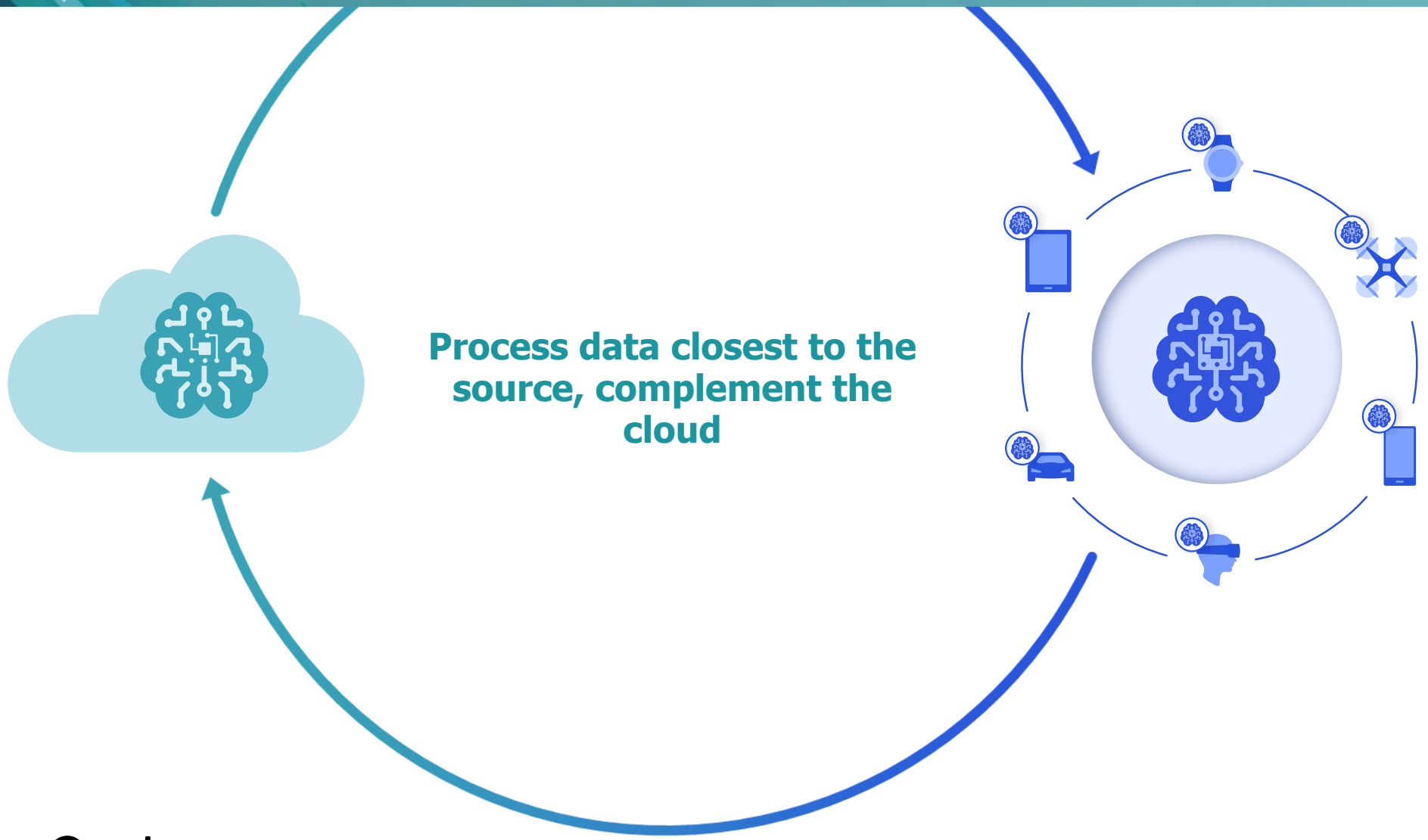
The Future of AI is Here Today: Deep Dive into Qualcomm's On-Device AI Offering

Vinesh Sukumar
Sr. Director, Product Management (AI/ML)
Qualcomm Technologies, Inc.

