Ziad Asghar

Vice President, Product Management, AI & Strategy

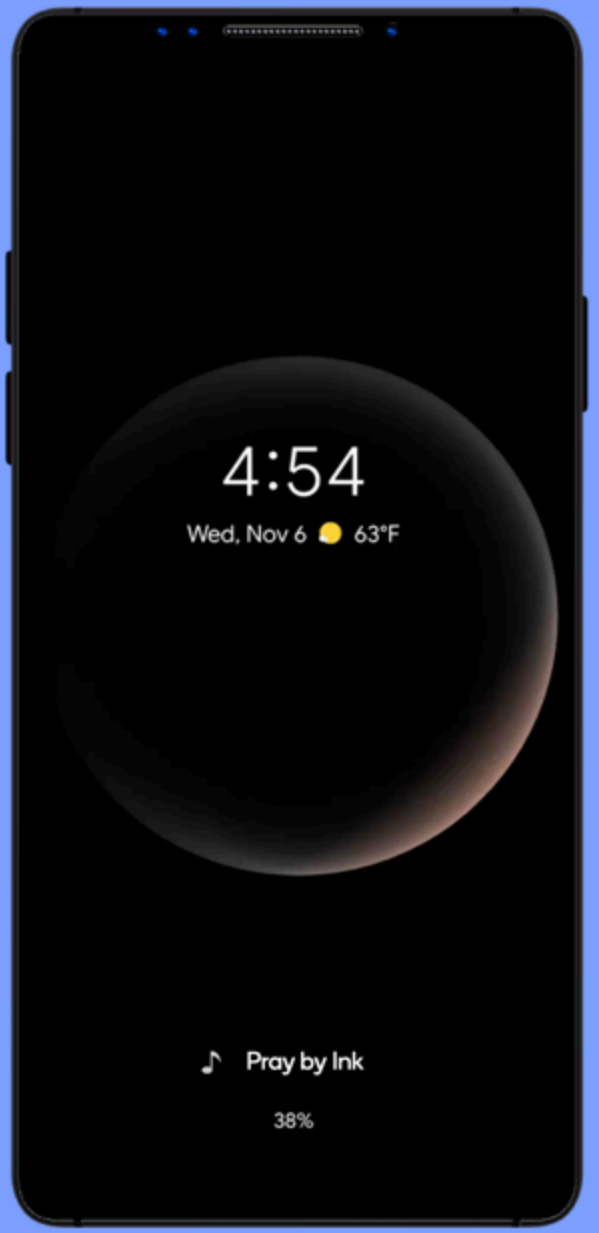
Qualcomm Technologies, Inc.


AI is in every aspect of our lives

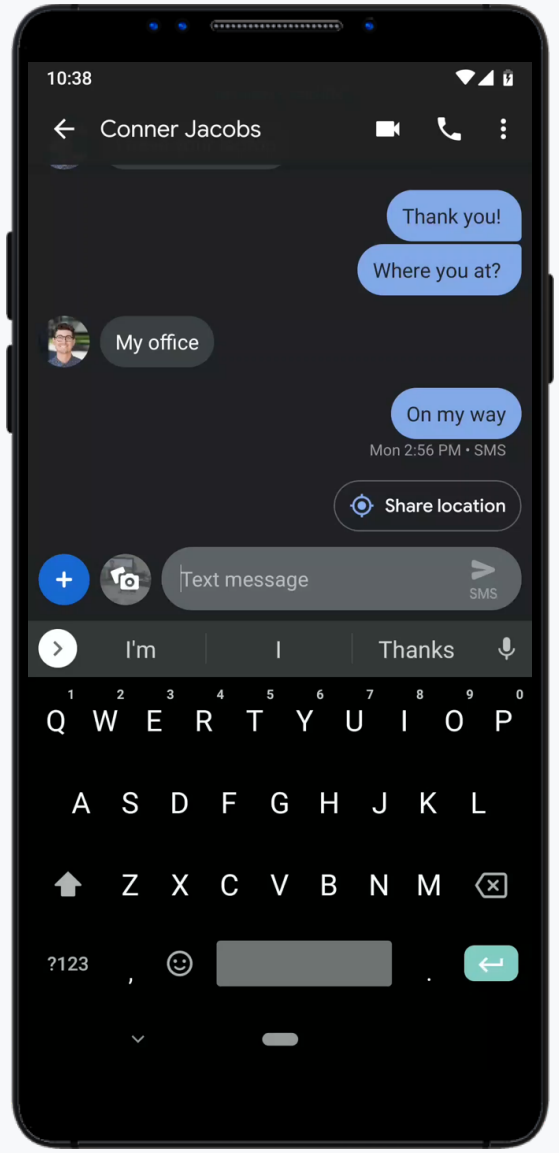
Conventional music detection

Music detection with on-device AI

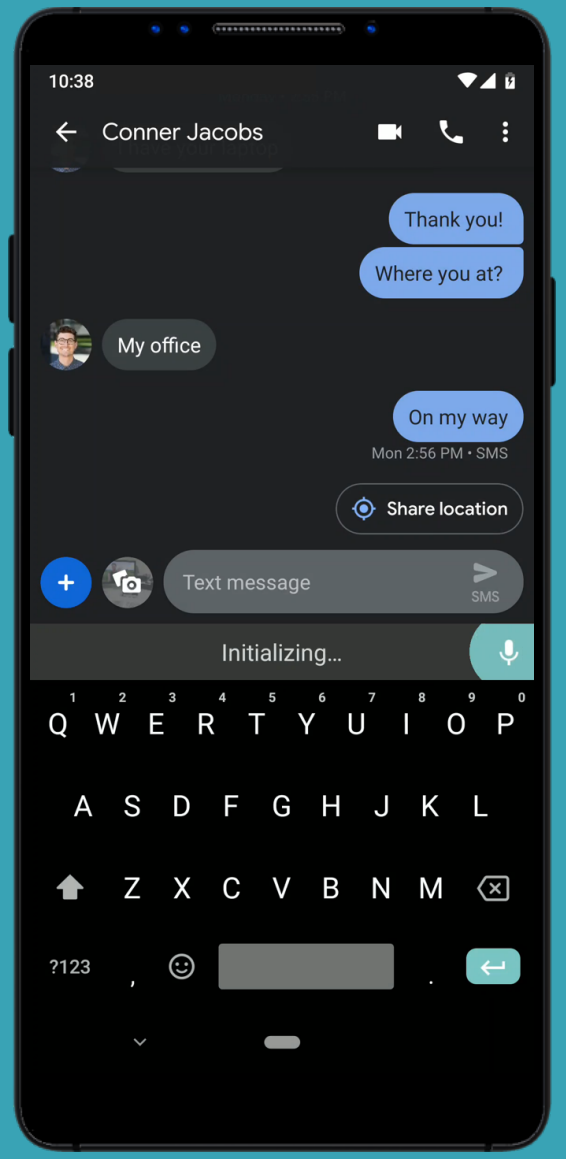
Always-on neural nets at ultra-low power



