Making AI ubiquitous

Qualcomm Technologies, Inc.
Devices, machines, and things are becoming more intelligent
Offering new capabilities to enrich our lives

**Reasoning**
Learn, infer context, and anticipate

**Perception**
Hear, see, monitor, and observe

**Action**
Act intuitively, interact naturally, and protect privacy
A world where virtually everyone and everything is intelligently connected
Edge cloud

On-device AI, processing, sensing, vision,... augmented by edge cloud

The intelligent wireless edge

Process data closest to the source to scale for massive amount of data and connected things

5G

Distributed autonomy

Privacy/security

New experiences

Immediacy

Processing over 5G

Efficiency

Customized/local value

Reliability

Private/public networks

Personalization
To scale, process massive amount of data close to source—on the device

In the past:
Cloud-centric AI
AI training and AI inference in the central cloud

Today:
Partially-distributed AI
Power efficient on-device AI interface

Future: Fully-distributed AI with lifelong on-device learning

1. Such as distributed/virtualized core, distributed packet gateway functionality for low latency, mobile edge compute, related to MEC Multi Access Edge Computing as defined by ETSI.
On-device intelligence is paramount

Process data closest to the source, complement the cloud
On-device intelligence is quickly gaining momentum. Key segments are expected to see full AI attach rates by 2025.

- **Mobile**: 2018 - 10%, 2025 - 100%
- **Automotive**: 2018 - 10%, 2025 - 100%
- **XR**: 2018 - 10%, 2025 - 100%
- **PCs/Tablets**: 2018 - 100%, 2025 - 100%
- **Smart speakers**: 2018 - 100%, 2025 - 100%

Source: Tractica, 2019
Mobile is becoming the pervasive AI platform

~7.8 Billion

Cumulative smartphone unit shipments forecast between 2018-2022

Source: IDC Aug. '18
Mobile scale changes everything

Bringing AI to the masses

Qualcomm

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
Shaping the future of transportation

Personalized driver settings
Driver awareness monitoring
Greater autonomous capabilities
Powering the factory of the future
AI for IoT across the home, industrial/enterprise, and Smart Cities

- More efficient use of energy and utilities
- Home hubs and smart appliances
- Sustainable cities and infrastructure
- Digitized logistics and retail
- Autonomous manufacturing and robotics
- Smart security for home and enterprise
- Smart displays and speakers
- Smarter agriculture
- Smarter agriculture
- Smarter agriculture
Power and thermal efficiency are essential for on-device AI

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 HW/SW and optimized 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
Consistent AI R&D investment is the foundation for product leadership

Qualcomm Artificial Intelligence Research is an initiative of Qualcomm Technologies, Inc. Qualcomm Snapdragon, Qualcomm Neural Processing SDK, Qualcomm Vision Intelligence Platform, Qualcomm AI Engine, Qualcomm Cloud A, and Qualcomm QCS400® are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
Leading research and development across the entire spectrum of AI

- Fundamental research:
  - G-CNN
  - Bayesian combinatorial optimization
  - Deep generative models
  - Deep transfer learning
  - Neural network compression

- Applied research:
  - Graph and kernel optimization
  - Machine learning training tools
  - Machine learning for graphics
  - CV DL for new sensors
  - Compute in memory
  - Source compression
  - Voice UI
  - Fingerprint
  - Video recognition & prediction
  - Power management
  - Hardware-aware deep learning
  - Hybrid reinforcement learning
  - Bayesian distributed learning
  - Hybrid reinforcement learning
  - Hybrid reinforcement learning
  - Hybrid reinforcement learning
Can we apply foundational mathematics of physics, like quantum field theory, to deep learning?
G-CNN

Video
Advancing fundamental AI research, such as generalized CNNs

Today’s deep learning

Traditional CNNs

Producing state-of-the-art results but... do not generalize input like rotations

- Translation works
- Rotation doesn’t work

(Convolutional neural networks would need to be retrained with new rotated images to determine new set of parameters—like filter weights)

Tomorrow’s deep learning

Gauge Equivariant CNNs

No matter how you rotate or move the object, the generalized model will still identify the object

- Like quantum field theory, to deep learning
- Applying foundational mathematics of physics

(Gauges Equivariant CNNs (G-CNN): Gauge equivariant CNN, Group, and Steerable CNN pioneered by Qualcomm AI Research do not need to be retrained)

Rotated objects and images applicable to drones, robots, cars, fisheye-lens cameras, VR, AR...
Unifying framework
Gauge equivariant CNN unify special cases like Group CNNs and Steerable CNNs, all pioneered by Qualcomm AI Research.
Robust performance, faster training, and fewer training examples required.

Equivariance
No matter how you rotate or move the object, it will still be identified.
G-CNN can generalize models for different symmetries – traditional CNNs must be retrained.

Broad societal benefits
Use cases like drones, robots, cars, XR, fisheye lenses, 3D gaming, …
But also areas like state-of-the-art accuracy on climate pattern segmentation.

Generalized geometry
Traditional CNNs work well on narrow field-of-view cameras, but fail on e.g. fish-eye cameras.
G-CNN can analyze image data on any curved space, from flat to spherical.

Pioneering deep learning research in generalized CNNs.
Trained neural network model

New input data

Inference output

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

Hardware awareness
CPU + GPU + DSP + AI Acceleration (scalar, vector, tensor)

Acceleration research
Such as compute-in-memory

Advancing AI research to increase power efficiency
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.
Qualcomm® Artificial Intelligence Engine

The hardware and software components for efficient on-device machine learning

4th Gen AI Engine

Mobile Apps

NN Frameworks

Runtime Software Frameworks

Libraries

Cores

Qualcomm Math Libraries

TensorFlow Lite

Google NN API

Qualcomm® Neural Processing SDK

OpenCL

Hexagon NN

Qualcomm® Kryo™ CPU

Qualcomm® Adreno™ GPU

Qualcomm® Hexagon™ DSP

Scalar

Vector

Tensor

Qualcomm Math Libraries, Qualcomm Artificial Intelligence Engine, Qualcomm Kryo, Qualcomm Adreno, Qualcomm Hexagon, and Qualcomm Neural Processing SDK are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
Qualcomm® Cloud AI 100

- Built on 7nm
- >350 TOPS Peak AI Performance
- Sampling 2nd half of 2019
Qualcomm® Neural Processing SDK
Software accelerated runtime for the execution of deep neural networks on device

Efficient execution on Snapdragon
• Takes advantage of Snapdragon heterogeneous computing capabilities
• Runtime and libraries accelerate deep neural net processing on all engines: CPU, GPU, and DSP with HVX and HTA

Model framework/Network support
• Convolutional neural networks and LSTMs
• Support for Caffe/Caffe2, TensorFlow, and user/developer defined layers

Optimization/Debugging tools
• Offline network conversion tools
• Debug and analyze network performance
• API and SDK documentation with sample code
• Ease of integration into customer applications

Available at: developer.qualcomm.com
Frameworks

- TensorFlow
- PyTorch
- ONNX
- Caffe2
- mxnet
- PaddlePaddle
- Chainer

OS

- Android
- Windows

Ecosystem

- Face Recognition
- Night Shot
- Super Resolution
- Noise Suppression
- Face Recognition
- Speech Recognition
- Object Detection
- Video Segmentation
- Bokeh

Devices

- Qualcomm AI Engine

Cognitive Toolkit
Foundational R&D

5G + AI technology leadership

Systems design expertise

Advanced silicon

Ecosystem investment

Uniquely positioned to power the intelligently connected future
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.
Thank you!

Follow us on:  

For more information, visit us at:  
www.qualcomm.com & www.qualcomm.com/blog

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

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

Qualcomm, Snapdragon, Hexagon, Adreno, and Kryo are trademarks of Qualcomm Incorporated, registered in the United States and other countries. 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 Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.