Center of Gravity Moving to the Edge…



Process data closest to the source, complement the cloud

Historically

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

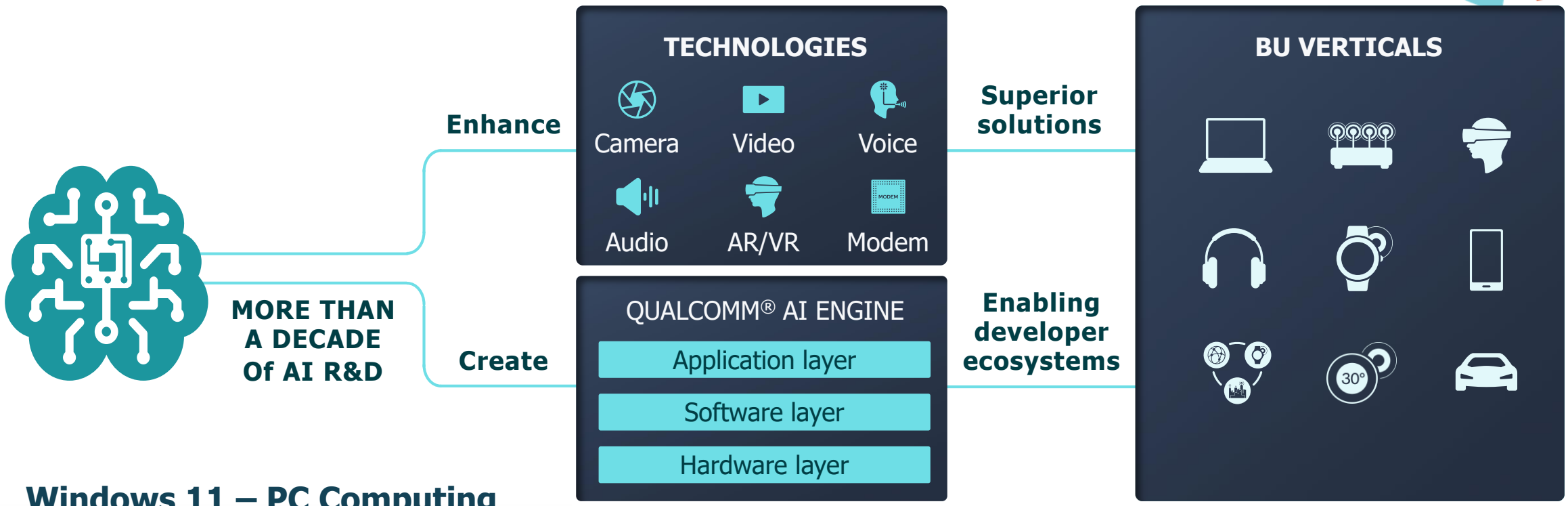
Increased Demand

- Autonomy
- Personalization
- Efficiency
- Security

Qualcomm

© 2022 Qualcomm Technologies, Inc.

At Qualcomm – AI Deployed Across Various Technologies & Verticals



Windows 11 – PC Computing

Devices will need a neural processing unit (NPU) to use these new Windows 11 features, which means they'll show up first on Lenovo's new ThinkPad X13s, which is powered by Qualcomm's Snapdragon 8cx Gen 3 compute platform.

Smart Camera – IoT Markets

Qualcomm debuts smart camera processor at ISC West

Qualcomm showcased a smart camera processor named QCS7230 at the International Security Conference and Exposition (ISC West) that expands its Vision Intelligence Platform portfolio to cities, enterprises, and public spaces.

ADAS Markets

Ferrari and Qualcomm team up for tech projects for road, racing cars

REUTERS

Qualcomm

Snapdragon, Qualcomm QCS7230, and Qualcomm AI Engine are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

© 2022 Qualcomm Technologies, Inc.

AI Applications

Qualcomm

AI Applications: Across Various Segments

Expanding beyond modalities of computer vision to linguistics, communication, commerce and language understanding

Mobile

AI Assisted Imaging
- AI 3A
- Scene-based Camera Selection

Image Understanding
- Face Detection / Tracking / Features
- Object Detection / Tracking
- Body Detection / Tracking / Pose
- Human Segmentation
- Sky Segmentation
- Multi-Class Segmentation
- Depth Estimation

Beautify / Augment
- Scene-based Image Enhancement

Image Processing
- AI based NR or Image SR
- Scene-based Camera Selection

Audio
- Real time language
- Natural language processing (NLP)

Modem
- Parameter optimization
- Robust sequence predictions

IoT

Robotics
- Autonomous navigation
- Obstacle Avoidance
- Picking and Sorting

Retail
- Visitor/Face/Gesture Recognition
- Object/People Detection and Counting
- Barcode decoding
- Empty shelf detection
- Dwell time

Transportation
- License plate recognition
- Face and facial landmark detection
- Drowsiness detection

Smart Devices
- Object/People detection
- Speaker detection
- Gun shot detection

Smart Buildings
- People Tracking
- Access Control

Manufacturing/Logistics
- Predictive maintenance
- Energy management with Asset demand

Compute

Productivity
- Background based noise cancellation on Audio (inbound and outbound)
- Segmentation/Blur/Super Resolution on Video
- Voice activation without keywords
- Face tracking
- Smart photo categorization

Privacy & Security
- Automatic screen unlock and login
- Privacy alert
- Guard mode

Content Creation & Gaming
- Gaming with gesture control
- Gaming with voice commands
- Intelligent highlight videos
- Game play improvement

Performance & Efficiency
- Power and Screen optimization

Cloud

Data Centers
- Natural language processing
- Computer vision
- Recommendation system

Edge Compute
- Theft detection
- Face/body/license plate detection / recognition
- Image classification and segmentation

Auto

IVI
- Occupancy monitoring system (OMS)
- Driver monitoring system (DMS)
- Surround perception
- Audio Command & Control

ADAS (Up to L4)
- Highway driving assist
  - Front collision warning
  - lane departure,
  - Traffic jam assist
  - Auto lane change
  - Auto lane merge
  - Traffic light recognition
  - Construction zones
  - Urban autonomous driving
- Parking assist
  - Person detection,
  - Perception
  - Valet parking
- Driver monitoring

