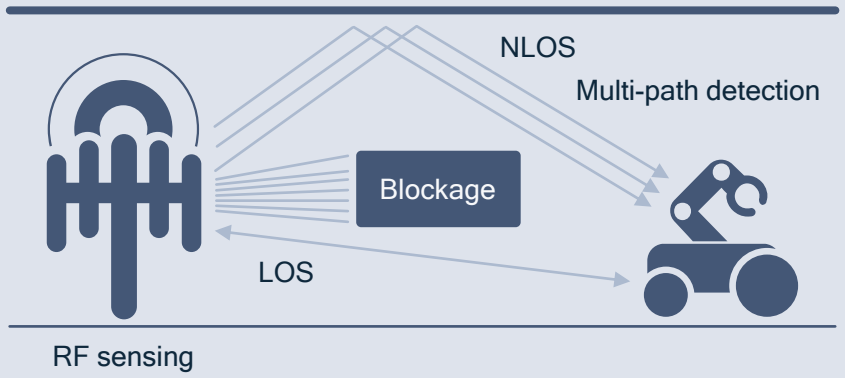
Optimizing transmit waveform to reduce peak-to-average power ratio (PAPR)

Allowing receiver to recover signal from a device operating in the non-linear PA region

1 Channel State Information

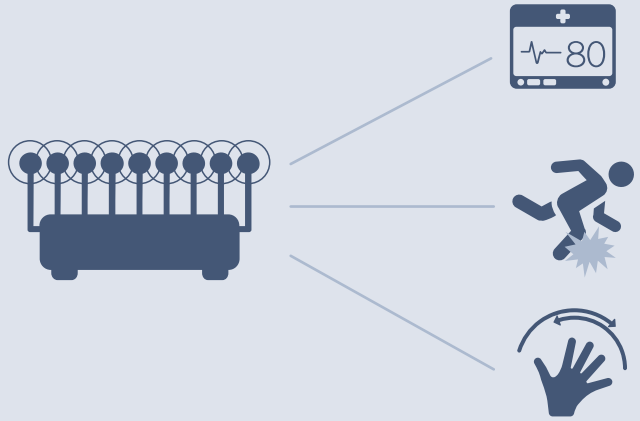
Demo: Enhanced 5G massive MIMO channel state feedback

More accurate device positioning



Learning device position over time without prior knowledge with RF sensing – complementing existing positioning methodologies[1]

Motion and gesture detection



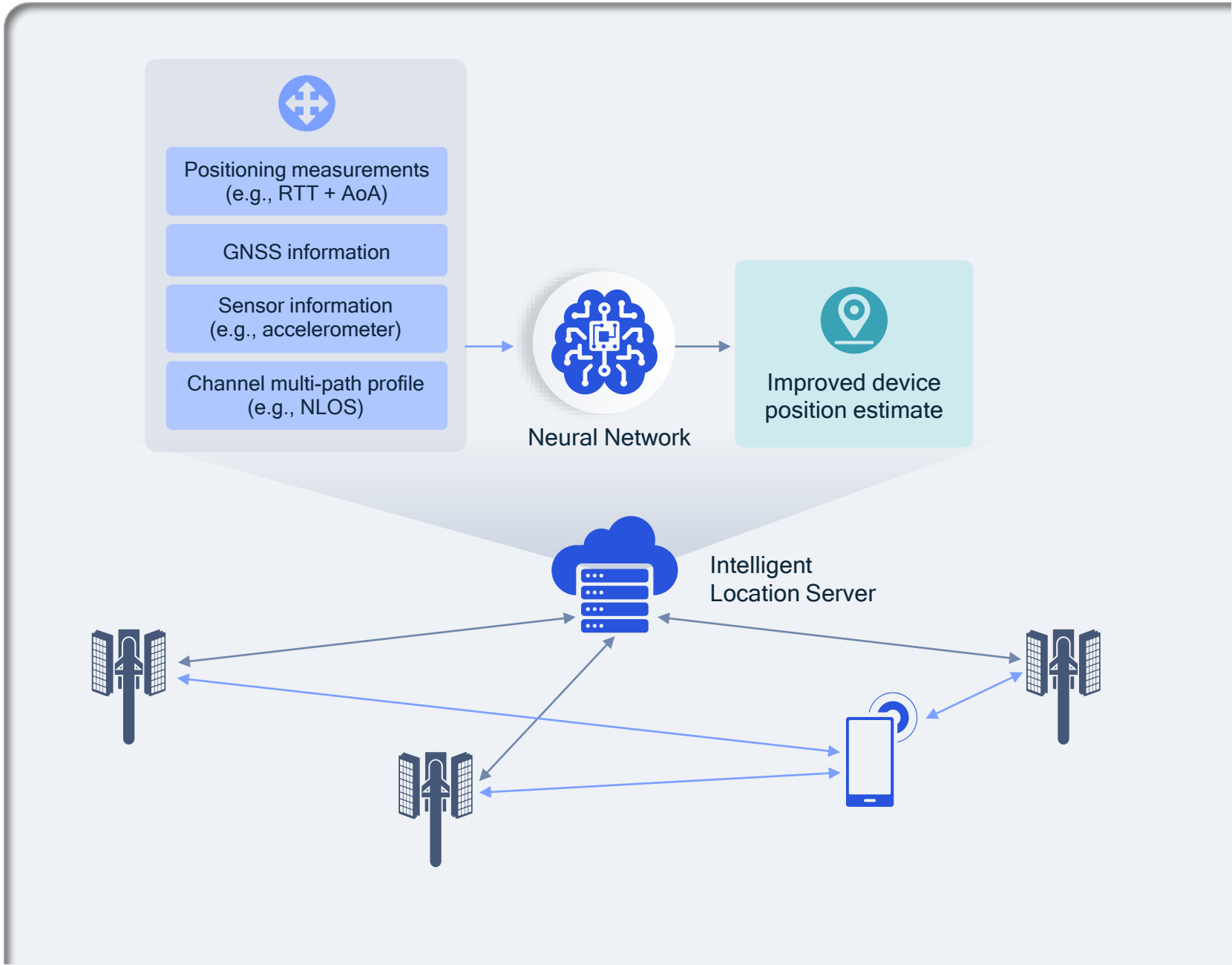
Sensing changes in environment to infer location and type of motion for a wide range of use cases (e.g., vital sign tracking, fall detection)

1. For example, Observed Time Difference of Arrival (OTDOA), Multiple Round Trip Time (Multi-RTT), Angle of Arrival (AoA)

Applying AI for contextual awareness and environmental sensing

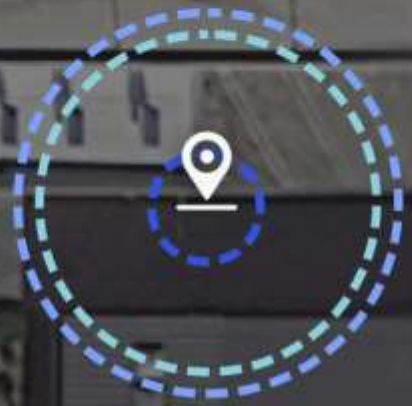
AI/ML for enhanced 5G positioning performance

5G positioning is supplemented by various assisting information, such as GNSS, multi-path profiles, and other sensors



For more challenging locations, utilizing GNSS data can further improve positioning accuracy