Cloud speech-to-text

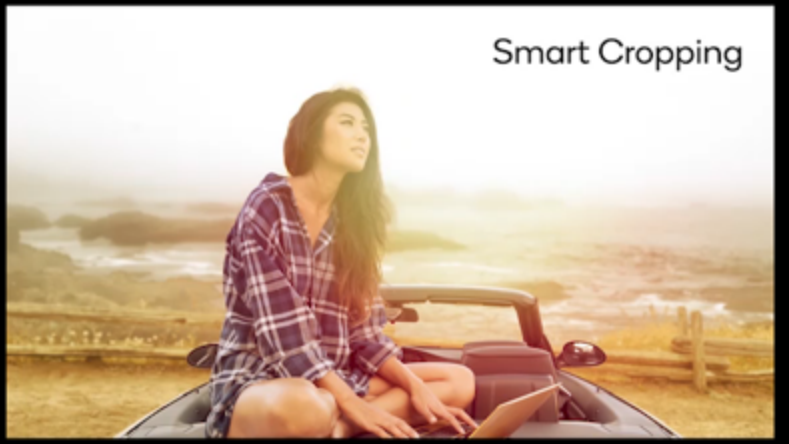


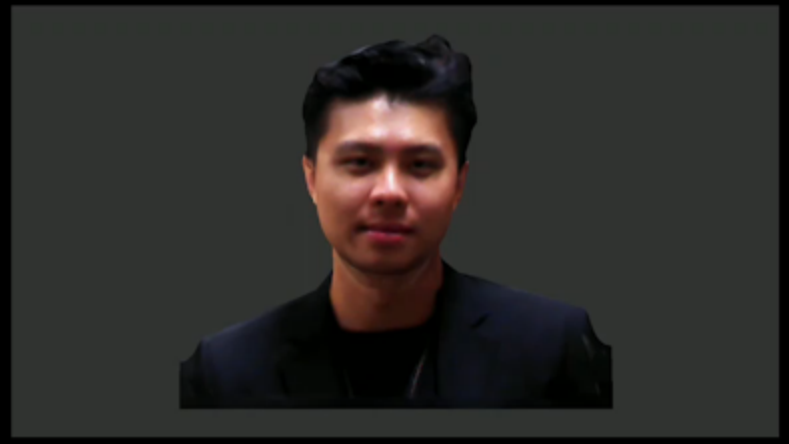
On device speech-to-text




eXtended Reality


5G
生意正在好轉
MAN
ENG
Qualcomm
snapdragon


Smart Cropping

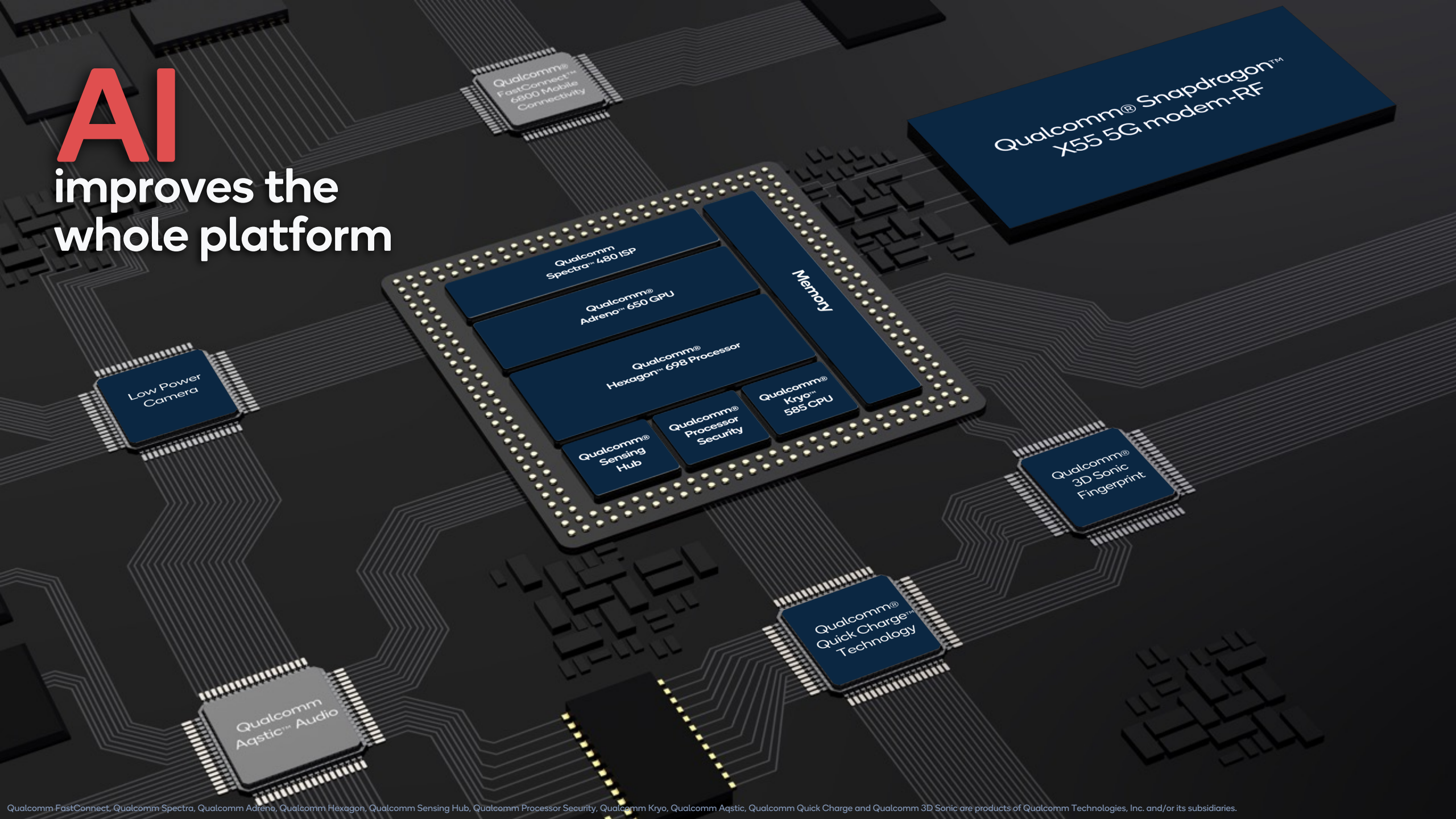


AI is happening as we “speak”

Software

Hardware

Developer tools



5th Generation Qualcomm® AI Engine

Adreno 650

New AI mixed precision instructions

2x higher TOPS*

16-bit and 32-bit FP

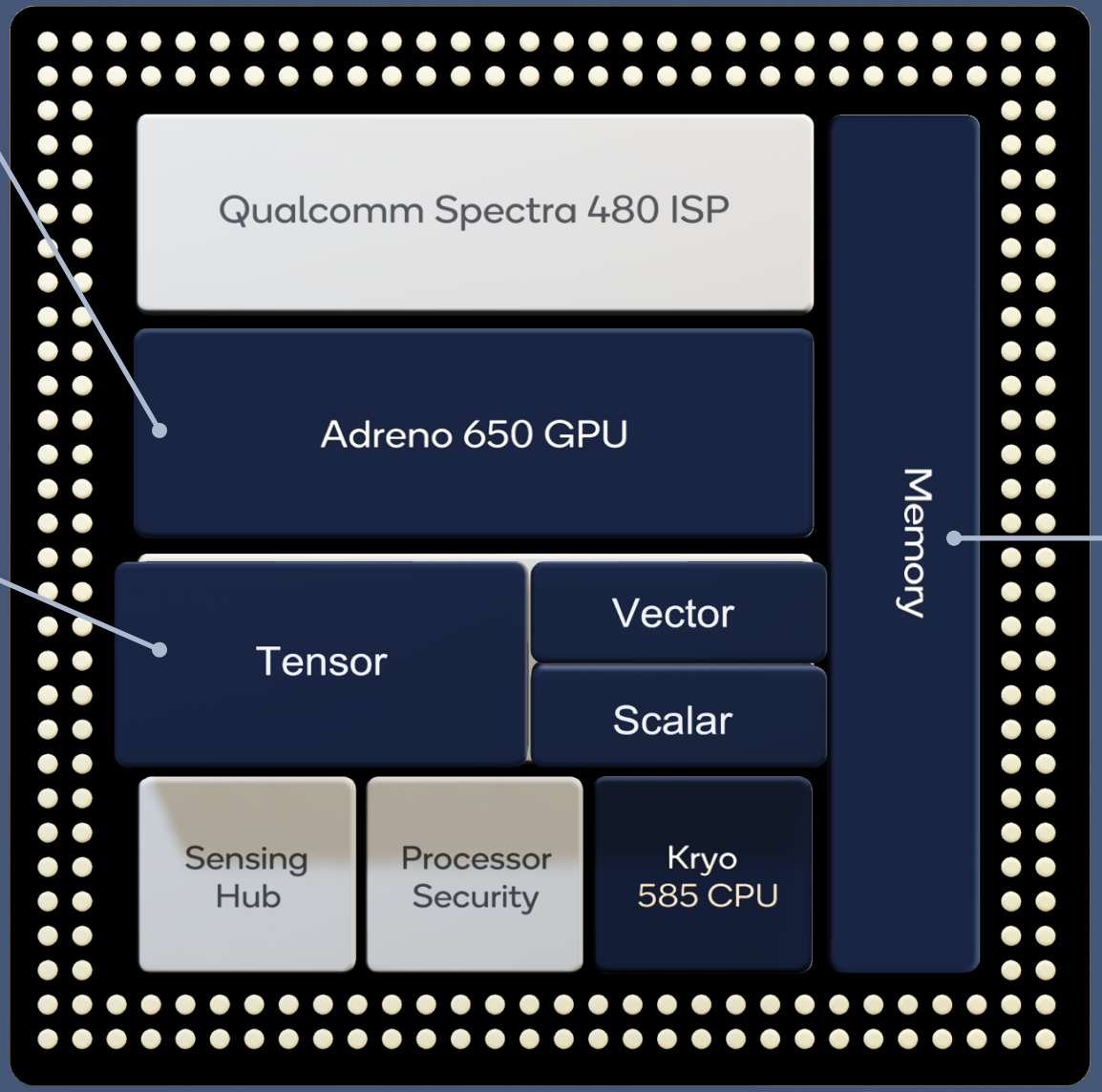
Hexagon 698

New Tensor Accelerator

- 4x higher TOPS*
- Up to 35% power savings*
- 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



LP-DDR5 memory

30% more bandwidth*

Improved AI processing

*Comparing to previous generation

Zero-order tensor
(a scalar)

First-order tensor
(a vector)

Second-order tensor
(a matrix)

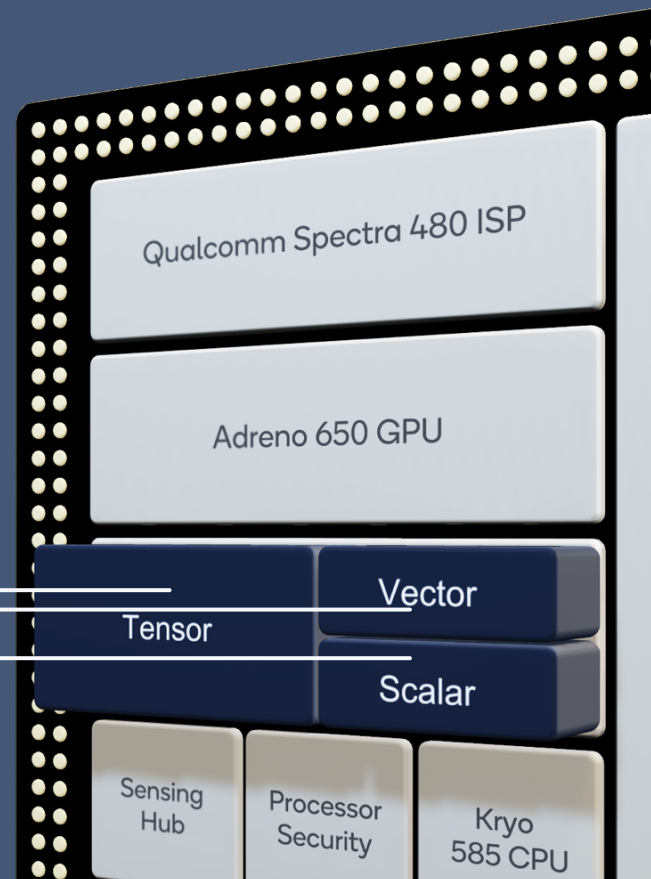
Third-order tensor
(a 3-dimensional array)

