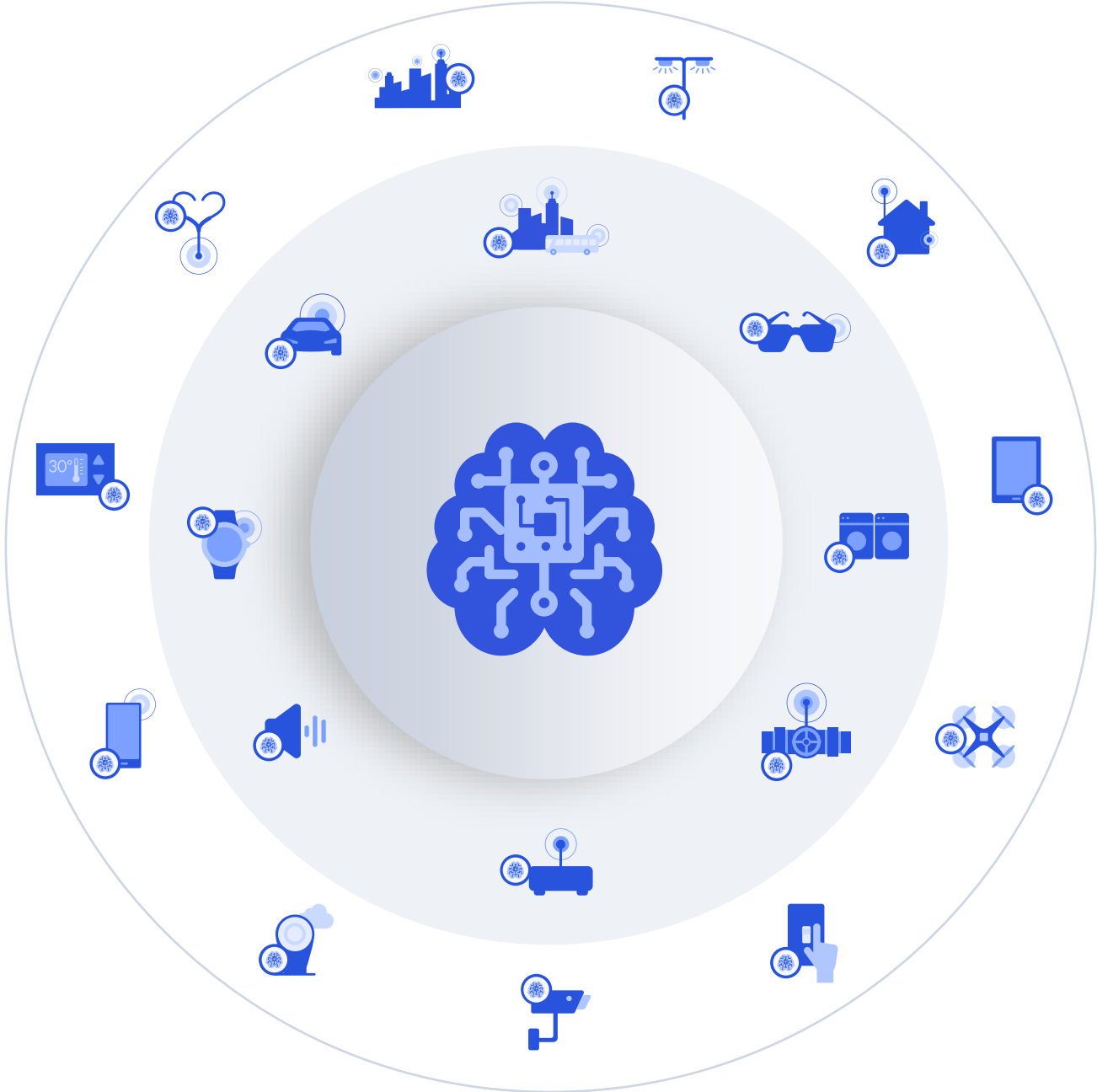


Making AI ubiquitous

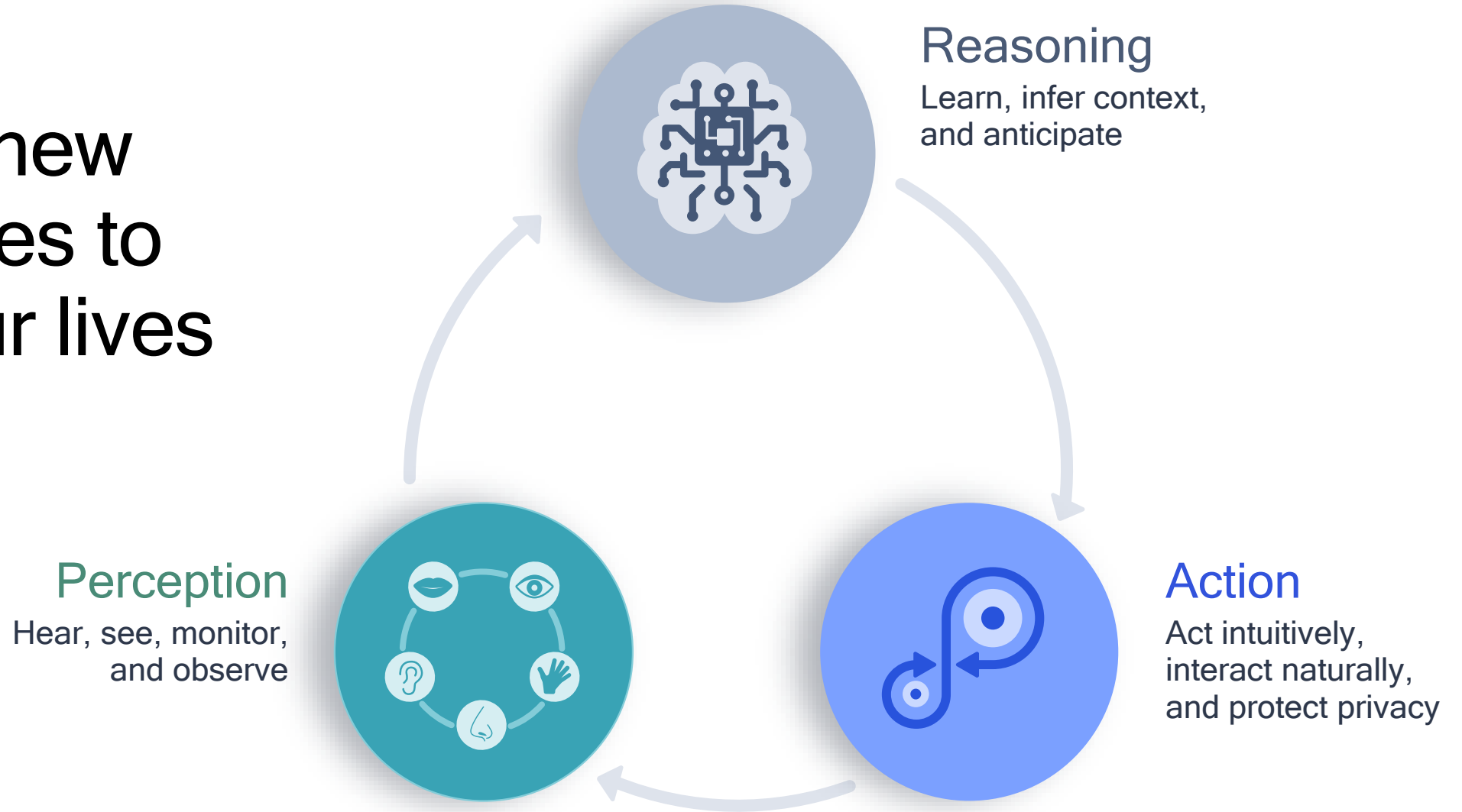
Qualcomm Technologies, Inc.