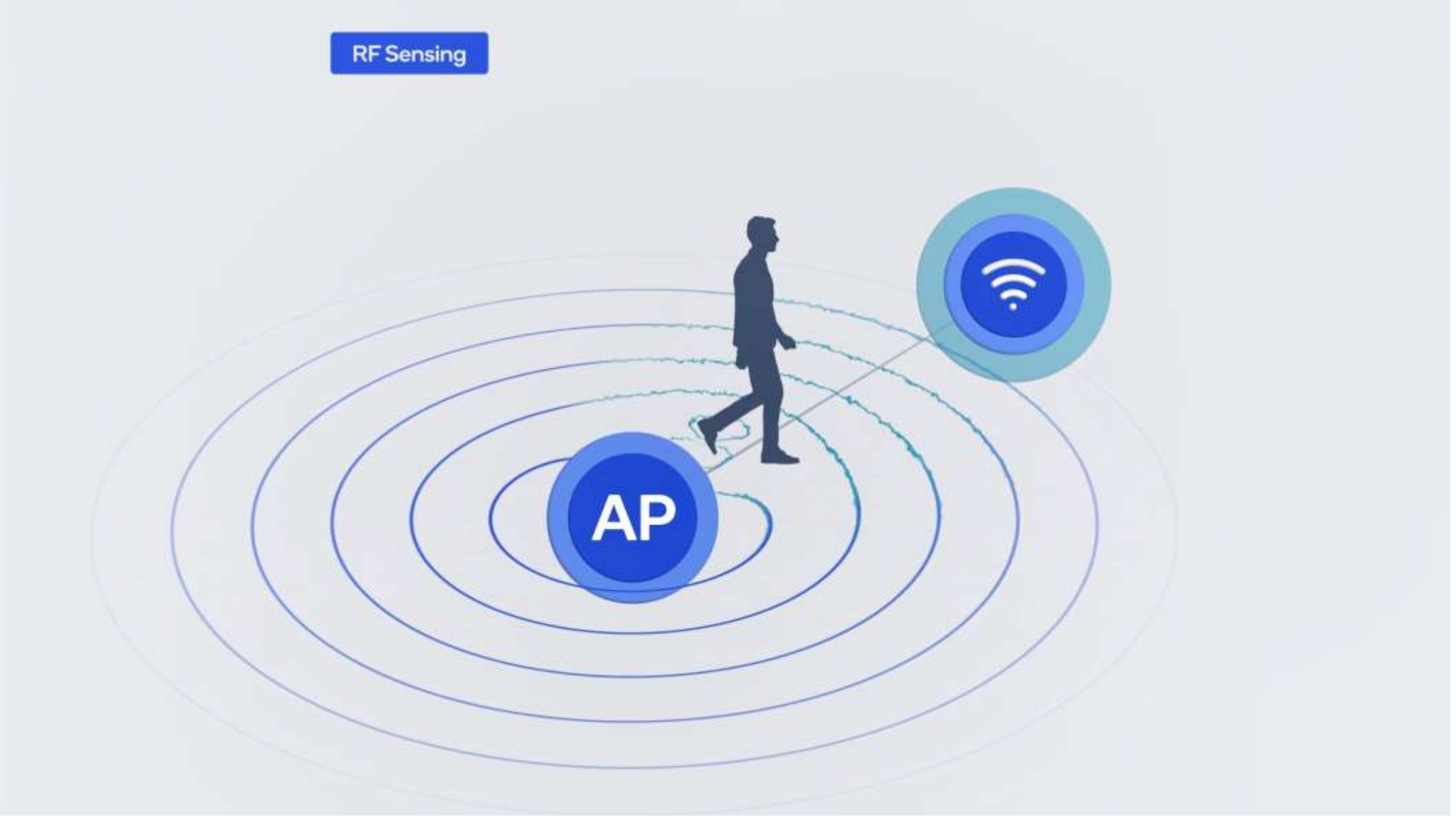
Device Location 2



Positioning Technique	Median Accuracy*
GNSS	8.37 m
5G (Rel-16)	7.75 m
ML-based Fusion (5G Rel-17+, GNSS)	2.57 m

*Based on one-shot positioning. Additional temporal processing and other sensor data assistance would further improve accuracy

Demo: Precise 5G positioning with machine learning fusion



Demo: Wireless AI-assisted indoor positioning

Leveraging unsupervised/weakly supervised learning – also applicable to 5G RF sensing (e.g., for positioning, motion and gesture detection)

AI enables intelligent 5G network management

Enhanced service quality

Better mobility management, user localization, and user behavior and demand prediction

Higher network efficiency

More efficient scheduling, radio resource utilization, congestion control and routing

Simplified deployment

More capable Self Organizing Networks (SON) for e.g., mmWave network densification

Improved network security

More effective detection and defense against malicious attacks by analyzing a massive quantity of data

Smart 5G mmWave densification

A more intelligent way to deploy 5G mmWave

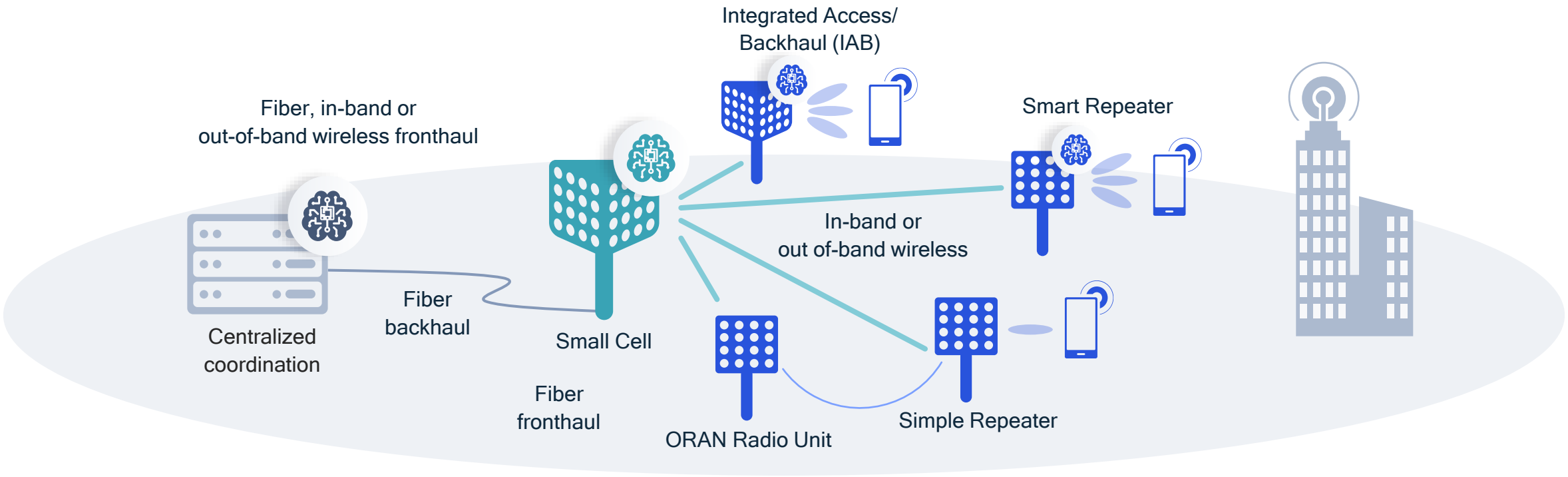
Create a digital twin of targeted deployment based on readily available sources/databases (e.g., Google Street/Aerial view, GIS, …)