Setup

Pooling 2 x 2 | Conv 5 x 5 x 64 | Pooling 2 x 2 | FC | FC

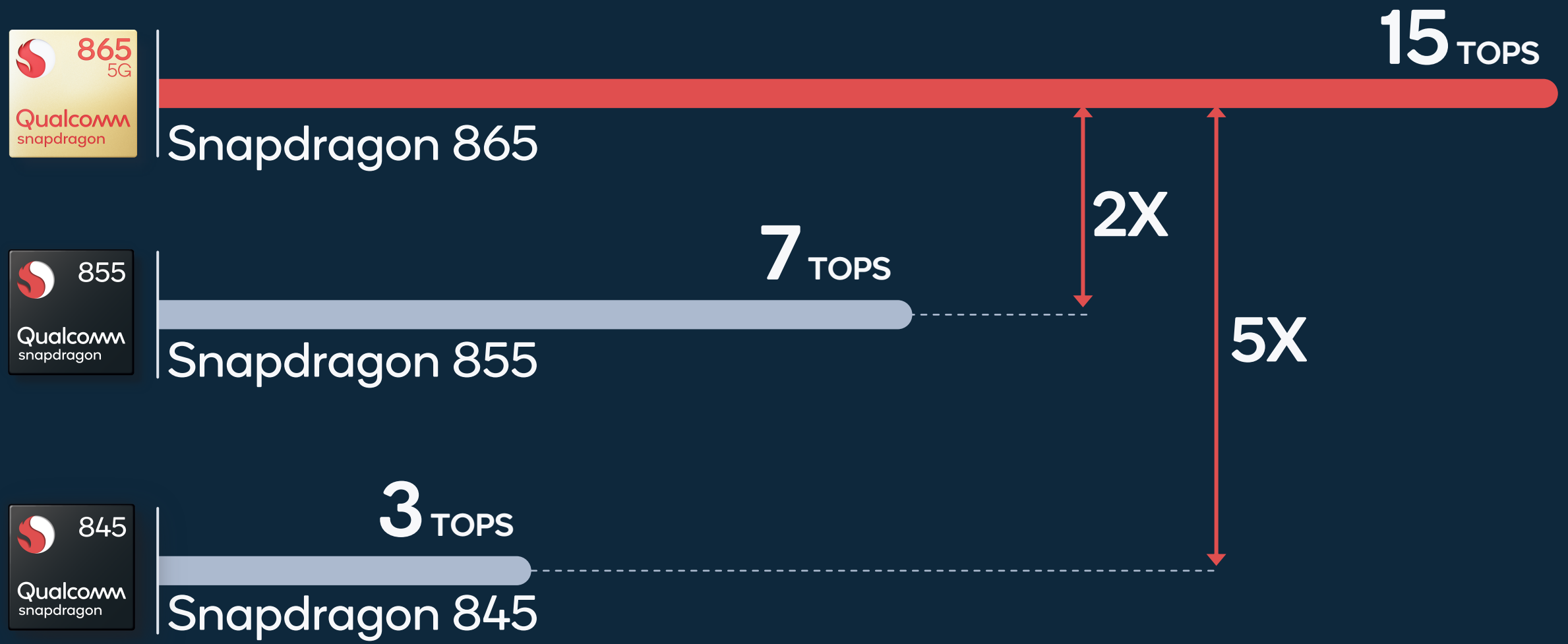
28 x 28 x 1 → 24 x 24 x 32 → 12 x 12 x 32 → 8 x 8 x 64 → 4 x 4 x 64 → 1024 → 10

Scalar processor | Vector processor | Tensor processor | Vector processor



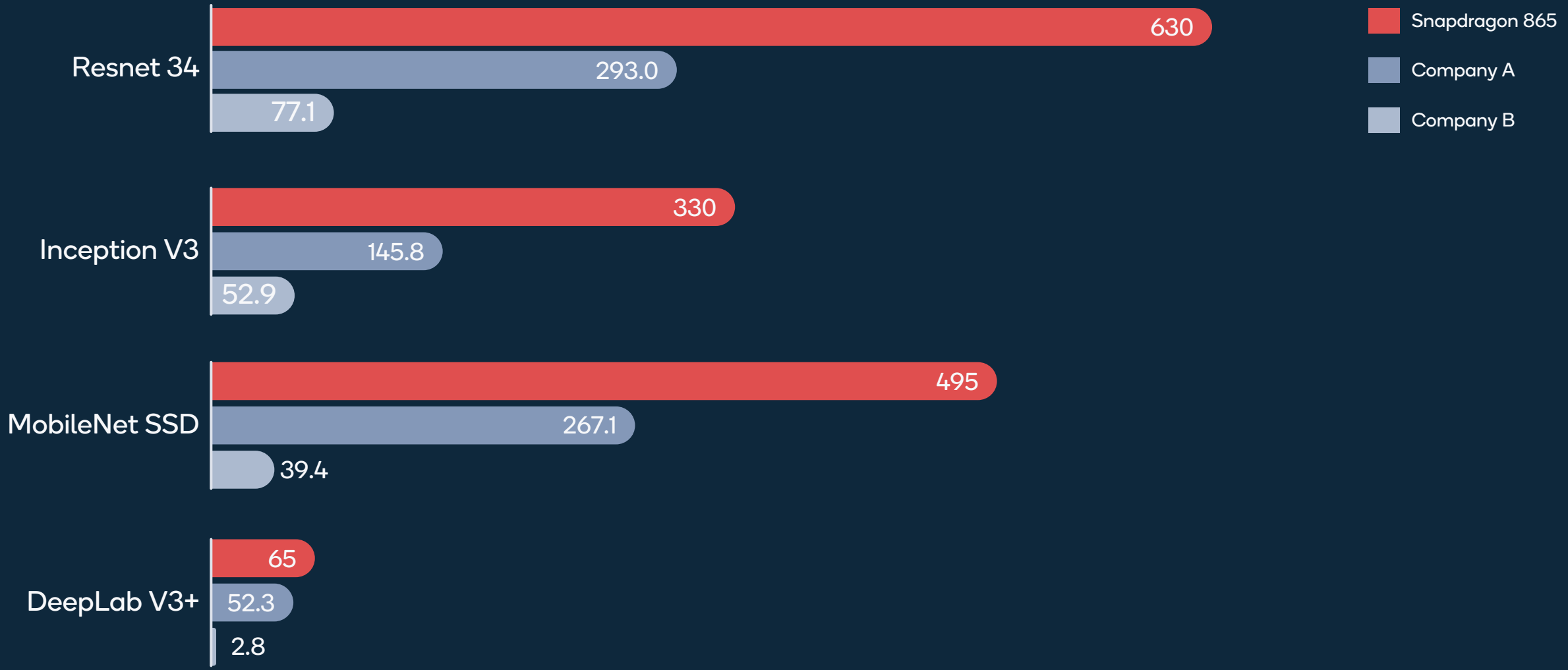
Trillion operations per second

Trillion operations per second



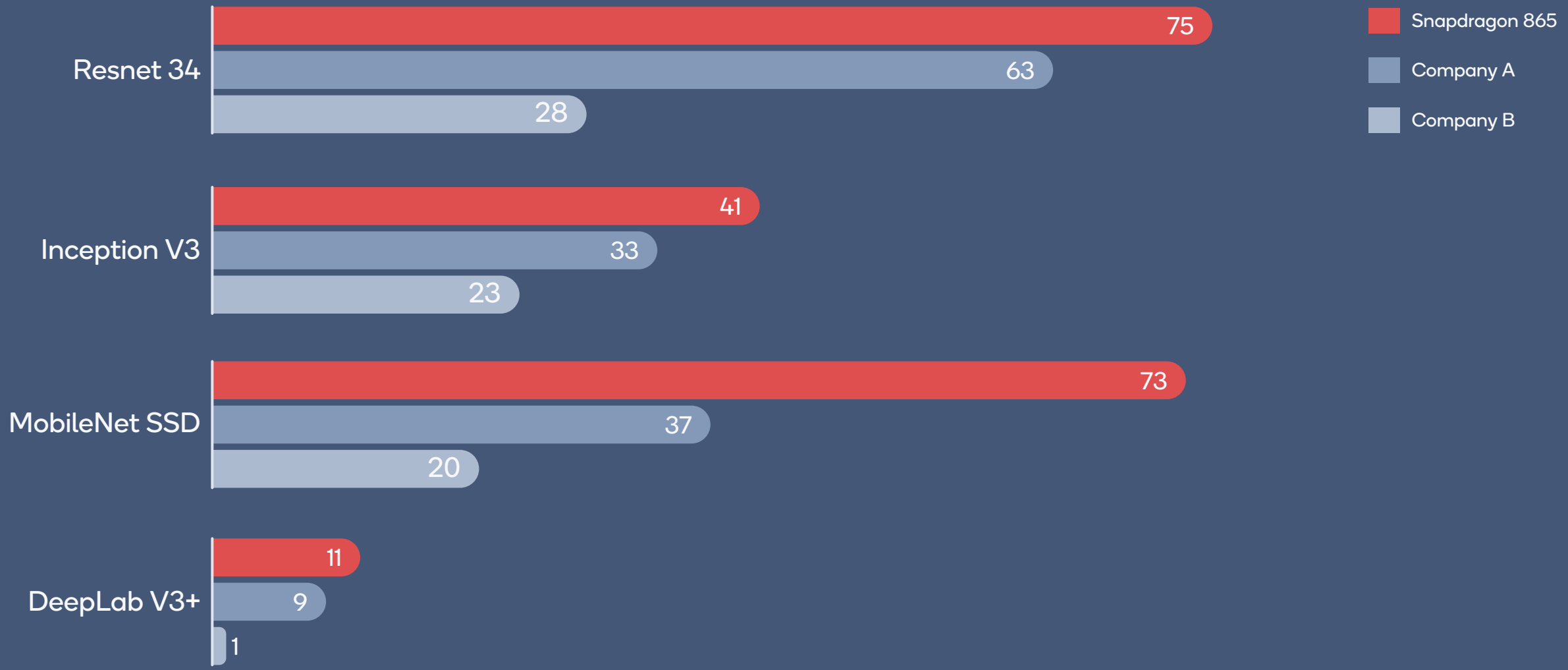
Peak performance on classification networks

