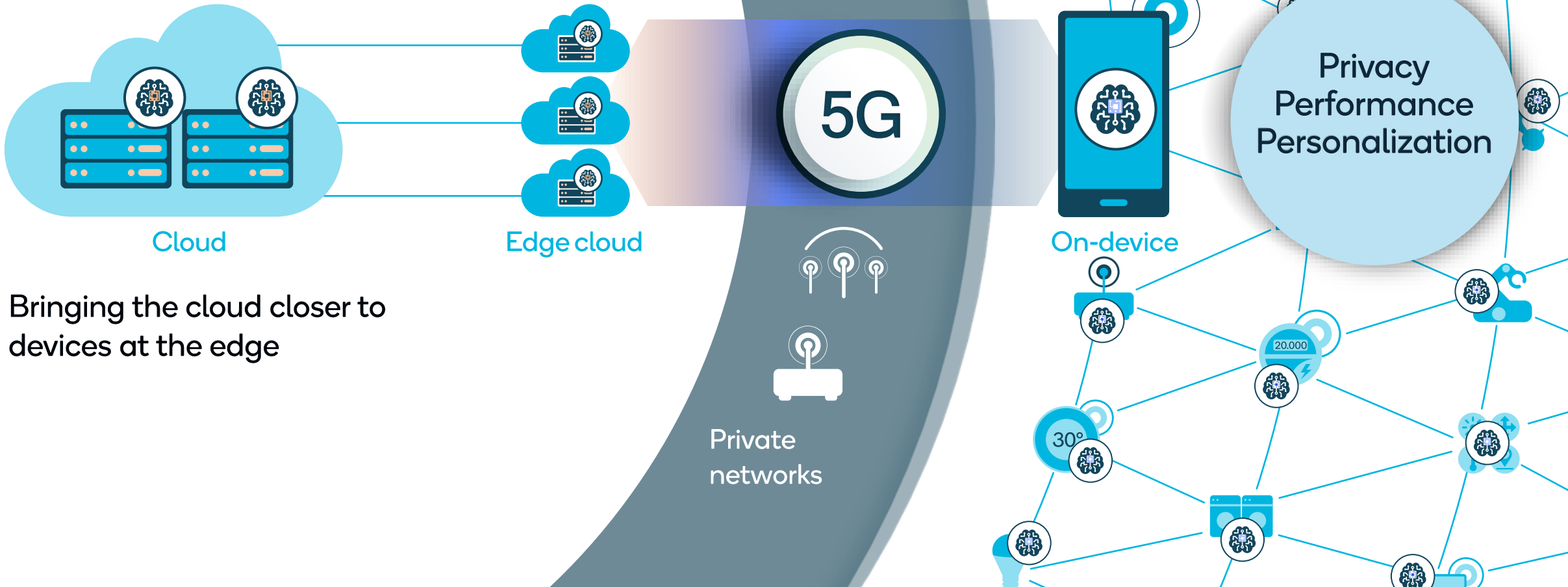
2020 embedded VISION summit®

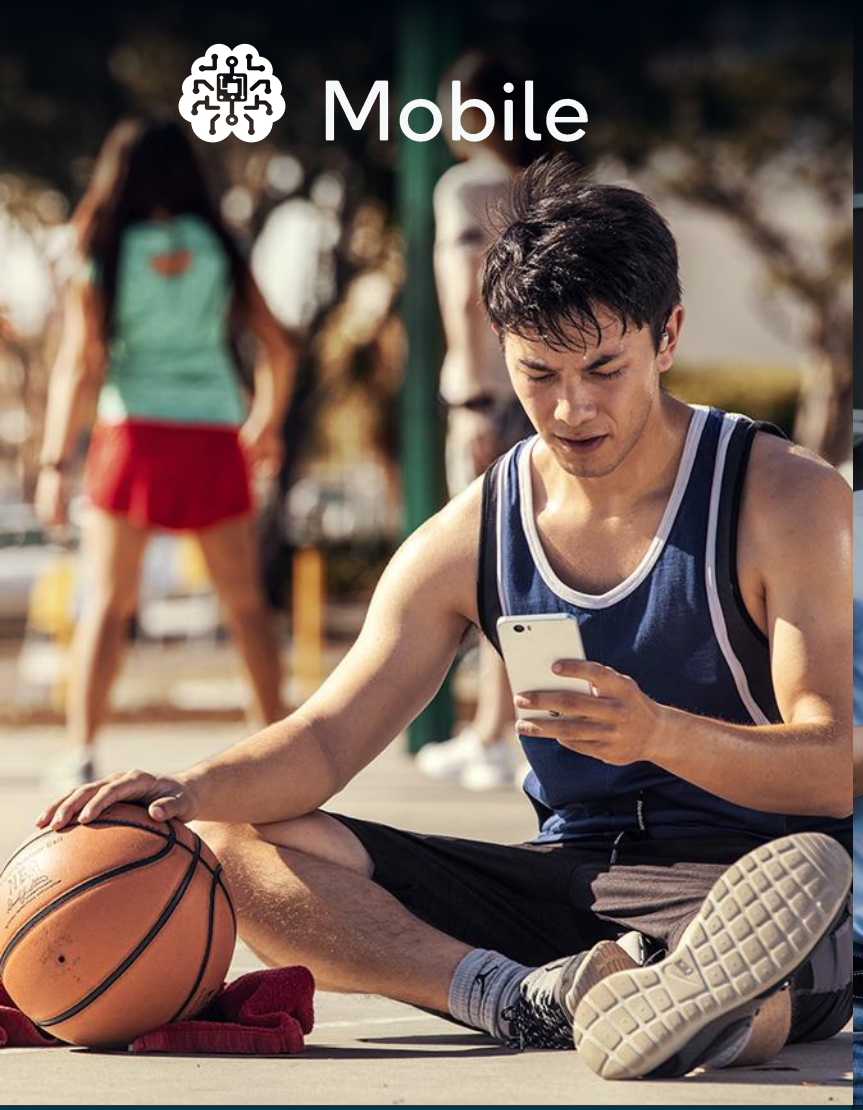
Qualcomm AI Leading the way with Distributed Intelligence

Ziad Asghar
VP, Product Management
Qualcomm Technologies, Inc.

We're creating a future of distributed intelligence



Bringing the cloud closer to devices at the edge



AI will drive transformation across industries

Powering the factory of the future

Shaping the future of transportation

Mobile –
THE most pervasive AI platform

>7.3 Billion

Cumulative smartphone unit shipments forecast between 2019–2023

>1 Billion

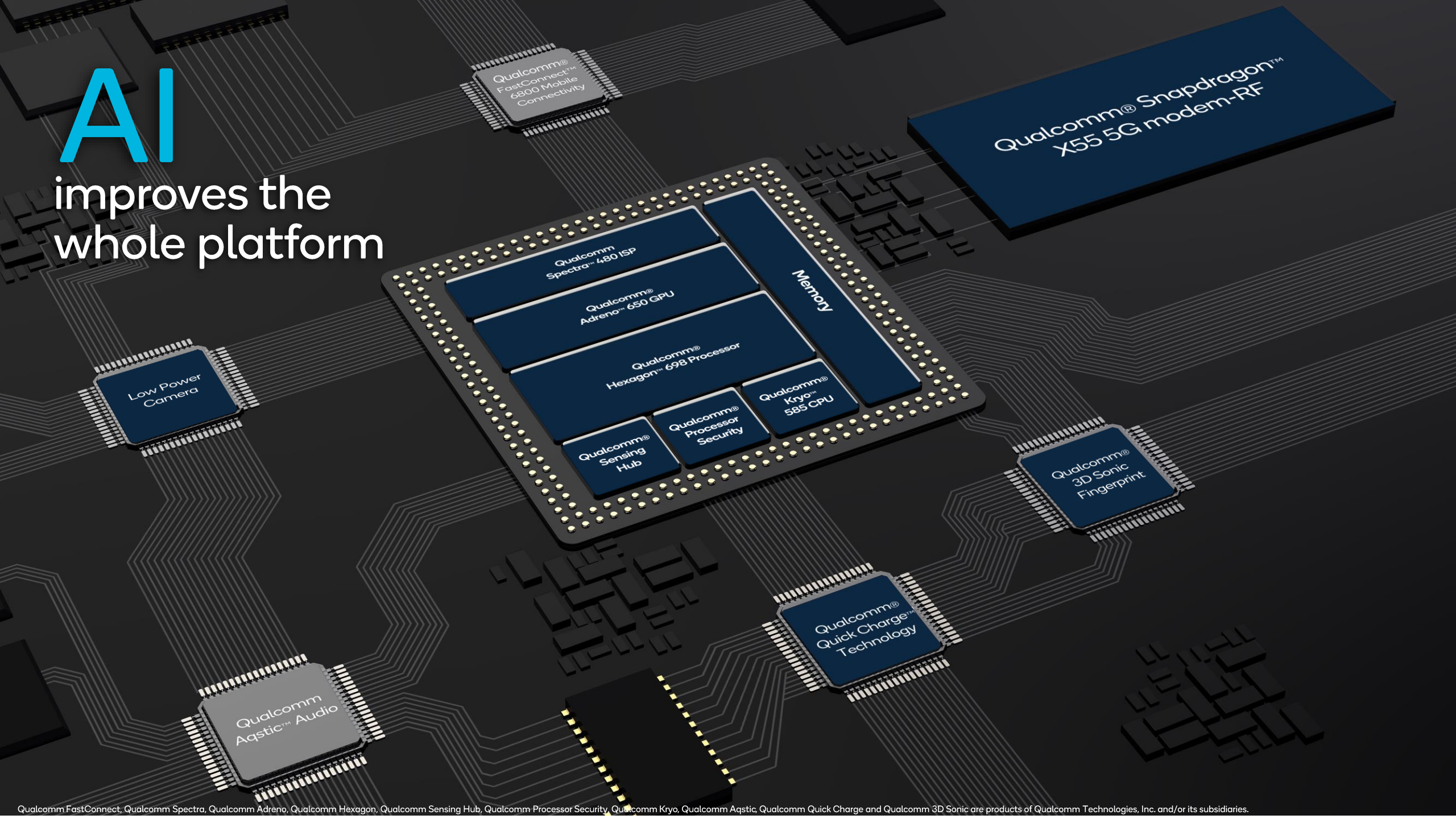
AI capable devices enabled with QTI technology