Reduce model complexity and balance tradeoffs through ML-based object recognition, clustering, and pruning

Optimize mmWave network deployment to provide focused capacity using diverse existing and new infrastructure (e.g., repeater, IAB,…)

Distributed topology enables more efficient deployments

Standardization in e.g., 3GPP, O-RAN Alliance



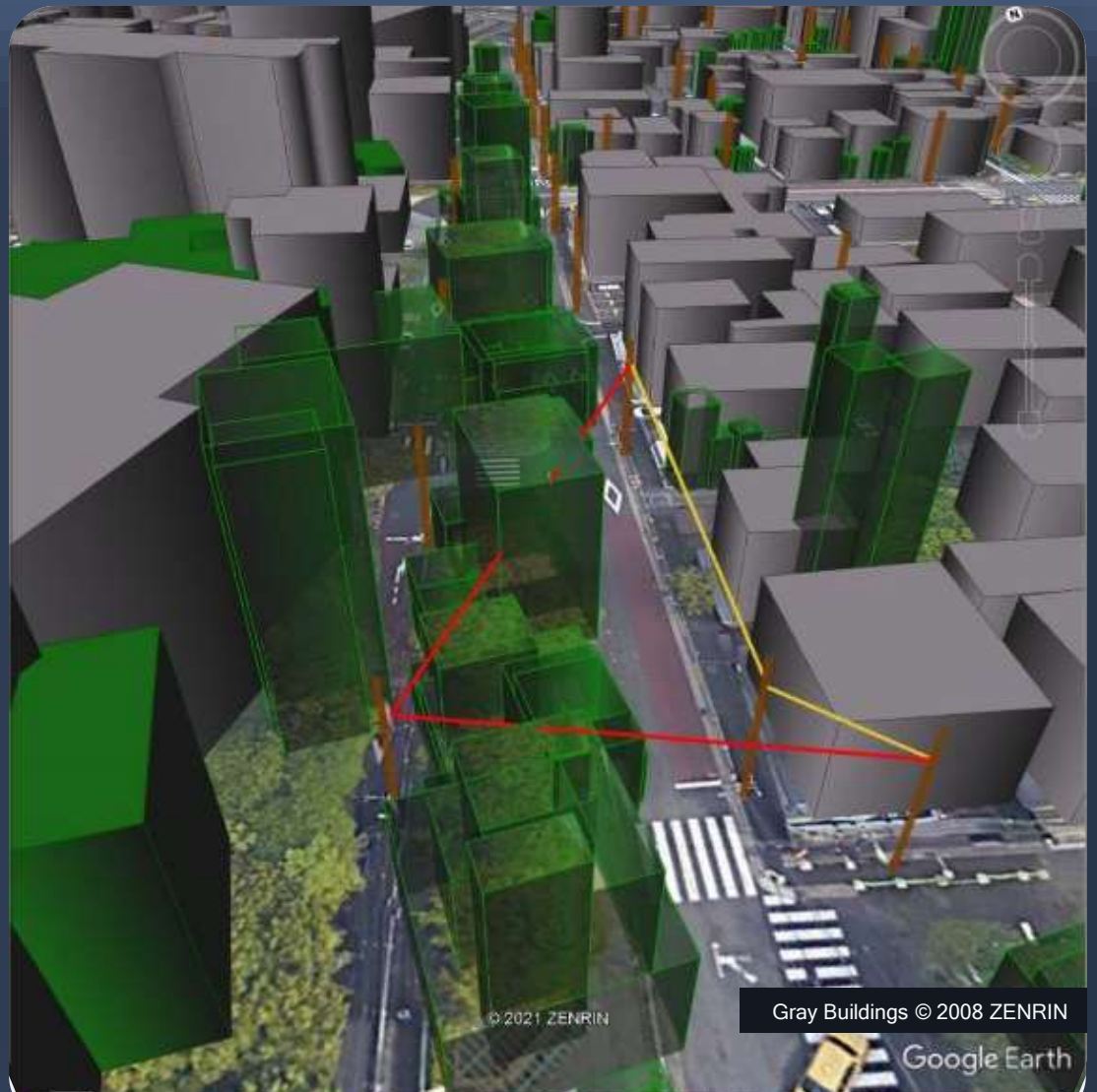
Diversity of node types e.g., small cells, IAB, repeaters, …