Innovation centered around energy efficiency & personalization

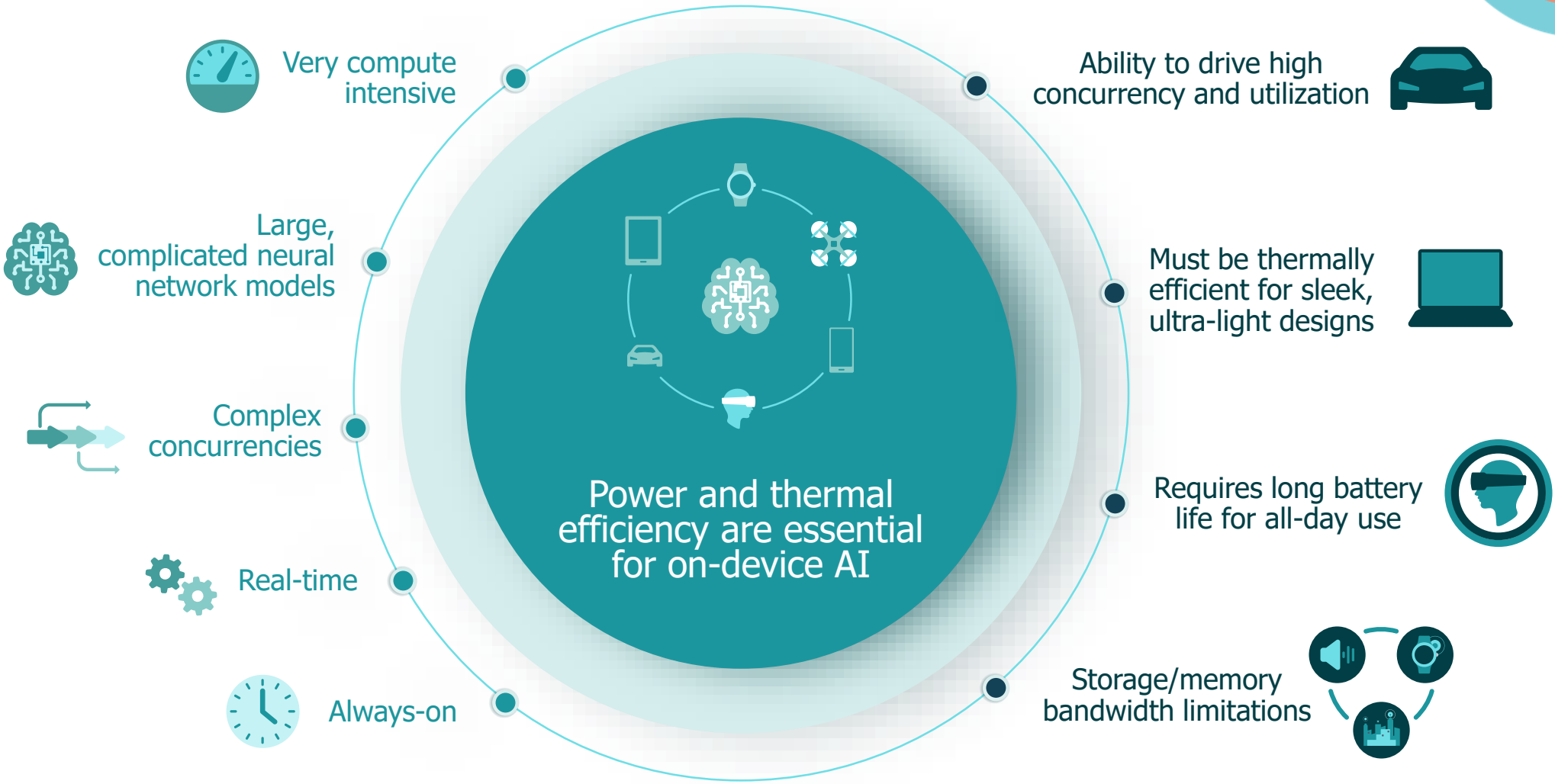
Innovation centered around making AI relevant in PC

Innovation centered around supporting high accuracy, latency & heterogeneity

Qualcomm

Few examples → Battery Life | Latency Focused | Throughput Heavy | Concurrently Enabled

Challenges of AI Applications



Lays the foundation for Qualcomm AI Hardware

Qualcomm

© 2022 Qualcomm Technologies, Inc.

AI Hardware

Qualcomm

Vision: Drive Leadership Capability Across All Markets

Performance

Invest in performance (Inf/Sec) and Power efficiency (Inf/Sec/W)

Scalability

Leverage existing AI HW engines to scale across various TDP points

Innovation

Feature innovation to drive leadership (Datatypes, Sparsity, Streaming…)