Devices, machines,
and things are becoming
more intelligent



Offering new capabilities to enrich our lives



Smartphone



Smart homes



Video conferencing



Autonomous vehicles



Smart factories



Extended reality



Smart cities



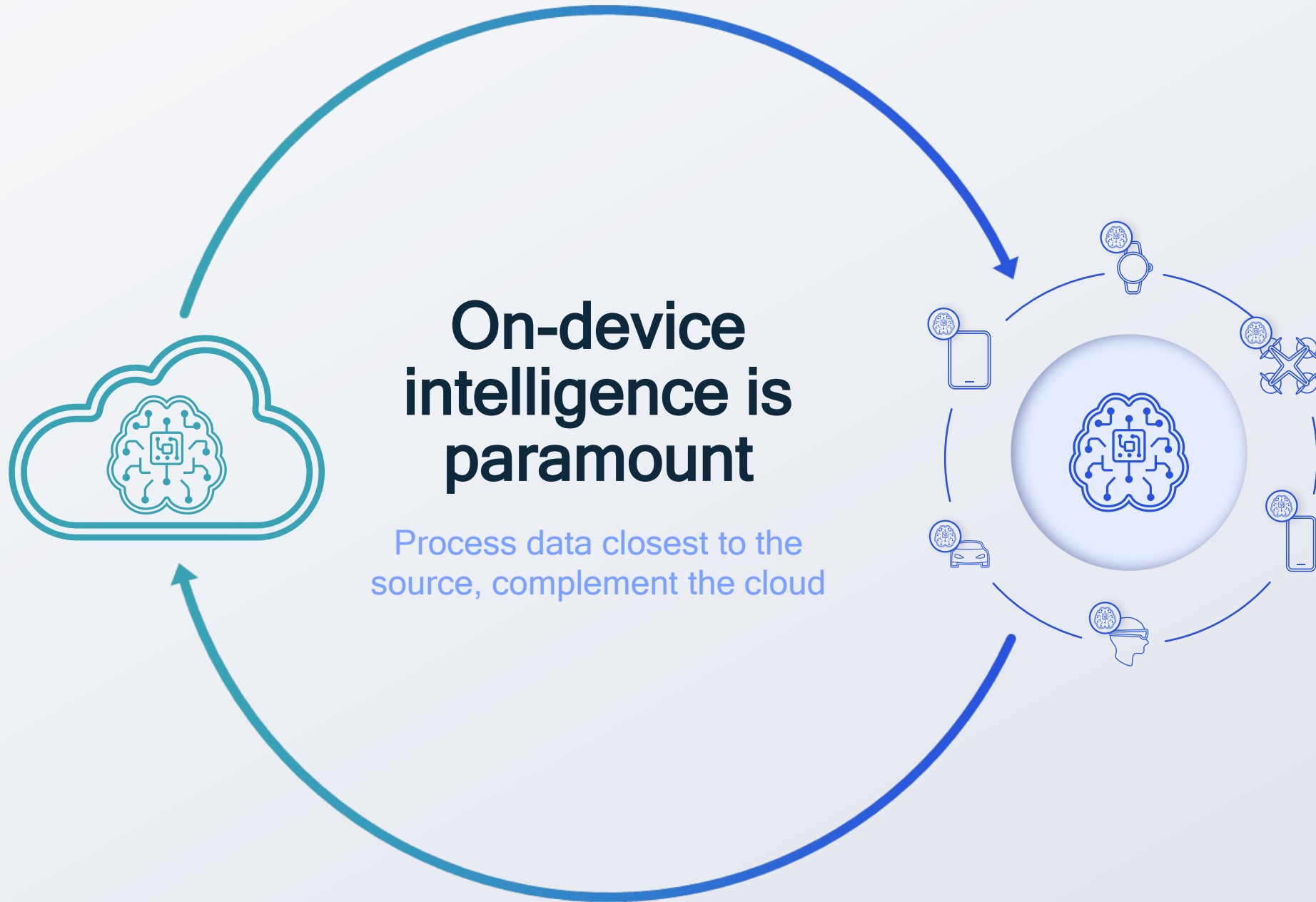
Video monitoring



The need for intelligent, personalized experiences powered by AI is ever-growing

A world where virtually everyone and everything is intelligently connected





Privacy

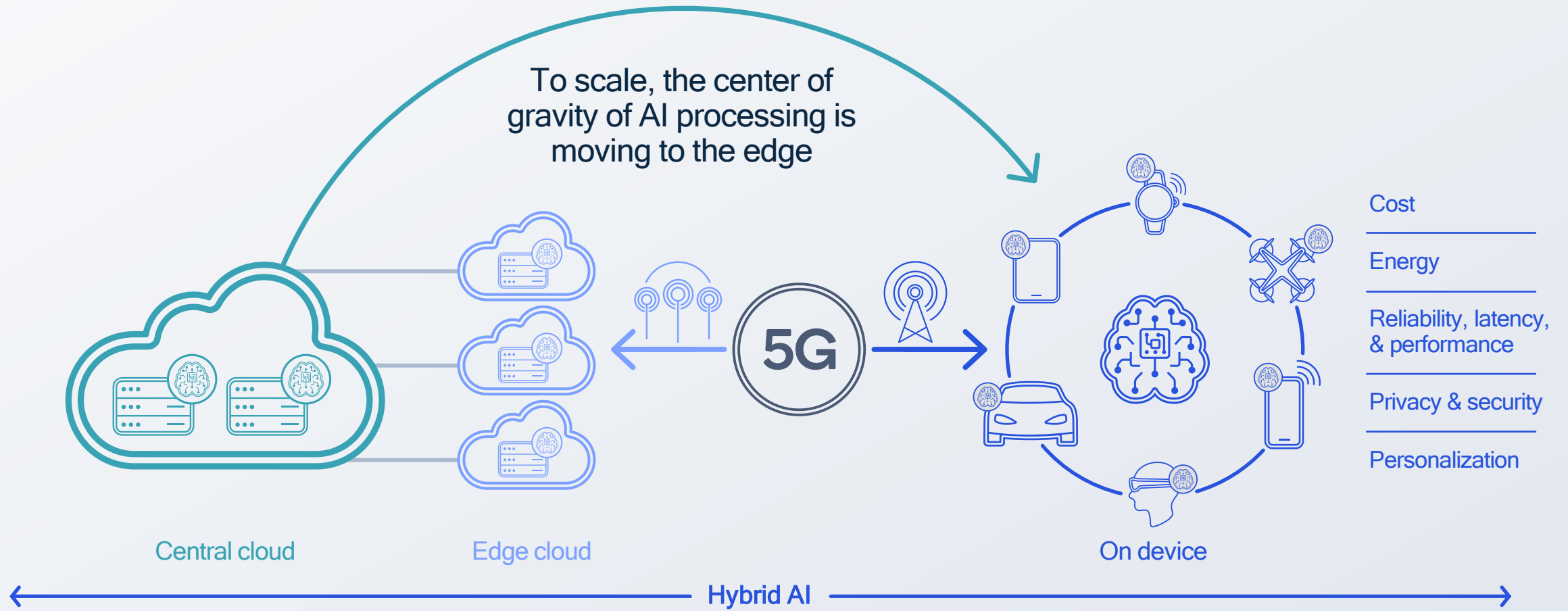
Reliability

Low latency

Cost

Energy

Personalization



We are leading the realization of the hybrid AI

Convergence of:

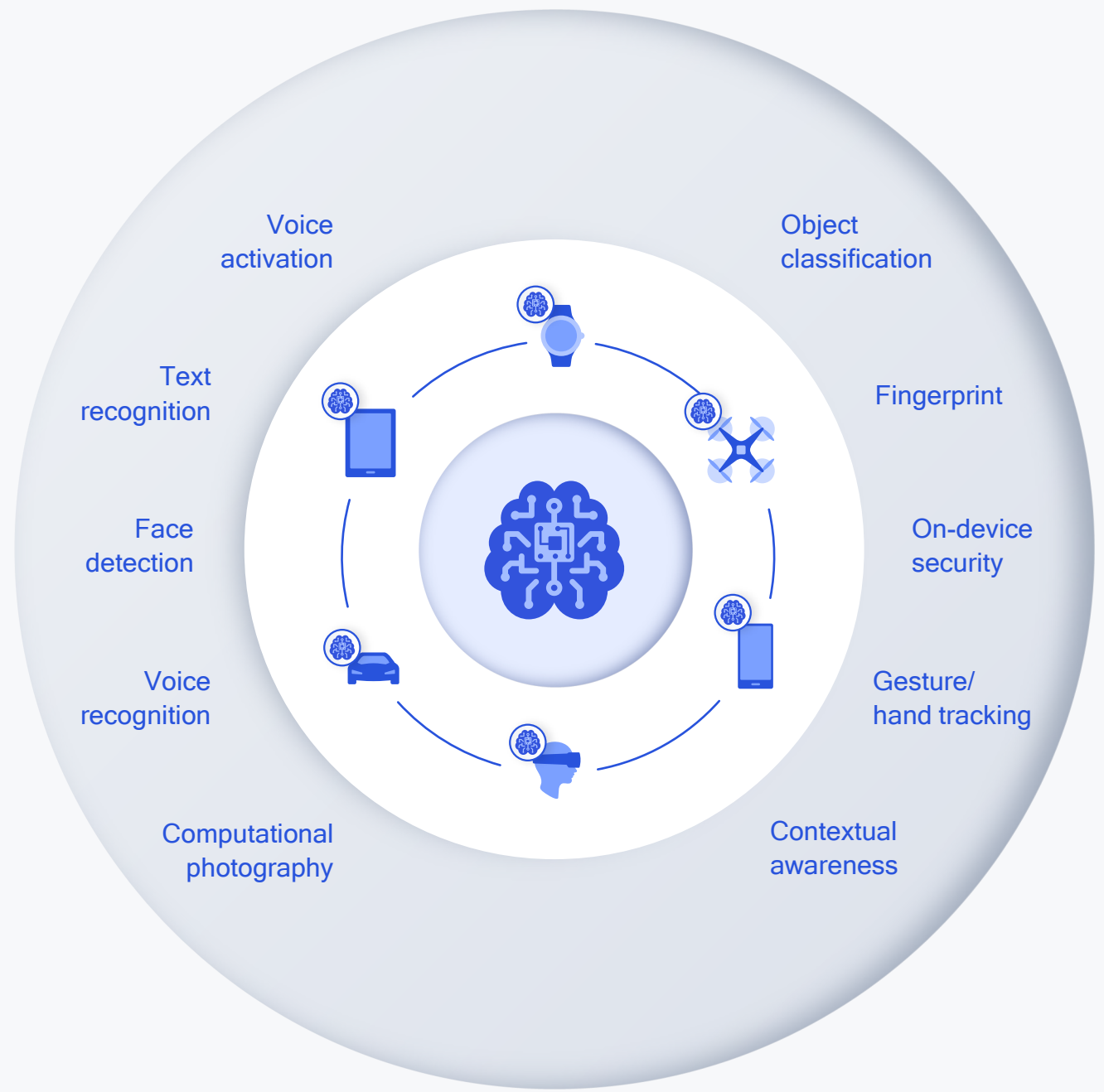
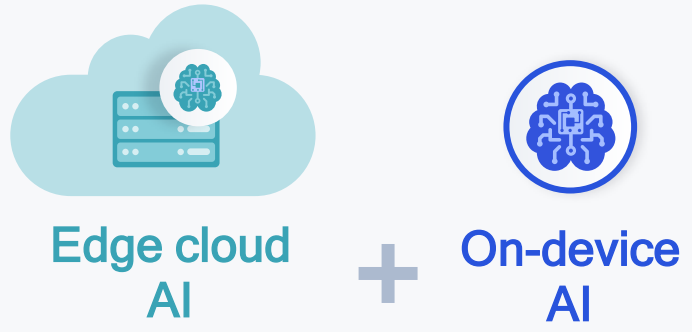
- Wireless connectivity
- Efficient computing
- Distributed AI