Diversity of interconnectivity e.g., fiber, out-of-band wireless, …

Many potential radio locations e.g., for different objectives

Creating a digital twin of targeted deployment using readily available databases

Example in Tokyo, Japan

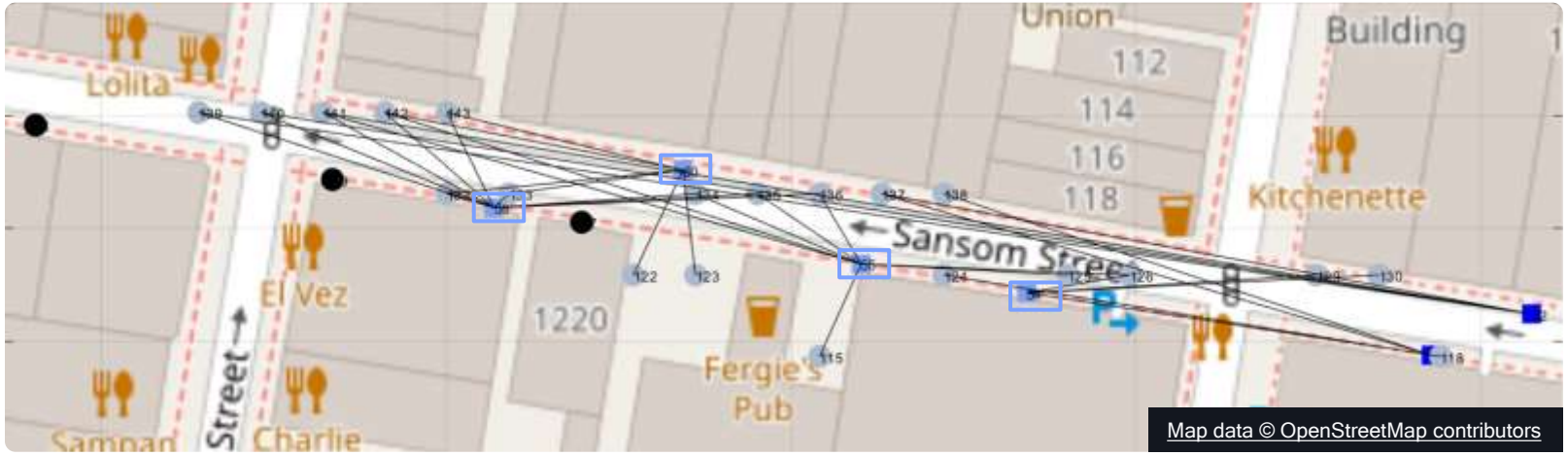


5G mmWave topology optimization example

Philadelphia design using small cells only

Two existing small cells serving a portion of device traffic

Four additional small cells to support DL/UL traffic of remaining devices



Existing small cells

Newly added small cells

Unused poles

Device hotspots

In-band access

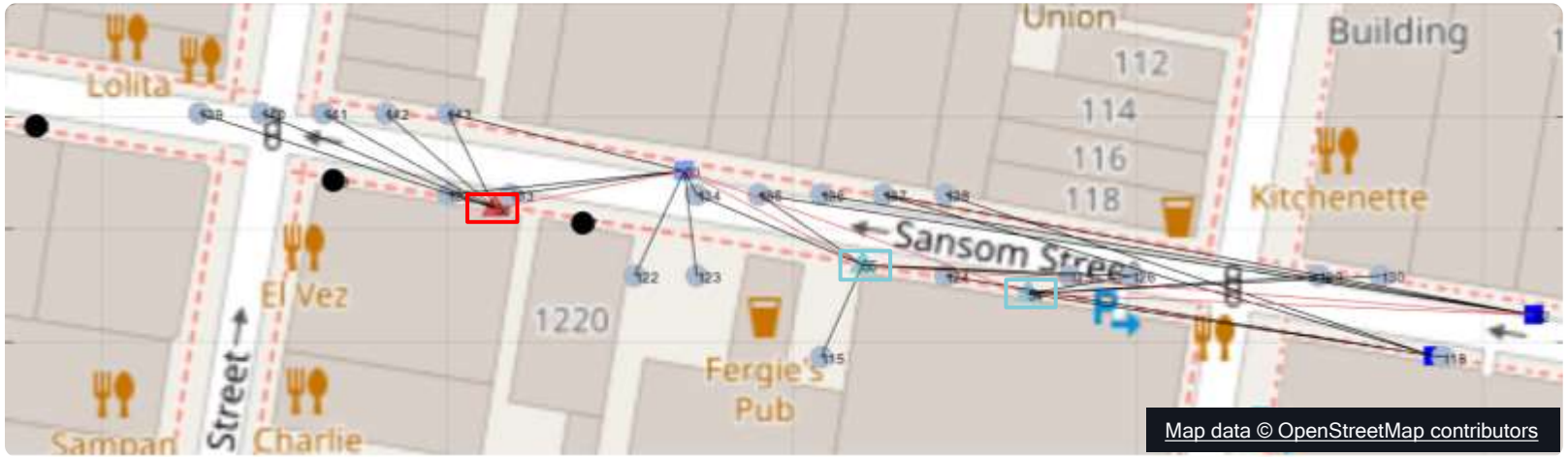
5G mmWave topology optimization example

Design with integrated access/backhaul, repeaters, small cells with in-band and out-of-band backhaul

Three new small cells are replaced with one IAB and two repeaters

One repeater is placed to bend signal around the corner to provide coverage

IAB now provides coverage to devices on the left, as new small cell cannot meet demand of those devices with its in-band bandwidth



Existing small cells

Newly added small cells

Unused poles

Device hotspots

Repeaters

Integrated access/ backhaul

In-band backhaul

Out-of-band backhaul

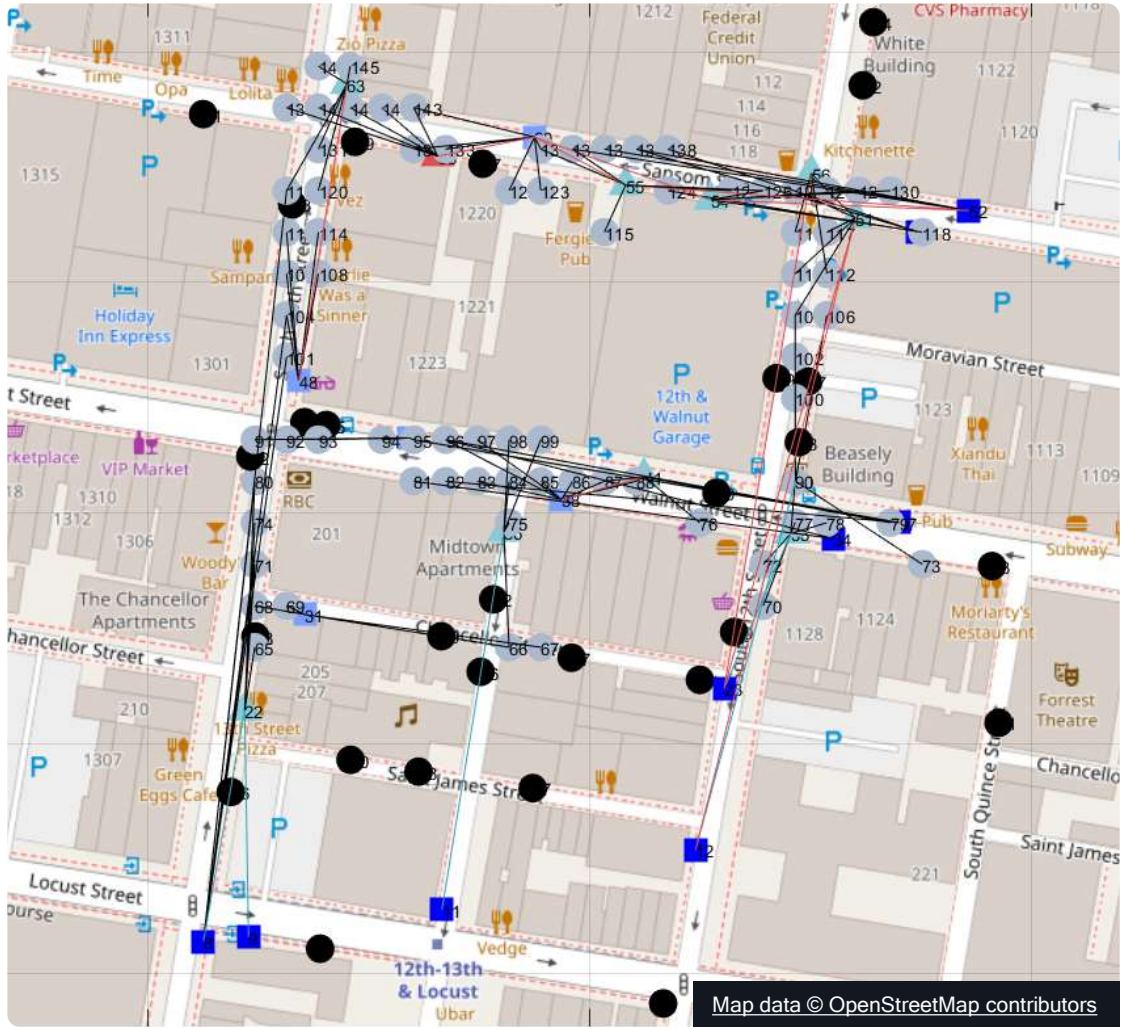
In-band access

Expanding design to larger deployment area with diverse mmWave topologies

Refreshed design based on new traffic requirements of the larger area

Additional small cells, repeaters, IABs to address increased traffic demand

Utilizing both in-band and out-of-band backhaul links for even more capacity



Existing small cells

Newly added small cells

Unused poles

Device hotspots

Repeaters

Integrated access/ backhaul

In-band backhaul

Out-of-band backhaul

In-band access

Machine learning

plays an expanding role in the evolution
of 5G towards 6G

Advancing 5G to fulfill its full promise

Enhanced mobile experiences, new capabilities, and expansion to diverse verticals