Co-Design

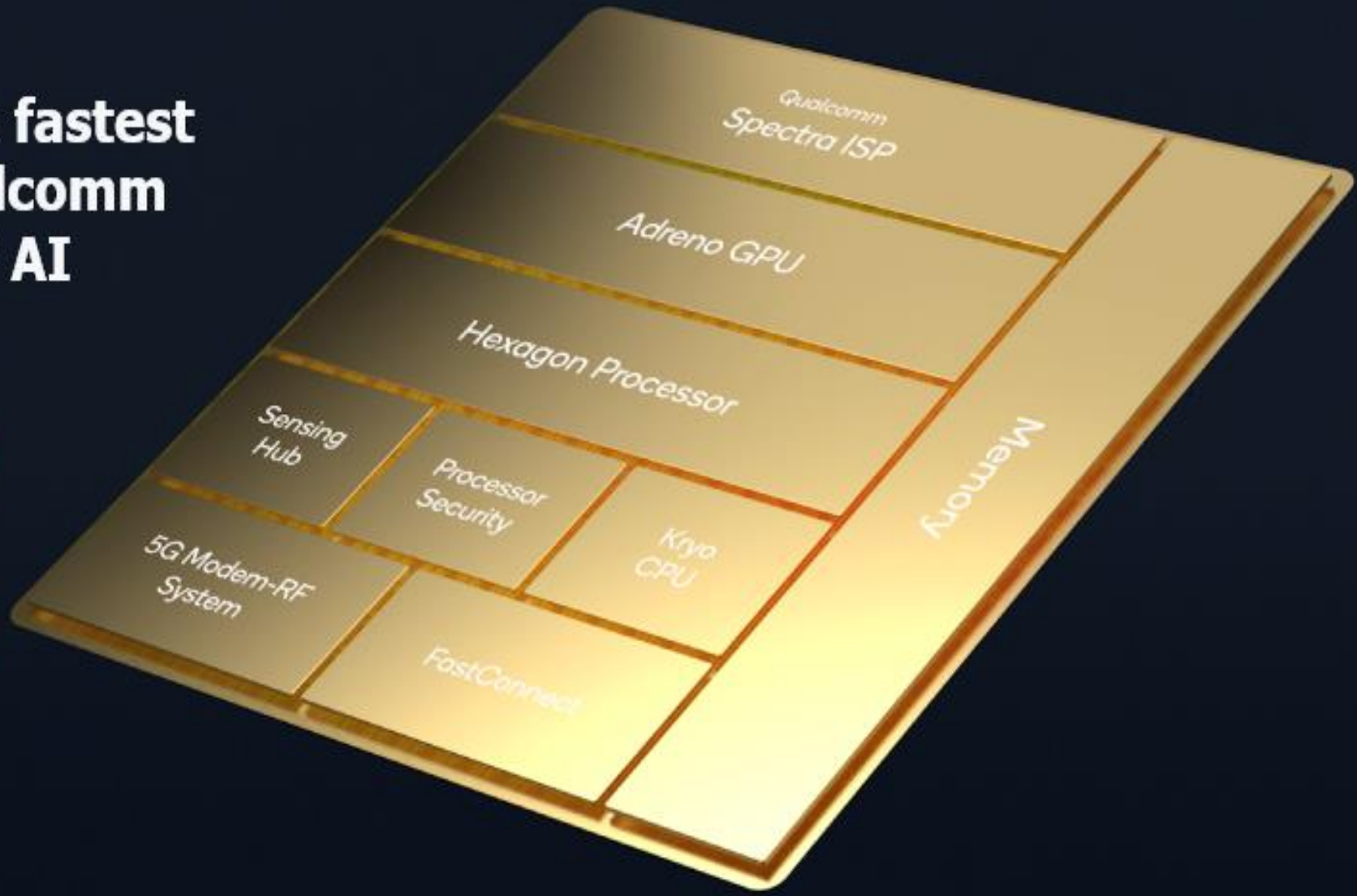
HW & SW co-design to make programmability easier (NAS, Compatibility)

Qualcomm

© 2022 Qualcomm Technologies, Inc.

7th Generation : Introducing Qualcomm AI Engine for All Verticals

The most powerful & fastest AI Engine from Qualcomm to run "Modern" AI



Qualcomm

© 2022 Qualcomm Technologies, Inc.