Unlocking the data that will fuel our digital future and generative AI

- 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



On-device intelligence is quickly gaining momentum

Key segments are expected to see full AI attach rates by 2025

10%

AI attach rate



2018

100%

AI attach rate



2025



Mobile



Automotive



XR



PCs



Smart speakers

Mobile is the
pervasive AI
platform

~8.6
Billion

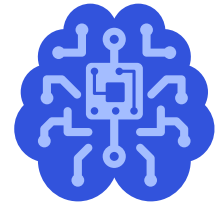
Cumulative smartphone
unit shipments forecast
between 2020-2025

Source: IDC June'21



Mobile scale
changes everything

Bringing AI to the masses



Qualcomm

Mobile computing



Smart cities



Smart homes



Automotive



Smartphones



Healthcare



Industrial IoT



Networking



Wearables



Extended reality



Rapid replacement cycles

Superior scale

Integrated/optimized technologies

AI offers enhanced experiences and new capabilities for smartphones

True personal assistance



Extended battery life



Enhanced connectivity



Superior photography



Natural user interfaces

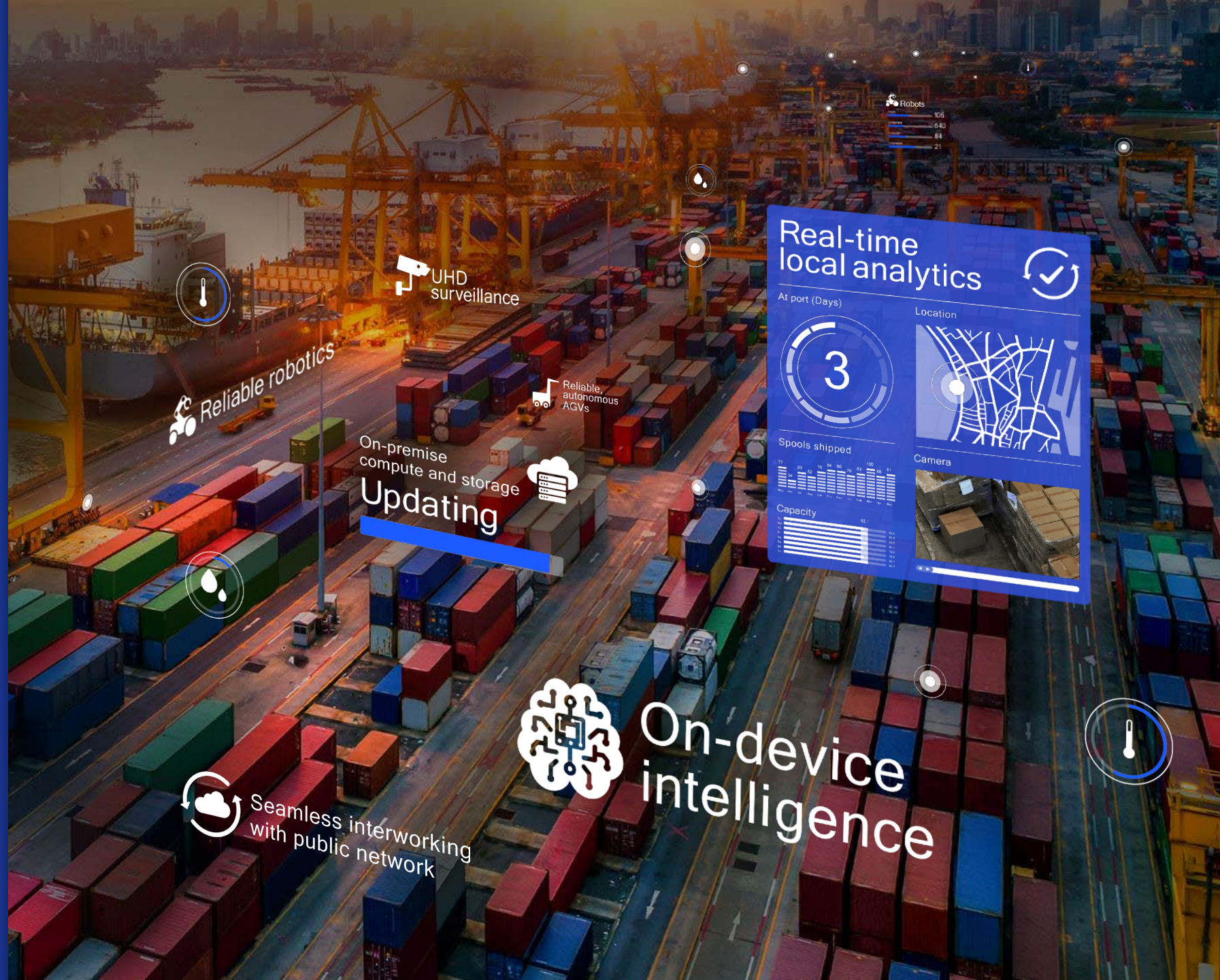


Enhanced security

A new development paradigm
where things repeatedly improve



AI will drive transformation across industries



UHD surveillance

Reliable robotics

Reliable autonomous AGVs

On-premise compute and storage

Updating

Real-time local analytics

At port (Days)

3

Location

Spools shipped

| | | | | | |
|----|----|----|----|-----|----|
| 77 | 85 | 92 | 98 | 100 | 81 |
| 34 | 52 | 78 | 94 | 95 | 81 |

Capacity

| | | | | | | | | | | | | | |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|---|
| 100 | 92 | 85 | 78 | 70 | 62 | 55 | 48 | 40 | 32 | 25 | 18 | 10 | 2 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|---|