Standardizing AI/ML in cellular communication systems

Broad range of work across standards and industry organizations

Developing AI use cases

Architectural framework for ML

Framework for evaluating intelligence level

Framework for data handling to enable ML

AI for autonomous and assisted driving

Developing an AI Mobile Device Requirement Spec (TS.47)

Focusing on AI mobile phone and tablet (may extend to IoT/wearable in future releases)

Data collection for network performance enhancements (RAN2/3 – Rel-16/17)

Study on AI/ML functional frameworks and use cases (RAN3 – Rel-17)

Network data analytic function for core AI/ML use cases (SA2 – Rel-16/17)

Management data analytic service, autonomous network (SA5 – R17)

Study on AI/ML model transfer performance requirements over 5G (SA1 – Rel-17)

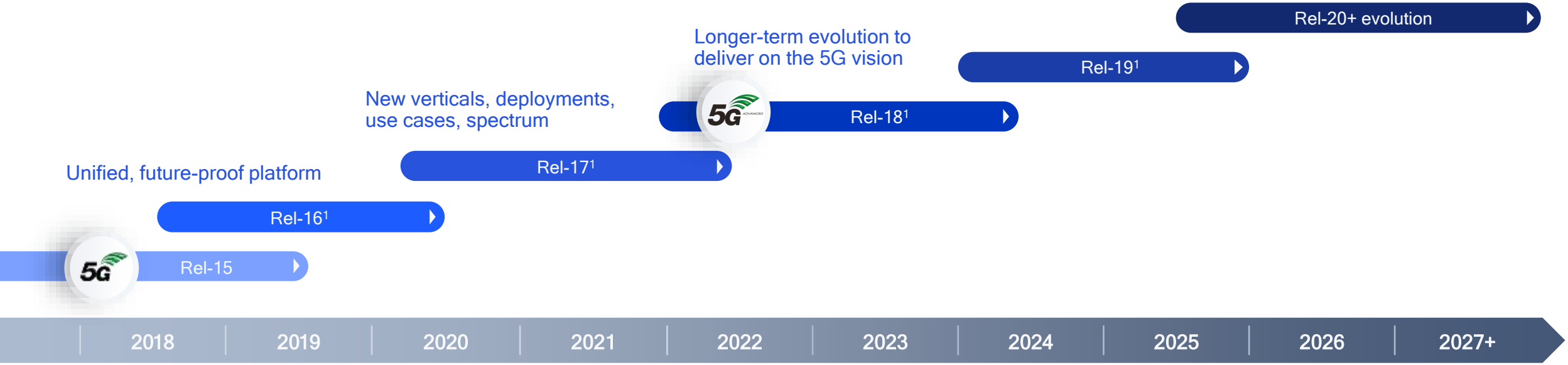
Defining reference architecture for Radio Intelligence Controller (RIC) and interfaces

Developing technical report: “AI/ML Workflow Description and Requirement”

Network automation and autonomy based on AI

Defining requirements and platform recommendations for reference implementation and interface standards

Driving the 5G technology evolution in the new decade



Rel-15 eMBB focus

- 5G NR foundation
- Smartphones, FWA, PC
- Expanding to venues, enterprises

Rel-16 industry expansion

- eURLLC and TSN for IIoT
- NR in unlicensed
- 5G V2X sidelink multicast
- In-band eMTC/NB-IoT
- Positioning

Rel-17 continued expansion

- Lower complexity NR-Light
- Higher precision positioning
- Improved IIoT, V2X, IAB, and more…

Rel-18+ 5G-Advanced

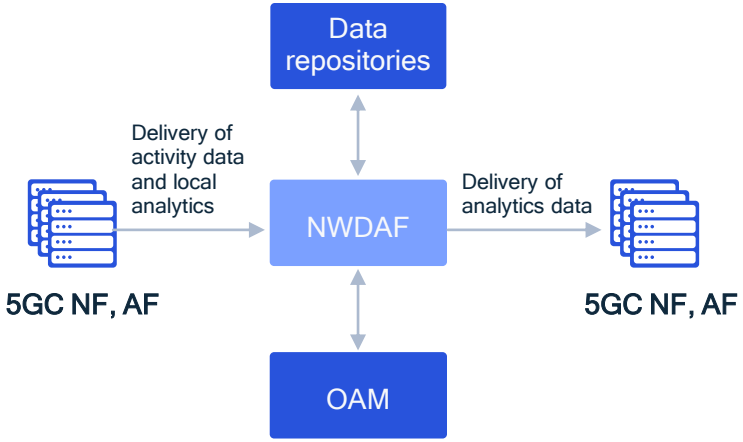
- Next set of 5G releases (i.e., 18, 19, 20, …)
- Potential projects in discussions
- Rel-18 expected to start in 2022

1. 3GPP start date indicates approval of study package (study item->work item->specifications), previous release continues beyond start of next release with functional freezes and ASN.1

Data collection for network performance enhancements

Part of 3GPP Release 16

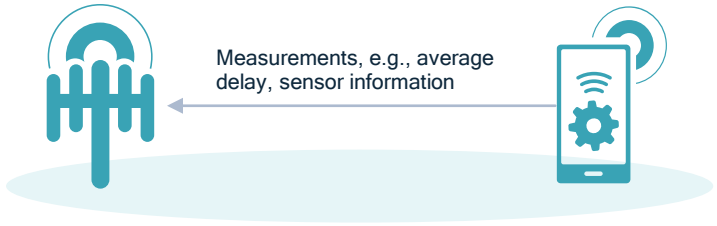
Enhanced Network Automation (eNA)



New enhanced core network function for data collection and exposure

Expanding NWDAF[1] from providing network slice analysis in Rel-15 to data collection and exposure from/to 5G core NF, AF, OAM[2], data repositories

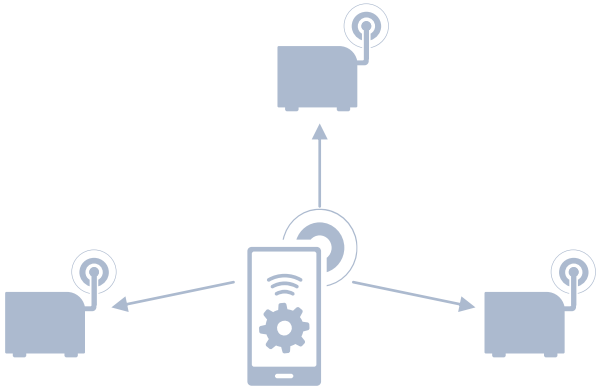
Minimization of Drive Testing (MDT)



Logged and immediate MDT, mobility history information, accessibility and L2 measurements[3]

Specifying features for identified use cases, including coverage, optimization, QoS verification, location information reporting, sensor data collection

Self Organizing Network (SON)



Mobility robust optimization (MRO), mobility load balancing (MLB), and RACH optimization

Specifying device reporting needed to enhance network configurations and inter-node information exchange (e.g., enhancements to interfaces like N2, Xn)

1 NWDAF – Network Data Analytics Function; 2 Network Function, Application Function, Operations Administration and Maintenance; 3 For standalone and dual connected 5G NR systems

Expanding 5G system support for wireless machine learning

Part of 3GPP Release 17

Enhancements for 5G network interfaces

Facilitating machine learning procedures such as model training and inference, as well as actions to enforce model inference output

Augmented network and device data collection

Supporting targeted applications (e.g., energy saving, load balancing, mobility management), operations enhancements, expanded use case[1]