7th Generation : Qualcomm AI Engine

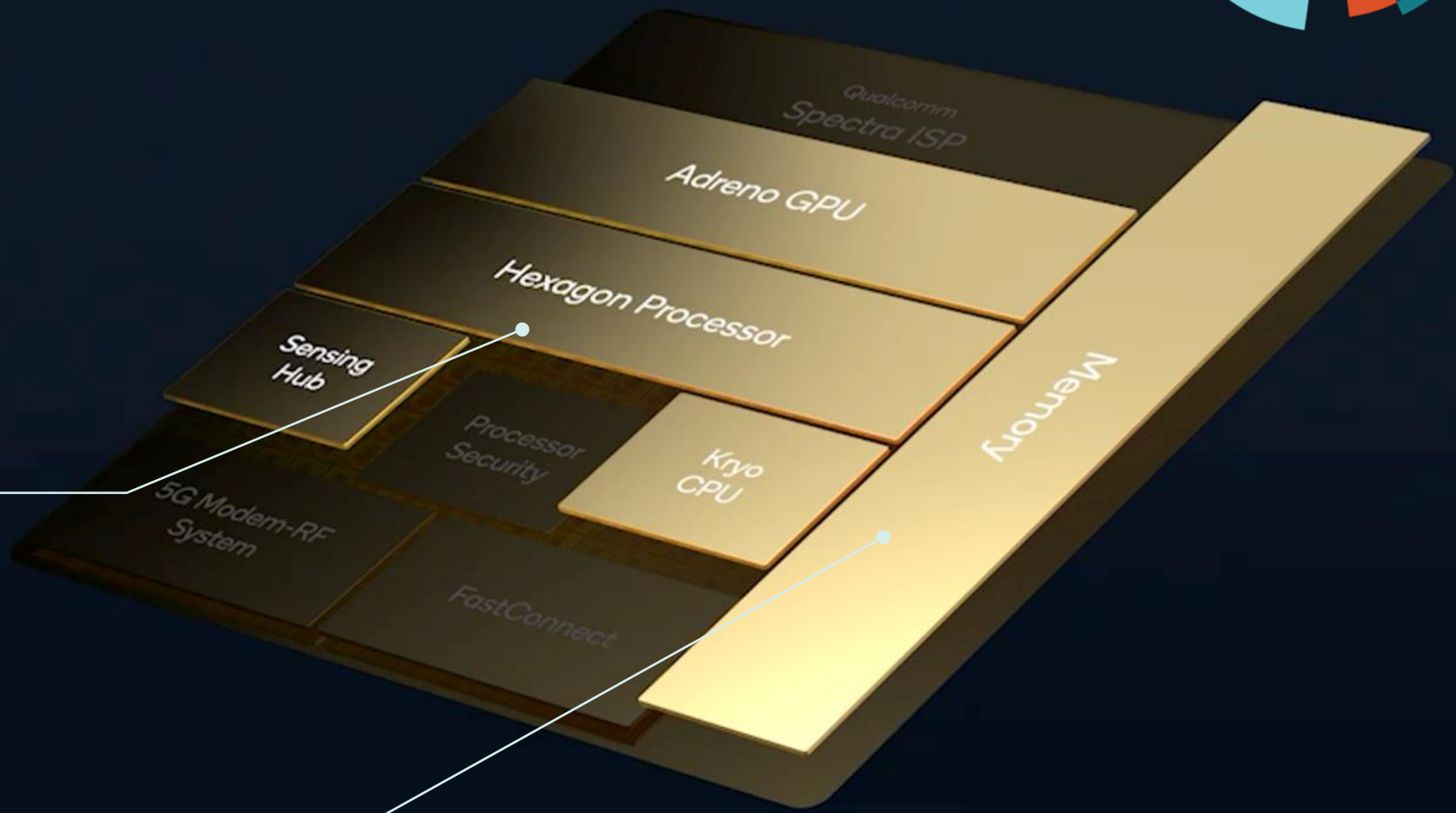
Snapdragon 8 Gen 1

2X

Computational performance

2X

Large-shared memory



Qualcomm

© 2022 Qualcomm Technologies, Inc.

Scaling – Dedicated Investments in AI HW Engines

Using Mobile Design as an Anchor Point

Investment in sparsity modules for supporting Auto ADAS usages

Optimize design for sustained low power with the ability to enable high concurrency

Investment in dedicated datatypes for higher performance and TCO (Total Cost of Ownership)

Qualcomm

© 2022 Qualcomm Technologies, Inc.

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

Performance Leadership – Using MLPerf ™ 2.0

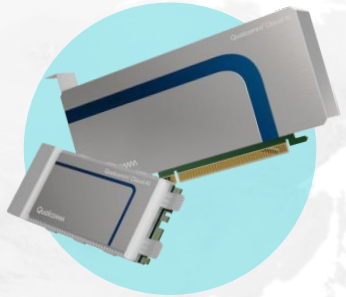
Qualcomm® 8C GEN 1 Mobile

Xiaomi Mi12 Platform

RESNET50	MOSAIC	BERT
2221 Inf/sec	752 Inf/sec	101 Inf/sec

Scaling from Mobile to Cloud..

Qualcomm® Cloud AI 100



RESNET50	16 75W cards in one server	20 server units in one server rack
	371,473 Inf/sec	7.4+M Inf/sec

Gigabyte Platform

8C GEN 1 achieves an average of about 26% better latency than Exynos devices across various categories

Cloud AI 100 servers achieve 3.7x higher rack-level ResNet-50 inference performance than Nvidia A100 servers

Qualcomm

© 2022 Qualcomm Technologies, Inc.
Qualcomm Cloud AI 100 is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

AI Software

Qualcomm

Vision: Accelerate AI Innovation and Solution Deployment

Performance

Accelerate “out of box” operator functionality and performance

Scalability

Ability to have programming consistency from Cloud to Edge

Tools

Accelerate AI Solution deployment with investment in Tools

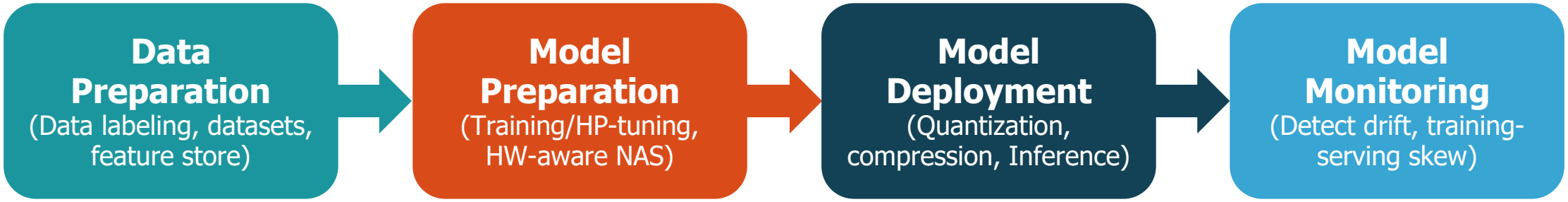
Innovation

Innovation to drive product leadership (Pre-emption, DFS, Multi Chaining)

Qualcomm

© 2022 Qualcomm Technologies, Inc.

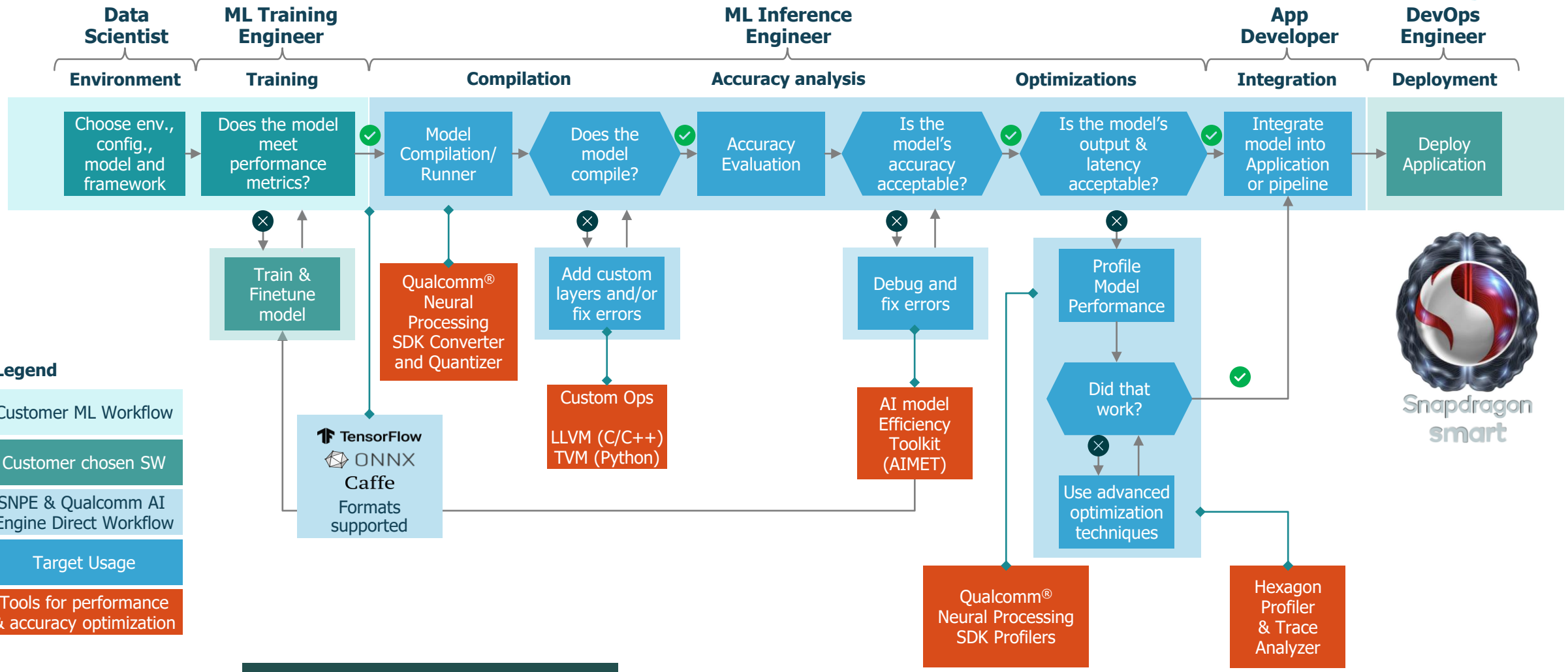
Qualcomm AI SW MLOps Cycle (Inception to Monitoring)



Focus for SW Discussion Today

© 2022 Qualcomm Technologies, Inc.

AI Software Workflow



Qualcomm

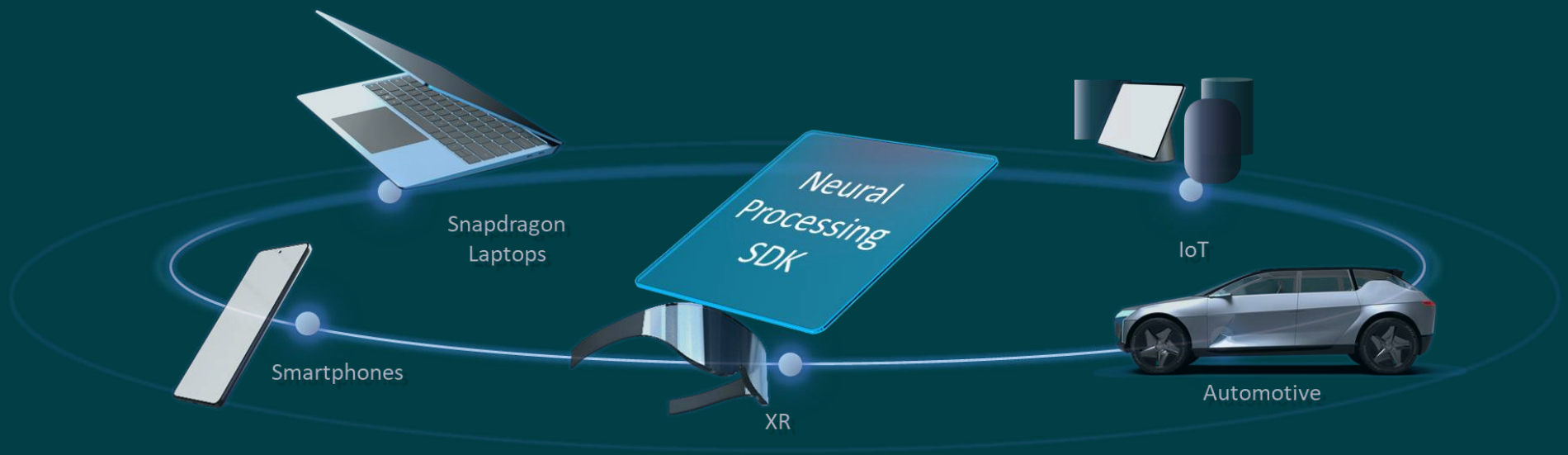
Qualcomm Neural Processing SDK and Qualcomm AI Engine direct are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

© 2022 Qualcomm Technologies, Inc.

STEP : 1 | NAS : Model Optimization Mapped to HW Intrinsics

Partnering with Google Vertex AI Team - Integrated into Qualcomm Software Stack

Snapdragon | Vertex AI NAS



How →

Search Space
Space of allowable architectures (Structure, operations, connectivity)
Search Algorithm
Sampling populations of good architecture candidates
Evaluation Strategy
Estimate performance of sampled architecture

Results →

- 8-10 % Accuracy Improvements
- ~20 to 30% Latency Reduction

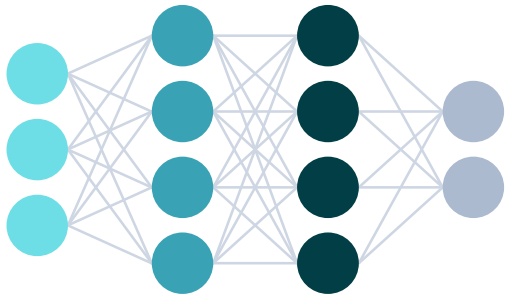
Note : DL architecture choices influence results

Qualcomm

© 2022 Qualcomm Technologies, Inc.

STEP : 2 AIMET : Quantize & Compress AI Models

Trained
AI model



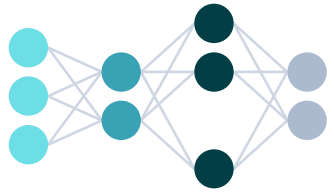
TensorFlow or PyTorch

AI Model Efficiency Toolkit
(AIMET)

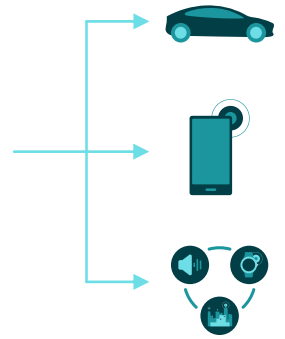
Compression

Quantization

Optimized
AI model



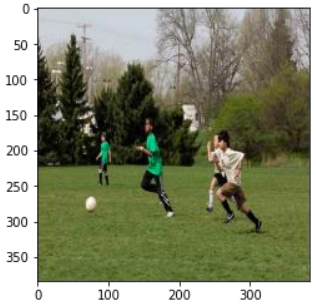
Deployed
AI model



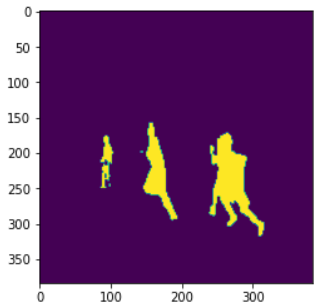
Why → Automated way of enabling reduction in precision of weights and activation while maintaining accuracy

State of the art network compression and quantization tools for various DL architectures (CNN, BERT, GAN's..)

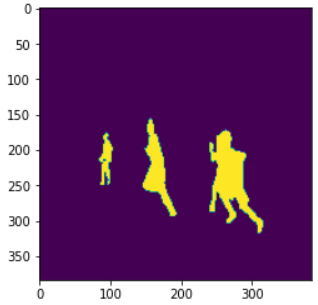
Results →



Original Image



FP16 Segmented Map



Quantized Segmented Map

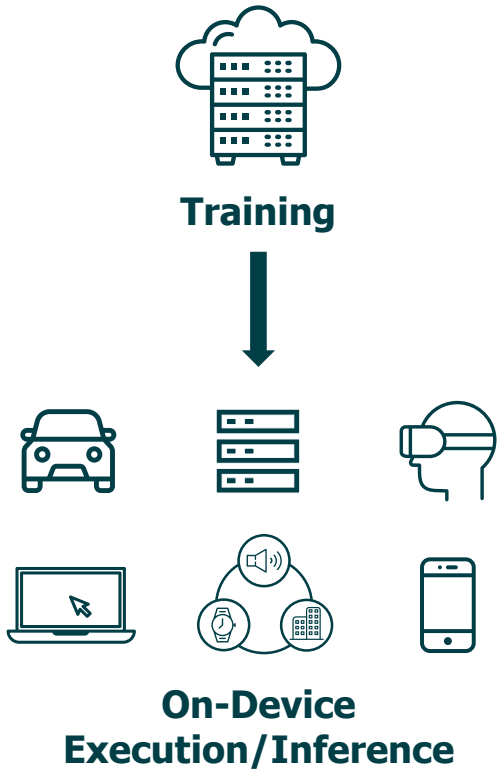
<0.5% Drop in Accuracy (Across various DL Models)
-> Using QAT and PTQ techniques from AIMET

Qualcomm

AIMET is a product of Qualcomm Innovation Center, Inc

STEP : 3 Run Time: Qualcomm AI Software Stack for Performance and Scalability Support

- Application Deployment



facebook | Preferred Networks | amazon | Microsoft | Baidu 百度 | Google

Caffe2 | PyTorch | Chainer | mxnet | Microsoft Cognitive Toolkit | PaddlePaddle | TensorFlow

.onnx | .onnx | .onnx | .onnx | .onnx | .onnx | .pb

Profiler

Debugger

Visualizer

Compilers

Performance — Scalability

Runtime framework

Qualcomm® Neural Processing SDK

PT Mobile

TF-Lite

NNAPI

ONNXRT

TF-Lite µ

Your–NN-Framework

Low level Library

Qualcomm Neural Network Library

QML | KERNELS | KERNELS | eAI

NEON | Open CL | Open CL

AI Engine

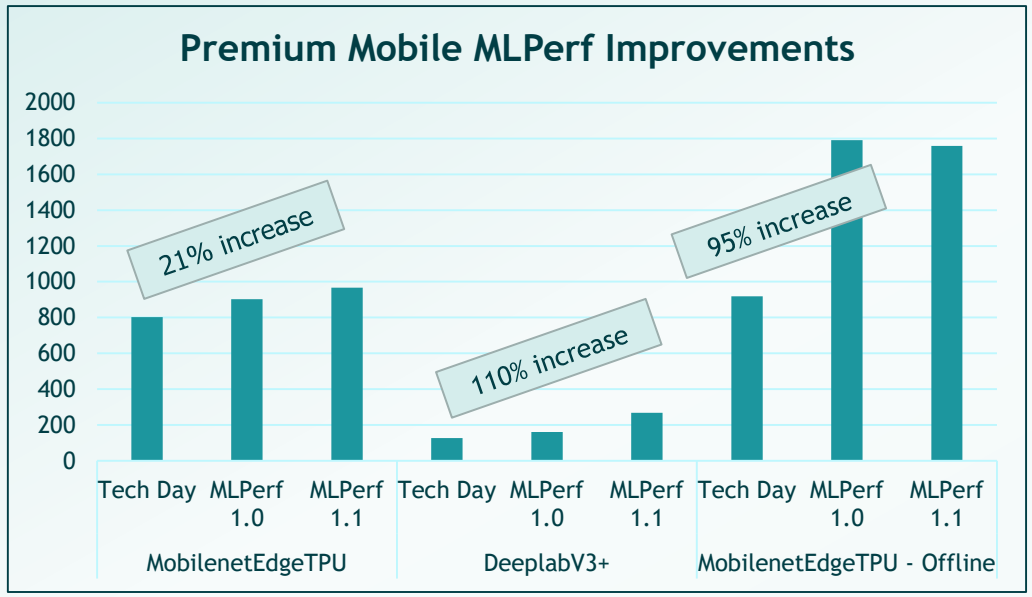
CPU | GPU | Hexagon Processor | Qualcomm Sensing Hub