Camera



On-device intelligence



Seamless interworking with public network



Boundless XR experiences





Shaping the future of transportation

Personalized driver settings

Driver awareness monitoring

Greater autonomous capabilities





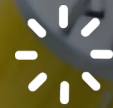
Powering the factory of the future



Surveillance



XR Guided execution



Ultra reliable,
low-latency wireless
connection



Dynamic factory
reconfigurability



5G NR
Private network



Real-time
supply chain
visibility



Predictive maintenance



Autonomous manufacturing and robotics



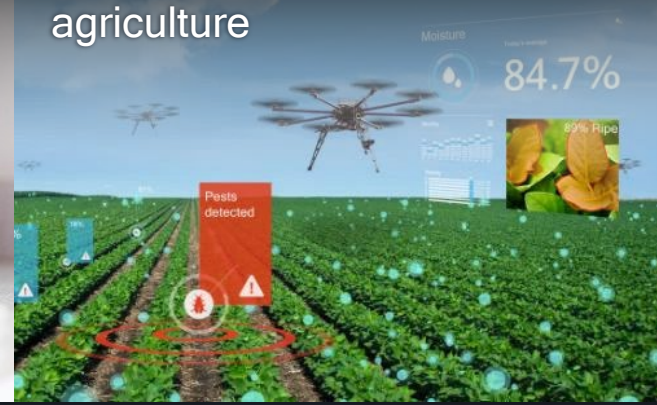
Smart security for home and enterprise



Smart displays and speakers



Smarter agriculture



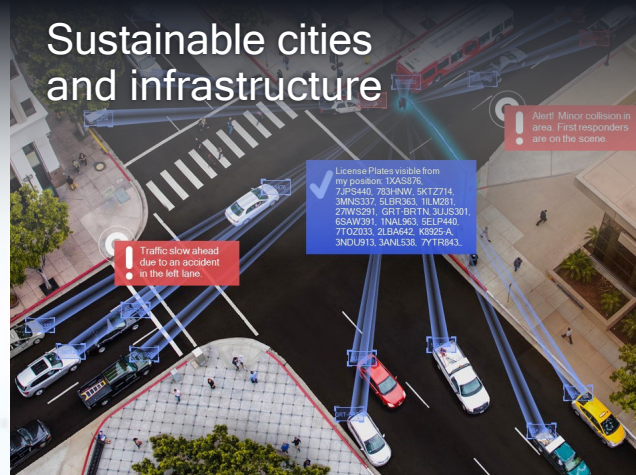
More efficient use of energy and utilities



Home hubs and smart appliances



Sustainable cities and infrastructure



Digitized logistics and retail





AI for IoT across the home, industrial/enterprise, and smart cities

The challenge of AI workloads

 Very compute intensive

 Large, complicated neural network models


 Complex concurrencies


 Real-time


 Always-on



Constrained mobile environment

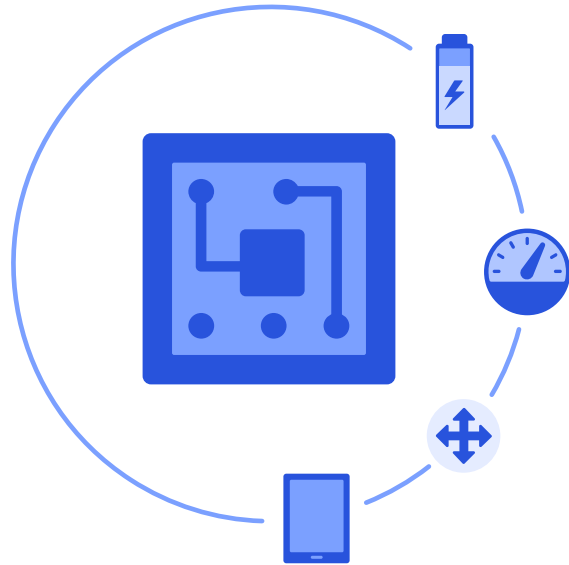
Must be thermally efficient for sleek, ultra-light designs 

Requires long battery life for all-day use 

Storage/memory bandwidth limitations 

Making power efficient AI pervasive

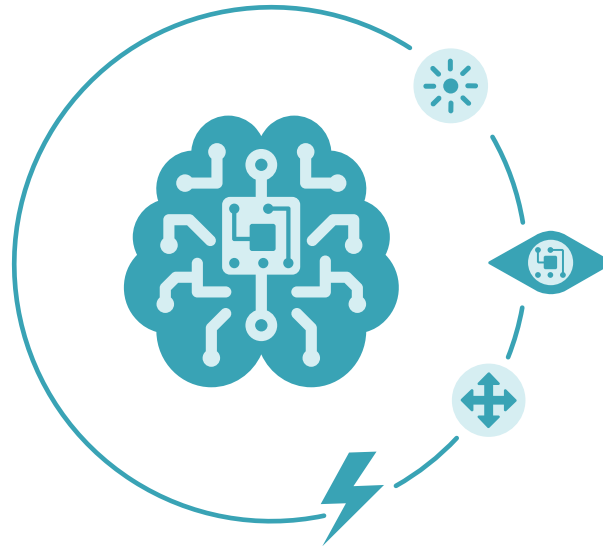
Focusing on high-performance hardware/software and optimized neural network design



Efficient hardware

Developing heterogeneous compute to run demanding neural networks at low power and within thermal limits

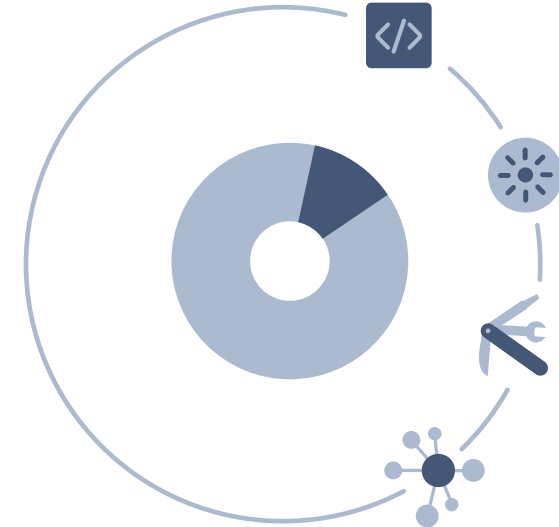
Selecting the right compute block for the right task



Algorithmic advancements

Algorithmic research that benefits from state-of-the-art deep neural networks

Optimization for space and runtime efficiency

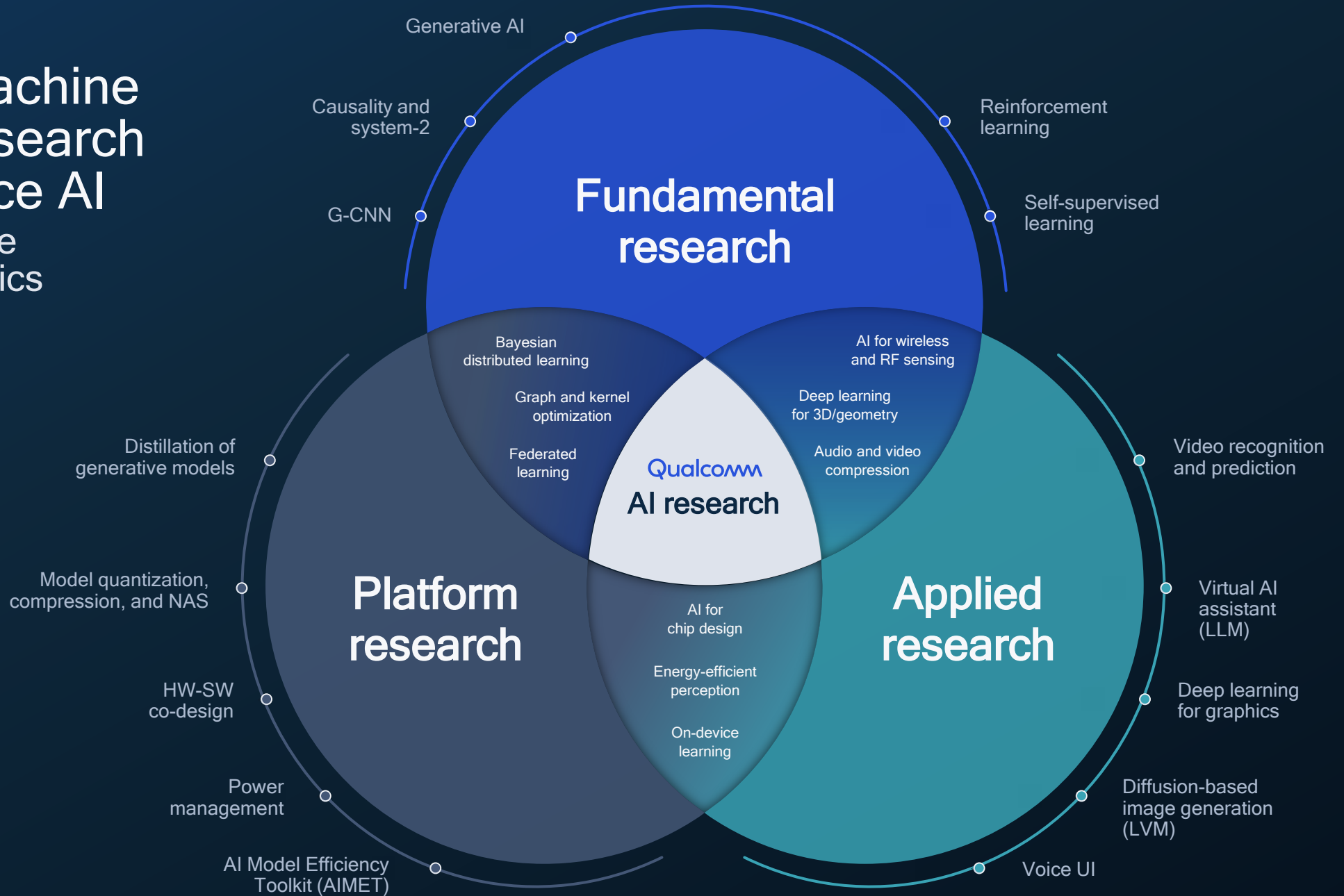


Software tools

Software accelerated run-time for deep learning

SDK/development frameworks

Leading machine learning research for on-device AI across the entire spectrum of topics