Support for over-the-top AI/ML services

Introducing new QoS (Quality of Service) definitions that are tailored for machine learning model delivery over 5G

1 Such as multicast, broadcast, V2X, sidelink, multi-SIM, RAN slicing, and more

Network architecture enhancements

Allowing for machine learning to run over different HW/SW and future RAN function split to improve flexibility and efficiency

AI/ML procedure enhancements

Optimizing system for model management, training (e.g., federated and reinforcement learning), and inference

Data management enhancements

Standardizing ML data storage/ access, data registration/ discovery, and data request/ subscription

New and expanded use cases

Supporting traffic/mobility prediction, coverage/capacity optimization, massive MIMO, SON, CSI feedback, beam management, and other PHY/MAC and upper layer improvements

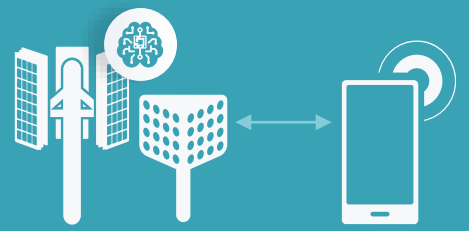
5G ADVANCED

5G Advanced (Rel-18+) targets to expand wireless machine learning to the end-to-end system across RAN, device, and air interface

Machine learning can bring continuous wireless enhancements

AI-native air interface design can enable continual system improvements in between major 3GPP releases through self-learning

No standardized improvement during nominal Work/Study Item phase towards subsequent release



Data-driven communication and network design

Data-driven system configuration provides end-to-end optimizations

Dynamic parameter adaptation based on fast machine learning algorithms

Neural network system design can customize to given wireless environment

Machine learning is a key technology vector on the path to 6G

Wireless AI/ML

Data-driven communication and network design, with joint training, model sharing and distributed interference across networks and devices

Scalable network architecture

Disaggregation and virtualization from cloud-to-edge, and use of advanced relay/mesh topologies to address growing demand

New radio designs

Waveform/coding for MHz to THz, intelligent surfaces, joint comms. and sensing, large-scale MIMO, advanced duplexing, energy-efficient RF

Communications resiliency

End-to-end configurable security, post quantum security, robust networks tolerant to failures and attacks

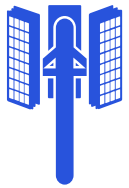
Coordinated spectrum sharing

New paradigms for more efficient use of spectrum, leveraging location /environmental awareness for dynamic/adaptive coordination

Merging of worlds

Physical, digital, virtual, e.g., ubiquitous, low-power sensing/monitoring, immersive interactions taking human augmentation to next level

Longer term R&D direction



Network

Link parameter prediction
Multi-cell interference learning
Mobility parameter prediction

Fully autonomous networks
Interference coordination/scheduling
Mobility handoff decisions

Air interface

Data-driven
propagation models

Data-driven optimization
of signaling, measurements,
and feedback

Devices

Predictive beam management
Channel state measurements
Device positioning

Joint sensing-communications
Dynamic ML model adaptation
Personalized lifelong learning

E2E approach with machine learning to improve 5G system performance and efficiency

Transition to ML data-driven air interface design and operation

Neural network air interface design for coding, waveform, and multiple access

Joint training, model sharing, and distributed inference across network & devices

Dynamic air interface operation and adaptation

Signal intelligence, baseband and medium access

ML-based channel feedback
Channel estimation & pilot optimization
MIMO detection
Link prediction & adaptation
Beam management and optimization
Spectrum sensing and sharing
Radio resource scheduling

Network intelligence and system optimization

Coverage and capacity optimization
Traffic and mobility prediction
Energy saving
Cooperative edge caching
Content-aware X-layer optimization
Enhanced personalized security
TCP optimization

Device intelligence and optimization

Digital front-end optimization
Antenna and RF optimization
Full duplex
Battery saving
Reflective intelligent surface

Vertical intelligence and other capabilities

High-precision positioning
Environmental sensing
Contextual awareness
Sensor fusion
Vehicular communication

Our AI research areas to advance wireless communication

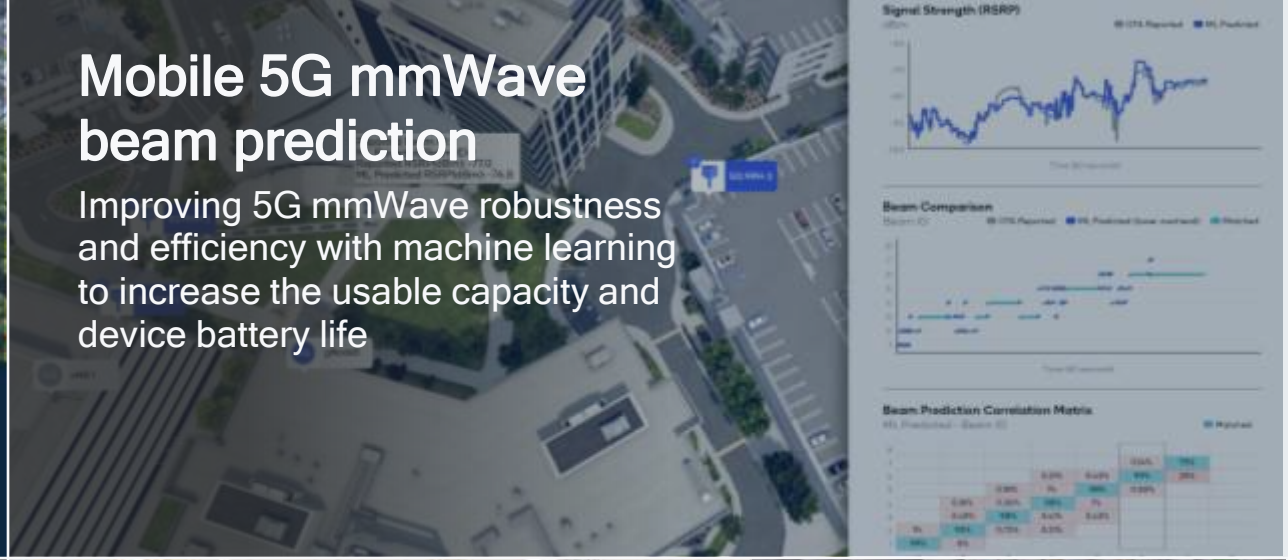
5G+AI

Enhanced 5G massive MIMO channel state feedback

Enabling efficient channel estimation for massive MIMO operations to improve user throughput and system capacity

Mobile 5G mmWave beam prediction

Improving 5G mmWave robustness and efficiency with machine learning to increase the usable capacity and device battery life