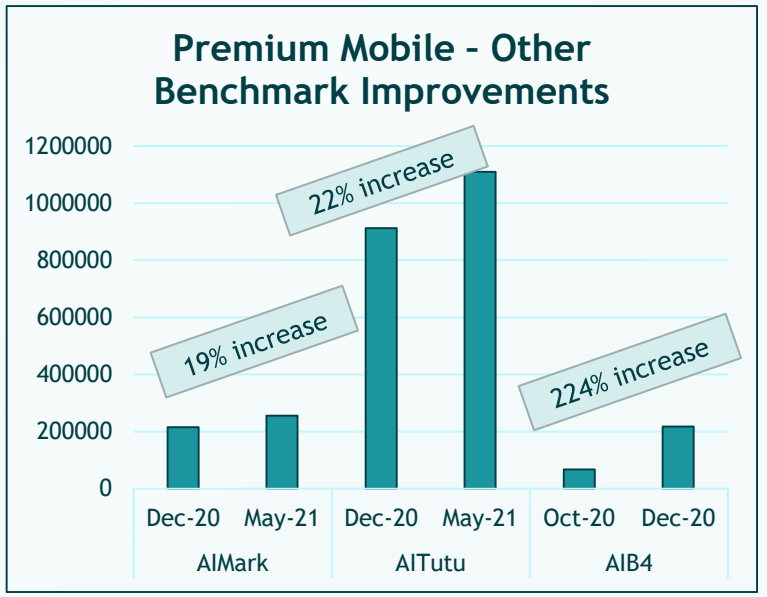
QNX | Ubuntu | Linux | Chrome | Windows | android

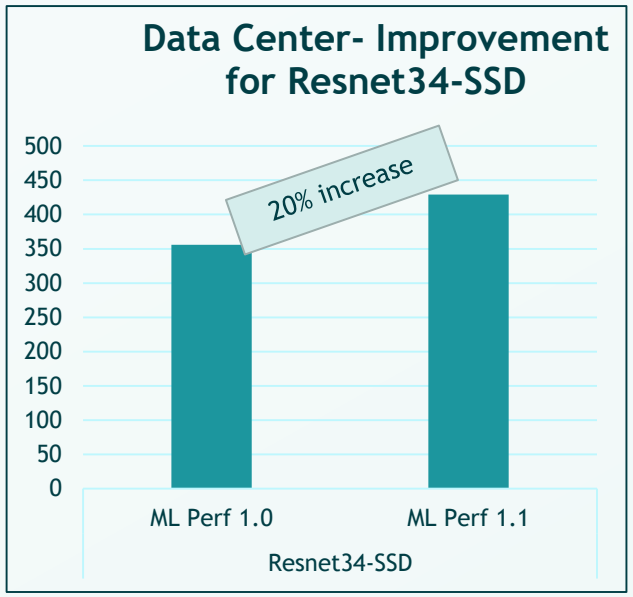
General purpose compute engines — AI engines

Qualcomm

STEP : 4 Continuous Improvements in SW Stack
- For Performance Leadership







- Meaningful improvements after initial chip shipment ; achieved through software investment / innovation

- **Consequences -**
  - Additional gains across the entire SoC portfolio for downstream chips and adjacent BUs from a common investment
  - Platform software that can be delivered incrementally to already released products can continue to improve
  - Improvements directly leveraged into next generation chips
  - Improvements in current generation can lead to relative performance gap compression with next-gen devices

Qualcomm

© 2022 Qualcomm Technologies, Inc.

Conclusions

Qualcomm

Conclusion

- AI Applications expanding beyond modalities of computer vision to linguistics, communication, commerce and language understanding
- With evolution of AI Applications across many verticals, continued push for innovation around latency, heterogeneity, concurrency and user personalization
  - This is putting a lot of emphasis for the need of custom HW modules in several BU verticals in Qualcomm
- Qualcomm silicon continues to show leadership in performance and energy efficiency in industry leading benchmarks
- High investment in software continues to accelerate AI solution deployment across all verticals

© 2022 Qualcomm Technologies, Inc.

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

Vinesh Sukumar
Senior Director, Product Management – AI/ML
vinesuku@qti.qualcomm.com

2022 Embedded Vision Summit

- *"Powering the Intelligent Connected Edge and the Future of On-Device AI"* **Ziad Asghar May 18 9:30 - 10:00 AM 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**
- *"Tools for Creating Next-Gen Computer Vision Apps on Snapdragon"* **Judd Heape May 18 10:50 - 11:20 AM PT**
- *"Seamless Deployment of Multimedia and Machine Learning Applications at the Edge"* **Megha Daga May 17 2:40 - 3:10 PM PT**

© 2022 Qualcomm Technologies, Inc.

Thank You

Qualcomm