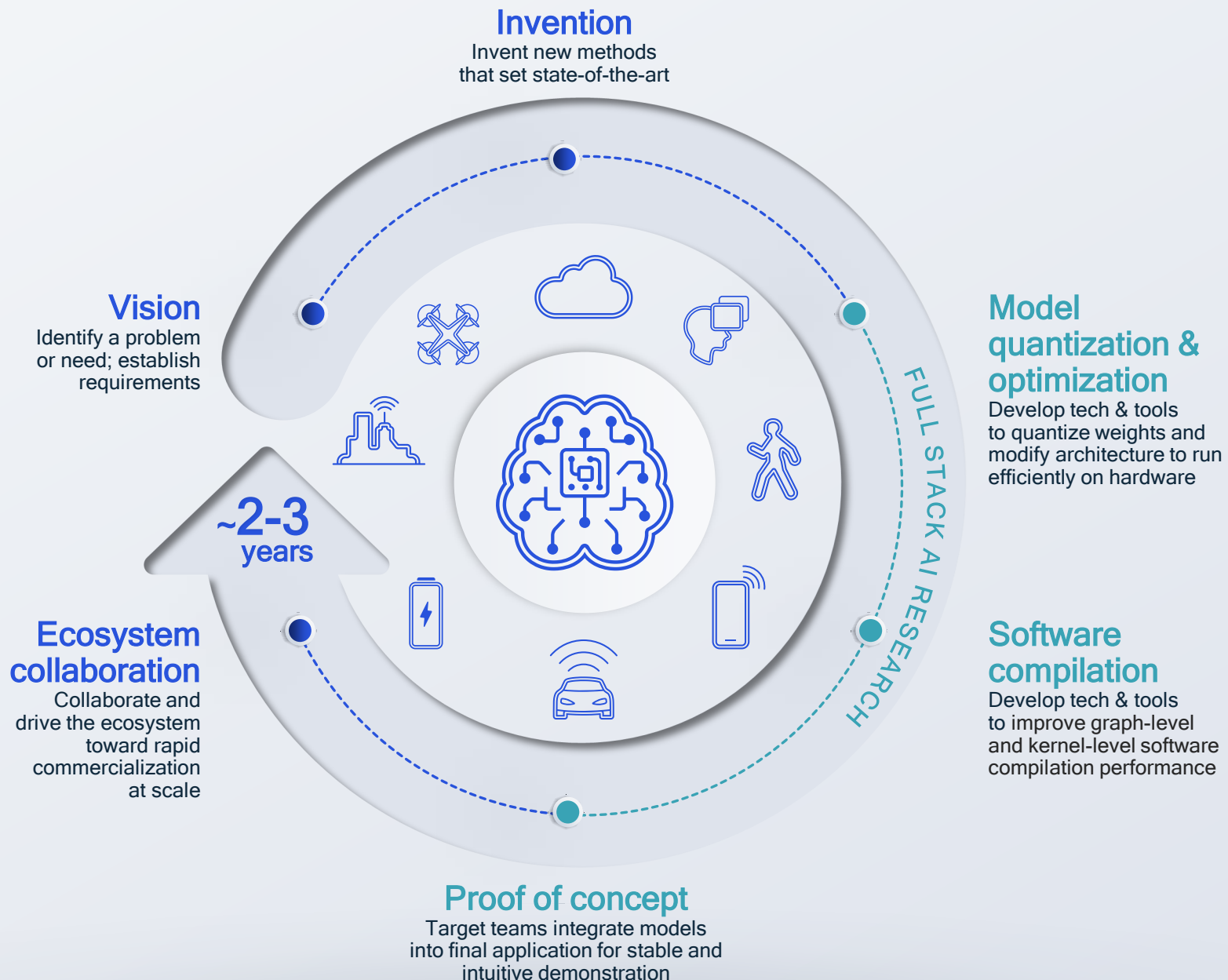


Full-stack AI research & optimization

Model, hardware, and software innovation across each layer to accelerate AI applications

Early R&D and technology inventions essential to leading the ecosystem forward

Transfer tech to commercial teams and influence future research with learnings from deployment



Model quantization

Invented the best techniques for fast deployment of 8-bit quantization



Best power-efficiency toolkit in the industry

On-device learning

Invented continuous learning techniques for SOTA on-device voice-UI



First demonstration of 30% improvement to keyword spotting

Federated learning

Invented methods for combining differential privacy and compression



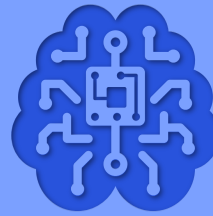
First end-to-end research software framework deployable on mobile

Video semantic segmentation

Top the Cityscape leaderboard with loss function innovation for boundary-awareness



First real-time SS at FHD on mobile



AI Firsts

Brought to you
by Qualcomm
AI Research

Group equivariant CNN

Pioneer for rotational equivariance; best paper at ICLR'18



First G-CNN segmentation for health on mobile

AI for wireless

Invented neural augmentation to enhance physical layer algorithms



First weakly supervised method for real-world passive RF sensing

Video super resolution

Full stack optimization for visual quality improvement at 4K resolution



First 4K SR at 100+ FPS on mobile

Neural video compression

Invented instance-adaptive for SOTA performance & new deployment scenarios



First real-time HD decoding on mobile

Advancing AI research to make edge AI ubiquitous

Power efficiency

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

On-device learning

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



Cloud



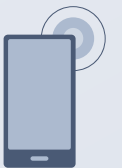
Edge cloud



IoT/IIoT



Automotive



Mobile/XR

Holistic model efficiency research

Multiple axes to shrink
AI models and efficiently
run them on hardware

Quantization

Learning to reduce
bit-precision while keeping
desired accuracy

Compilation

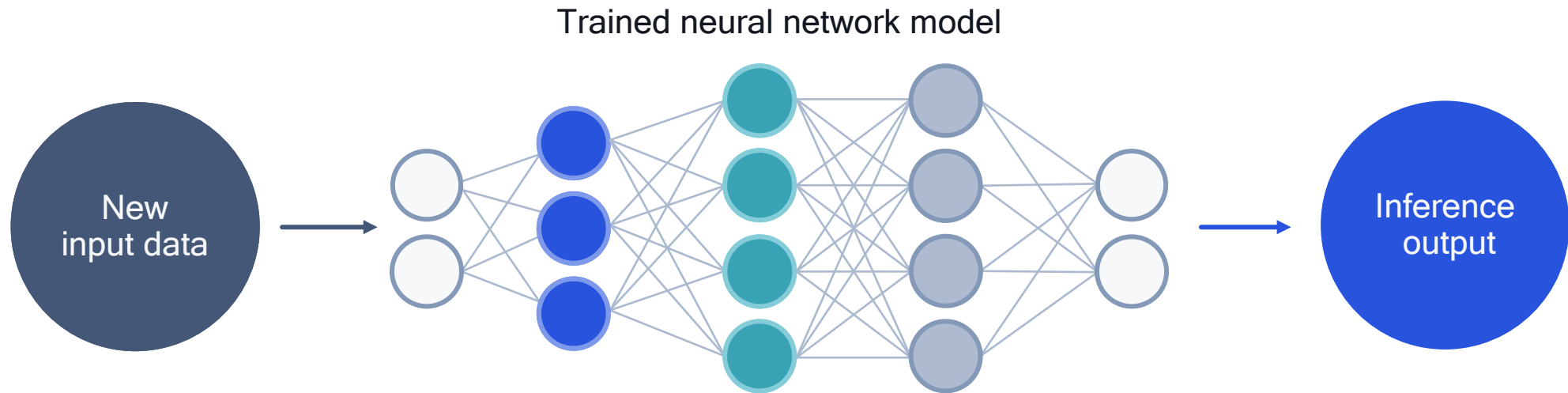
Learning to compile
AI models for efficient
hardware execution

Conditional compute

Learning to execute only parts
of a large inference model
based on the input

Neural architecture search

Learning to design smaller
neural networks that are
on par or outperform
hand-designed
architectures on
real hardware



