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

<table>
<thead>
<tr>
<th>Mobile</th>
<th>AR/VR</th>
<th>IoT/Camera</th>
<th>IoT/Speakers</th>
<th>Automotive/ADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face ID/Auth</td>
<td>Gesture Tracking</td>
<td>Video Summarization</td>
<td>Multi Keyword Detection</td>
<td>Vehicle Detection</td>
</tr>
<tr>
<td>Super Resolution</td>
<td>3D Reconstruction</td>
<td>Age/Ethnicity/Emotion Detection</td>
<td>ECNS</td>
<td>Pedestrian Detection</td>
</tr>
<tr>
<td>ASR</td>
<td>Object Detection</td>
<td>Zone Intrusion</td>
<td>Voice Biometrics</td>
<td>Path Planning</td>
</tr>
<tr>
<td>Voice Translation</td>
<td>Scene Segmentation</td>
<td>Food Classification</td>
<td>ASR</td>
<td>Traffic Sign Recognition</td>
</tr>
<tr>
<td>Night Shot</td>
<td>Split Rendering</td>
<td>Face Detect/Reco</td>
<td>Audio Context Detection</td>
<td>Speech/Speaker Recognition</td>
</tr>
<tr>
<td>Portrait Backlight</td>
<td>Hand Tracking</td>
<td>Object Avoidance</td>
<td>Speech Synthesis</td>
<td>Distracted Driver Alerts</td>
</tr>
<tr>
<td>Call Noise Reduction</td>
<td>Body Pose Tracking</td>
<td>Multi-object Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call Voice Enhancement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Best-in-class AI performance at the edge
AI improves the whole platform
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\textsuperscript{1}
16-bit and 32-bit FP

Hexagon 698

New Tensor Accelerator
• 4x higher TOPS\textsuperscript{1}
• Up to 35% power savings\textsuperscript{2}
• 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
Trillion operations per second

- Snapdragon 865+ (15 TOPS)
- Snapdragon 855 (7 TOPS, 2X improvement)
- Snapdragon 845 (3 TOPS, 5X improvement)
Antutu Top 10 Best Performing Android Flagship Phones, August 2020

<table>
<thead>
<tr>
<th>Phone</th>
<th>Processor</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xiaomi 10 Ultra</td>
<td>Snapdragon 865 5G</td>
<td>646730</td>
</tr>
<tr>
<td>iQOO5</td>
<td>Snapdragon 865 5G</td>
<td>637616</td>
</tr>
<tr>
<td>ROG game 3</td>
<td>Snapdragon 865 5G</td>
<td>629245</td>
</tr>
<tr>
<td>Tencent Black Shark 3S</td>
<td>Snapdragon 865 5G</td>
<td>620311</td>
</tr>
<tr>
<td>vivo X50 Pro+</td>
<td>Snapdragon 765 5G</td>
<td>609396</td>
</tr>
<tr>
<td>OPPO Find X2 Pro</td>
<td>Snapdragon 865 5G</td>
<td>607276</td>
</tr>
<tr>
<td>OPPO Find X2</td>
<td>Snapdragon 865 5G</td>
<td>600976</td>
</tr>
<tr>
<td>Xiaomi 10 Pro</td>
<td>Snapdragon 865 5G</td>
<td>600618</td>
</tr>
<tr>
<td>iQOO Neo3</td>
<td>Snapdragon 865 5G</td>
<td>599955</td>
</tr>
<tr>
<td>Redmi K30 Pro</td>
<td>Snapdragon 865 5G</td>
<td>595738</td>
</tr>
</tbody>
</table>
Always-on use cases require a different approach to AI
Best-in-class Performance per unit power is key for edge devices
AI Engine Power Efficiency improvement
Hexagon Tensor Accelerator

Snapdragon 865

Snapdragon 855

Snapdragon 845

Performance/Watt

0 1 2 3

1.5X
3X
New Qualcomm® AI Model Efficiency Toolkit

FP32

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

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

© 2020 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

Increase in performance per watt with quantization

© 2020 Qualcomm
Enabling new experiences
Real time language translation

User 1 speech (L1)
Mic 1
Mic 2
Qualcomm Aqstic audio Codec
A
EC/NS
Audio DSP
Translated & encoded speech
L1
L2
Qualcomm Hexagon Vector eXtensions
ASR—Recognizes speech & speech to text
NMT—Understands & translates
TTS—Text to speech
Modem
RFFE
To User 2 (L2)
User 2 translated speech (L2)
D
Qualcomm Aqstic audio Codec
A
Speech enhancement
L1
L2
Hexagon Vector eXtensions
ASR
NMT
TTS
Modem
RFFE
From User 2 (L2)

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

Qualcomm Aqstic and Qualcomm Hexagon are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
XR gaming

AI improves gaming experience on the device

**Display**
- Visual style transfer

**Mechanics**
- AI opponent
- AI game assistant

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

**Audio**
- AI audio effects
- AI placed audio

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

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

**Neural Processing SDK**
- Hexagon
- CPU
- GPU
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
We’re creating a future of distributed intelligence
• 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=CjwKCAjwi9z6BRAYEiwAmo64LfQjU8qH8TxqKTM2PZQp8JibXrjev85wLfKFknJnS_b494vZ7e_WhoCPQkQAvD_BwE

• Qualcomm Platform Solution Ecosystem:
  https://www.qualcomm.com/support/qan/platform-solutions-ecosystem

• GitHub AIMET:
  https://github.com/quic/aimet
Thank you