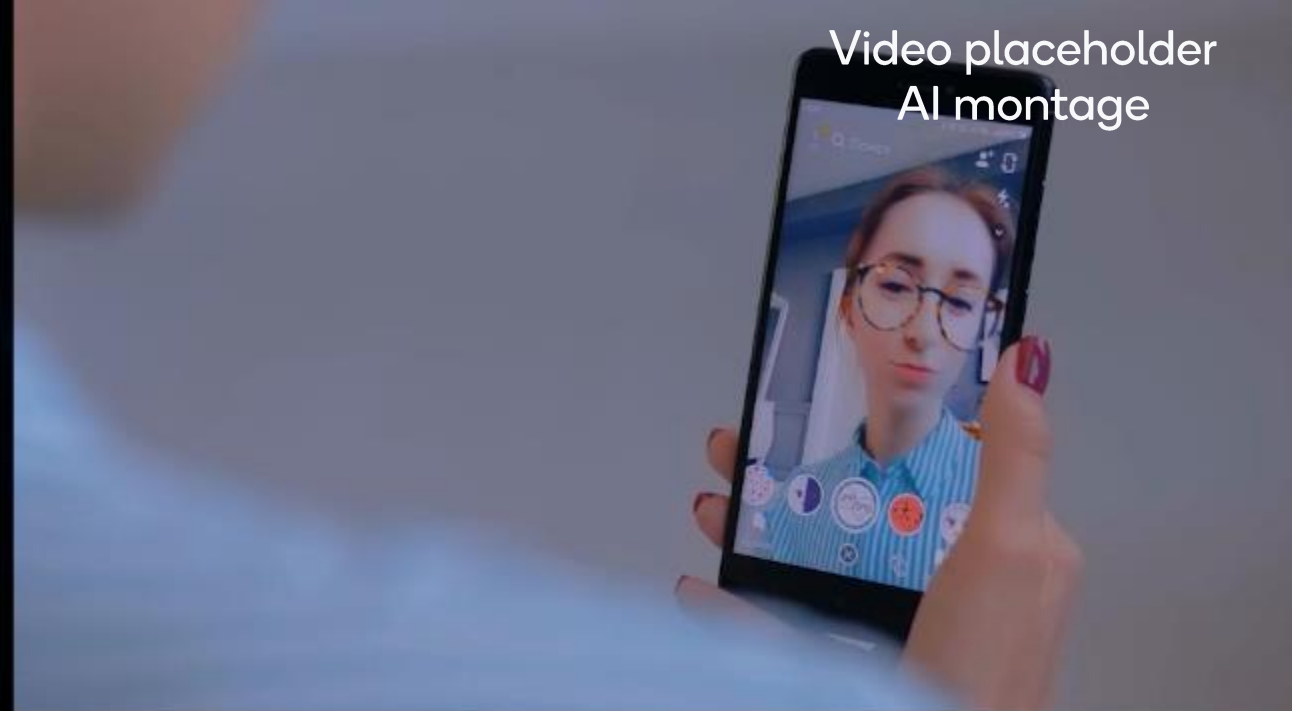


Video placeholder  
AI montage



1024x2048 @ 31.25 FPS

CLASSES

- Road
- Sidewalk
- Building
- Wall
- Fence
- Pole
- Traffic light
- Traffic sign
- Vegetation
- Terrain
- Sky
- Person
- Rider
- Car
- Truck
- Bus
- Train
- Motorcycle
- Bicycle