Compression

Learning to prune model while keeping desired accuracy

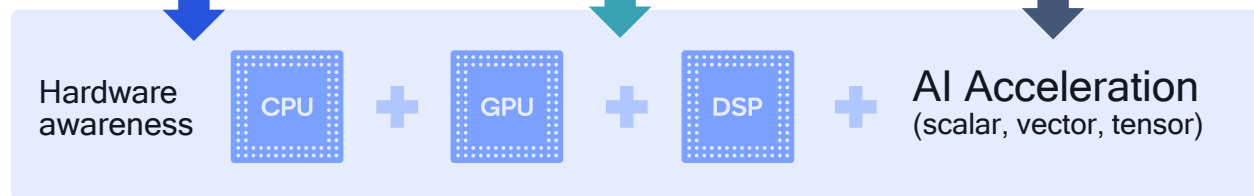
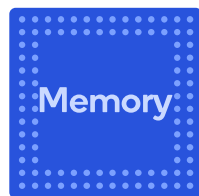
Quantization

Learning to reduce bit-precision while keeping desired accuracy

Compilation

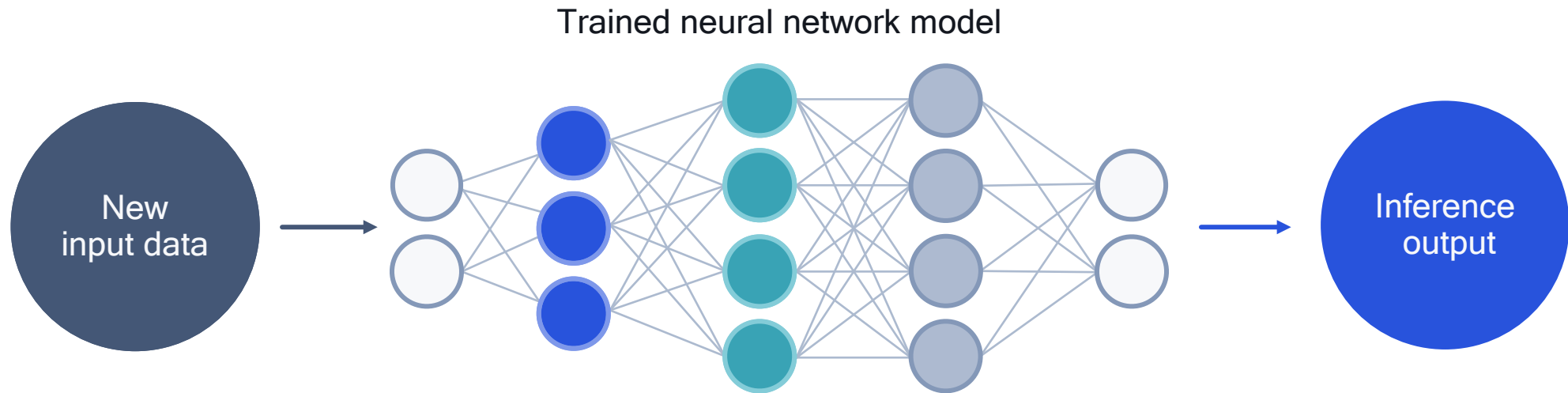
Learning to compile AI models for efficient hardware execution

Applying AI to optimize AI model through automated techniques



Acceleration research
Such as compute-in-memory

Advancing AI research to increase power efficiency



Compression
Learning to prune model while keeping desired accuracy

Quantization
Learning to reduce bit-precision while keeping desired accuracy

Compilation
Learning to compile AI models for efficient hardware execution

Applying AI to optimize AI model through automated techniques

Recent examples

3x

Compression with less than 1% loss in accuracy¹

Up to **16x**

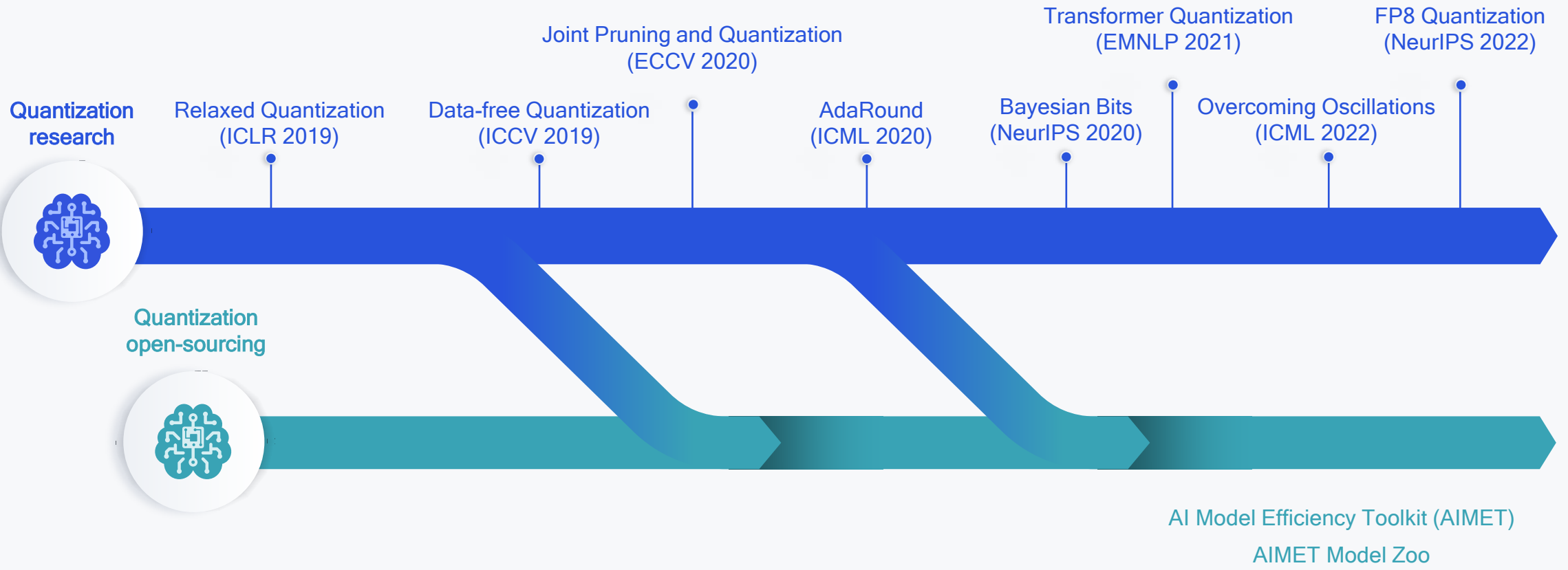
Perf. per watt improvement from savings in memory and compute²

4x

Performance improvement over TensorFlow Lite³

Advancing AI research to increase power efficiency

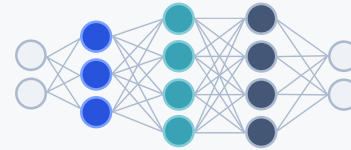
1: With both Bayesian compression and spatial SVD with ResNet18 as baseline. 2: For a quantized INT8 model vs a FP32 model that is not quantized. 3: On average improvement of tested AI models.



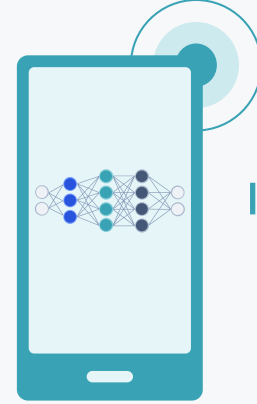
Leading AI research and fast commercialization

Driving the industry towards integer inference and power-efficient AI

What is on-device learning?