(Inf/sec)



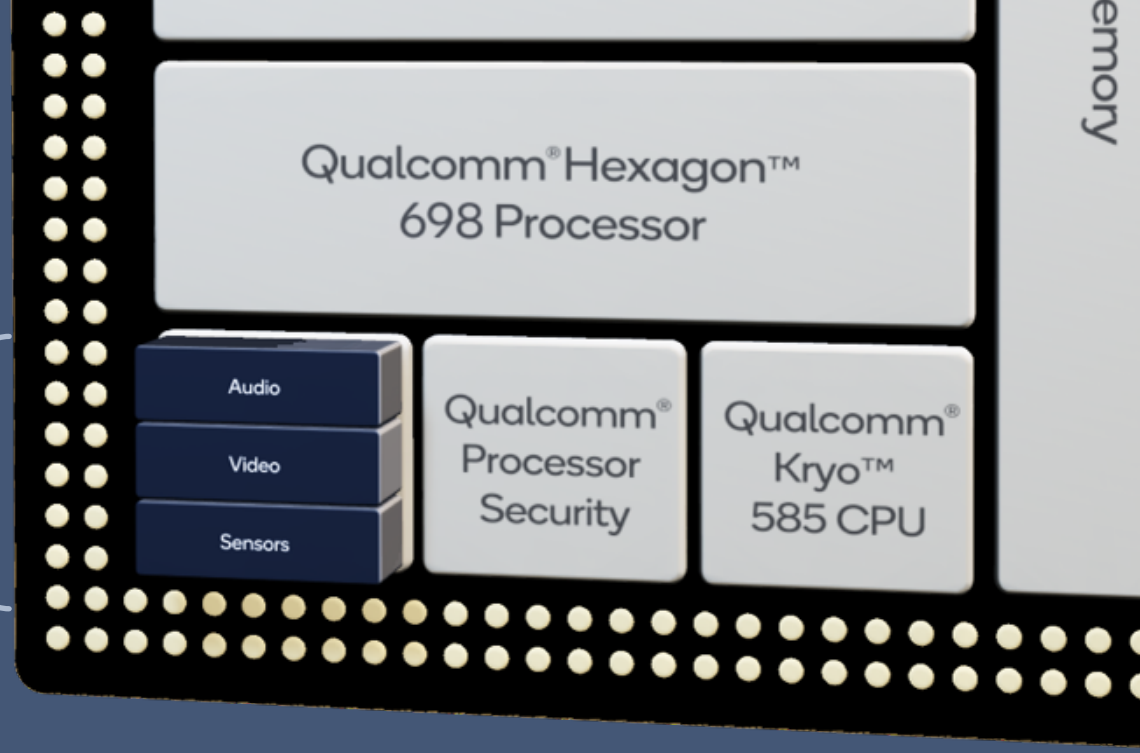
Power consumption

(Inf/Watt)