同时根据每次训练收集的数据与选手共同成长

# Qualcomm® AI Innovation Challenge

TensorFlow  
Lite

oppo

Testin云测  
助力产业智能化



IVOC  
中国智谷(重庆)科技园



EXTREME VISION  
极视角

CSDN

ThunderSoft®

创业邦  
CYZONE



# Qualcomm AI Innovation Challenge

A stage for developers to unleash the potential of Qualcomm AI Engine

主办方: Qualcomm 开源技术合作伙伴: TensorFlow Lite

联合主办方: CATZ IVAC 高通 OPPO Testin云测  
CSDN ThunderSoft 创业邦

协办方: 重庆经开区, Qualcomm 中国, 中科院联合创新中心



Video placeholder  
AI Innovation Challenge

## 20+ Hero Apps



## 4 Educational Webinars for Developers



## 12,000+ Developers Touchbased

Qualcomm Spectra<sup>™</sup>  
580 ISP

Qualcomm<sup>™</sup> Hexagon<sup>™</sup>  
780 Processor

---

# Hardware

---

Qualcomm<sup>™</sup>  
Sensing  
Hub

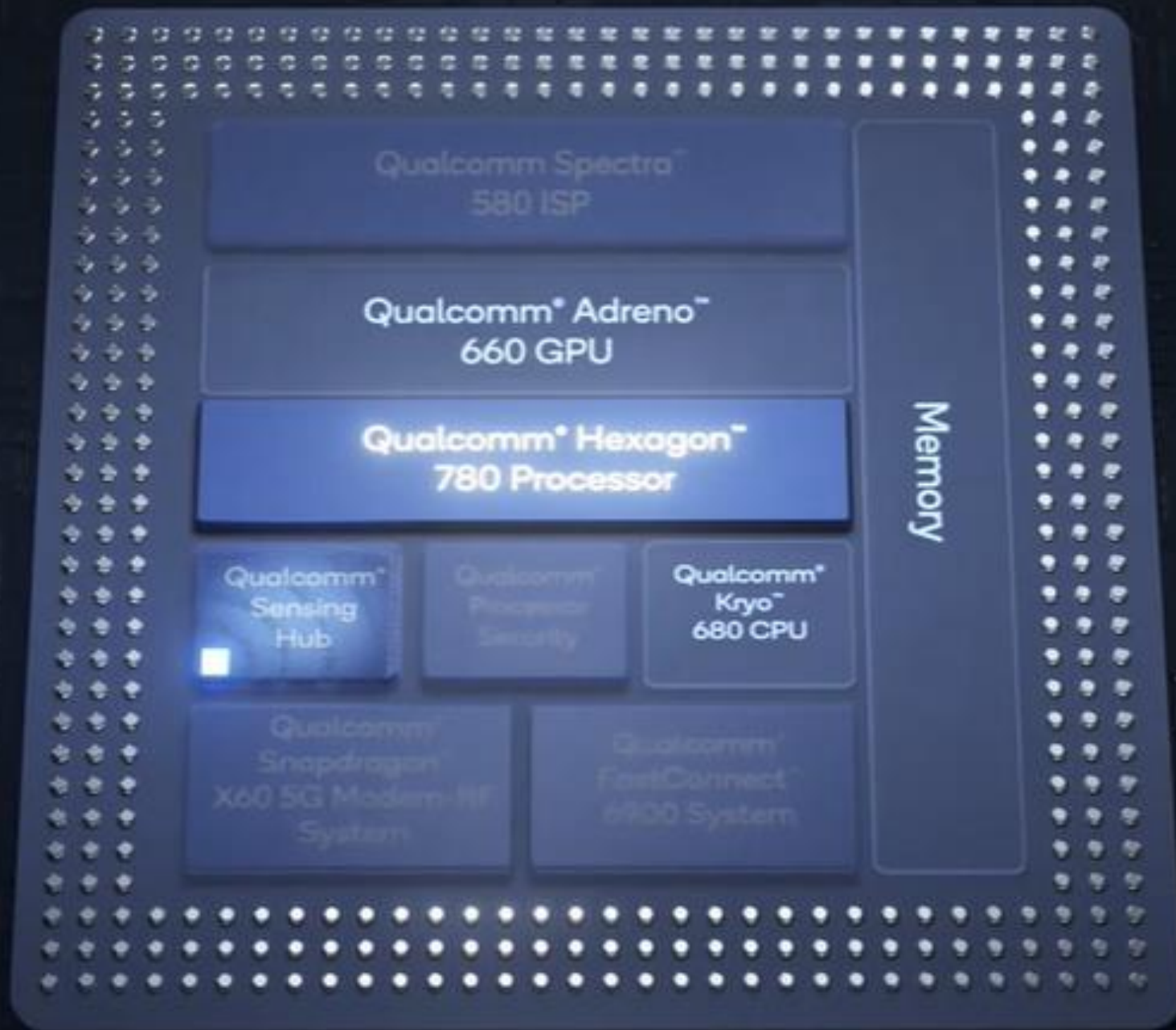
Qualcomm<sup>™</sup>  
Processor  
Security