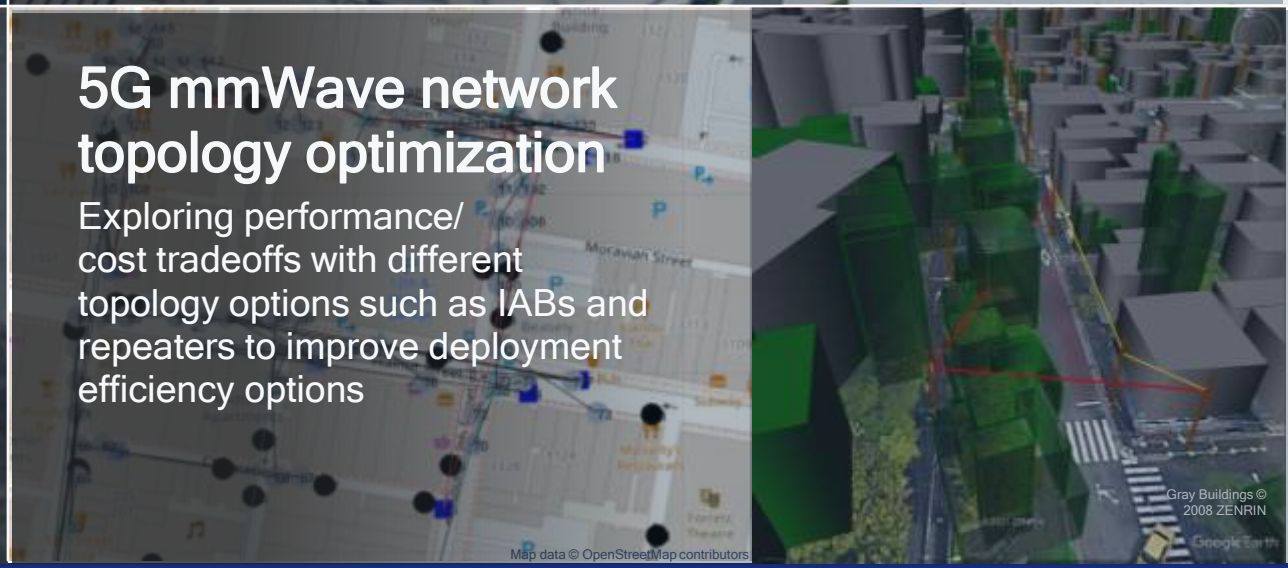


Precise 5G positioning with sensor fusion

Combining 5G positioning measurements with GNSS, multi-path profiles and sensor inputs to improve accuracy

5G mmWave network topology optimization

Exploring performance/ cost tradeoffs with different topology options such as IABs and repeaters to improve deployment efficiency options



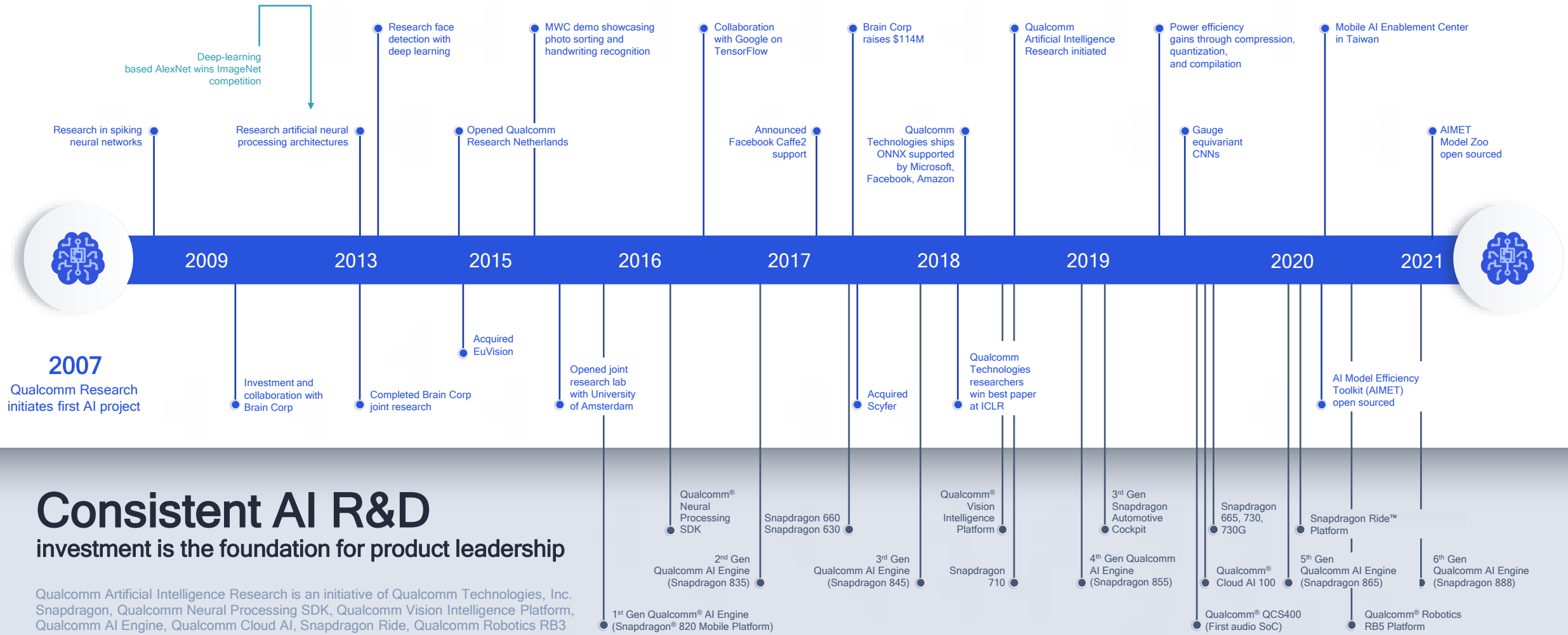
Qualcomm

Showcasing advanced wireless machine learning at MWC21

5G OTA prototypes and system simulations

Our AI leadership

Over a decade of cutting-edge AI R&D, speeding up commercialization and enabling scale



Consistent AI R&D

investment is the foundation for product leadership

AI Highlights

2nd gen Qualcomm® Sensing Hub

Dedicated AI accelerator

First to support TensorFlow Micro

Industry leading AI use cases

Super movie with Tetras.AI

Snapchat lenses accelration

NLP with Hugging Face

Skin condition detection with trinamiX

Qualcomm Neural Processing SDK & AI Model Efficiency Toolkit

New features and improvements

Qualcomm AI Engine direct

Easier and faster access to the entire AI Engine

Qualcomm® Hexagon™ 780 Processor

Fused AI Accelerators

- Tensor - 2X compute capacity
- Scalar - 50% performance improvement
- Vector – Support for additional data types