Source: IDC Mobile Phone Tracker 1Q2019

AI use cases in key segments

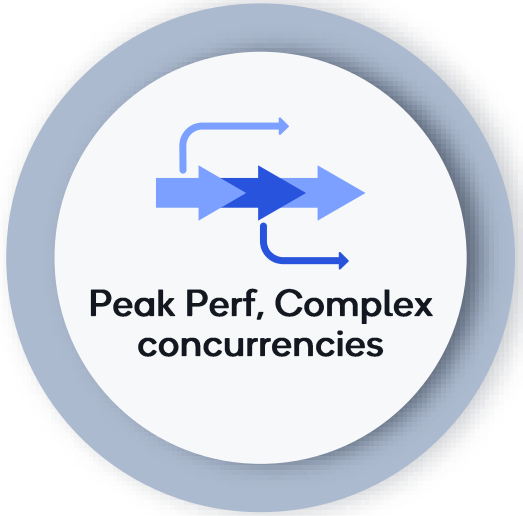
Mobile	AR/VR	IoT/Camera	IoT/Speakers	Automotive/ADAS
Face ID/Auth	Gesture Tracking	Video Summarization	Multi Keyword Detection	Vehicle Detection
Super Resolution	3D Reconstruction	Age/Ethnicity/ Emotion Detection	ECNS	Pedestrian Detection
ASR	Object Detection	Zone Intrusion	Voice Biometrics	Path Planning
Voice Translation	Scene Segmentation	Food Classification	ASR	Traffic Sign Recognition
Night Shot	Split Rendering	Face Detect/Reco	Audio Context Detection	Speech/Speaker Recognition
Portrait Backlight	Hand Tracking	Object Avoidance	Speech Synthesis	Distracted Driver Alerts
Call Noise Reduction	Body Pose Tracking	Multi-object Detection		
Call Voice Enhancement				

Best-in-class AI performance at the edge



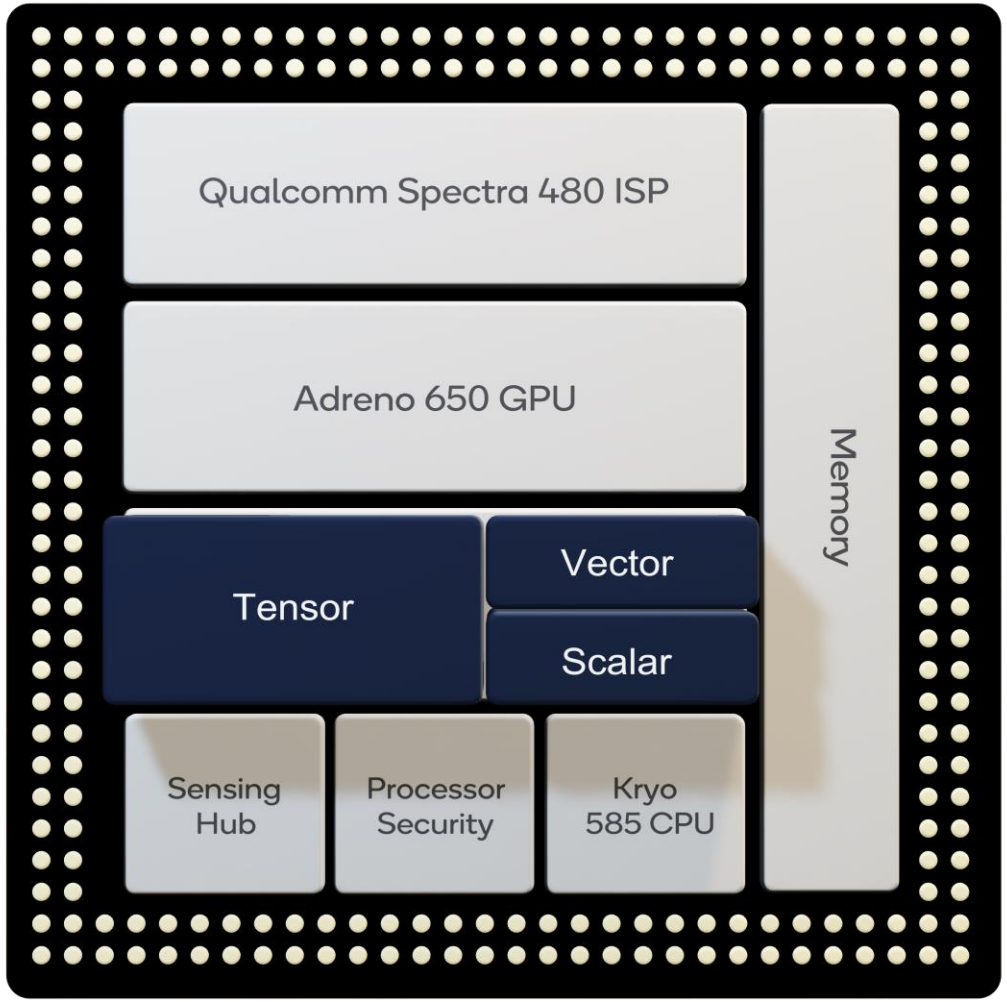
Best-in-class AI performance at the edge

Use cases driving increase in peak performance



5th Generation

Qualcomm® AI Engine



Qualcomm AI Engine is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

Adreno 650

New AI mixed precision instructions

2x higher TOPS[1]

16-bit and 32-bit FP

Hexagon 698

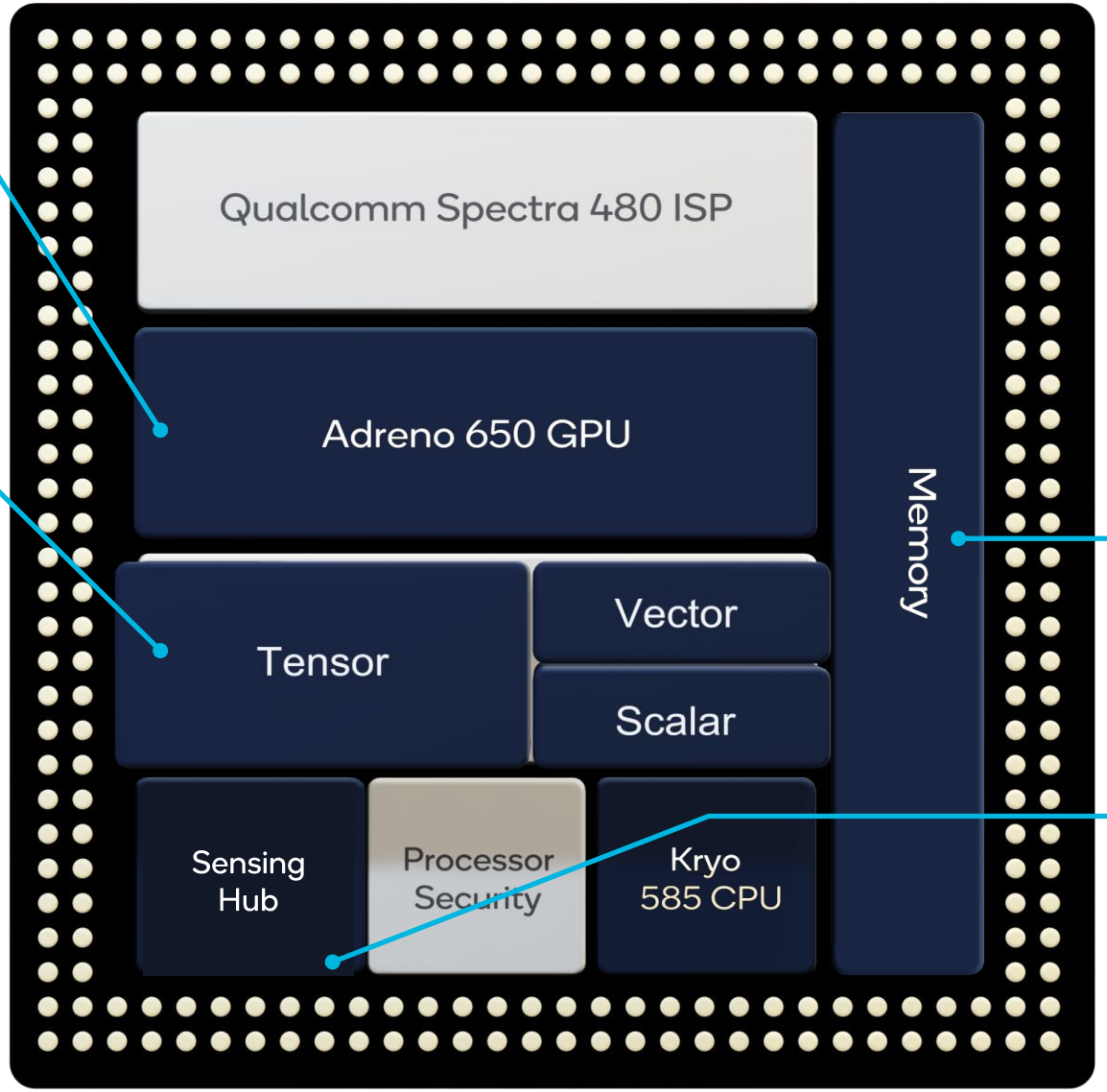
New Tensor Accelerator

- 4x higher TOPS[1]
- Up to 35% power savings[1]
- 8-bit and 16-bit INT

Deep Learning Bandwidth Compression

- Up to 50% lossless Compression
- Frees up bandwidth for other parts of the SoC
- Saves power due to reduced memory transfers

1. Comparing to previous generation



LP-DDR5 memory

30% more bandwidth*

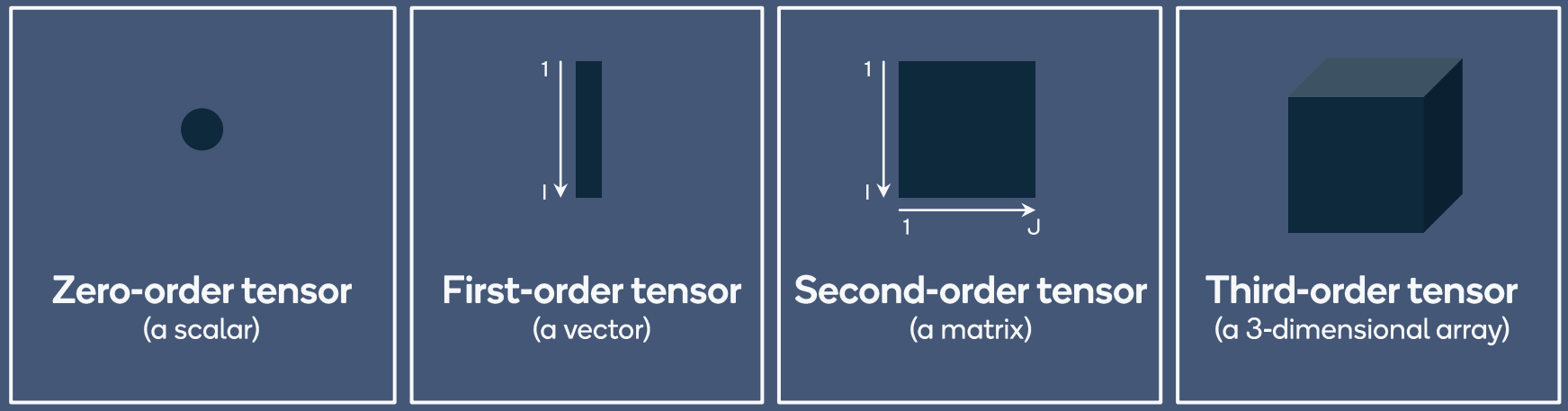
Improved AI processing

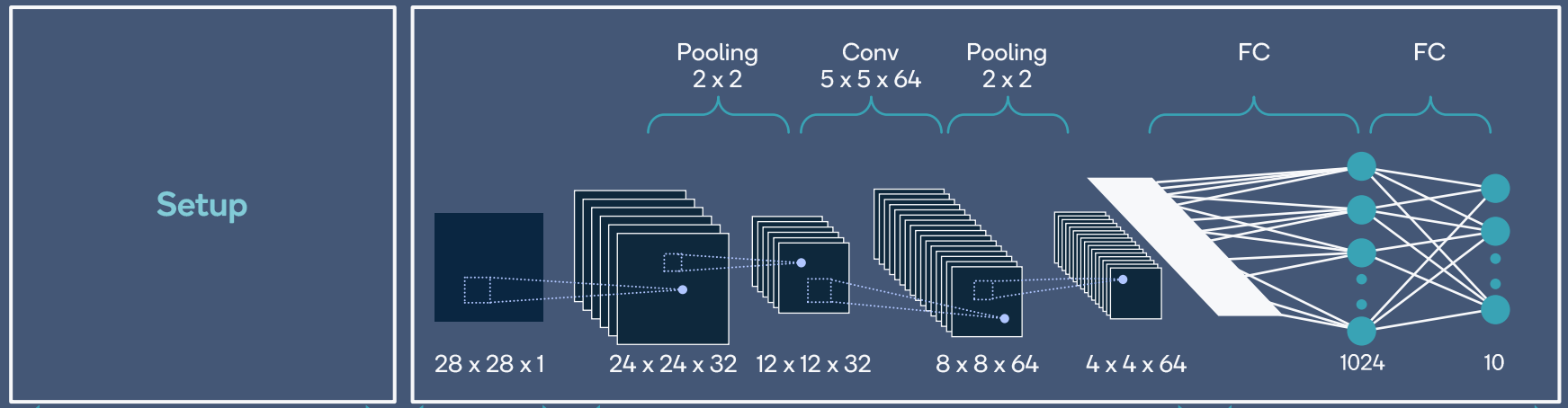
Sensing Hub

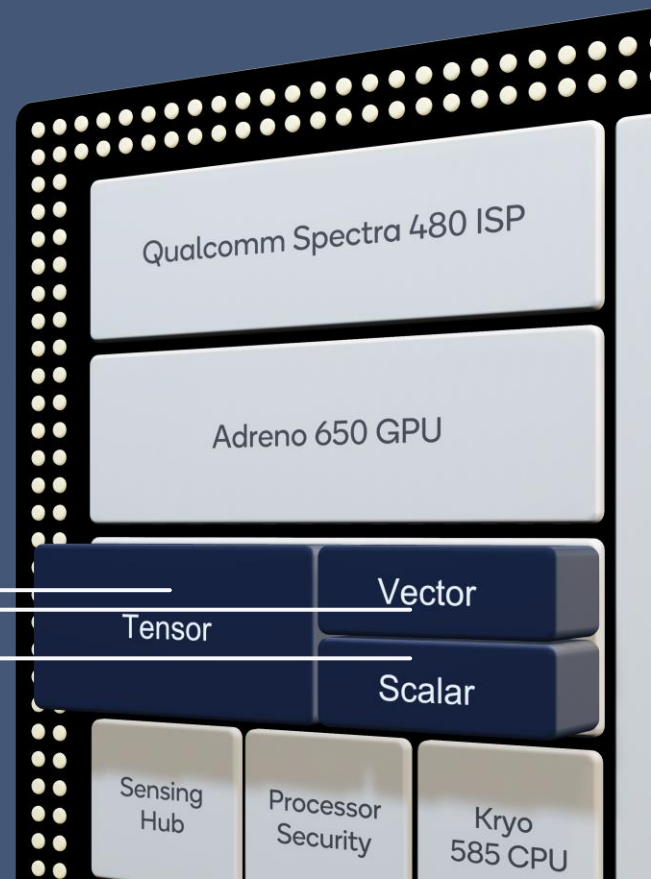
<1mW camera

<1mA voice multi-word wakeup

Scalable sensor framework

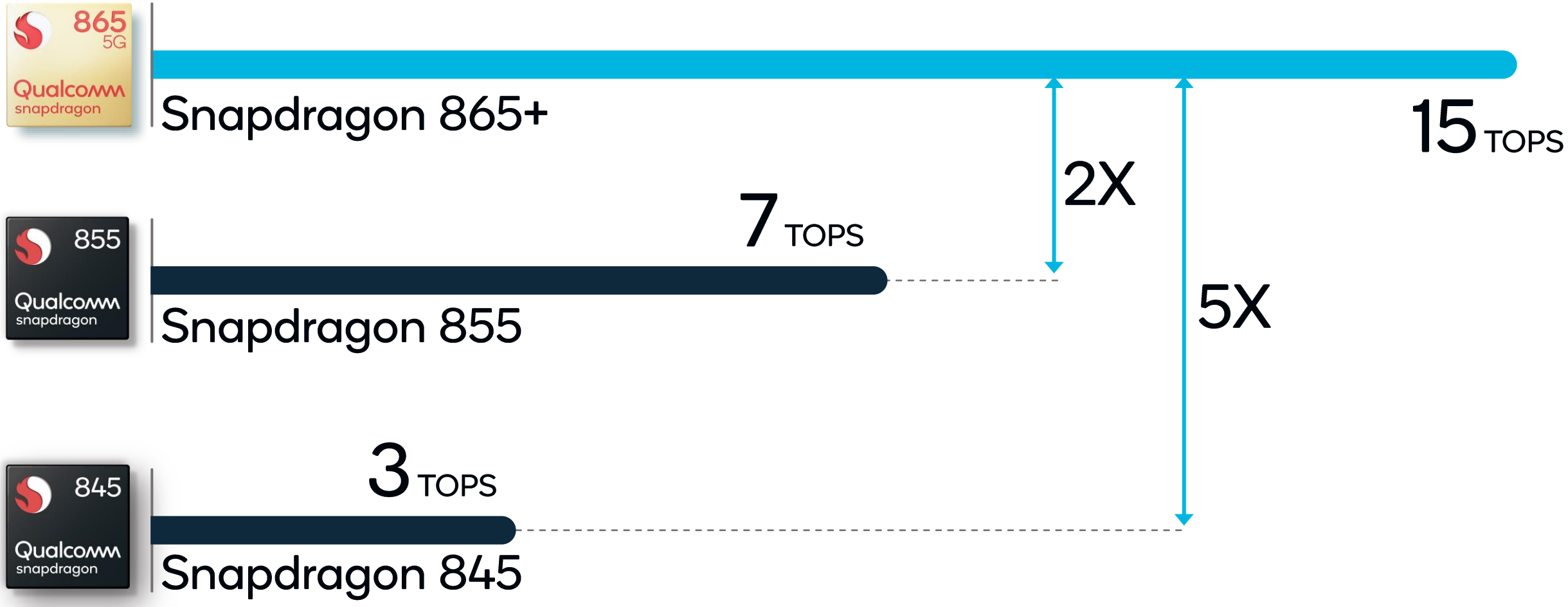






Qualcomm

Trillion operations per second



Antutu Top 10 Best Performing Android Flagship Phones, August 2020

2020 embedded VISION summit

Qualcomm snapdragon

Qualcomm Neural Processing SDK

TensorFlow Lite

Android Neural Networks API

Runtime

Qualcomm

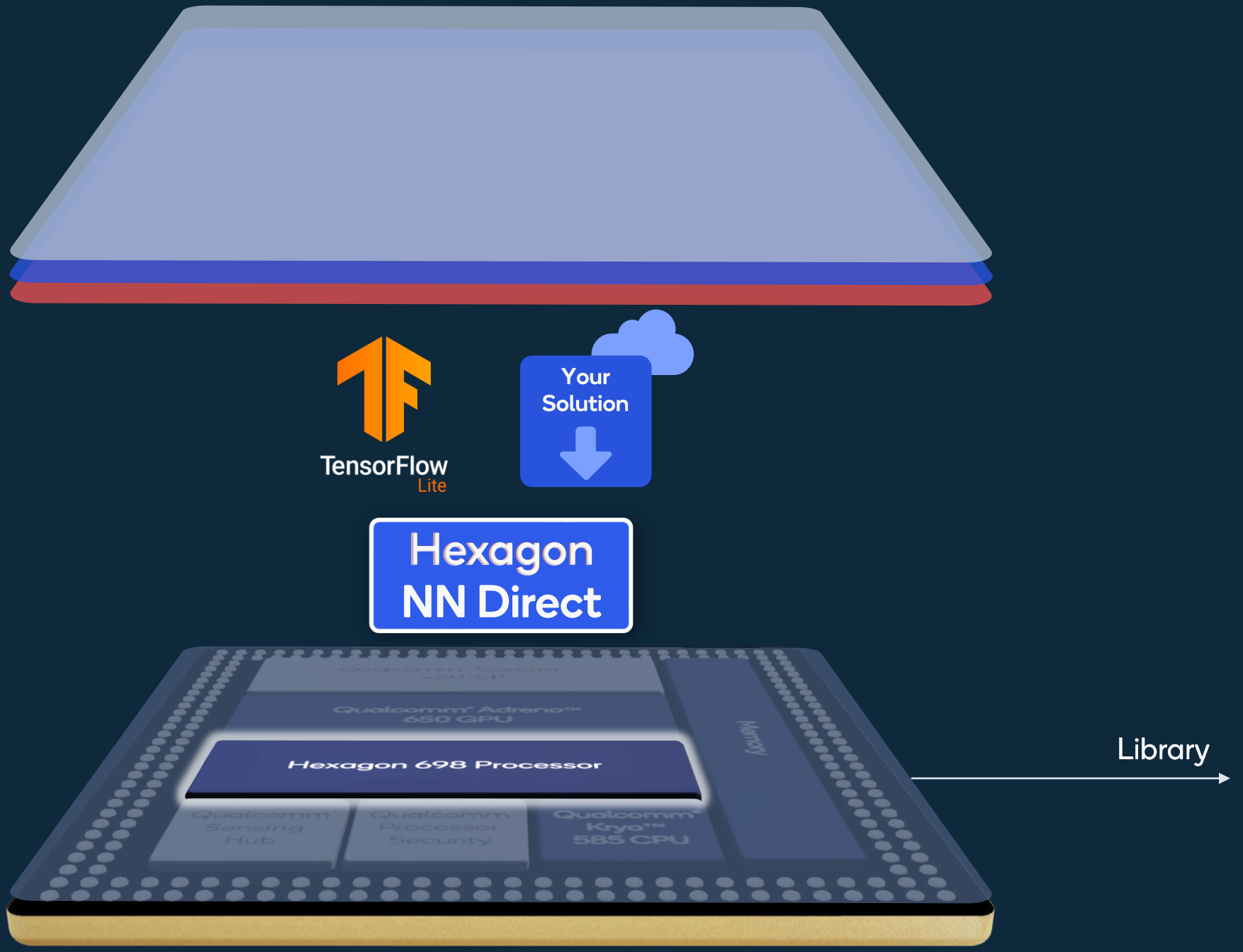
Qualcomm Neural Processing SDK

- Improved performance
- Data free quantization
- Support for more network models
- Graph analysis

3-5x
Speed improvements on Hexagon processor

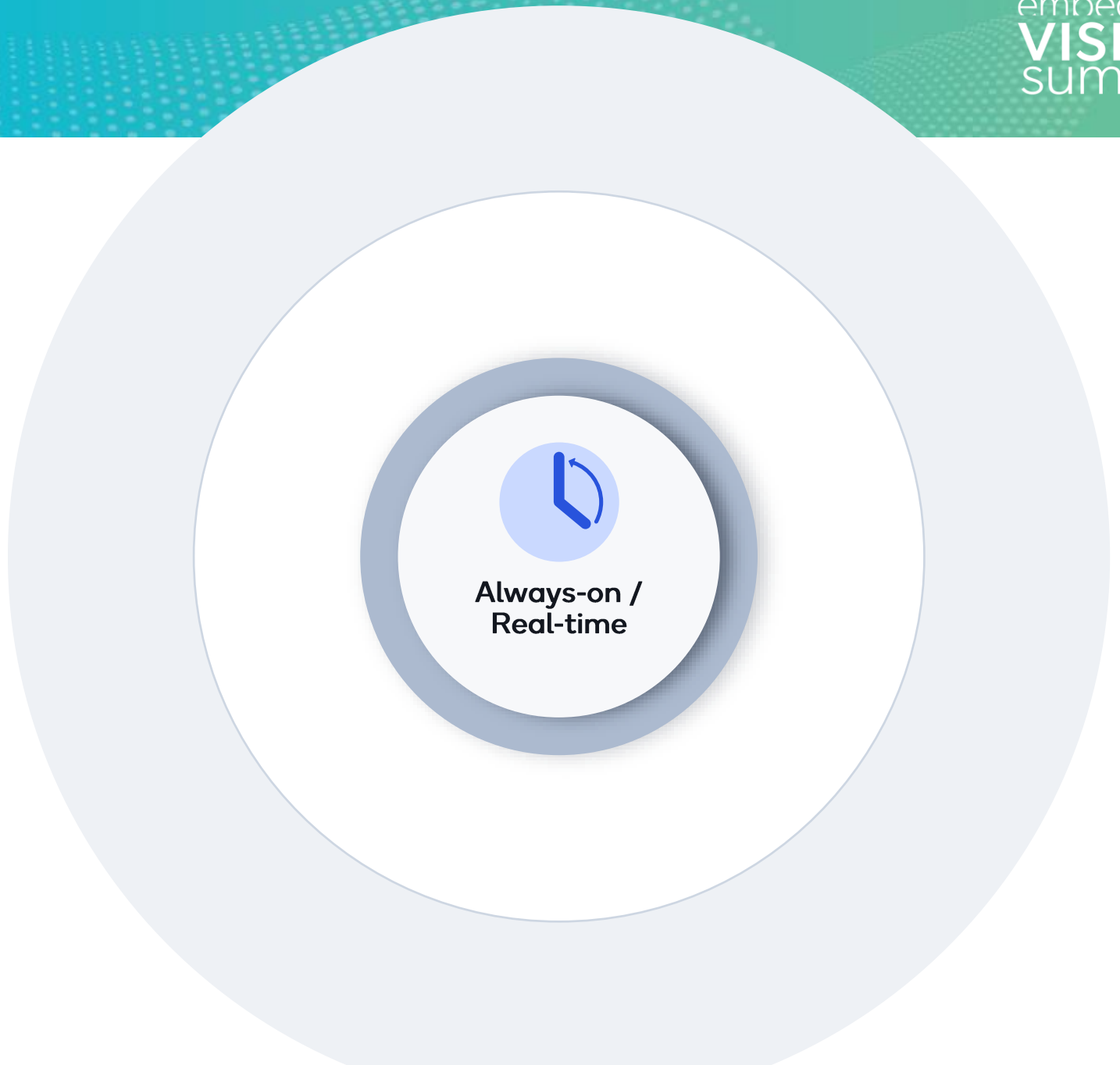
Android Neural Networks API

- 3x number of accelerated operators
- Shipping with Android 10
- Optimized for Google speech recognition and Google Lens
- Wide adoption from developers



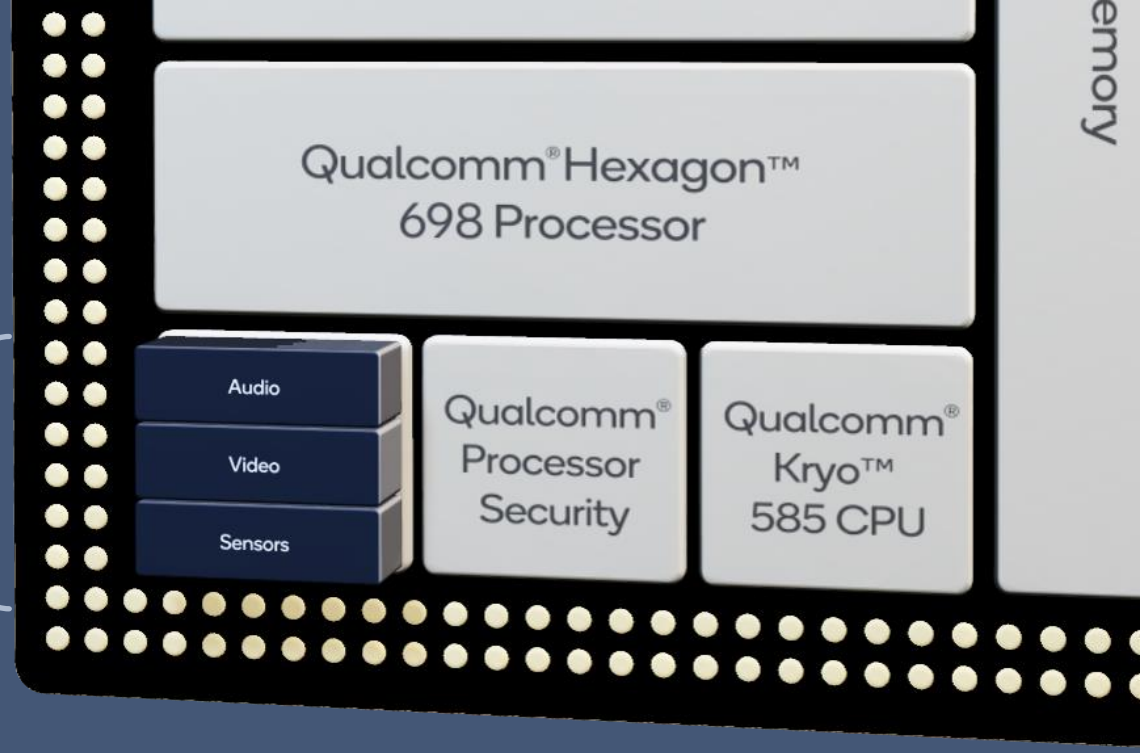
Qualcomm

Always-on use cases require a different approach to AI



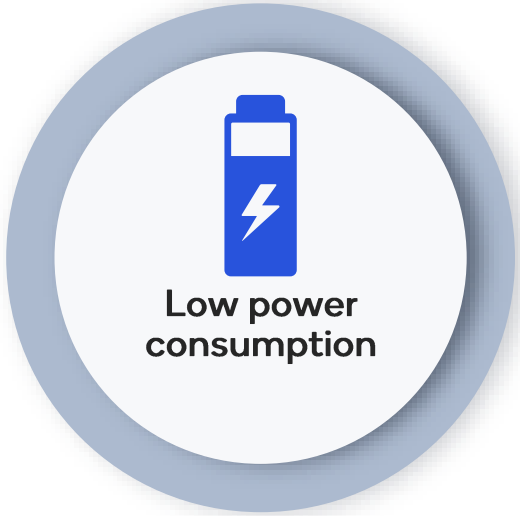
Qualcomm Sensing Hub

<1mA



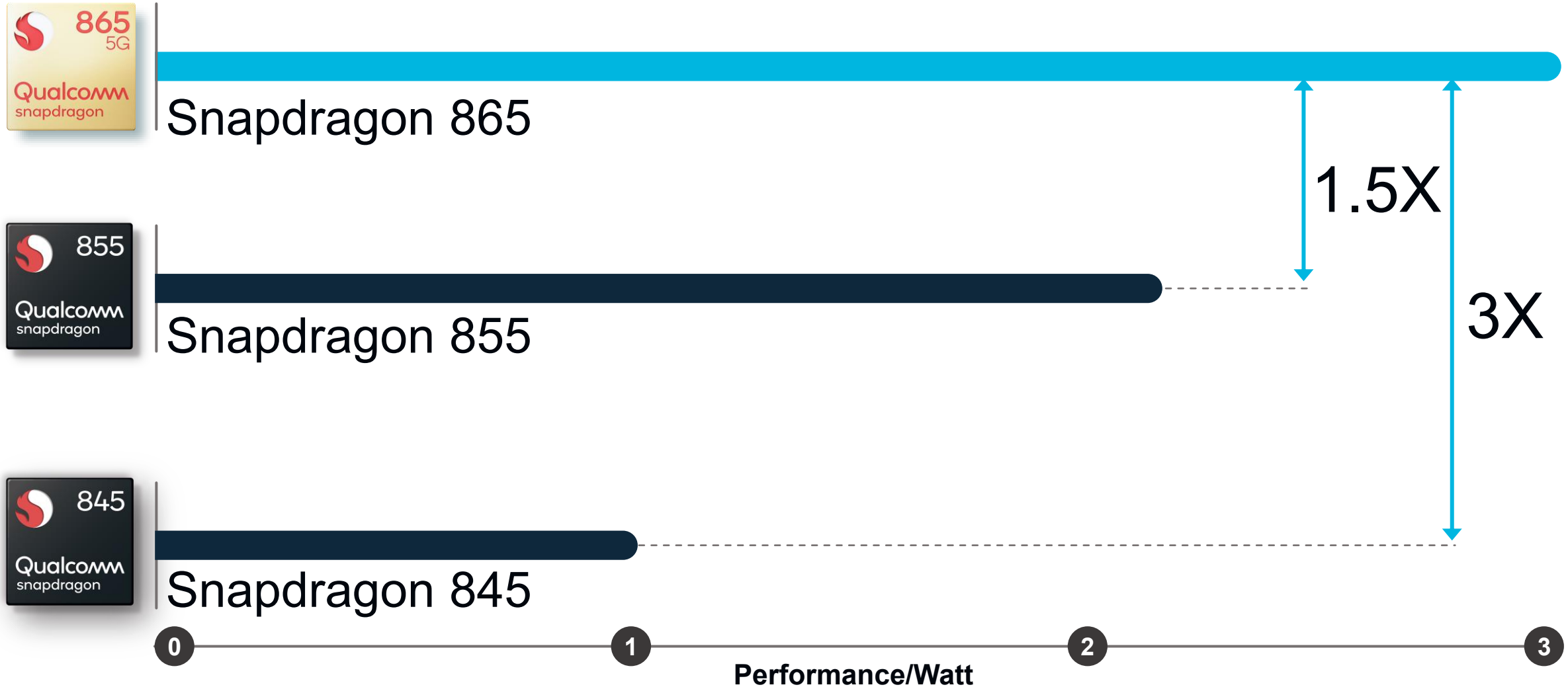
Safety | Health & Fitness | Security | Context | Low-power camera

Best-in-class Performance per unit power is key for edge devices

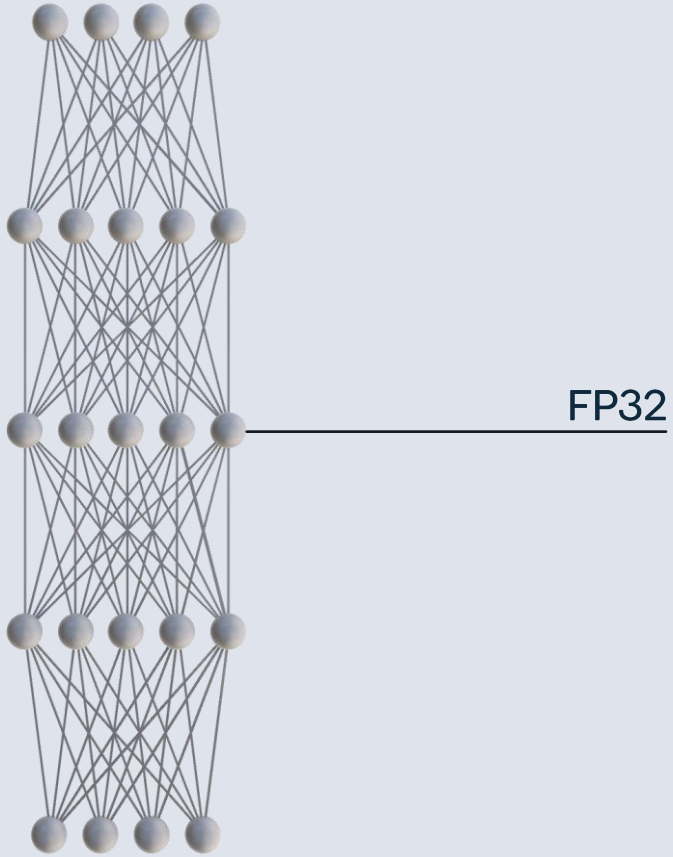


AI Engine Power Efficiency improvement

Hexagon Tensor Accelerator



New Qualcomm® AI Model Efficiency Toolkit



865 5G

Qualcomm snapdragon

Qualcomm

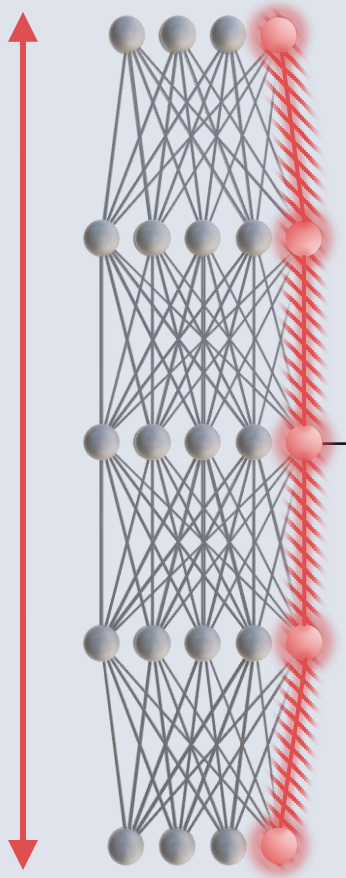
Qualcomm AI Model Efficiency Toolkit is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

New Qualcomm® AI Model Efficiency Toolkit

Model compression

Spatial SVD

Bayesian compression



FP32

3x
Compression with less than 1% loss in accuracy*

865 5G
Qualcomm snapdragon

Qualcomm

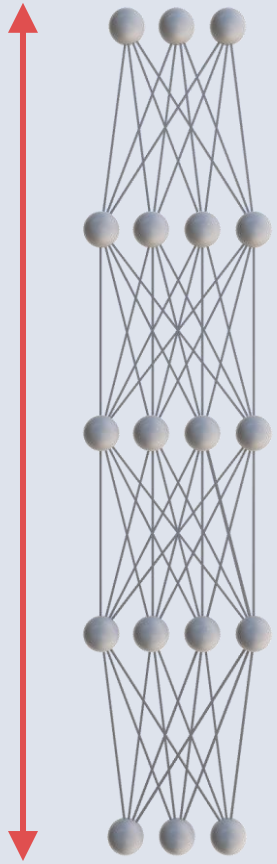
*Comparison between baseline and compression with both Bayesian compression and spatial SVD. Example uses ResNet18 as baseline.

New Qualcomm® AI Model Efficiency Toolkit

Model compression

Data free quantization

Quantization aware training



FP32

3x
Compression with less than 1% loss in accuracy*

865 5G
Qualcomm snapdragon

Qualcomm

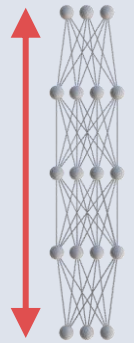
*Comparison between baseline and compression with both Bayesian compression and spatial SVD. Example uses ResNet18 as baseline.

New Qualcomm® AI Model Efficiency Toolkit

Model compression

Data free quantization

Quantization aware training



INT8

>4x

Increase in performance per watt with quantization

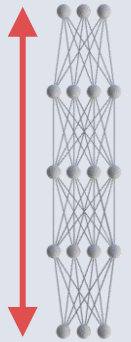
865 5G

Qualcomm snapdragon

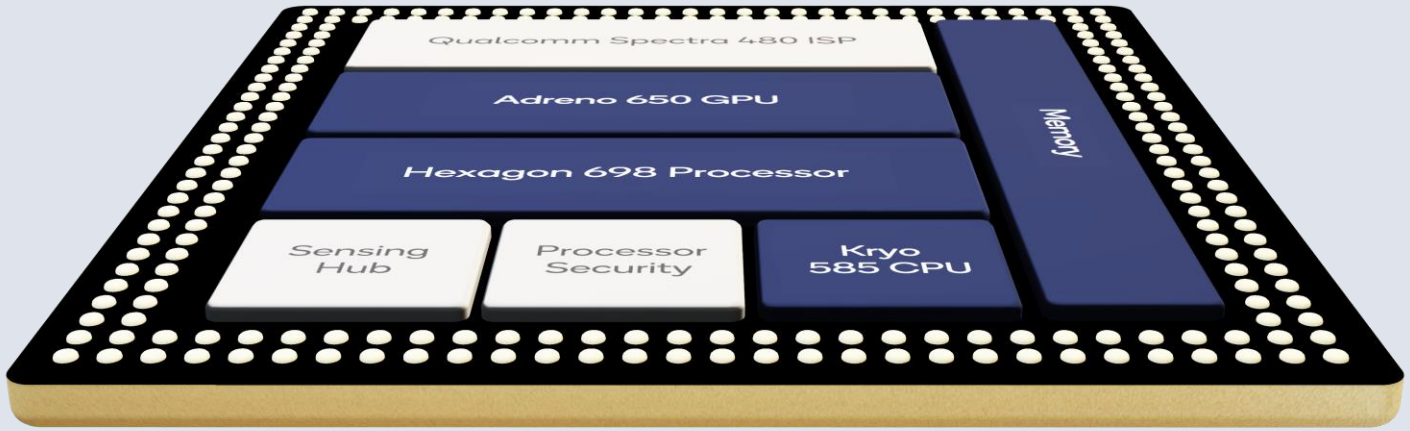
Qualcomm

3x
Compression with less than 1% loss in accuracy*

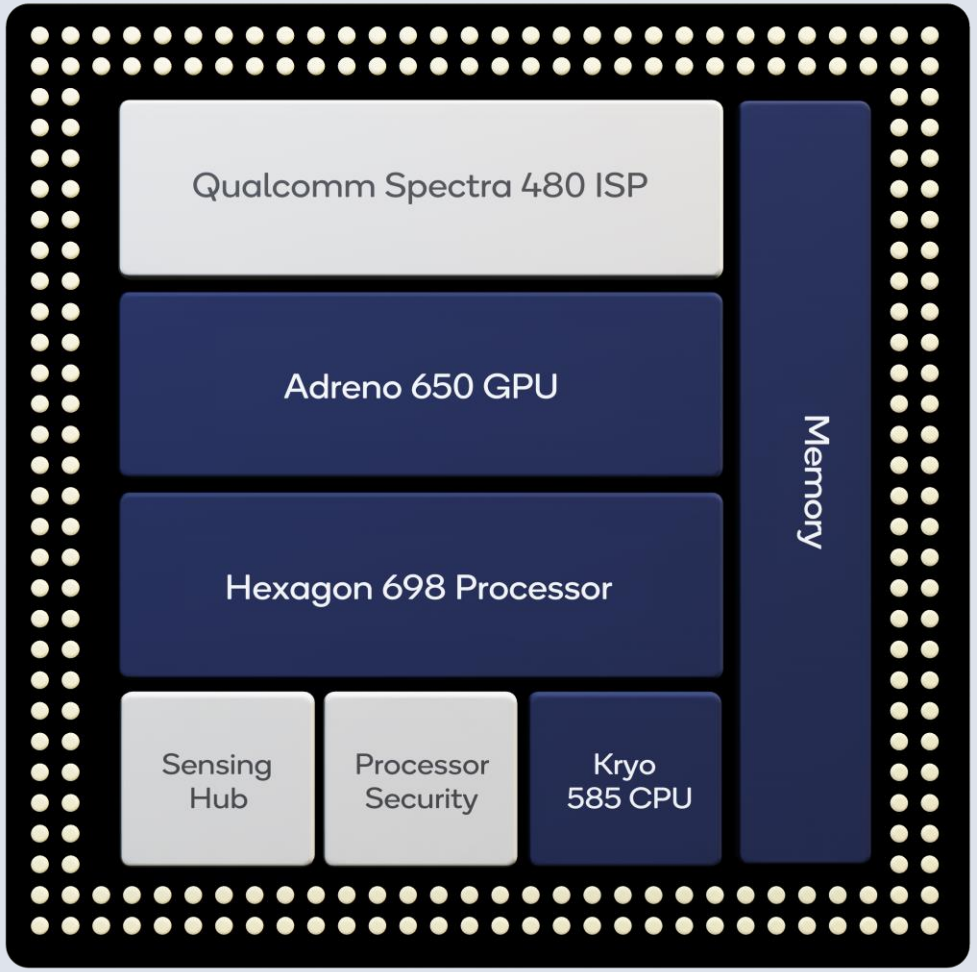
>4x
Increase in performance per watt with quantization



INT8



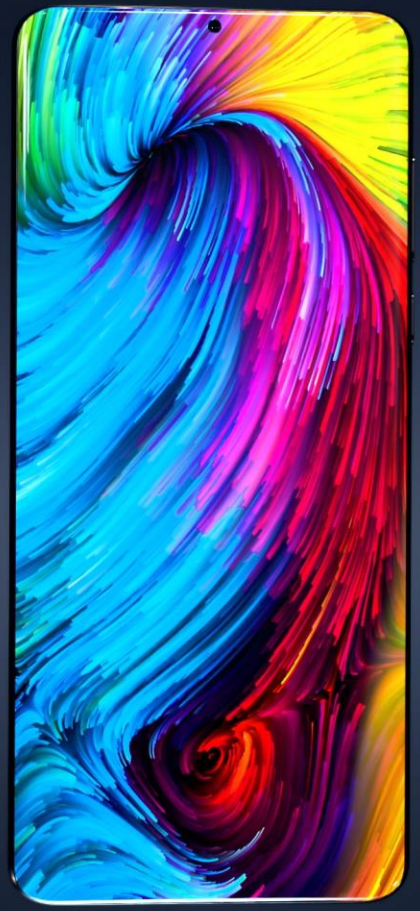
Qualcomm



Qualcomm

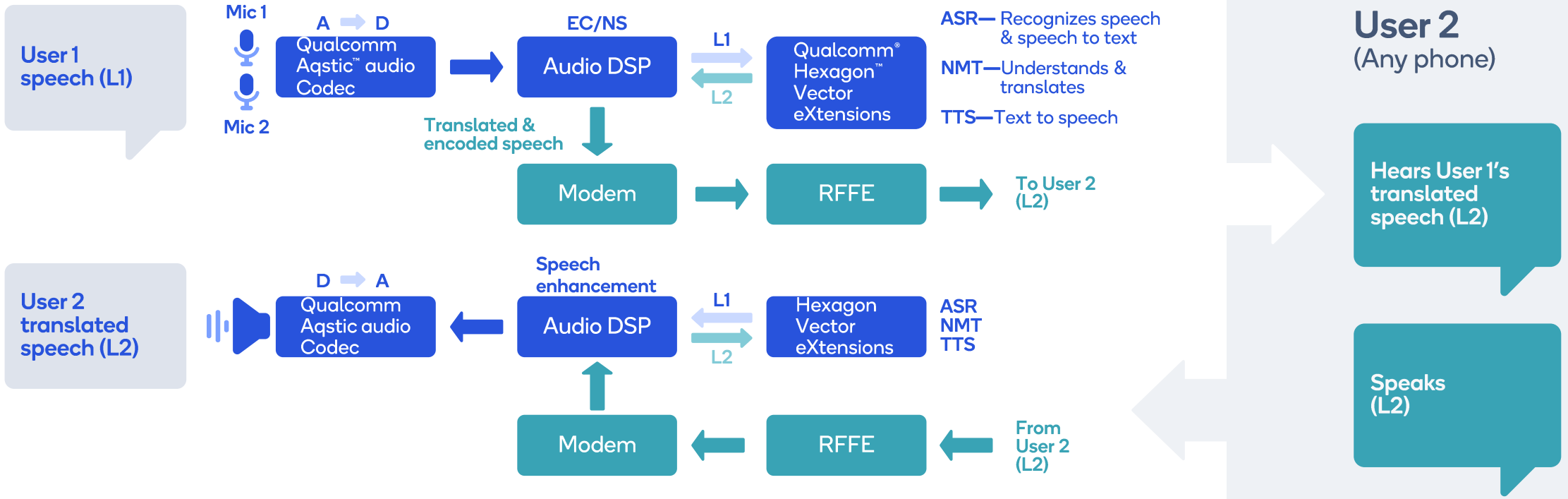

Qualcomm


有道 youdao



Enabling new experiences

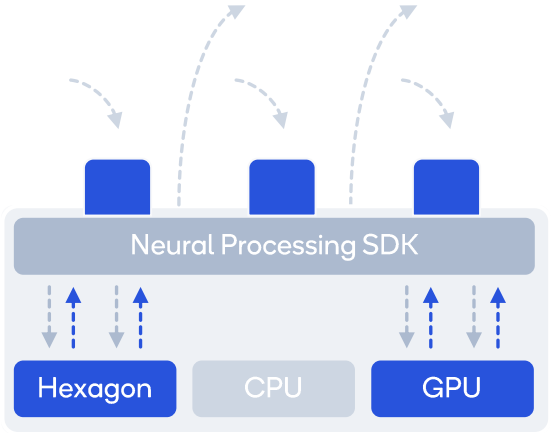
Real time language translation



Requirements

1. Different types of NNx (LSTM/RNN,GRU/GLU, Transformer/BERT)— Ops & Topology
2. Low-latency (real time operation)
3. Power consumption, performance, concurrency and thermal

XR gaming



Control
- Hands tracking
- Eye tracking
- Body motion
- Object tracking

Display
- Visual style transfer

Audio
- AI audio effects
- AI placed audio

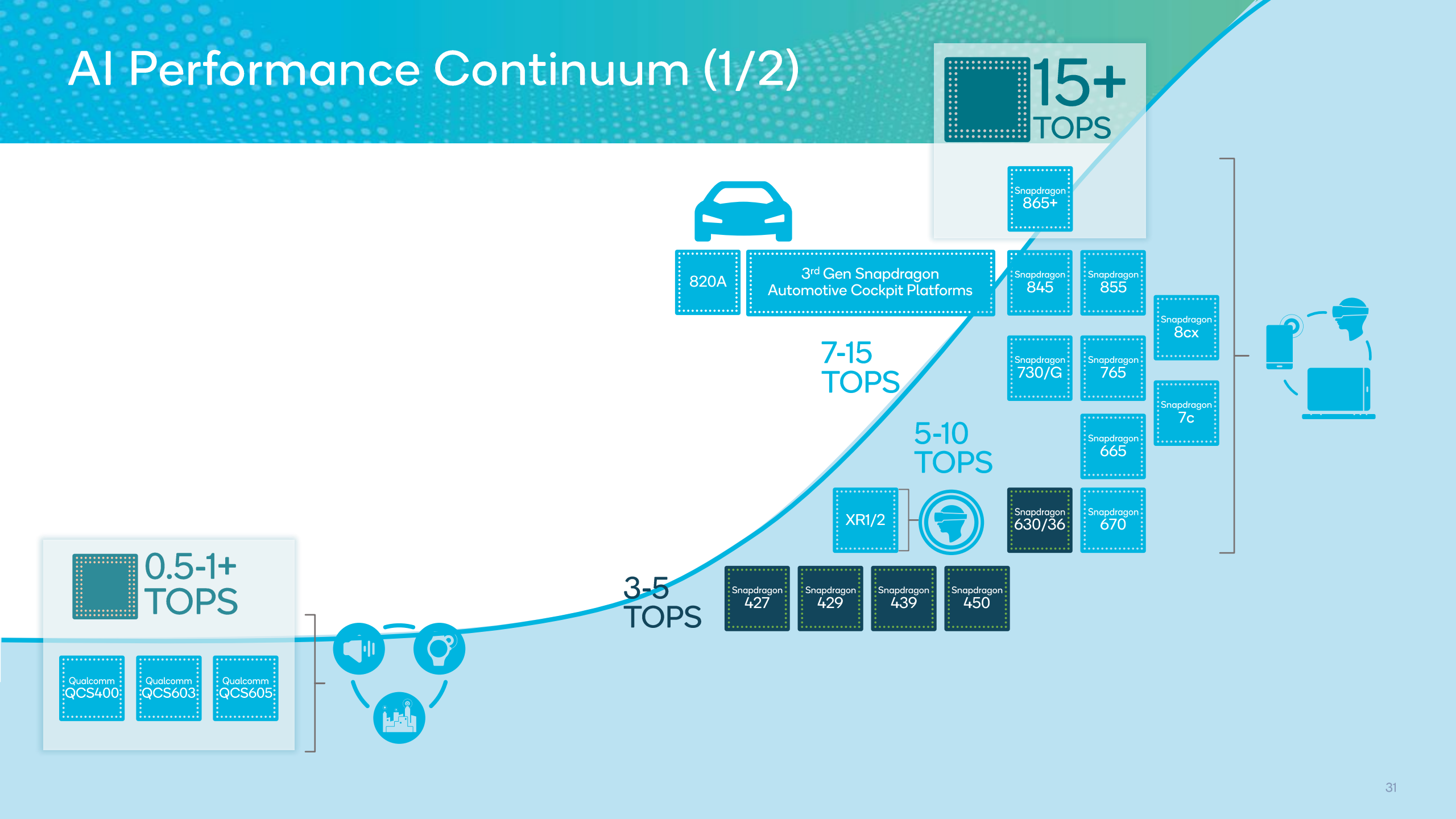
Mechanics
- AI opponent
- AI game assistant

Hearing
- No howling
- Speech recognition & NLP
- Your voice selection

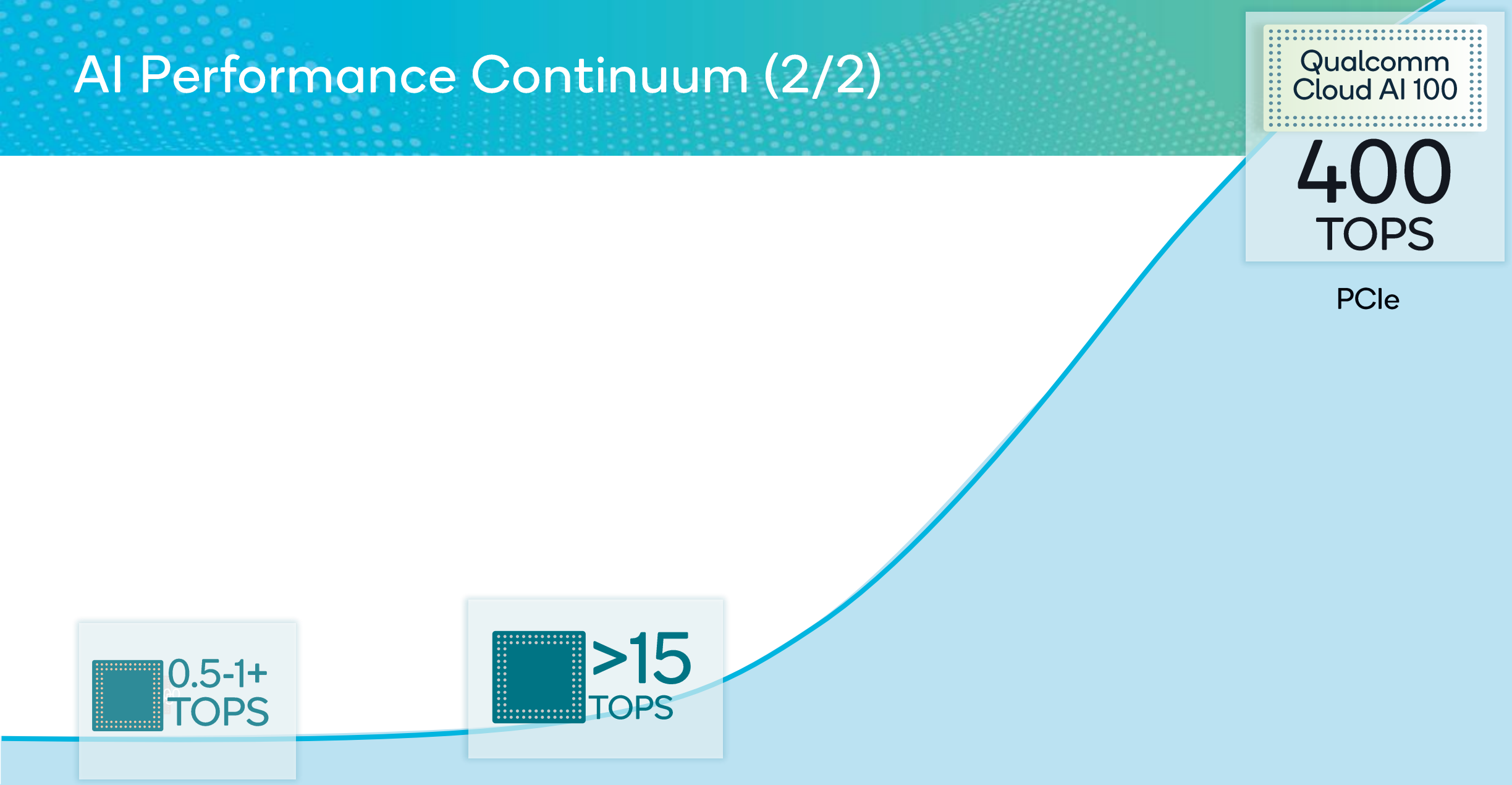
XR World
- People/object detection
- Scene segmentation
- Pose estimation
- Intent detection

AI improves gaming experience on the device

AI Performance Continuum (1/2)



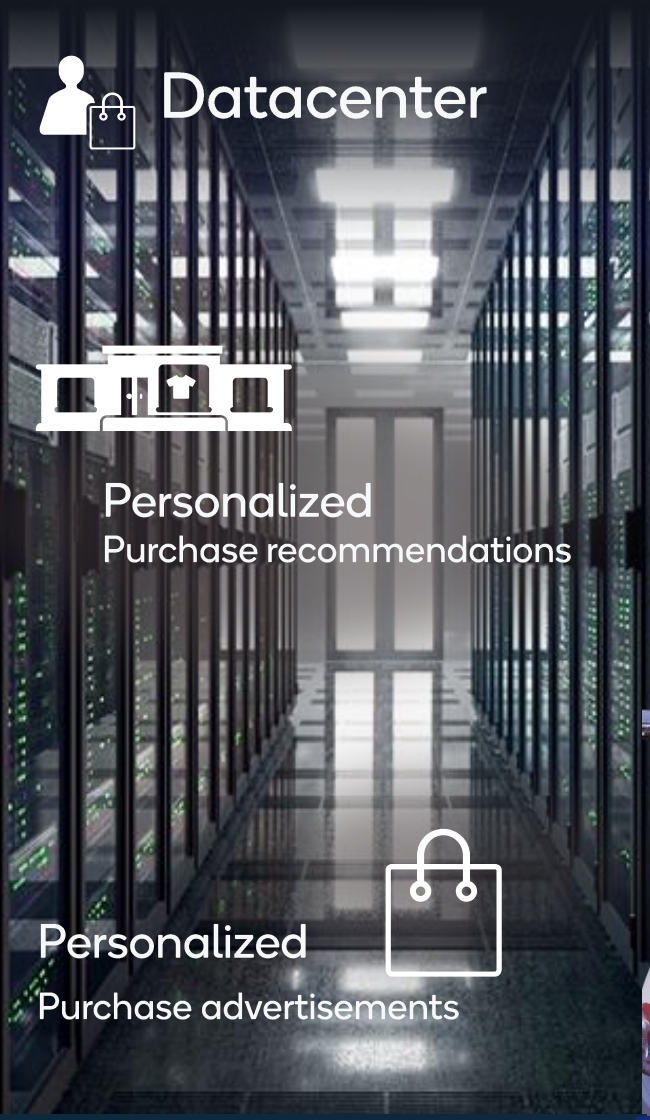
AI Performance Continuum (2/2)

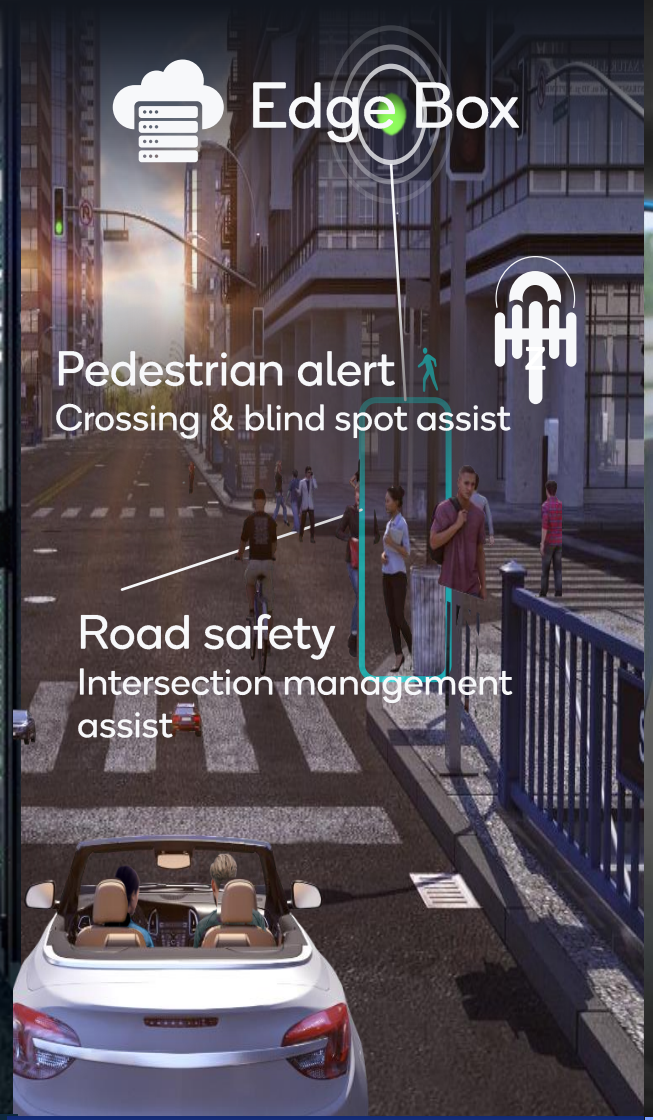


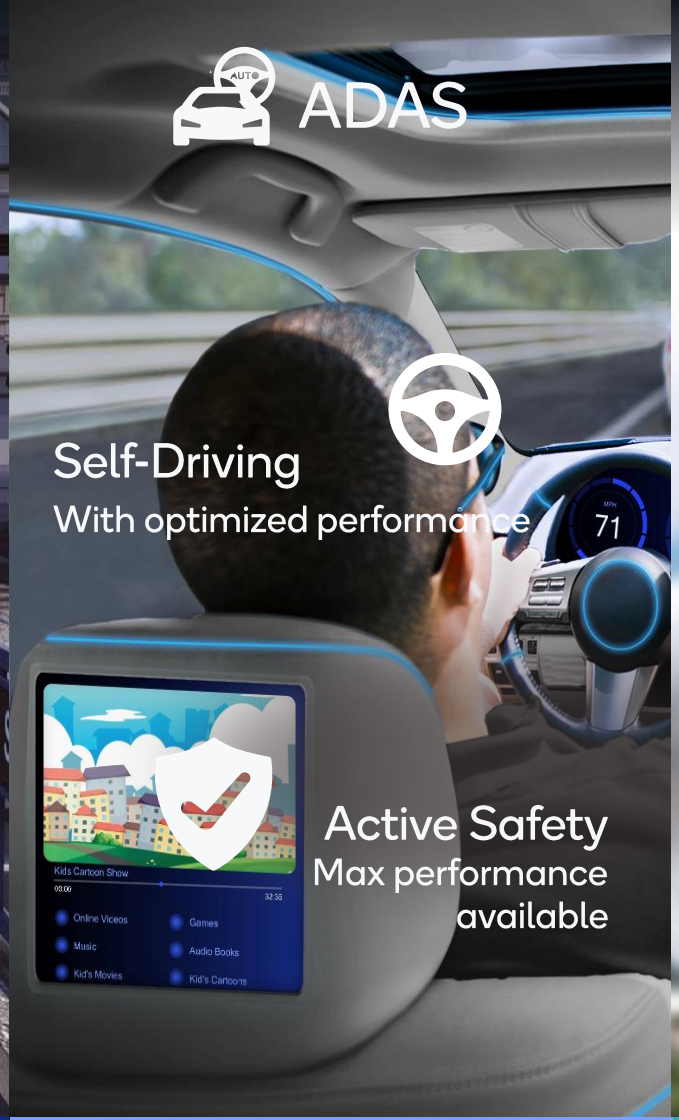
Cloud AI 100 addressing edge-to-cloud industries

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







Powering the shopping of the future

Pioneering safety

Shaping the future of transportation

Reinventing the communication experience

We're creating a future of distributed intelligence

Resource Slide

- Qualcomm AI page:

https://www.qualcomm.com/invention/artificial-intelligence

- Qualcomm AI research:

https://www.qualcomm.com/invention/artificial-intelligence/ai-research?cmpid=fofyus193556&gclid=CjwKCAjw19z6BRAYEiwAmo64LfQjU8vqH8TxqKTM2PZQp8JibXrjev85wLfKFknJnS_b494yZ7e_WhoCPQkQAvD_BwE

- Qualcomm Platform Solution Ecosystem:

https://www.qualcomm.com/support/qan/platform-solutions-ecosystem

- GitHub AIMET:

https://github.com/quic/aimet

Thank you

Qualcomm