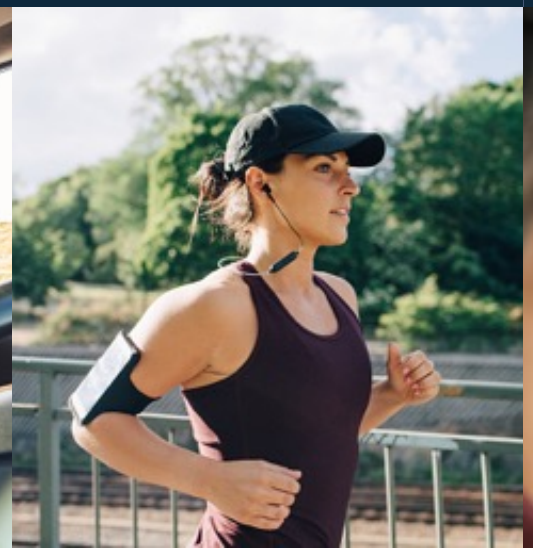


Qualcomm Sensing Hub

<1mA



Safety	Health & Fitness	Security	Context	Low-power camera

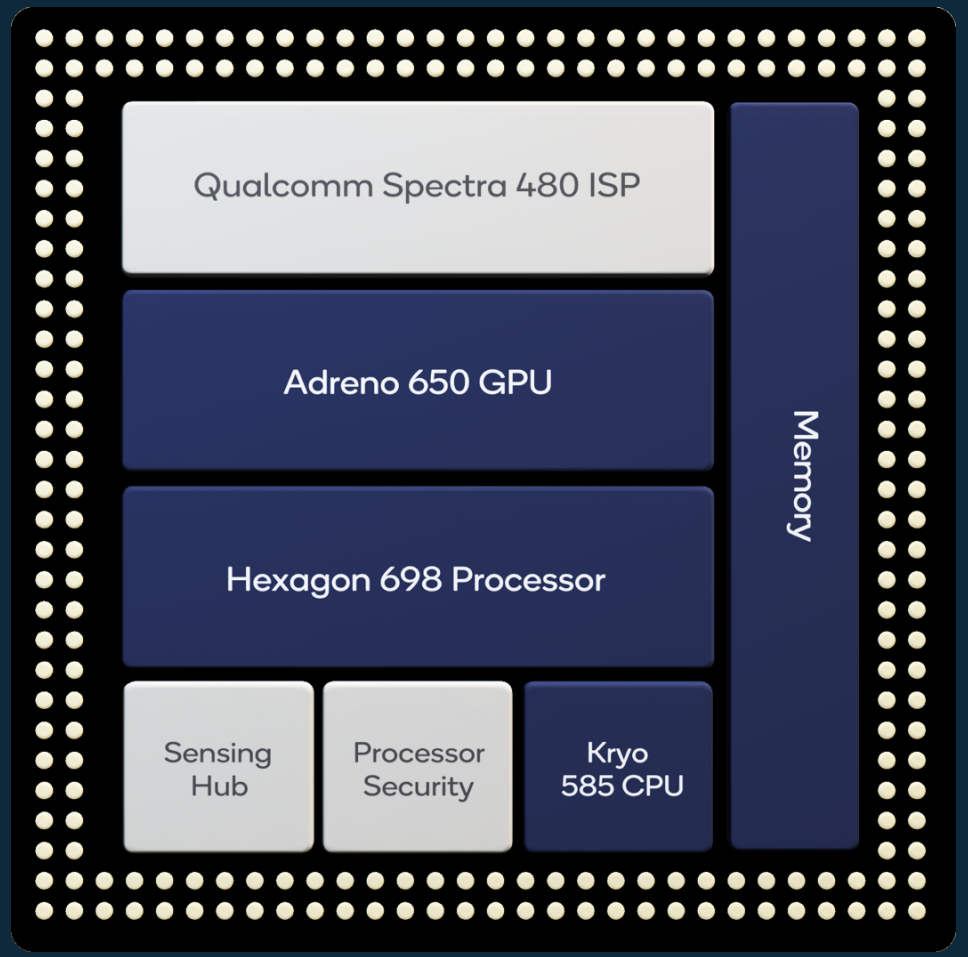


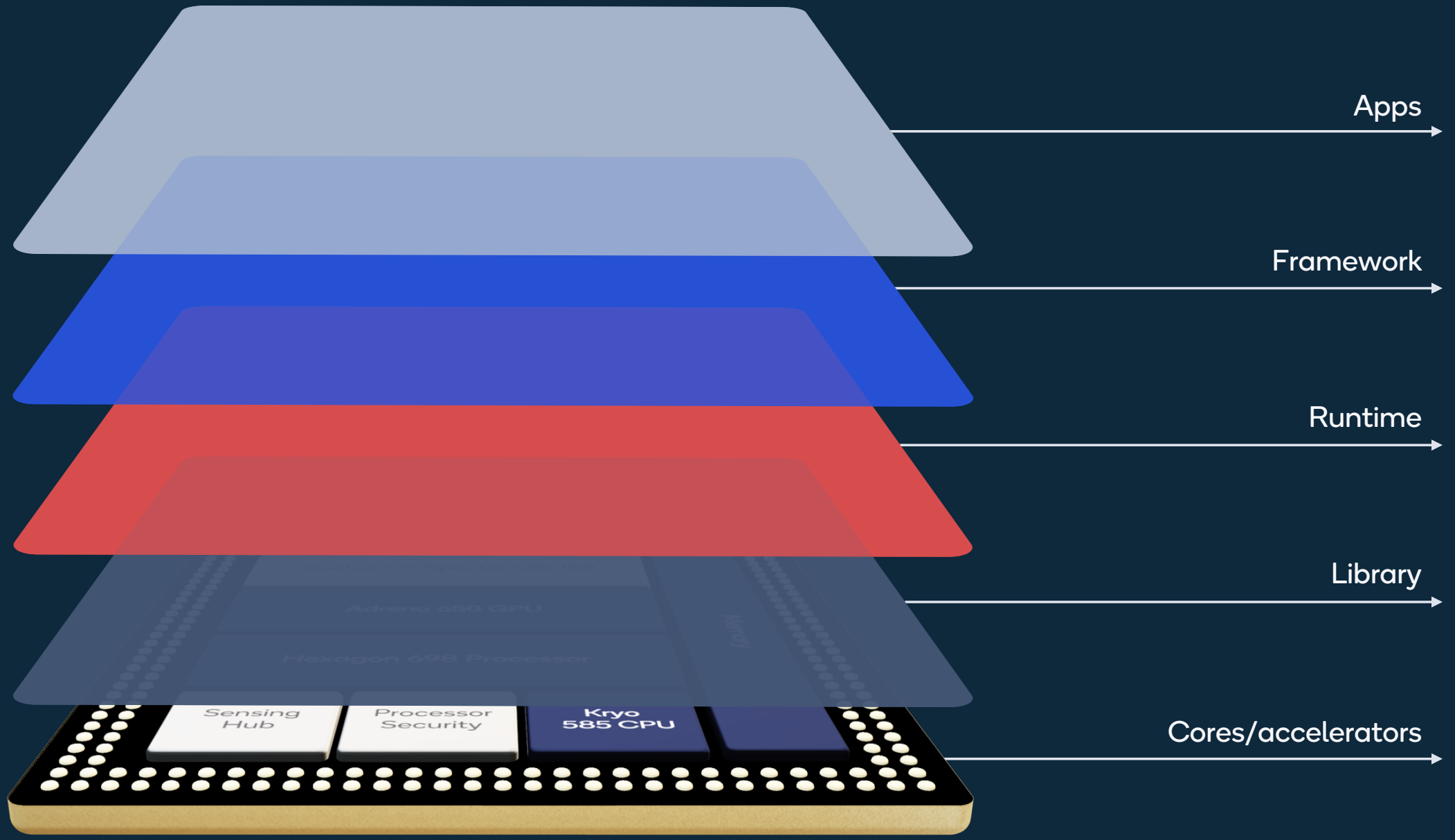
Hardware

Software

Developer tools


Qualcomm
Jeff Gehlhaar
Vice President, Technology, AI Software
Qualcomm Technologies, Inc.


Qualcomm Spectra 480 ISP
Adreno 650 GPU
Hexagon 698 Processor
Sensing Hub
Processor Security
Kryo 585 CPU
Memory

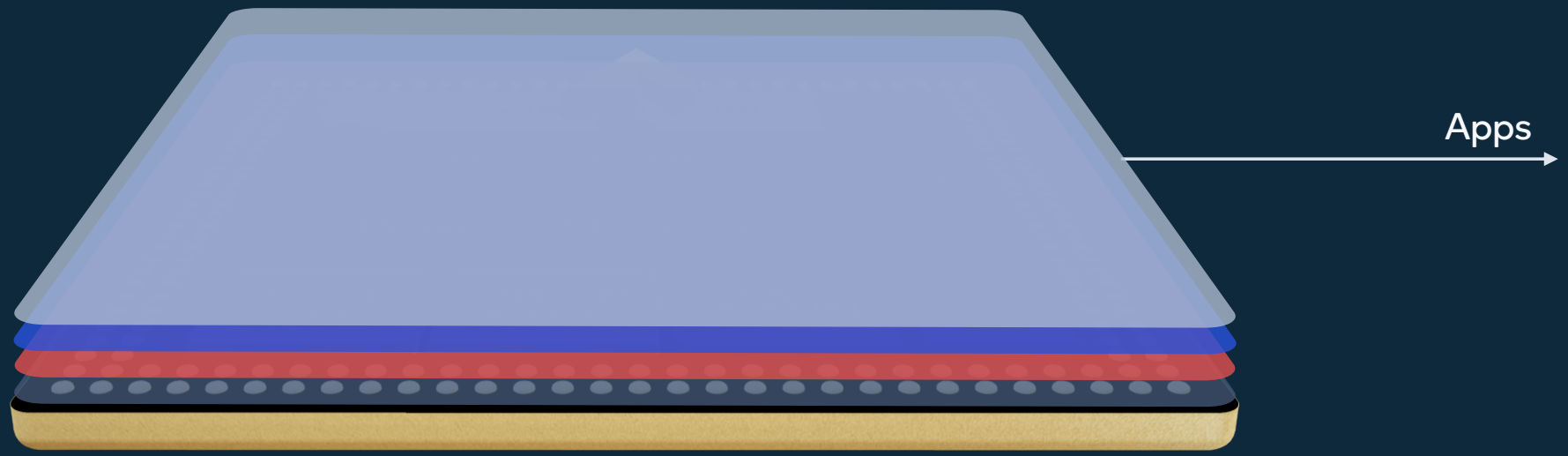

Apps
Framework
Runtime
Library
Cores/accelerators
Sensing Hub
Processor Security
Kryo 585 CPU

ByteDance Google ArcSoft Morpho 有道 youdao Audio Analytic

Qeexo anyvision A BETTER TOMORROW AISPEECH SEQURETEK SIMPLIFY SECURITY ELEVOC loom.ai Truepic

XPERI sensory babblelabs Mobius Labs SYNTIANT ultraleap

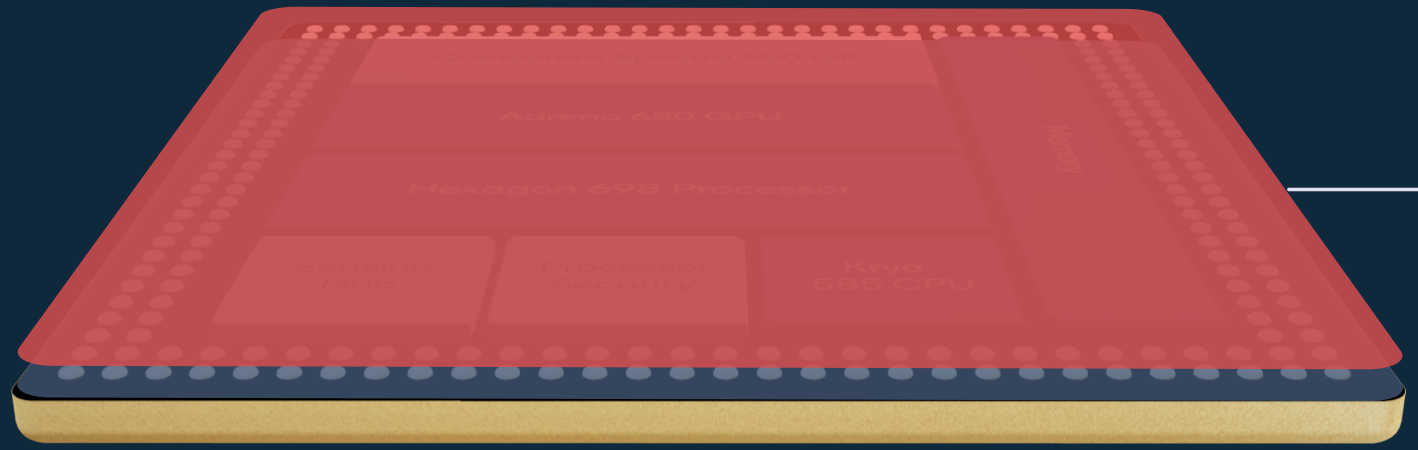
iQIYI 爱奇艺 AAC TECHNOLOGIES amn Tencent 腾讯 NAVER whale