Qualcomm<sup>™</sup>  
Kryo  
680 CPU

Qualcomm<sup>™</sup>  
Snapdragon<sup>™</sup>  
X60 5G Modem-RF  
System

Qualcomm<sup>™</sup>  
FastConnect<sup>™</sup>  
6900 System

Memory





# 6<sup>th</sup> Generation Qualcomm<sup>®</sup> AI Engine



# Hexagon 780 Processor





# Fused

AI accelerator

Qualcomm Spectra™  
580 ISP

Qualcomm® Hexagon™  
780 Processor

Qualcomm®  
Sensing  
Hub

Qualcomm®  
Processor  
Security

Qualcomm®  
Kryo™  
680 CPU

Qualcomm®  
Snapdragon™  
X60 5G Modem-RF  
System

Qualcomm®  
FastConnect™  
6900 System

Memory



865  
5G

Qualcomm  
snapdragon

Adreno 650 GPU

Memory

Tensor

Vector

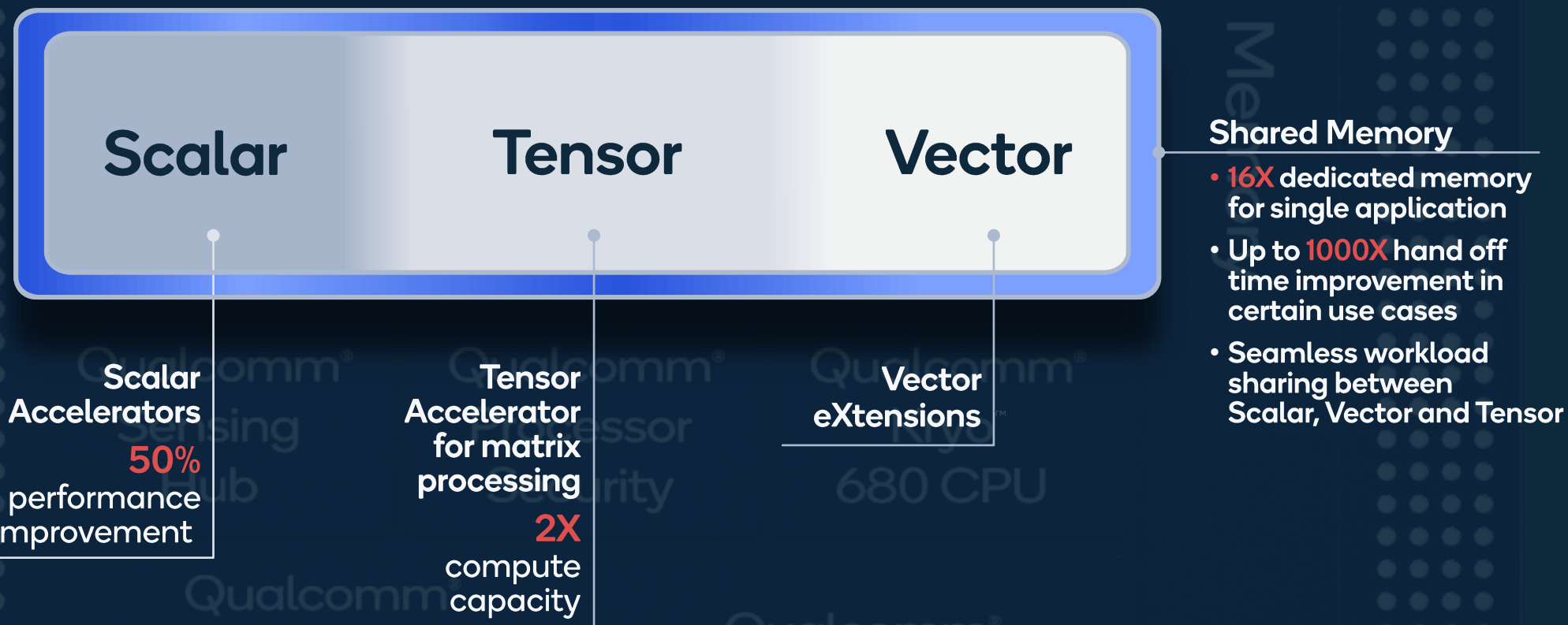
Scalar

Sensing  
Hub

Processor  
Security

Kryo  
585 CPU

# Hexagon 780 Processor



\*Comparing to previous generation

## 43% faster AI performance

### New instructions:

4-input mixed precision dot product

Wave Matrix Multiply for 16/32-bit floating point

### Fused AI accelerator:

Up to **3X**  
Performance per watt

Tensor **2X**  
compute capacity

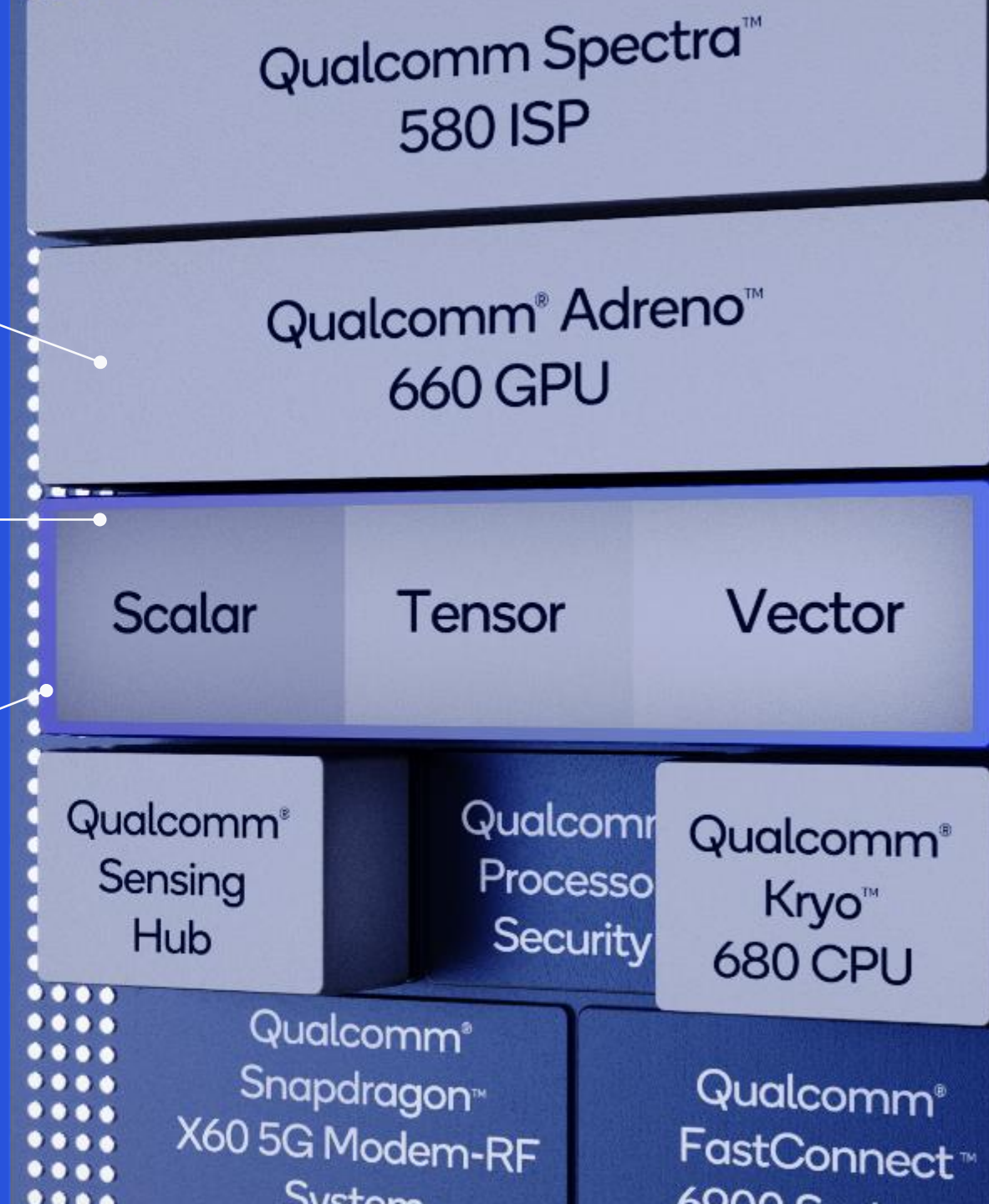
Scalar **50%**  
performance improvement

Vector support for additional data types

### Shared Memory:

**16X**  
dedicated memory

Up to **1000x**  
hand off time improvement in certain use cases



Scalar

Tensor

Vec

# Fused AI Accelerator

# Trillion operations per second



26 TOPS

Snapdragon 888



15 TOPS

Snapdragon 865

# Peak performance on classification networks

Inference: (inf/s)

Resnet 50

Snapdragon 888

670



Company A

310

Company B

230

MobileNet V2

1,110

770

550

VGG19

320

150

70

Inception V3

600

260

144



# Power consumption

Efficiency: (inf/W)

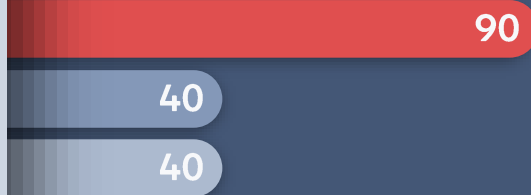
Resnet 50



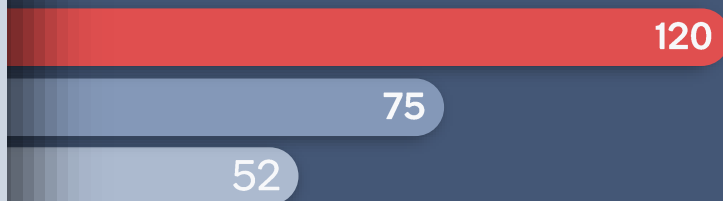
MobileNet V2

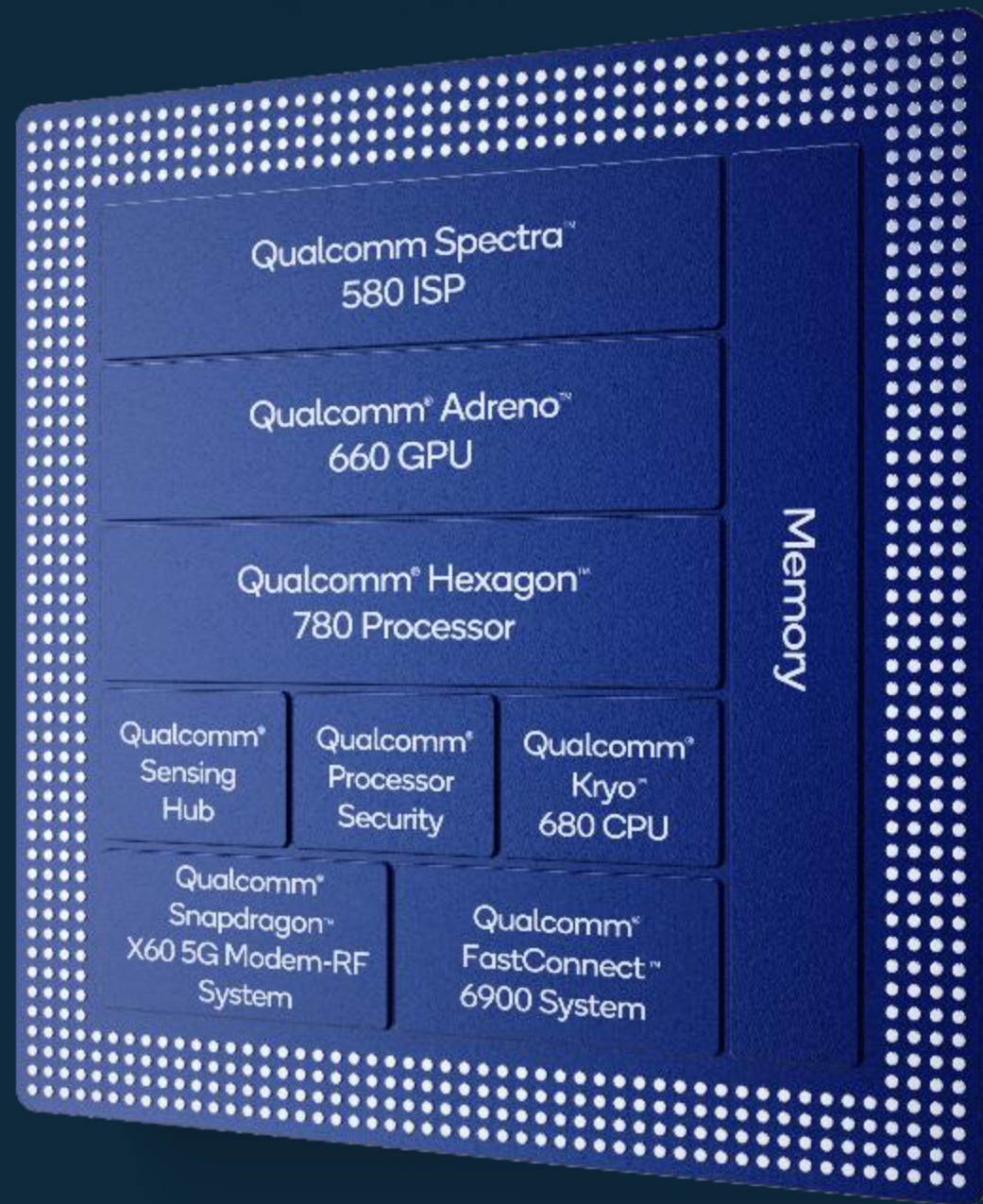


VGG19



Inception V3





**2<sup>nd</sup> Gen**  
**Qualcomm<sup>®</sup>**  
**Sensing Hub**

Qualcomm<sup>®</sup> Hexagon<sup>™</sup>  
780 Processor

Qualcomm<sup>®</sup>  
Sensing  
Hub

Qualcomm<sup>®</sup>  
Processor  
Security

Qualcomm<sup>®</sup>  
Kryo  
680 Core

Qualcomm<sup>®</sup>  
Snapdragon<sup>™</sup>  
X60 5G Modem-RF  
System

Qualcomm<sup>®</sup>  
FastConnect  
6900 System

# Qualcomm® Hexagon™ 780 Processor

Qualcomm®  
Sensing  
Hub

Qualcomm®  
Processor  
Security

Qualcomm®  
Kryo  
680 C

Qualcomm®  
Snapdragon™  
X60 5G Modem-RF  
System

Qualcomm®  
FastConnect  
6900 System

Always-on low-power  
dedicated hardware  
AI processor

# AI performance

Qualcomm  
Sensing  
Hub

With 2<sup>nd</sup> Gen  
Qualcomm Sensing Hub

1st Gen  
Qualcomm Sensing Hub

5X

80%

Task Reduction  
offload from  
Hexagon Processor\*

\*Based on INT8, GOPS/s, LPI performance

Screen wake

Ambient audio /  
audio event detection

Lift detection



TensorFlow  
Micro

2nd Gen  
Qualcomm  
Sensing Hub

Car crash detection

Activity recognition

Earthquake detection



## Hot word detection on Snapdragon 888 using TensorFlow Micro

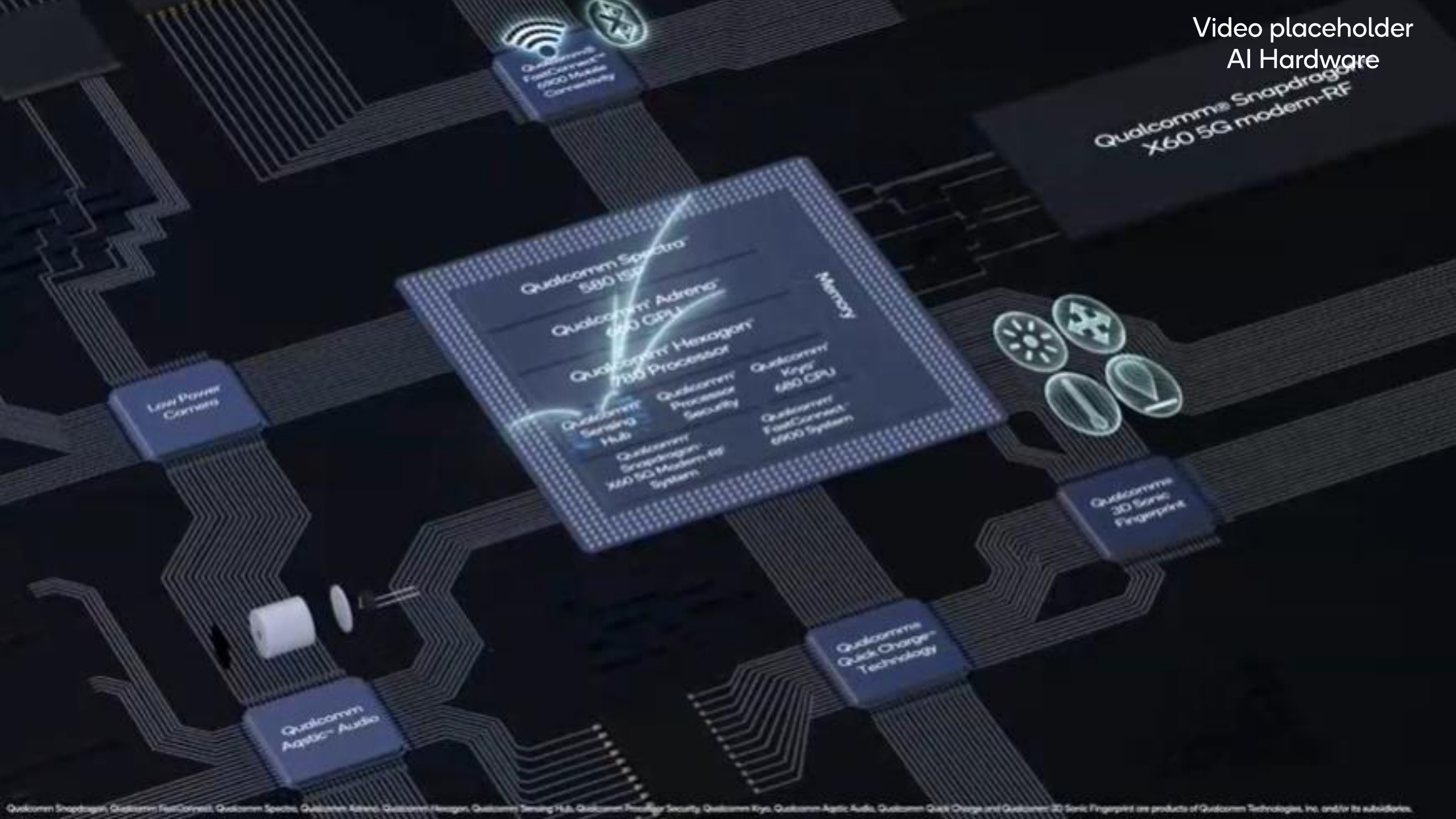


TensorFlow  
Micro

2nd Gen  
Qualcomm  
Sensing Hub

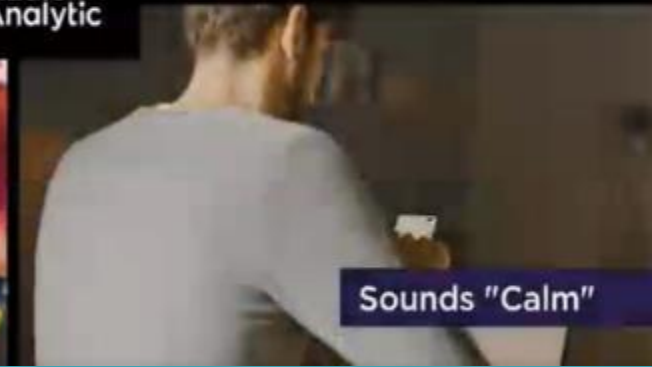
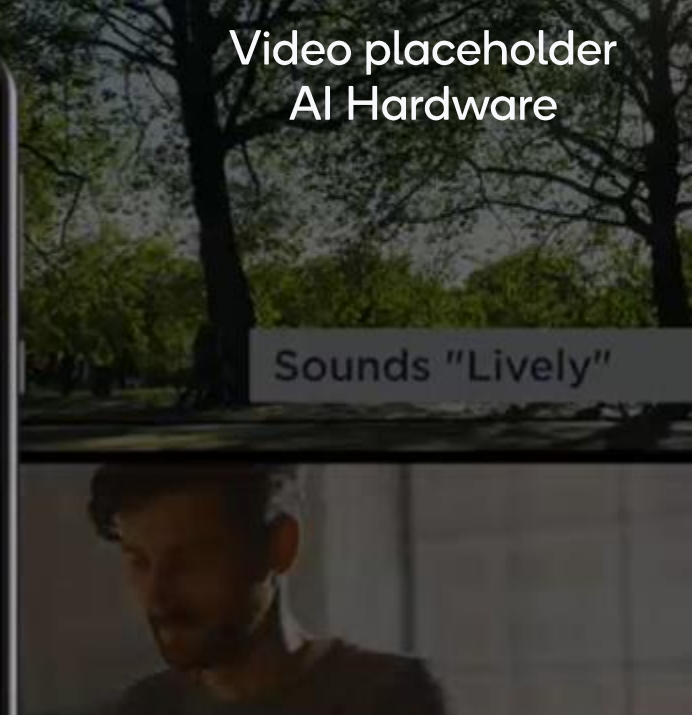
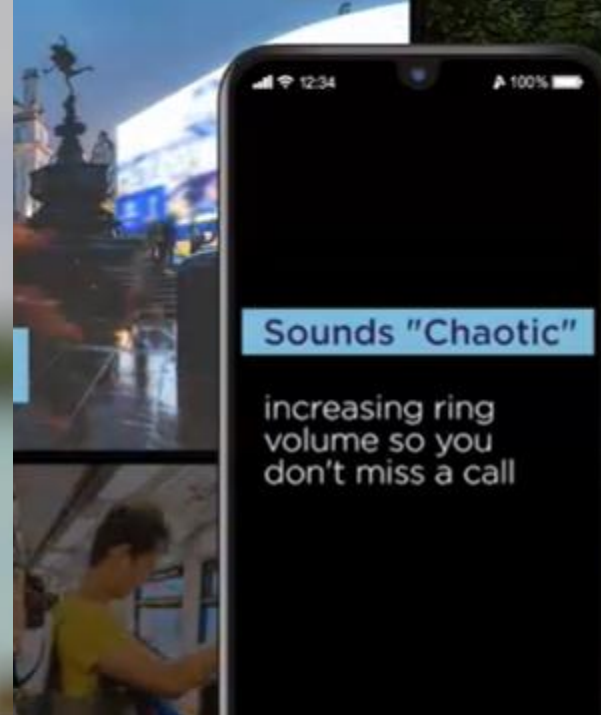
38%  
Task offload from  
Hexagon processor

Video placeholder  
AI Hardware





Video placeholder  
AI Hardware



Video placeholder  
AI Hardware

Qualcomm  
snapdragon



trinamix

A brand of  
BASF – We create chemistry





# Software



**Ever**  
**on-device**  
**AI SDK**

# Qualcomm® Neural Processing SDK

Powering over

**500M+**

Android Devices

# Hexagon NN direct

on Qualcomm® Snapdragon™ 865





Qualcomm  
Neural Processing SDK



Android Neural  
Networks API



Qualcomm<sup>®</sup>  
AI Engine direct





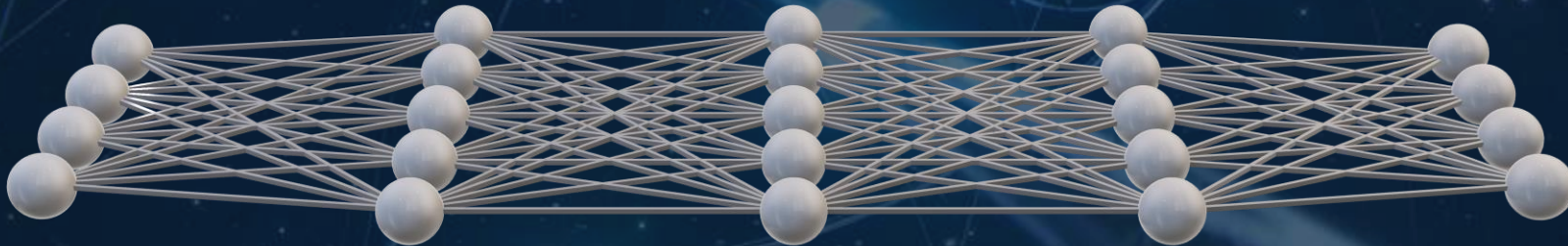
# Qualcomm<sup>®</sup> AI Engine direct

Accessibility for everyone

Unified AI API  
available across  
Qualcomm AI Engine  
(Hexagon, CPU, GPU)

Compatible with  
5<sup>th</sup> gen Qualcomm AI Engine

Modularity & Extensibility  
Per accelerators & operation





Qualcomm  
snapdragon





# User defined operators

---



operators

---

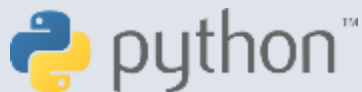
OpenCL

Hexagon  
SDK



Qualcomm  
AI Engine  
direct

Custom operators  
efficiently written in

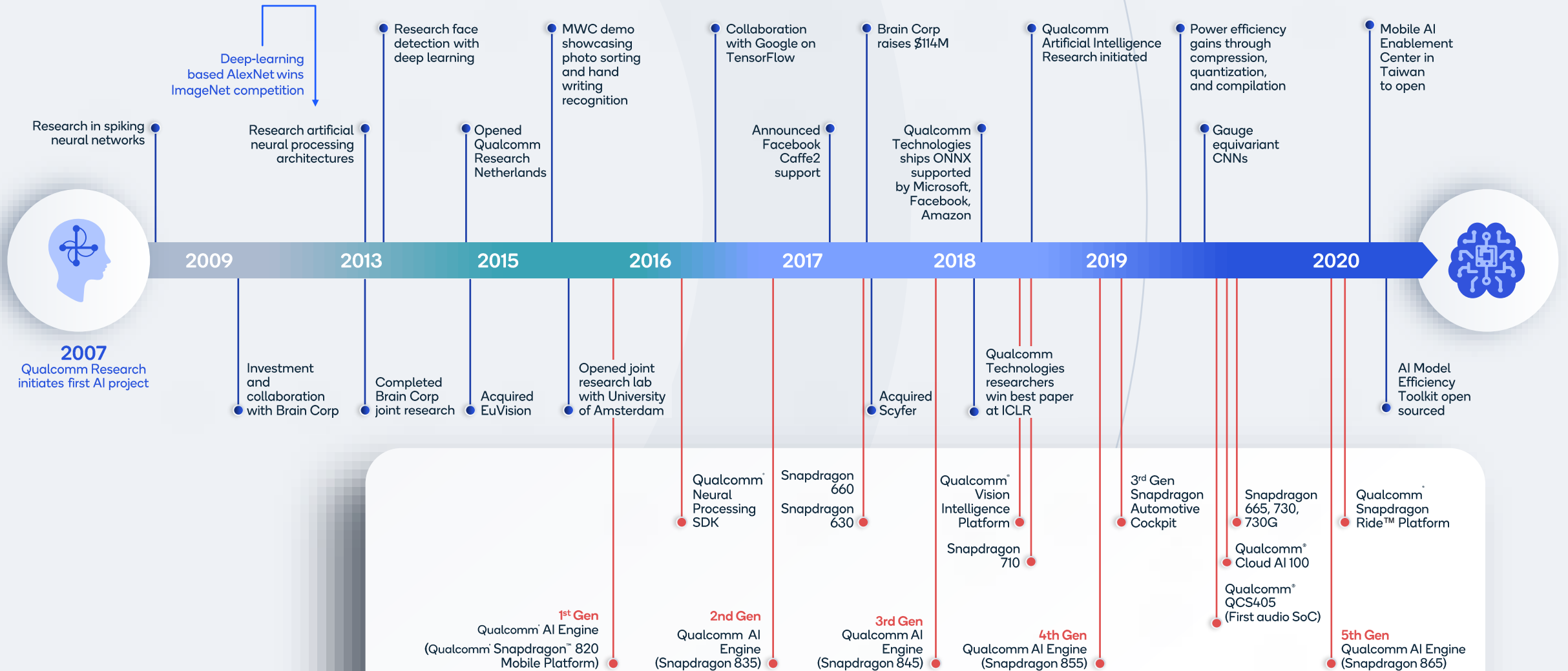


```
1 def quantized_add_generic(size, a, b, a_offset, b_offset, a_mult, b_mult, output, target, ctx):
2     # Construct the TVM computation.
3     A_offset = tvm.var('A_offset', dtype='uint8')
4     B_offset = tvm.var('B_offset', dtype='uint8')
5     A_mult = tvm.var('A_mult', dtype='uint16')
6     B_mult = tvm.var('B_mult', dtype='uint16')
7
8     N = tvm.var("N")
9     A = tvm.placeholder((N,), name='A', dtype='uint8')
10    B = tvm.placeholder((N,), name='B', dtype='uint8')
11
12    C = tvm.compute((A.shape),
13                  lambda i: ((A[i].astype('int32') - A_offset.astype('int32')) * A_mult.astype('int32')) +
14                            ((B[i].astype('int32') - B_offset.astype('int32')) * B_mult.astype('int32')), name='C')
15
16    # Create the schedule.
17    s = tvm.create_schedule(C.op);
18    px, x = s[C].split(s[C].op.axis[0], nparts=1)
19    s[C].bind(px, tvm.thread_axis("pipeline"))
20
21    # Construct the callable object "func" corresponding to the computation.
22    func = tvm.build(s, [A, B, C, N, A_offset, B_offset, A_mult, B_mult], target, name='qadd_tvm')
23
24    func(tvm.ndarray.array(a, ctx=ctx), tvm.ndarray.array(b, ctx=ctx), output,
25         size, a_offset, b_offset, a_mult, b_mult)
```

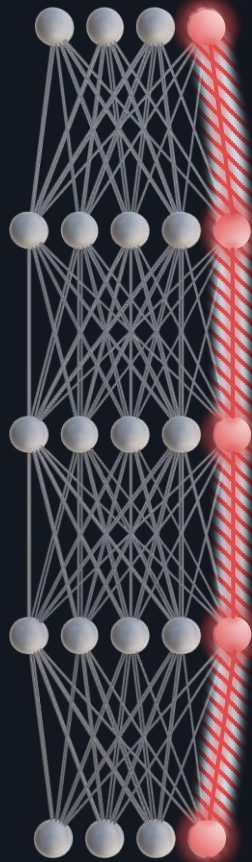
**13**  
years

of  
**research**  
to  
**product**

# 13 years of research to product



# AI Model Efficiency Toolkit

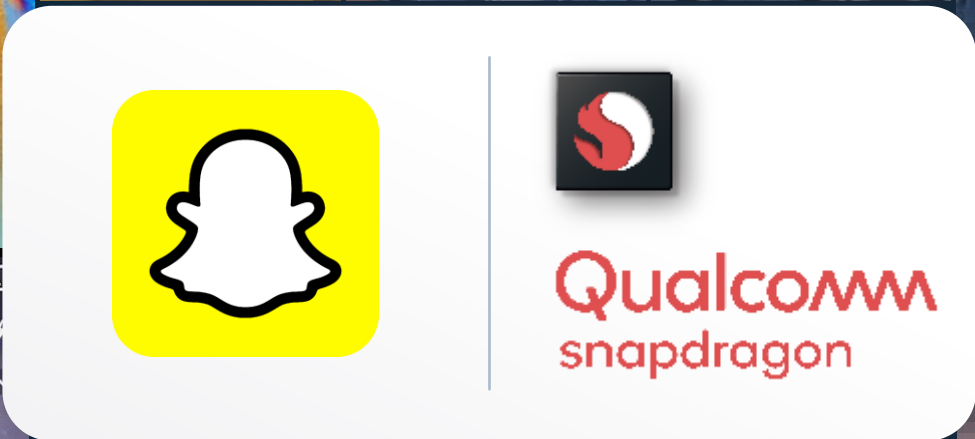
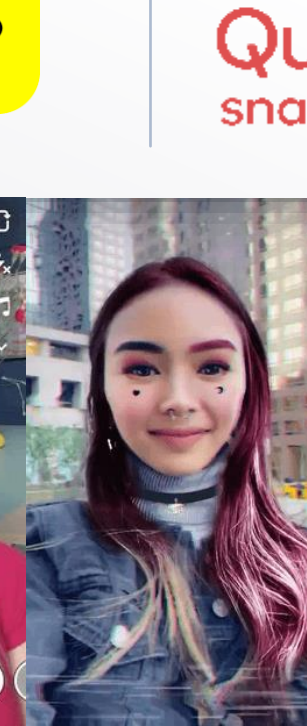
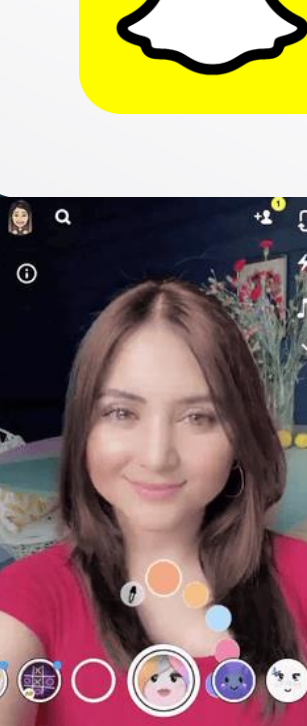
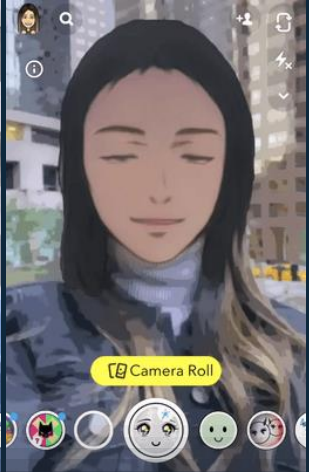
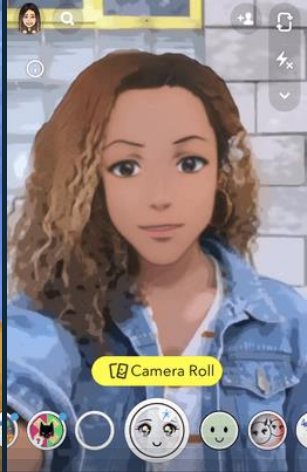
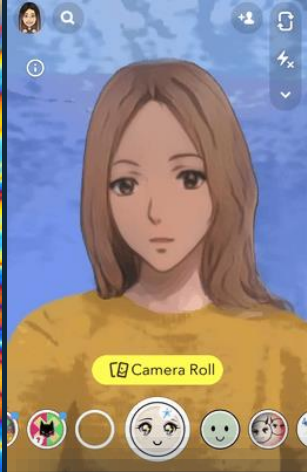


Improved/Robust  
quantization for  
INT16,8,4

Quantization  
aware training  
with range  
learning

Mix precision  
support

Opensource



## 2nd gen Qualcomm Sensing Hub



Dedicated AI accelerator  
First to support TensorFlow Micro

## Hexagon 780 Processor



Fused AI Accelerators

- **Tensor** - 2X compute capacity
- **Scalar** - 50% performance improvement
- **Vector** – Support for additional data types

3X performance per watt improvement  
16X dedicated memory  
Up to 1000X hand off time improvement in certain use cases

## 6th gen Qualcomm AI Engine



26 TOPS

# AI Highlights



## Qualcomm Neural Processing SDK & AI Model Efficiency Toolkit

New features and improvements

## Qualcomm AI Engine direct



Easier and faster access to the entire AI Engine

## TVM Opensource



More efficient coding


## Industry leading AI use cases



Super movie with Tetras.AI  
Snapchat lenses acceleration  
NLP with Hugging Face  
Skin condition detection with triniMiX



# Thank you

Follow us on:    

For more information, visit us at:

[www.qualcomm.com](http://www.qualcomm.com) & [www.qualcomm.com/blog](http://www.qualcomm.com/blog)

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

©2018-2020 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm, Snapdragon, Snapdragon Ride, Qualcomm Spectra, Adreno, Hexagon and Kryo are trademarks or registered 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.