Deploy



Inference

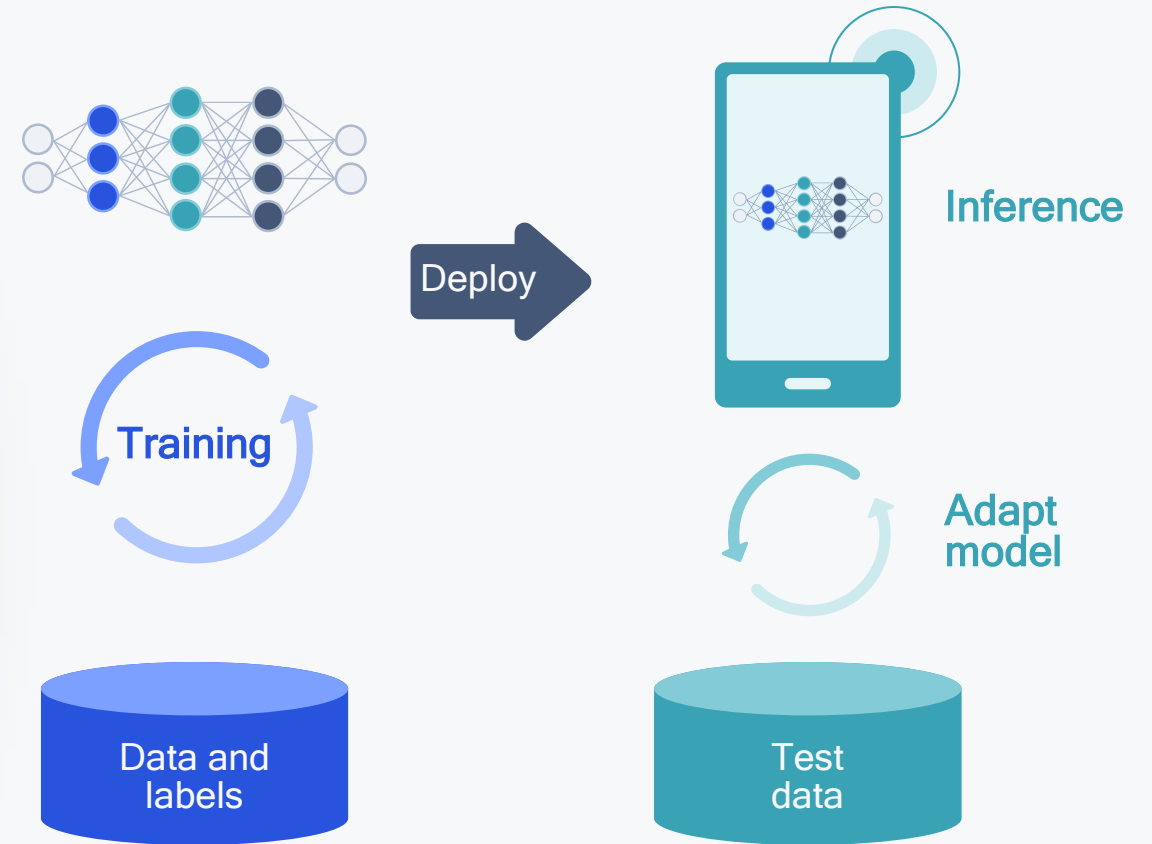


Offline training
A model is trained in the cloud with data reflecting the target application

On-device learning
Modifying model after deployment based on the test environment

On-device learning offers several benefits

- Continuous learning
- Personalization
- Data privacy
- Scale



With offline training, the test data can differ from training data (domain shift, distribution shift, anomalies) and may even change continuously

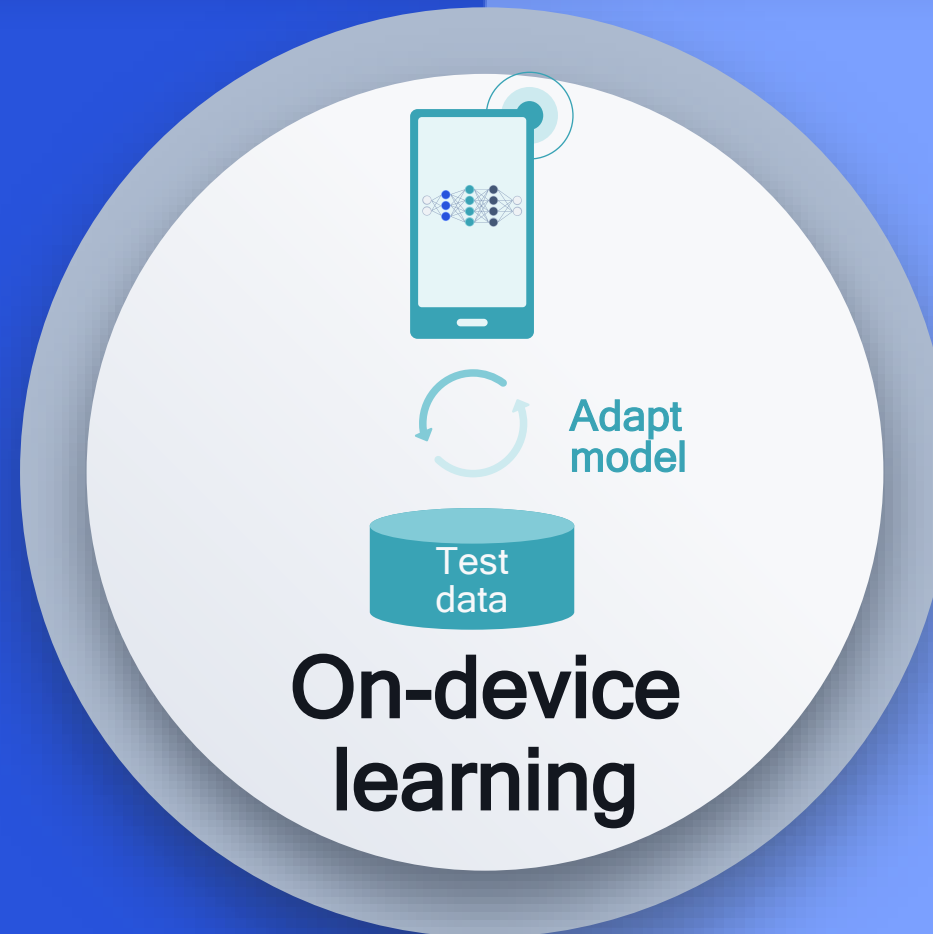
On-device learning can help to improve and maintain accuracy when original pre-trained model cannot generalize well

Overcoming challenges to achieve on-device ML benefits

Important considerations for on-device learning to achieve benefits for different use cases

Benefits

- Better examples than training dataset
- Ability to run with smaller models that adapt to the target data
- Preservation of privacy during model development



Challenges

- Local data can be limited, e.g., noisy labels and class imbalance
- Overfitting or catastrophic forgetting
- Limited compute, storage, and/or power
- Adversarial attacks to training
- Federated learning communication overhead

Our AI research areas address the key deployment challenges of on-device learning