PyTorch
TensorFlow
ONNX
Framework

Qualcomm
snapdragon

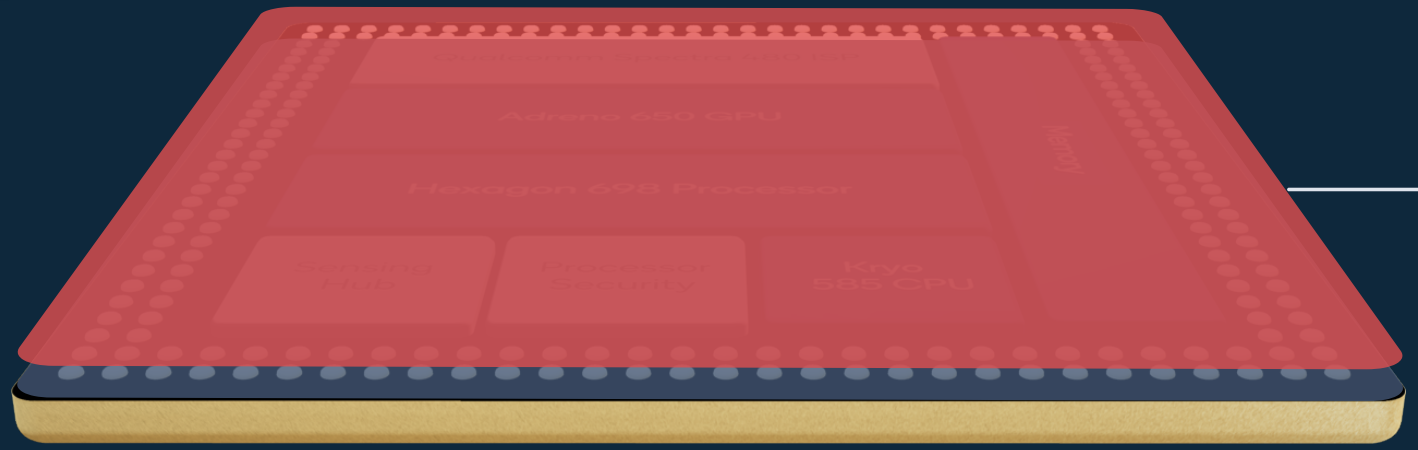
Qualcomm
Neural Processing SDK



Qualcomm
snapdragon

Qualcomm
Neural Processing SDK

Android Neural
Networks API



Runtime

Qualcomm Neural Processing SDK

Improved performance

Data free quantization

Support for more network models

Graph analysis

3-5x

Speed improvements on Hexagon processor

Android Neural Networks API

3x number of accelerated operators

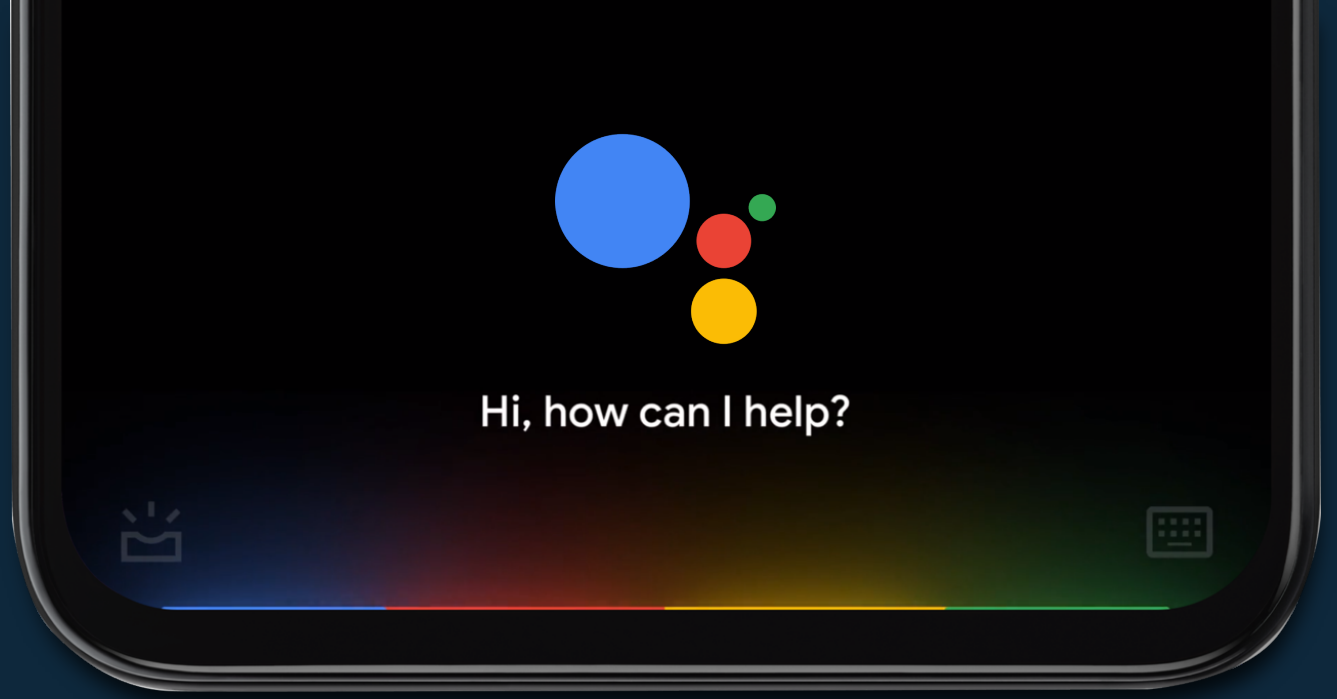
Shipping with Android 10

Optimized for Google speech recognition and Google Lens

Wide adoption from developers

Android Neural Networks API

Moving ASR for Google Assistant from CPU to Hexagon processor



3x
Power savings

30%
Lower latency

*Based on forward projections running in Android NN

Current operator count

160+

operators

User defined operators

operators

OpenCL

Hexagon SDK

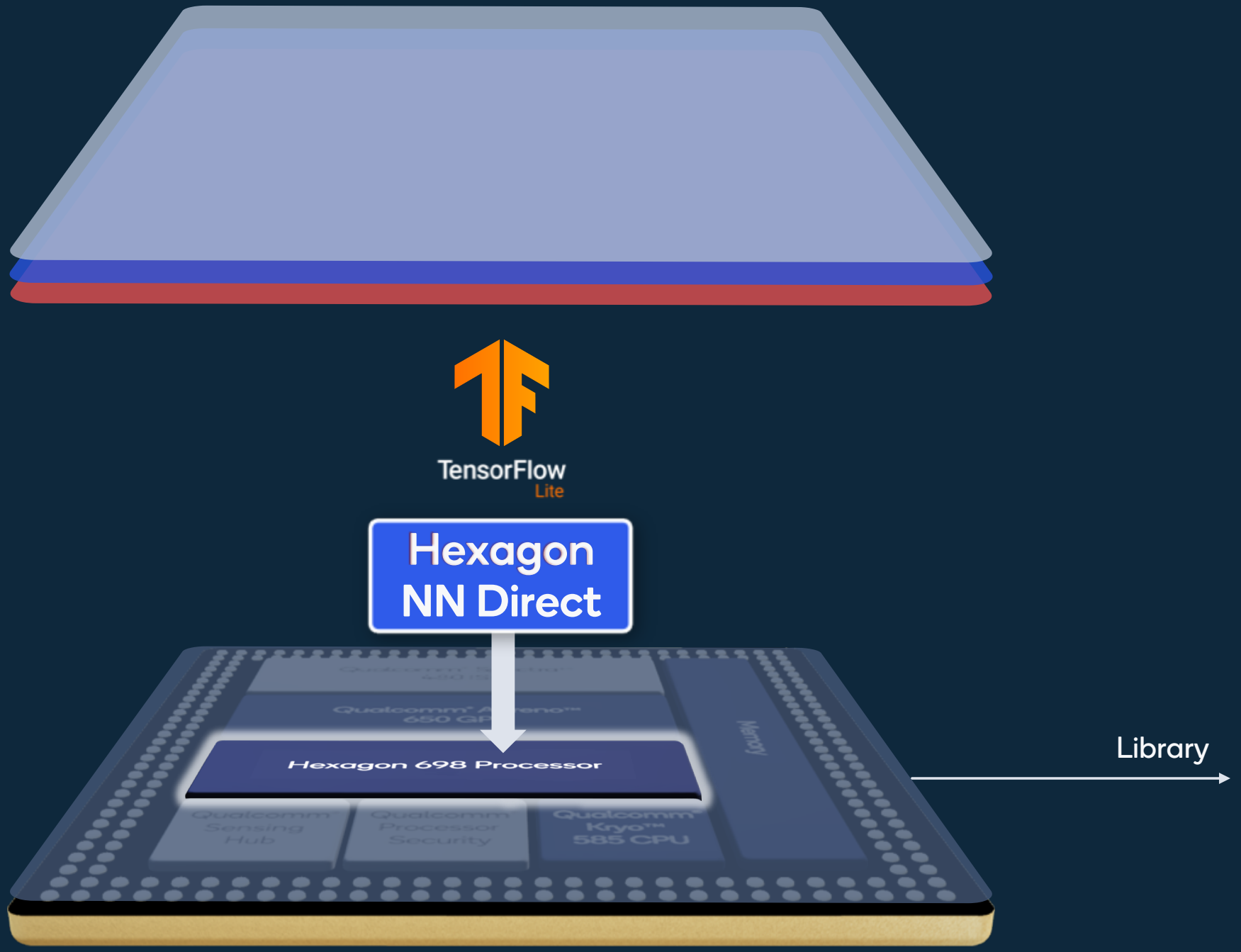
Qualcomm
snapdragon

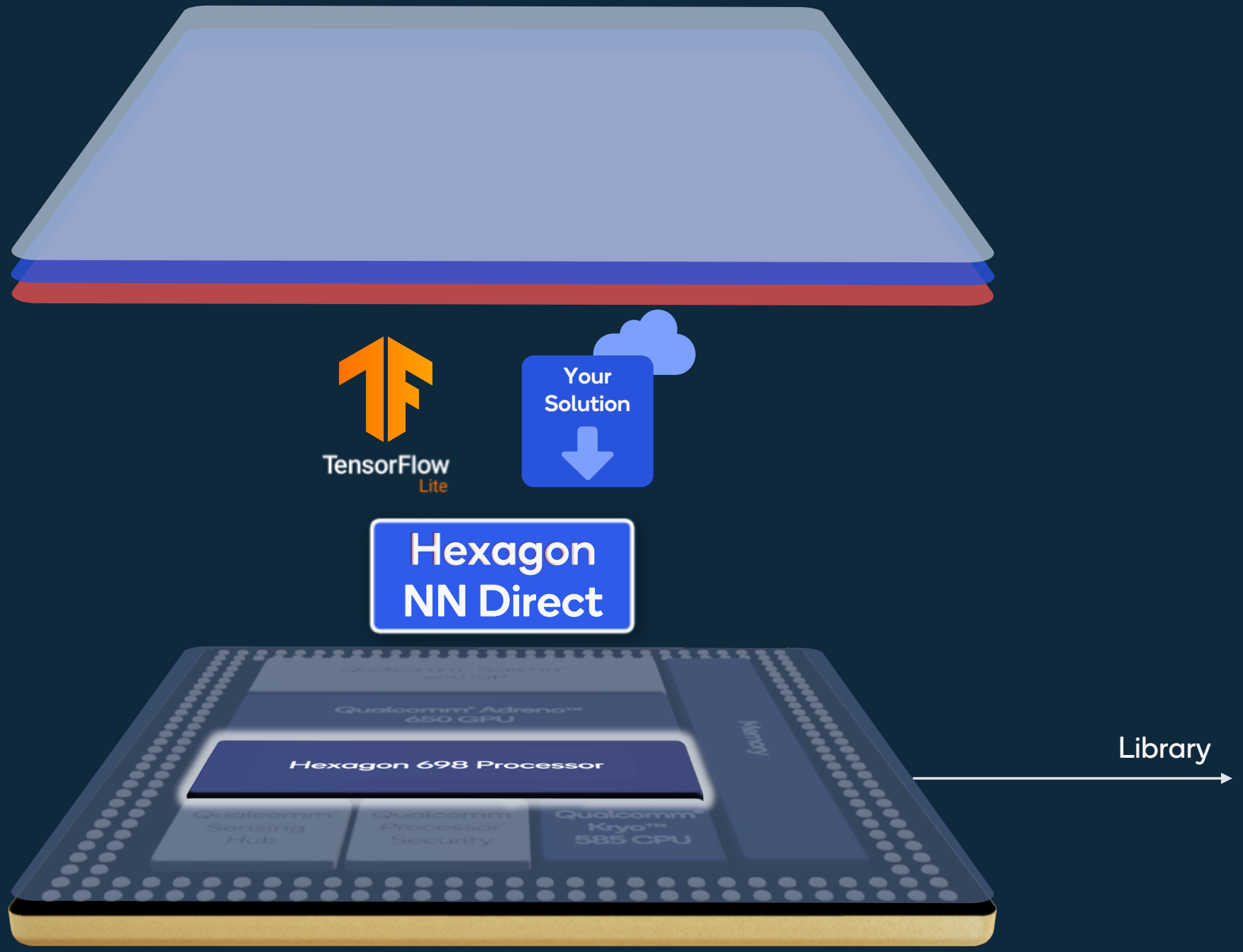
Qualcomm
Neural Processing SDK

TensorFlow
Lite

Android Neural
Networks API

Runtime


TensorFlow
Lite
Hexagon
NN Direct
Hexagon 698 Processor
Library

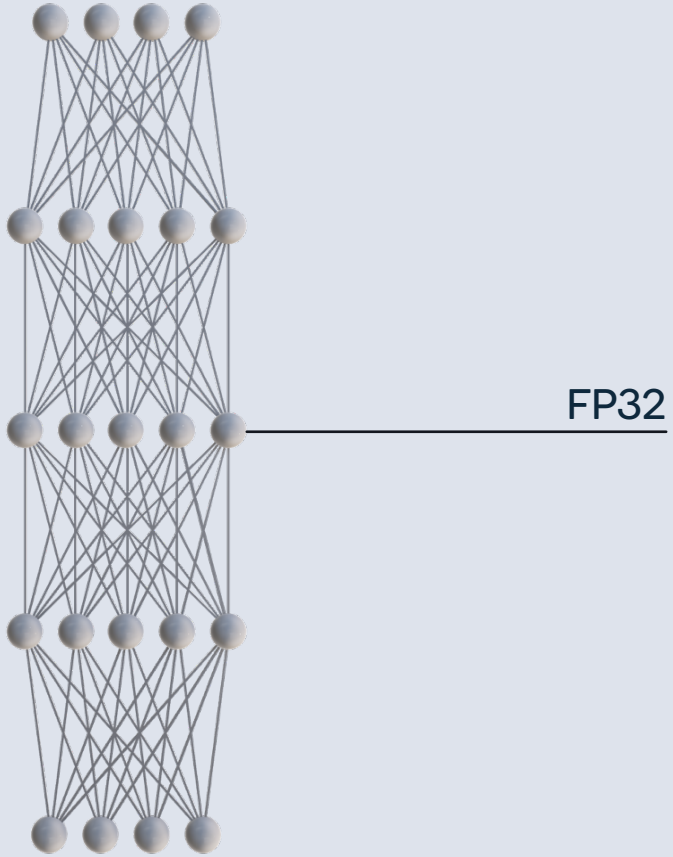

TensorFlow
Lite
Your
Solution
Hexagon
NN Direct
Qualcomm® Adreno™
650 GPU
Hexagon 698 Processor
Memory
Qualcomm
Sensing
Hub
Qualcomm
Processor
Security
Qualcomm
Kryo™
585 CPU
Library

Hardware

Developer tools

Software

New Qualcomm® AI Model Efficiency Toolkit

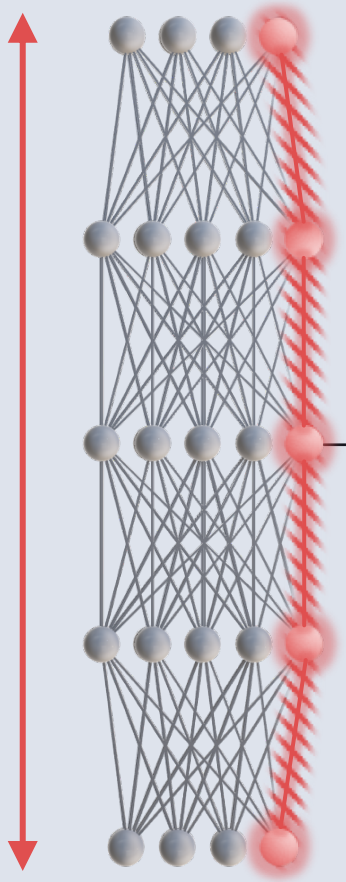


New Qualcomm® AI Model Efficiency Toolkit

Model compression

Spatial SVD

Bayesian compression



FP32

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

865 5G
Qualcomm snapdragon

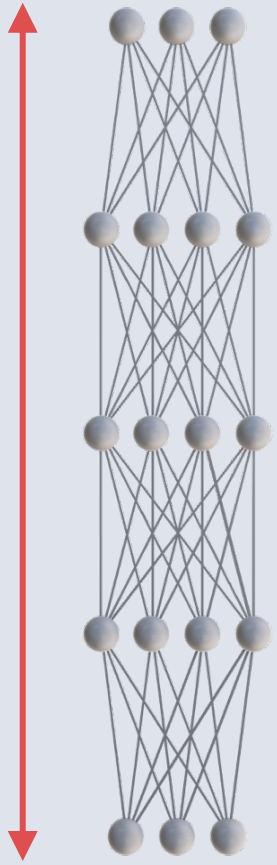
*Comparison between baseline and compression with both Bayesian compression and spatial SVD. Example uses ResNet18 as baseline.

New Qualcomm® AI Model Efficiency Toolkit

Model compression

Data free quantization

Quantization aware training



FP32

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

865 5G
Qualcomm snapdragon

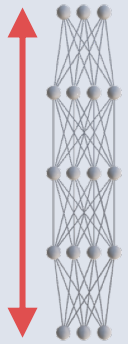
*Comparison between baseline and compression with both Bayesian compression and spatial SVD. Example uses ResNet18 as baseline.