3X performance per watt improvement

16X dedicated memory

Up to 1000X hand off time improvement in certain use cases

6th gen Qualcomm AI Engine

26 TOPS

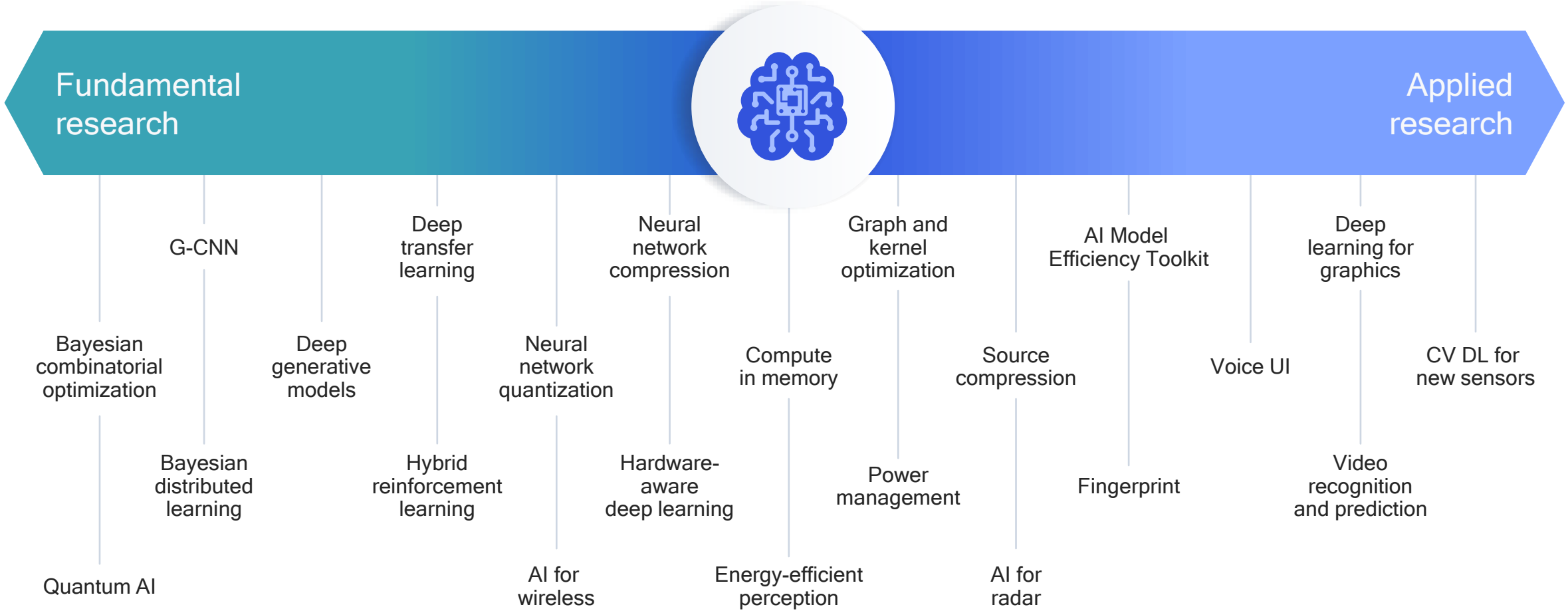
TVM Opensource

More efficient coding

*Compared to previous generations

Leading research and development

Across the entire spectrum of AI



Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

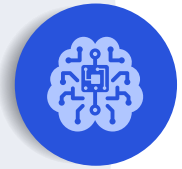
Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion



Reasoning

Scene understanding, language understanding, behavior prediction

Action

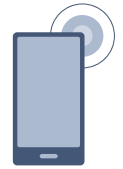
Reinforcement learning for decision making

Edge cloud

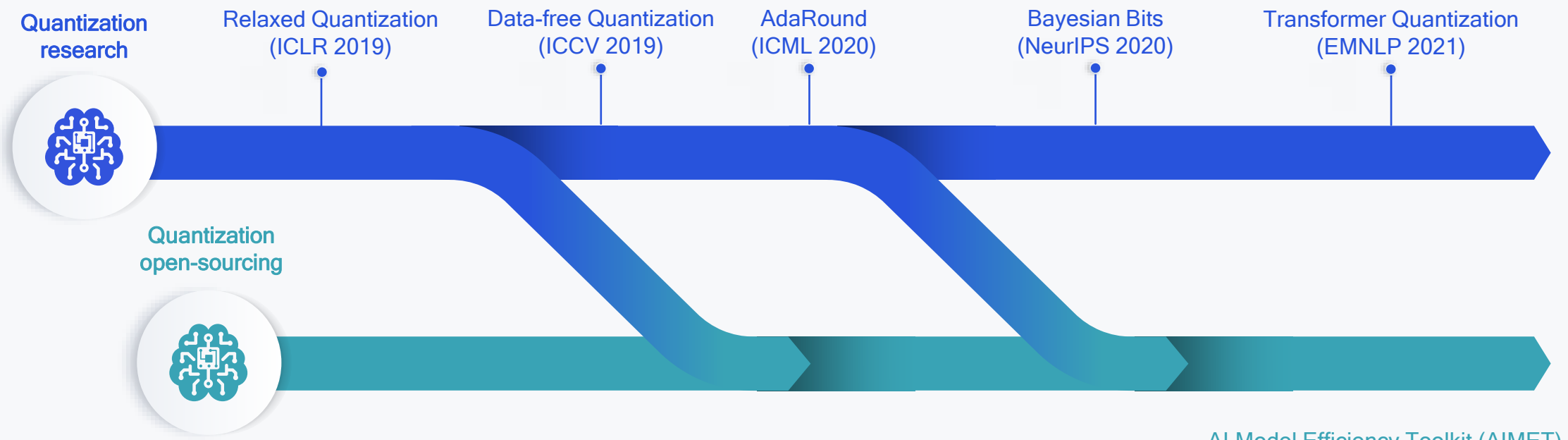
Cloud

IoT/IIoT

Automotive



Mobile



AIMET and AIMET Model Zoo are products of Qualcomm Innovation Center, Inc.

Leading AI research and fast commercialization

Driving the industry towards integer inference and power-efficient AI

AIMET makes AI models small

Open-sourced GitHub project that includes state-of-the-art quantization and compression techniques from Qualcomm AI Research

Trained
AI model

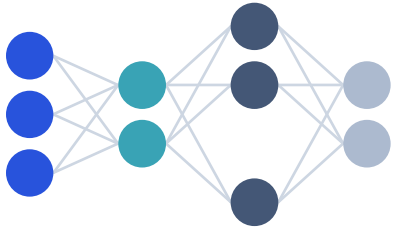
TensorFlow or PyTorch

AI Model Efficiency Toolkit
(AIMET)

Compression

Quantization

Optimized
AI model



Deployed
AI model

If interested, please join the AIMET GitHub project: https://github.com/quic/aimet

Features:

- State-of-the-art network compression tools
- State-of-the-art quantization tools
- Support for both TensorFlow and PyTorch
- Benchmarks and tests for many models
- Developed by professional software developers

AIMET

Providing advanced model efficiency features and benefits

Benefits


Lower power


Lower memory bandwidth


Maintains model accuracy


Lower storage


Higher performance


Simple ease of use

Features

Quantization

State-of-the-art INT8 and INT4 performance

Post-training quantization methods, including Data-Free Quantization and Adaptive Rounding (AdaRound) – *coming soon*

Quantization-aware training

Quantization simulation

Compression

Efficient tensor decomposition and removal of redundant channels in convolution layers

Spatial singular value decomposition (SVD)

Channel pruning

Visualization

Analysis tools for drawing insights for quantization and compression

Weight ranges

Per-layer compression sensitivity

AIMET Model Zoo

Accurate pre-trained 8-bit quantized models

Image classification

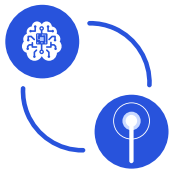
Semantic segmentation

Super resolution

Object detection

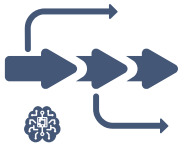
Pose estimation

Speech recognition



5G and AI are two synergistic, essential ingredients that are fueling future innovations

Applying AI techniques to solve difficult wireless challenges and deliver new values



Machine learning plays an expanding role in the evolution of 5G towards 6G

Qualcomm

The essential role of AI in the 5G future

How machine learning is accelerating wireless innovations in the new decade and beyond

Qualcomm

Thank you

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog