Powering the Connected Intelligent Edge and the Future of On-Device AI

Ziad Asghar
Vice President, Product Management
Qualcomm Technologies Inc.

Parietal
Touch

Occipital
Vision

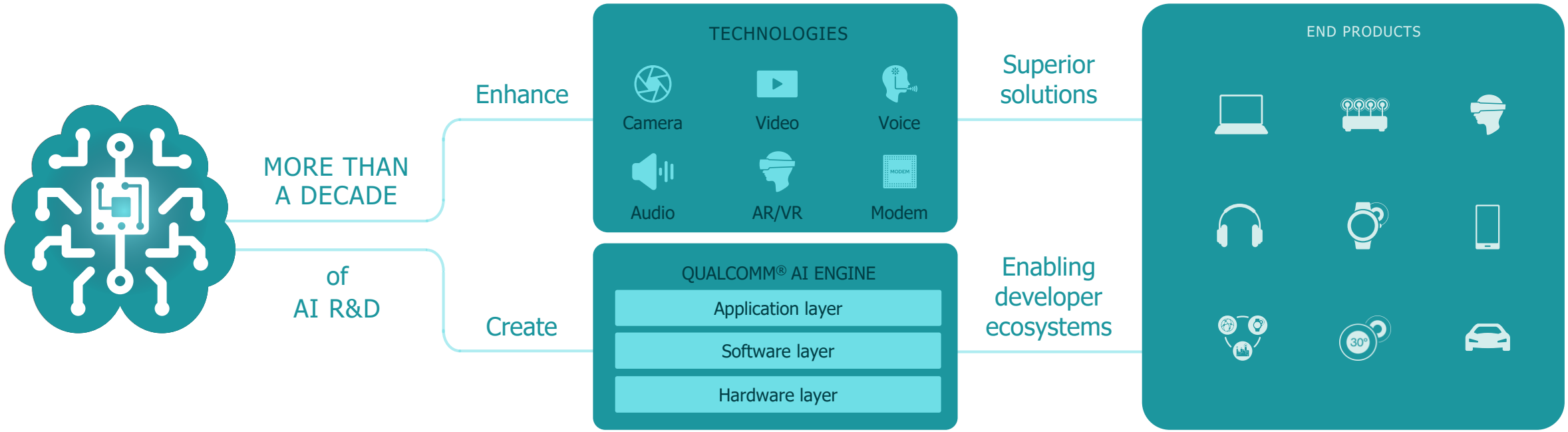
Cerebellum
Coordination

Frontal
Speech

Temporal
Facial Recognition
Hearing

	Modern Human Brain	Smartphone Chip
Evolution (Years)	~300,000	~12
Active Users	~7.5 Billion	~6.5 Billion
Avg. Weight	~4.5 kilograms	1-2 grams
Typical Power	~20 Watts	>5 Watts
Speed	10-15 KHz	1-30 GHz
Connections	~100 Billion Neurons	~10 Billion Transistors

We Apply AI Broadly Across our Business



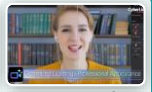
AI Use Cases

Productivity

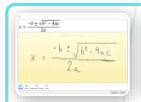
CNN based

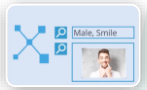

Noise Suppression


Background Blur


Face Beautify

NLP/NLU based


Ink to Text


Semantic Search


Deep Chart


Text & Sentiment Analysis


Whiteboarding


Designer Ideas

Content Creation/ enhancements


Detail Enhancement


RAW Processing


Colorize Photo


Style Transfer


Intelligent Zoom


Scene Adjust


De-Noising removal

GAN's based


Super Resolution


Exposure Compensation


Deep Fake

Battery life | Latency focused | Throughput heavy | Concurrently enabled

Streaming


Virtual Avatar


Background Replacement


Auto Subtitling (S2T)


Character Animation


Celebrity recognition


Recording Assistant

Gaming


Anti Cheat


Game Highlights

RL + LSTM based


ML Agents


Gaming AI Control Center


Auto Subtitling

Auto Markets


Object Detection


Vehicle Detection


Traffic Sign Recognition

Always ON


Face Detection

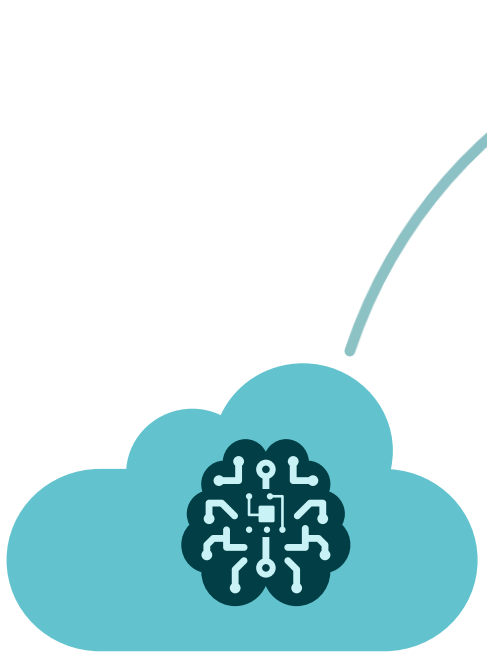
Commerce


Recommendations

Security

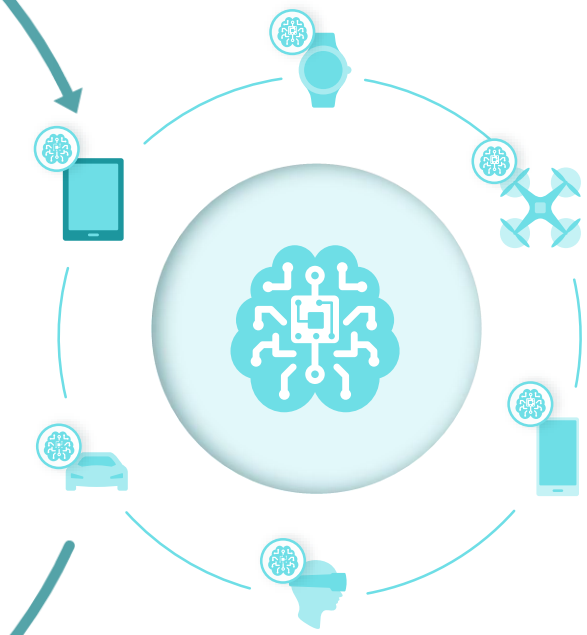

Defender

Center of Gravity Moving to the Edge…



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

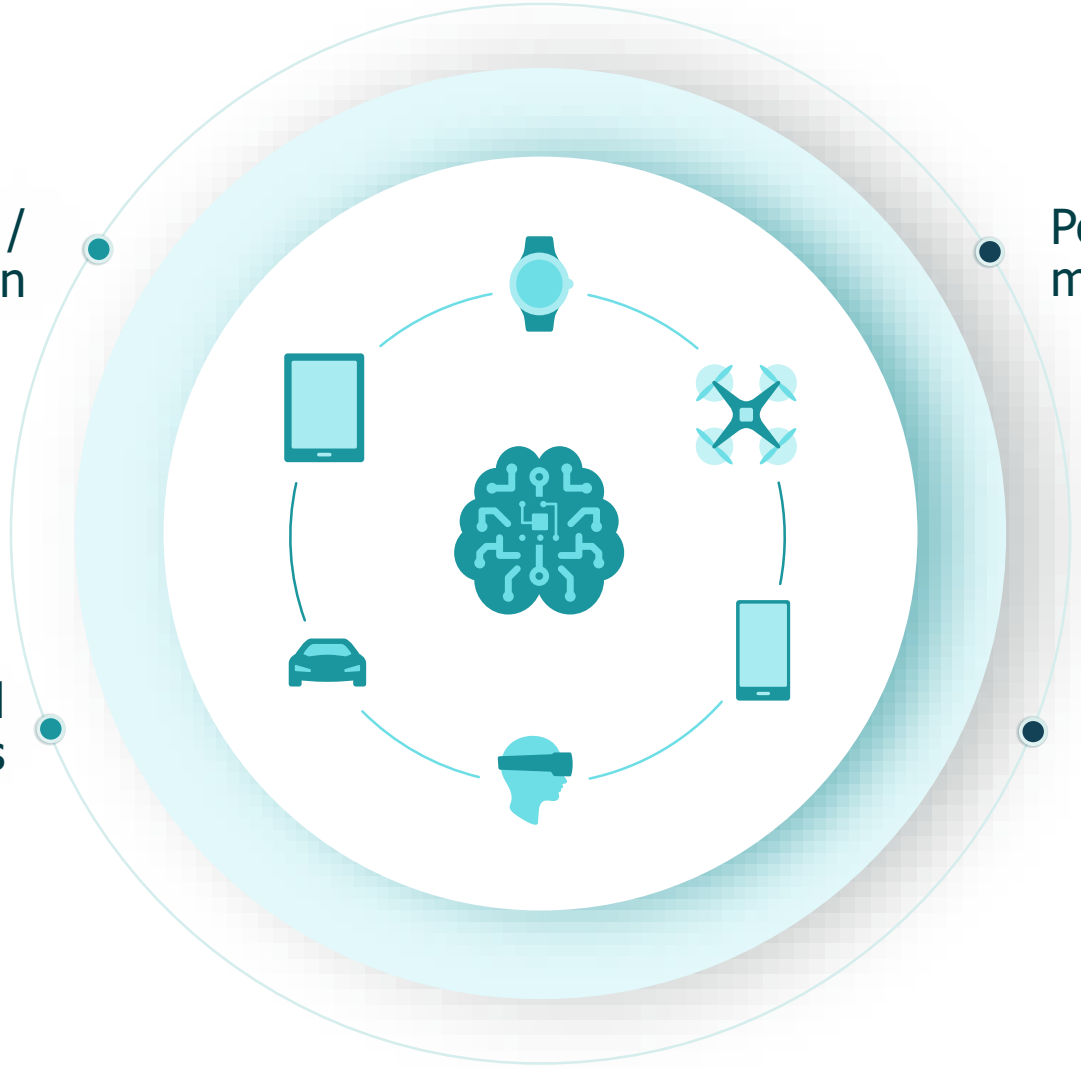
Processing data closer to devices at the edge drives new system values (e.g., lower latency, enhanced privacy)

Past	Today	Future
Cloud-centric AI	Partially-distributed AI	Fully-distributed AI
AI training and inference in the central cloud	Power-efficient on-device AI inference	With lifelong on-device learning

On-device AI Is Challenging

Data Privacy / Personalization

Power in a constrained mobile environment



Large complicated neural network models

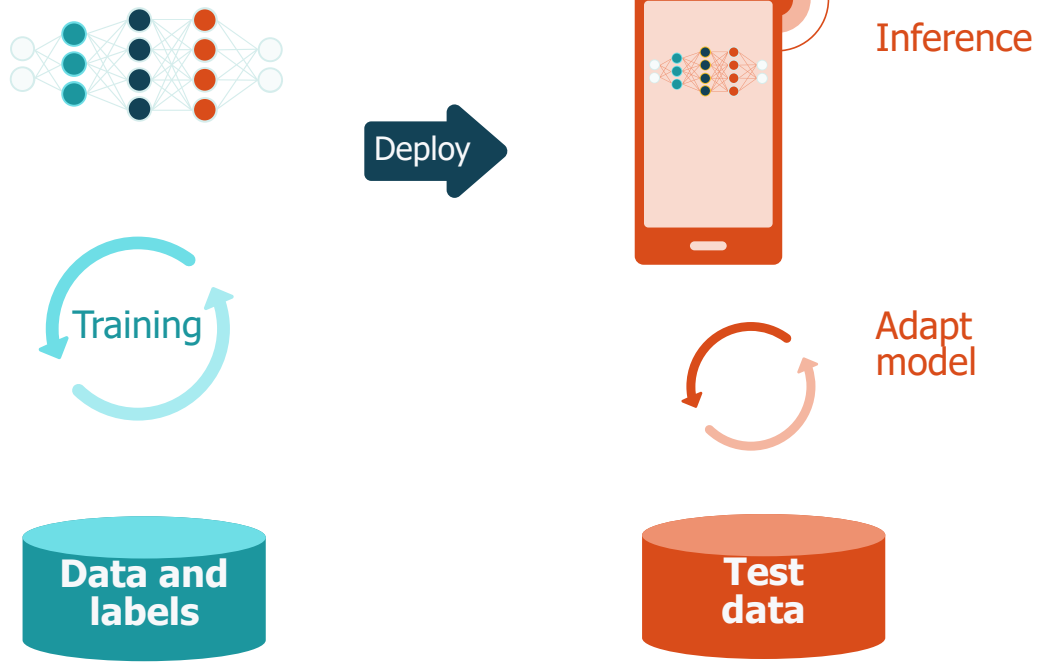
Leveraging across segments

Data Privacy / Personalization

Qualcomm

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

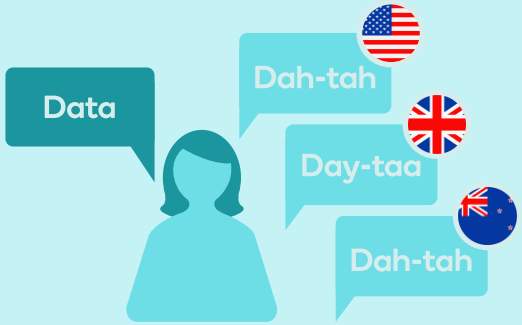
On-device learning can help to improve and maintain accuracy when original pre-trained model cannot generalize well

Few-shot Learning for Increased Personalization

Improving keyword spotting (KWS) performance of outlier users through on-device learning

Keyword spotting

Identify when a keyword is spoken using always-on ML



Keyword spotting challenge

- In practice, it is hard to collect all types of accented utterance
- The KWS model may not be sensitive to users' accents and have poor performance for outliers

Keyword spotting solution

- Locally adapt the model to user enrollments
- Personalize the model at enrollment time

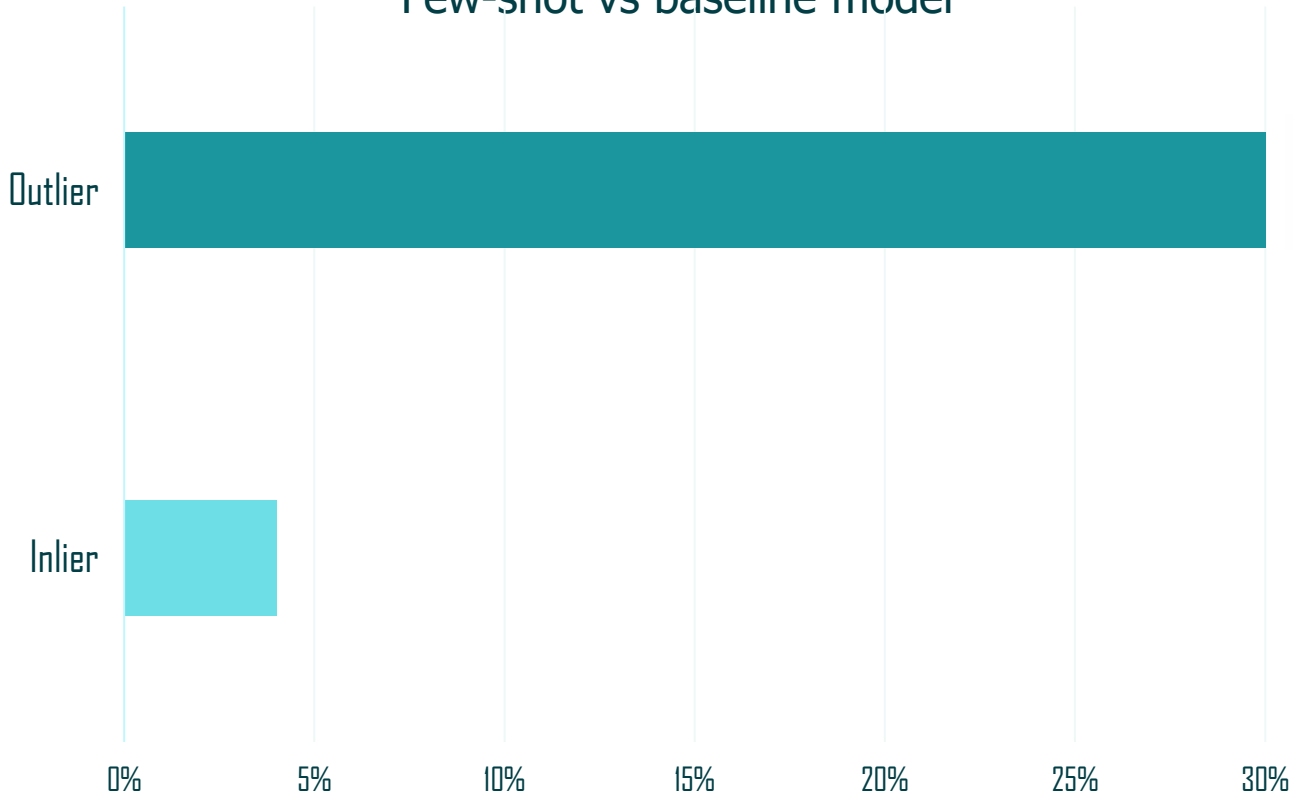
Detection rate for outlier users is over 30% worse, on average

Few-shot Learning for KWS Improves Performance

Personalization improvements across the board but particularly for outliers

Average detection rate improvement

Few-shot vs baseline model



Up to **30%** Improvement in detection rate*

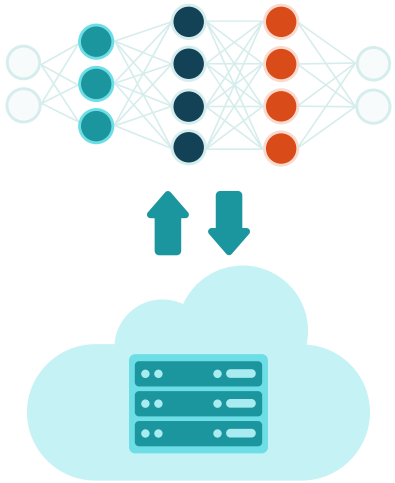
Qualcomm AI Research internal results.
For the few-shot results, performance is the average of all the locally adapted models.

Federated Learning for Global Adaptation

Federated learning brings on-device learning to new level

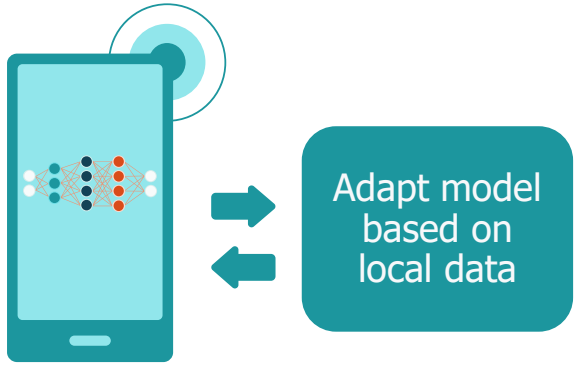
Adaptation on the device, once or continuously, locally and/or globally for continuous model enhancement

Offline learning



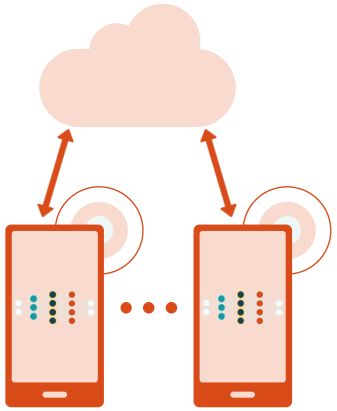
Data

On-device learning



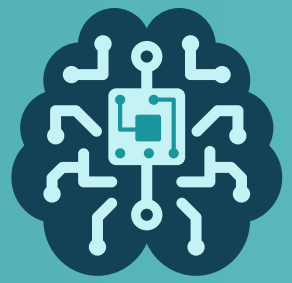
Locally adapt once to a few samples (e.g., few shot learning) or continuously (e.g., unsupervised learning)

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



Large Complicated Neural Networks Models

Qualcomm

BERT → DistilBERT

- **BERT:** Use bidirectional learning to help understand the context of each word from left and right—but this comes with a challenge
  - Number of parameters: 110M (Base version) & 340M (Large version)

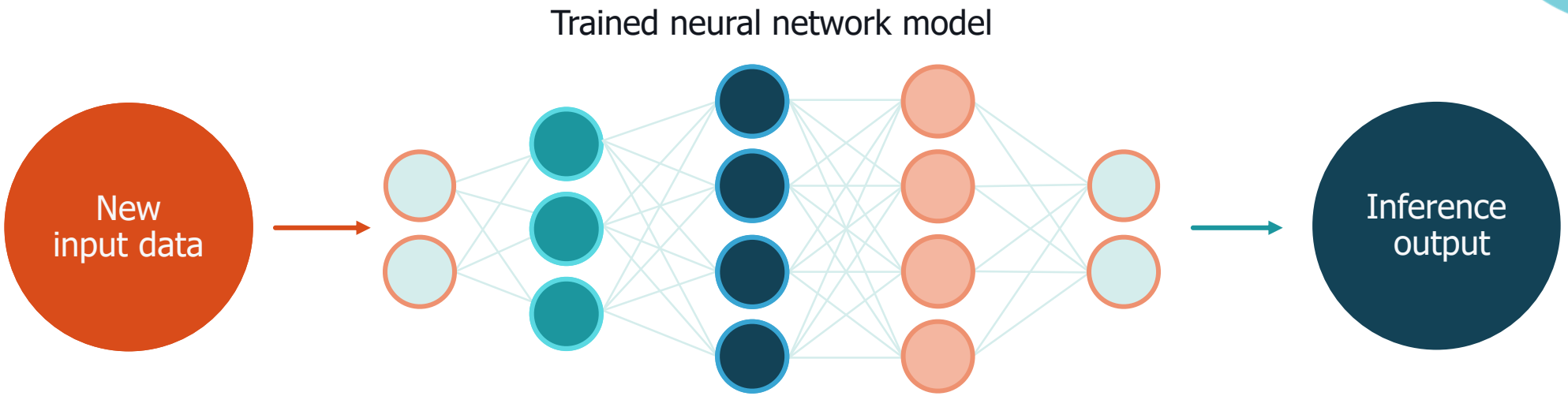
  ➔ Has good accuracy but comes at high computational cost not affordable across all platforms

- **DistilBERT:** Working with partners, look at enabling language understanding broadly where compute is challenged

  ➔ Has 95% accuracy of BERT (Base) with the half the parameter footprint

At Qualcomm, we support both, depending on the business vertical needs

Advancing AI to Optimize Model



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

Pushing the Limits of What's Possible with Quantization

Data-free quantization

How can we make quantization as simple as possible?

Created an automated method that addresses bias and imbalance in weight ranges:

- ☑ No training
- ☑ Data free

AdaRound

Is rounding to the nearest value the best approach for quantization?

Created an automated method for finding the best rounding choice:

- ☑ No training
- ☑ Minimal unlabeled data

Bayesian bits

Can we quantize layers to different bit widths based on precision sensitivity?

Created a novel method to learn mixed-precision quantization:

- ☑ Training required
- ☑ Training data required
- ☑ Jointly learns bit-width precision and pruning

SOTA 8-bit results

Making 8-bit weight quantization ubiquitous

<1% Accuracy drop for MobileNet V2 against FP32 model

Data-Free Quantization Through Weight Equalization and Bias Correction (Nagel, van Baalen, et al., ICCV 2019)

SOTA 4-bit weight results

Making 4-bit weight quantization ubiquitous

<2.5% Accuracy drop for MobileNet V2 against FP32 model

Up or Down? Adaptive Rounding for Post-Training Quantization (Nagel, Amjad, et al., ICML 2020)

SOTA mixed-precision results

Automating mixed-precision quantization and enabling the tradeoff between accuracy and kernel bit-width

<1% Accuracy drop for MobileNet V2 against FP32 model for mixed precision model with **computational complexity equivalent to a 4-bit weight model**

Bayesian Bits: Unifying Quantization and Pruning van Baalen, Louizos, et al., NeurIPS 2020)

Qualcomm

SOTA: State-of-the-art

Data Free Quantization Results in AIMET

Post-training technique enabling INT8 inference with very minimal loss in accuracy

DFQ example results

% Reduction in accuracy between FP32 ad INT8

<1%	<1%	<1%
MobileNet-v2 (top-1 accuracy)	ResNet-50 (top-1 accuracy)	DeepLabv3 (mean intersection over union)

DFQ: data free quantization

W4A8 Results Look Promising

Increase in performance per watt from savings in memory and compute[1]

Model	FP32	INT4 Accuracy	Comments
ResNet50	76.1%	75.4%	Using Post-training Quantization
ResNet18	69.8%	69%	
EfficientNet-Lite	75.3%	74.3%	
Regnext	78.3%	77.2%	
Mobilenet-v2	71.7%	71.3%	Using Quantization Aware Training

With better PTQ and QAT techniques, increasingly more models will be able to use W4A8, resulting in better energy efficiency

Power in a Constrained Mobile Environment

Qualcomm

Leading Techniques to Efficiently Quantize AI Models

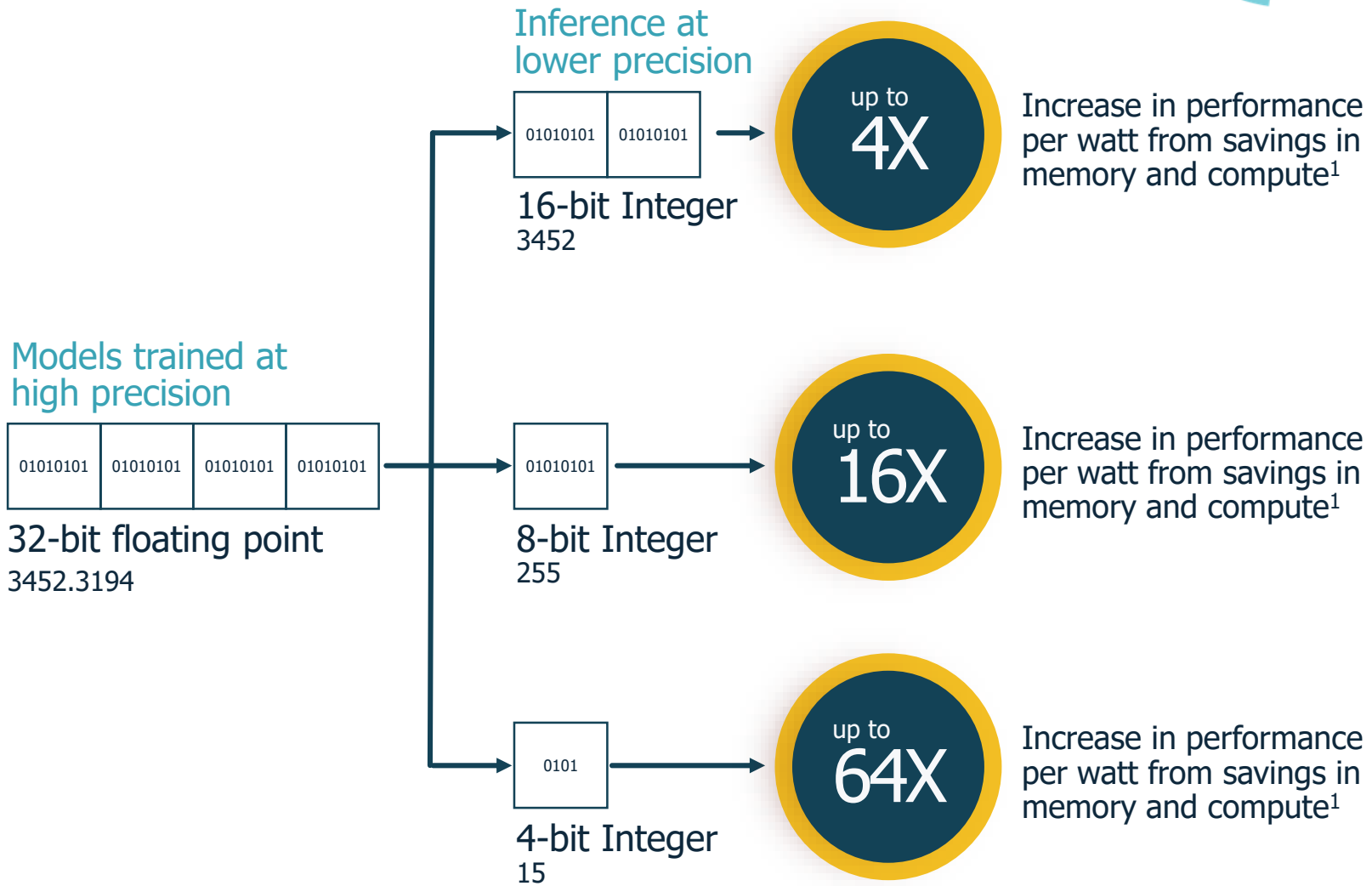
Automated reduction in precision of weights and activations while maintaining accuracy

Promising results show that low-precision integer inference can become widespread

Virtually the same accuracy between a FP32 and quantized AI model through:

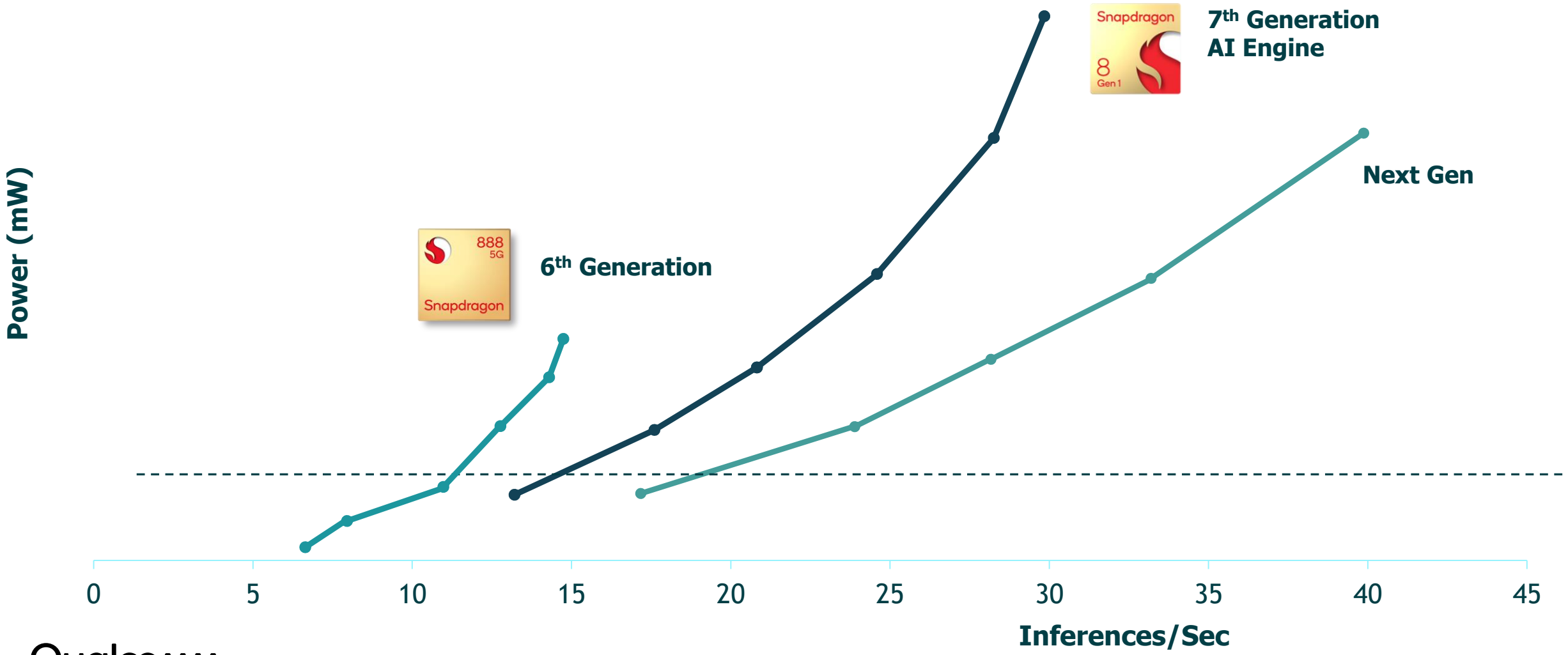
- Automated, data free, post-training methods
- Automated training-based mixed-precision method

Significant performance per watt improvements through quantization



1: FP32 model compared to quantized model

Improving AI Performance/W Across Generations



Snapdragon

Google Cloud
Vertex AI NAS

Search space

Set of operations and how they are connected to form valid network architectures

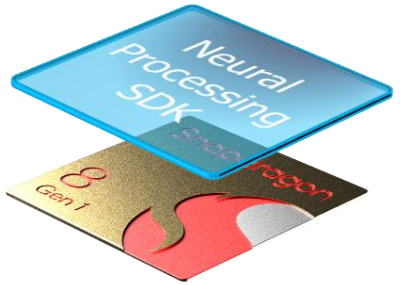
Search algorithm

Method for sampling a population of good network architecture candidates

Evaluation strategy

Method to estimate the performance of sampled network architectures

Vertex AI NAS
Qualcomm Neural Processing SDK
Integration

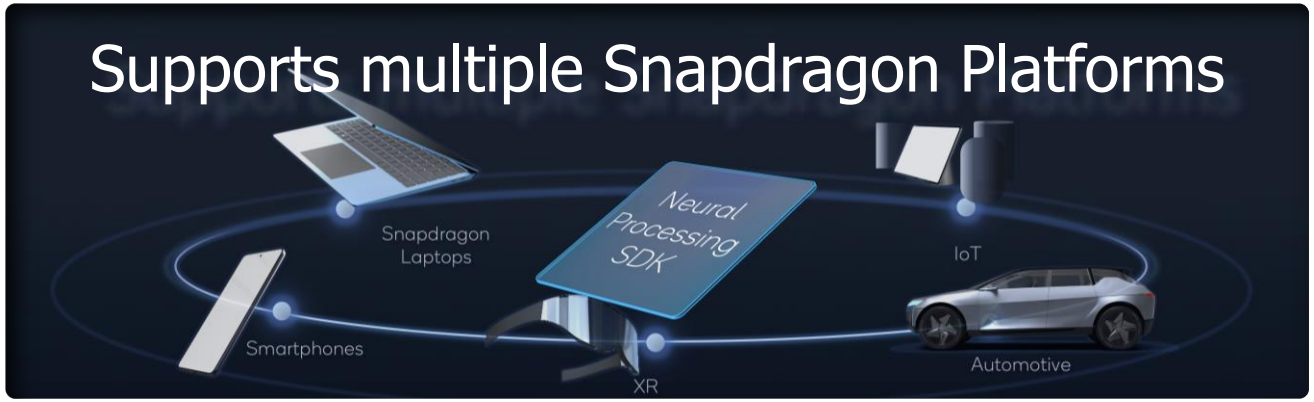


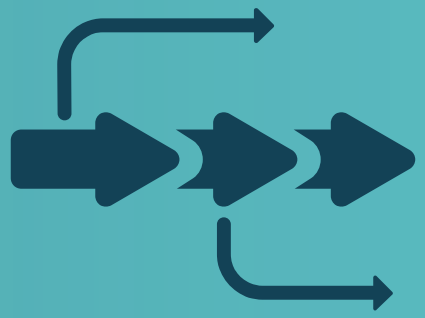
High Accuracy Smaller Models

Low Latency Higher Performance

8-10% Accuracy Improvements

~20 to 30% Latency Reduction





Leveraging Across Segments

Qualcomm

Optimizing and Deploying State-of-the-art AI Models for Diverse Scenarios at Scale Is Challenging

Neural network complexity

Many state-of-the-art neural network solutions are large, complex, and do not run efficiently on target hardware

Neural network diversity

For different tasks and use cases, many different neural networks are required

Device diversity

Deploying neural networks to many different devices with different configurations and changing software is required

Cost

Compute and engineering resources for training plus evaluation are too costly and time consuming

One Technology Roadmap that Scales to Address all Growth Vectors

Edge AI

Camera

GPU
Graphics

CPU
Processing

Connectivity

On-device intelligence

High performance, low power

Everything wireless

HIGH COMPLEXITY

LOW COMPLEXITY

AI Challenges: Support structure across BU's

Performance optimization points :

Innovative form factors are being constantly designed across many BU verticals and as such, one of the challenges is the ability to drive AI performance optimization (FPS or Latency or FPS/W) across multiple power envelopes

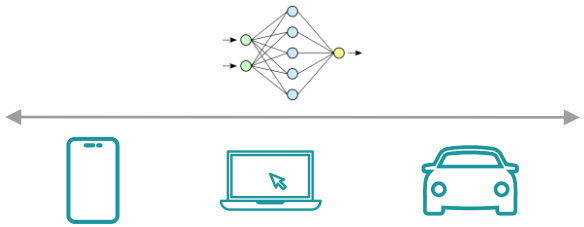
- **500 mW or less**
- **2 W or less**
- **5 W of less**
- **15 W to 200 W** (Varying TDP points)

Support for different DL architectures and operators :

AI based applications are quite widespread from Image quality related (Mobile) to productivity (Compute) to assistance and monitoring (Auto ADAS) markets. This stretches QCOM's internal ecosystem to support :

Challenging DL architectures:

- Generative models (Mobile) to transformer models (Compute) to Lidar models (Auto ADAS) which demands constant investment in compilers, tools, operators and other SW modules



Desired feature support :

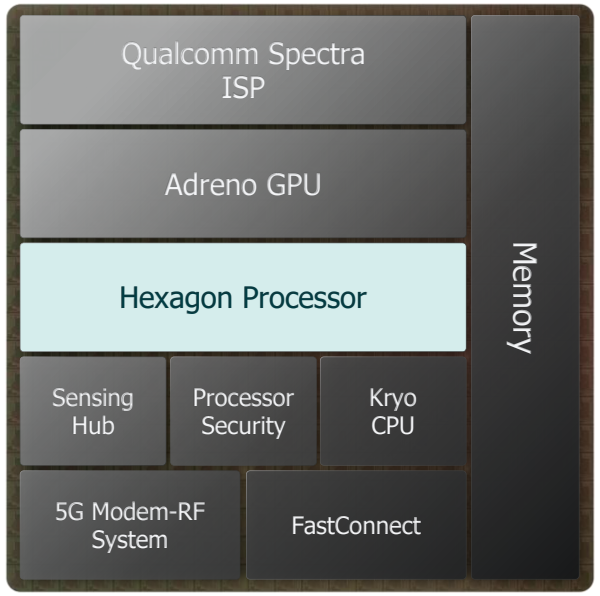
Ability to drive innovation using AI has seen increased traction in the ecosystem but the need for various feature support varies by BU vertical

- Support for high concurrency (For Auto and XR verticals)
- Support for newer data types (For Data center verticals)
- Support for application scalability (For Mobile and Compute verticals)

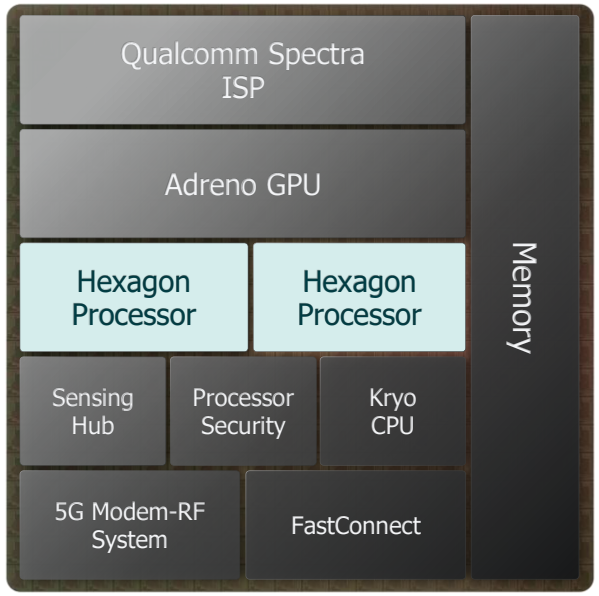
Adjacent Markets – Current Scale Model for Supporting AI

Hardware must scale to offer optimized hardware across product lines

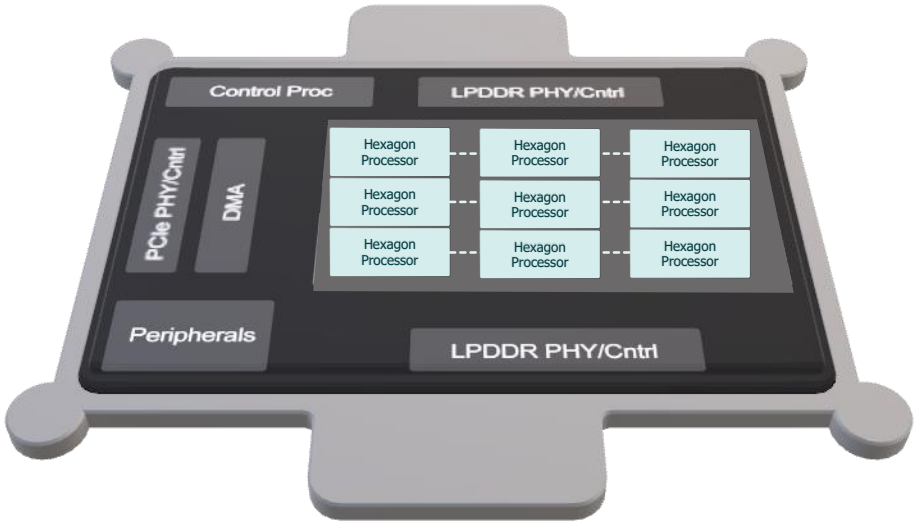
Anchor point at Mobile



Dual AI cores at Compute



Multiple AI cores at Cloud/Edge/Auto



Scaling AI HW for for different markets and AI needs

Tools + Compilers

AIMET

TVM

Models

ResNet | SSD | MobileNet | Mobile BERT | VDSR | DeepLab

Applications

Snapdragon

Qualcomm®
Neural Processing SDK

TensorFlow Lite

Android Neural
Networks API

Frameworks

ONNX

Runtime

SDKs

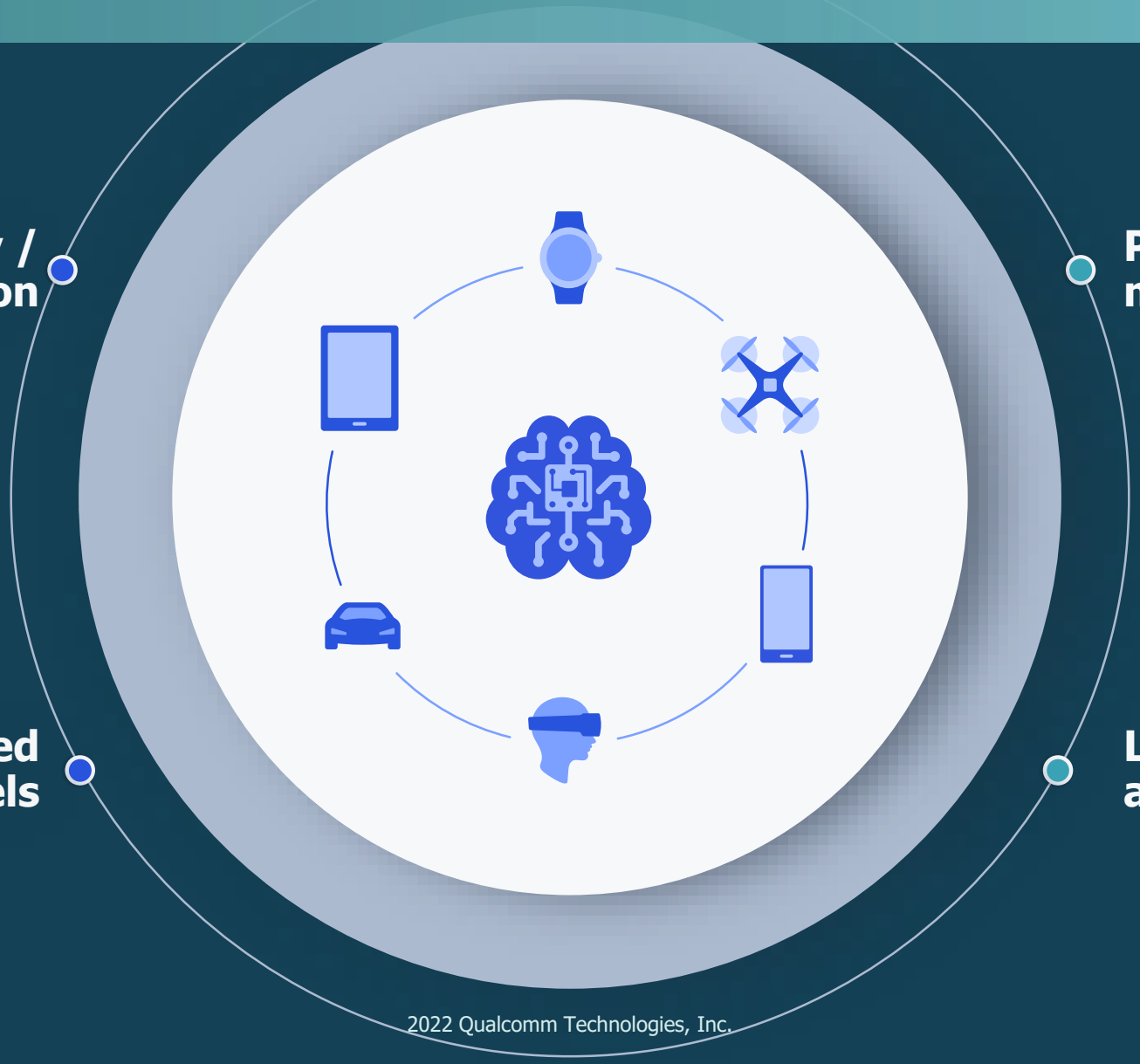
Qualcomm AI software stack scales across Products lines

Supporting every AI software layer from applications to the metal

Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc. and/or its subsidiaries

Qualcomm

On-device AI: Challenges & Proposed Solutions



Data Privacy / Personalization

On-Device learning
Federated Learning
Few shot learning

Large complicated neural network models

Compression & Quantization

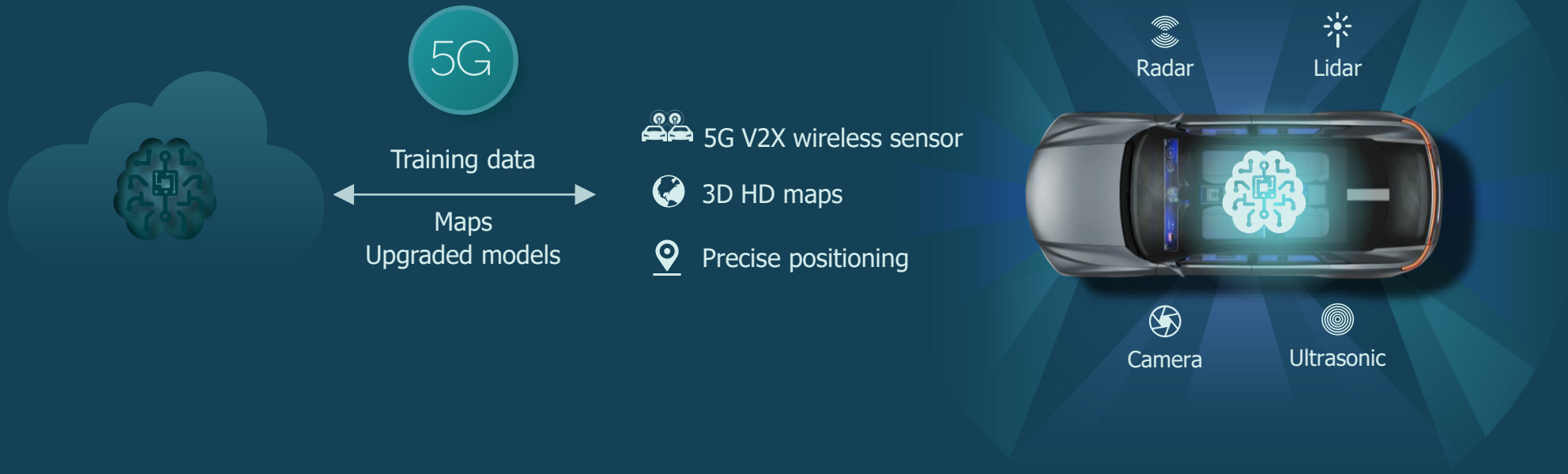
Power in a constrained mobile environment

Quantization for better efficiency
Improving Perf/W
NAS

Leveraging across segments

Single AI Stack
Scalable HW

Automotive On-device AI



Car makes local decisions

Latency

Reliability

Efficiency

Metaverse to Introduce New Challenges

AI Ecosystem Partners & Use Cases

Members of:
Qualcomm
Platform Solutions
Ecosystem

ArcSoft, Tetras.AI, JIIGAN 极感科技, BlinkAI, IMMERVISION, Almalence Innovative Imaging Technologies, Black Sesame Technologies

XPERI, Mobius Labs, vidhance, NEXOPTIC, mobilize, trinamiX, loom.ai

ultraleap, AMS, ellipticlabs, Myelin Foundry, TikTok, 有道 youdao 网易旗下搜索, TRUEPIC

Audio Analytic, ELEVOC 听得见的智能, Tencent, 思必驰 AISPEECH, babblelabs, Adobe, morpho

Camera Effects
Augmented Reality Lenses
Camera filters/lenses
Style Transfer
2D/3D Avatars
Beautification
Creative Movie Effects
3D Depth/Reconstruction

Language Processing
Keyword Detection
Automatic Speech Recognition
Transcription
Language Translation

Gestures/Proximity
Ultrasound-based Presence det.
Hand Gestures

Content Attribution
Image/Video Authenticity

Context
Contextual Stickers
Contextual Sound

Camera/Video IQ
Low-light imaging
Wide angle imaging
Intelligent Video Zoom
Semantic Segmentation
HDR , Denoise
Portrait enhancements
SAT
Video Stabilization
Bokeh
Digital Zoom (SR)
Scene/Object Detection
Remosaic/Demosaic

Speech Pre-Processing
Noise suppression

Biometrics
2D/3D Face Authentication
Voice Print

Gaming
Super Resolution
AI Agents
Anti-howling

Video Post Processing
AI Upscaling

Resources

Qualcomm Mobile AI

Mobile AI | On-Device AI | Qualcomm®

Qualcomm & Google NAS

Qualcomm Technologies and Google Cloud Announce Collaboration on Neural Architecture Search for the Connected Intelligent Edge | Qualcomm

Ziad Asghar
Vice President, Product Management
zasghar@qti.qualcomm.com

2022 Embedded Vision Summit

- *"Seamless Deployment of Multimedia and Machine Learning Applications at the Edge"* **Megha Daga May 17 2:40 - 3:10 PM PT**
- *"Tools for Creating Next-Gen Computer Vision Apps on Snapdragon"* **Judd Heape May 18 10:50 - 11:20 AM PT**
- *"The Future of AI is Here Today: Deep Dive into Qualcomm's On-Device AI Offerings"* **Vinesh Sukumar May 18 12:00 - 12:30 PM PT**
- *"A Practical Guide to Getting the DNN Accuracy You Need and the Performance You Deserve"* **Felix Baum May 18 2:40 - 3:10 PM PT**

Qualcomm

Thank You

Qualcomm