Few-shot learning

How to adapt the model to a few labeled samples



Continuous learning with unlabeled data

How to use unlabeled data to do unsupervised learning



Federated learning for global adaptation

How to implement federate learning at scale and address deployment challenges



Low-complexity on-device learning

How to implement on-device learning to improve efficiency

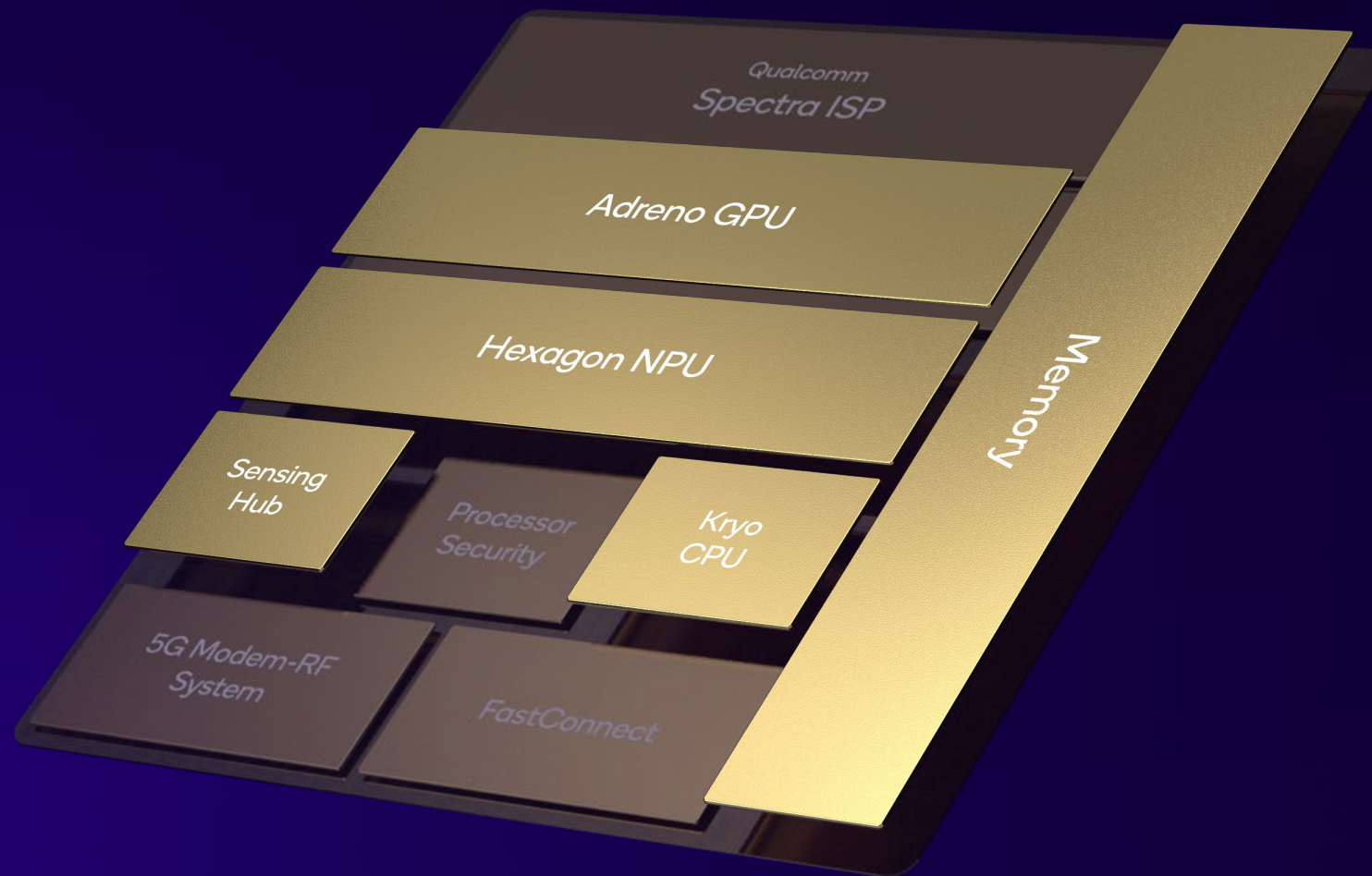


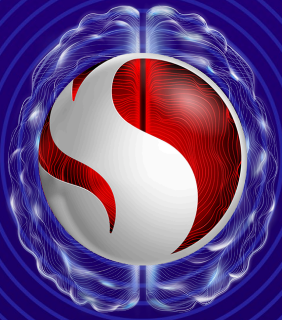
Qualcomm AI Stack

Qualcomm AI Studio



Qualcomm[®] AI Engine





Snapdragon
smart

<1 sec
per image

World's fastest
Stable Diffusion
and ControlNet

On-device
personalization

Qualcomm® Sensing Hub

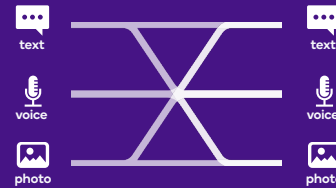


Qualcomm
AI Stack

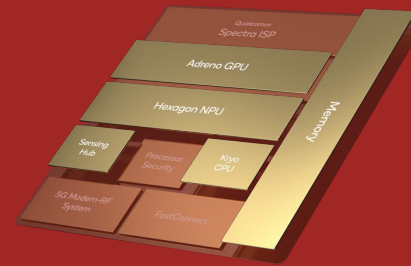
Fully optimized models

PyTorch ExecuTorch

First to support
Multi-modality
gen AI models



Upgraded
Qualcomm®
Hexagon™ NPU
designed for
gen AI



◆ Text Prompt. |



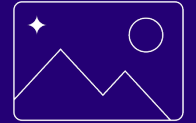
"Voice prompt."

Up to

20 tokens per second

Meta Llama 2 and
Baichuan powered
AI assistants

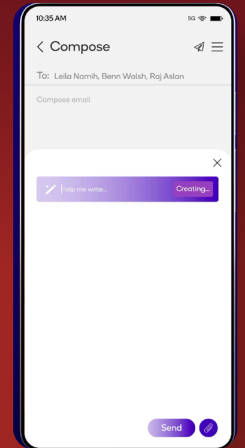
/ Photo Prompt /



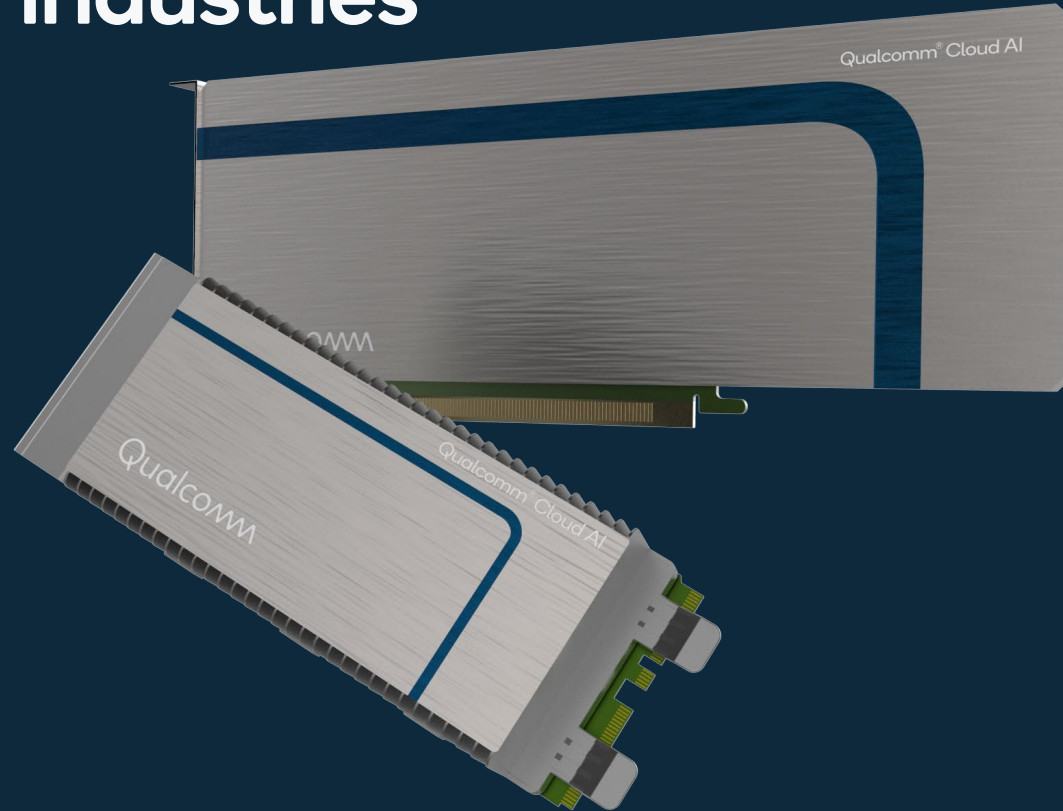
Up to

10B

parameter support

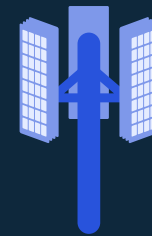


Qualcomm Cloud AI 100 addressing edge-to-cloud industries



Data Center/
Cloud Edge

5G
Edge Box



5G
Infrastructure



Datacenter



Personalized
Purchase recommendations



Personalized
Purchase advertisements

Pioneering
safety



Edge Box

Pedestrian alert
Crossing & blind spot assist



Road safety
Intersection management
assist



Powering the shopping
of the future



5G Infra



Reinventing the
communication
experience



Hardware Architecture

- Up to 400 TOPS
- Power
 - DM.2e @ 15W
 - DM.2 at 25W
 - PCIe/HHHL @ 75W
- AI Core (AIC) - Up to 16 cores
- Precision – INT8, INT16, FP16, FP32
- On-die SRAM – Up to 144 MB
- 4x64 LPDDR4x (2.1GHz) with inline ECC
 - Up to 32GB on card DRAM
- PCIe Gen 3/4 - Up to 8 lanes

On-device & hybrid AI are critical for Gen AI to scale

Benefits

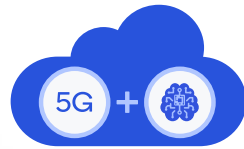
- Cost
- Energy
- Reliability, performance, and latency
- Privacy and security
- Personalization



Qualcomm



Foundational
R&D



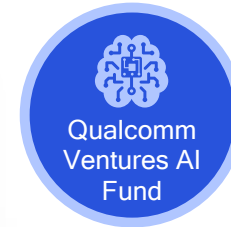
5G + AI
technology
leadership



Systems
design
expertise



Advanced
silicon



Ecosystem
investment



Uniquely positioned to enable
intelligence computing everywhere



Qualcomm

Intelligence is becoming more distributed, with power-efficient on-device AI complementing the cloud

Mobile is democratizing AI and bringing it to new frontiers

Qualcomm Technologies is well positioned to provide superior AI solutions and make AI ubiquitous

Connect with us



www.qualcomm.com/research/artificial-intelligence



www.qualcomm.com/news/onq



[@QCOMResearch](https://twitter.com/QCOMResearch)



www.youtube.com/c/QualcommResearch



www.slideshare.net/qualcommwirelessevolution

Thank you

Qualcomm

Follow us on: [in](#) [twitter](#) [instagram](#) [youtube](#) [facebook](#)

For more information, visit us at:

qualcomm.com & qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018-2023 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm, Snapdragon, Adreno, Hexagon, Kryo, FastConnect, Snapdragon Ride, and Qualcomm Spectra are trademarks or registered trademarks of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.