New Qualcomm® AI Model Efficiency Toolkit

Model compression

Data free quantization

Quantization aware training



INT8

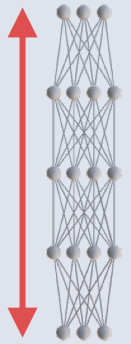
>4x

Increase in performance per watt with quantization

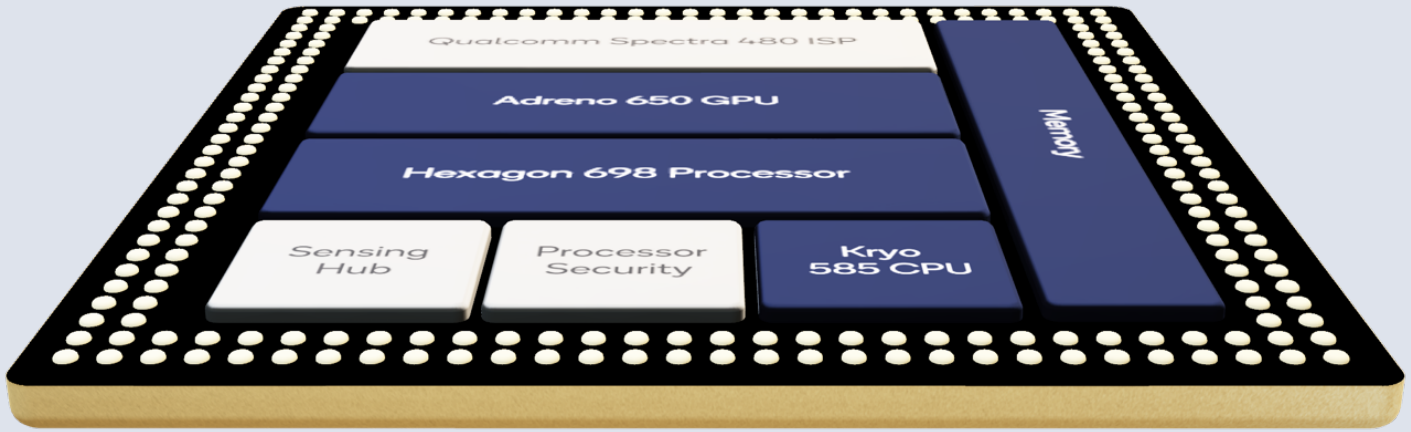
Qualcomm snapdragon 865 5G

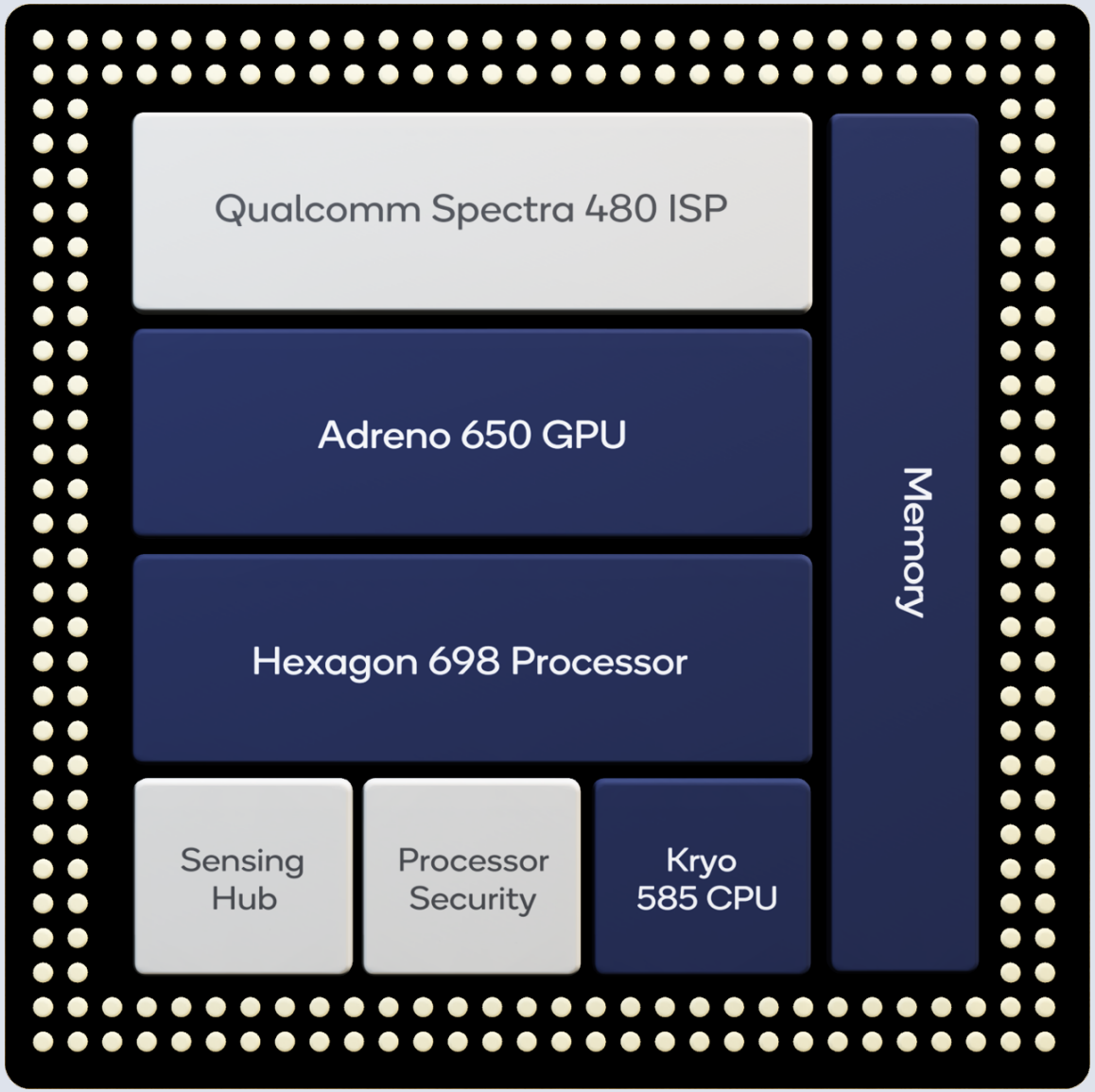
3x
Compression with less than 1% loss in accuracy*

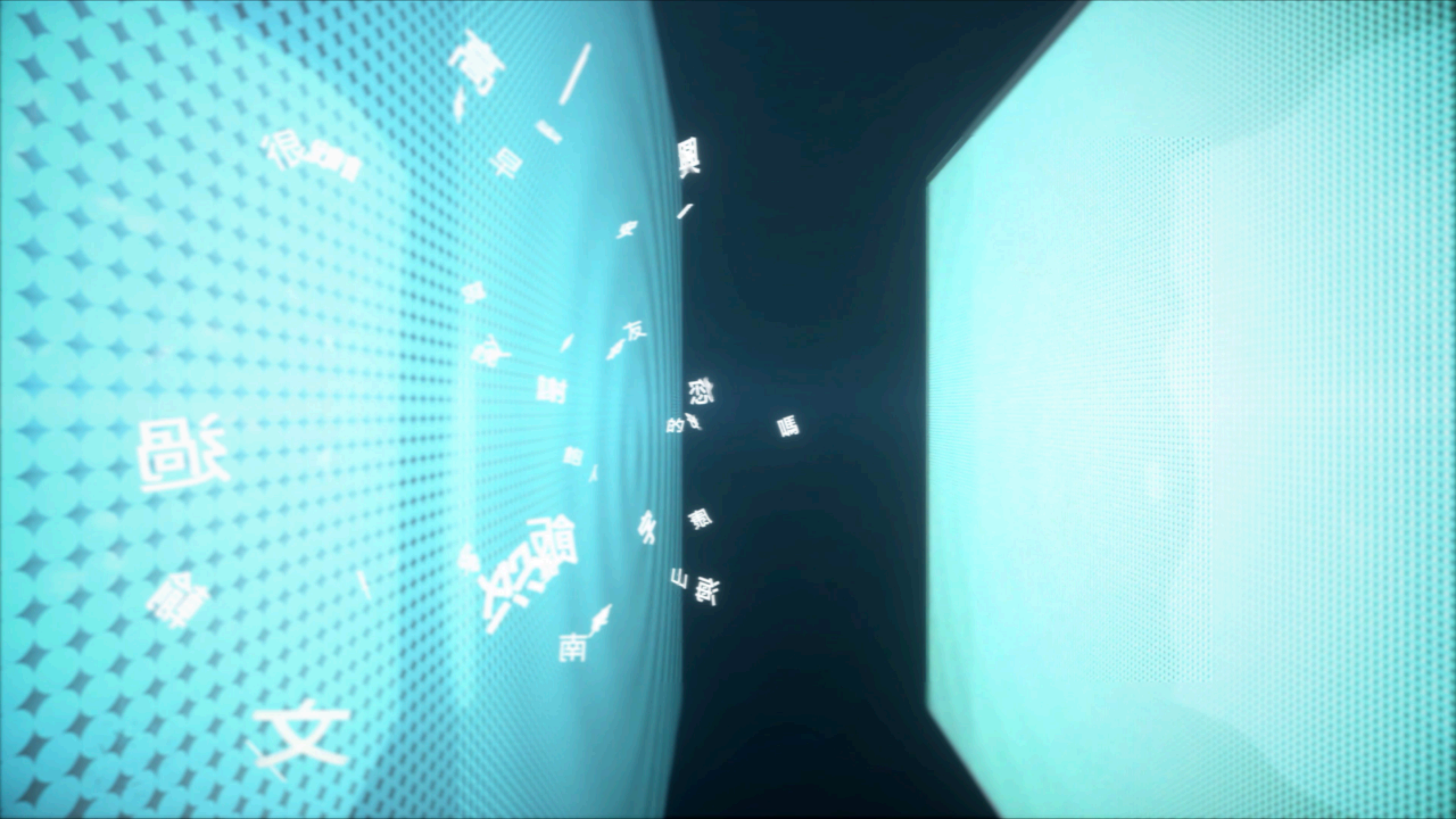
>4x
Increase in performance per watt with quantization



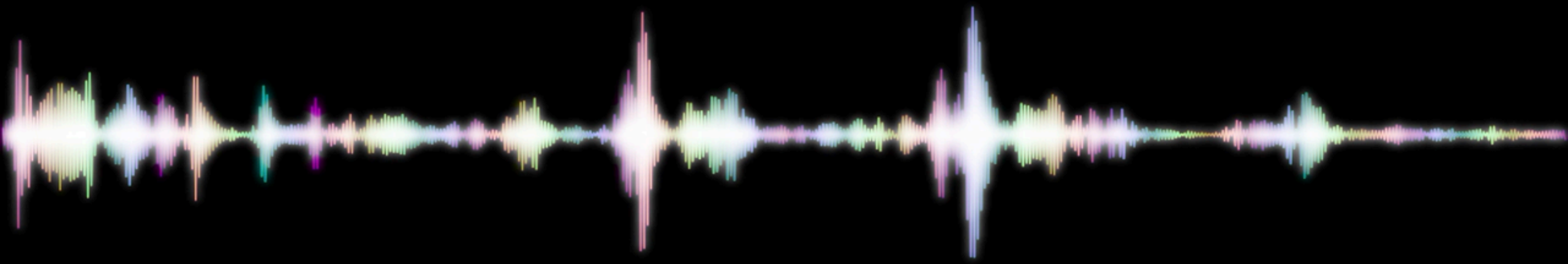
INT8




Qualcomm Spectra 480 ISP
Adreno 650 GPU
Hexagon 698 Processor
Sensing Hub
Processor Security
Kryo 585 CPU
Memory



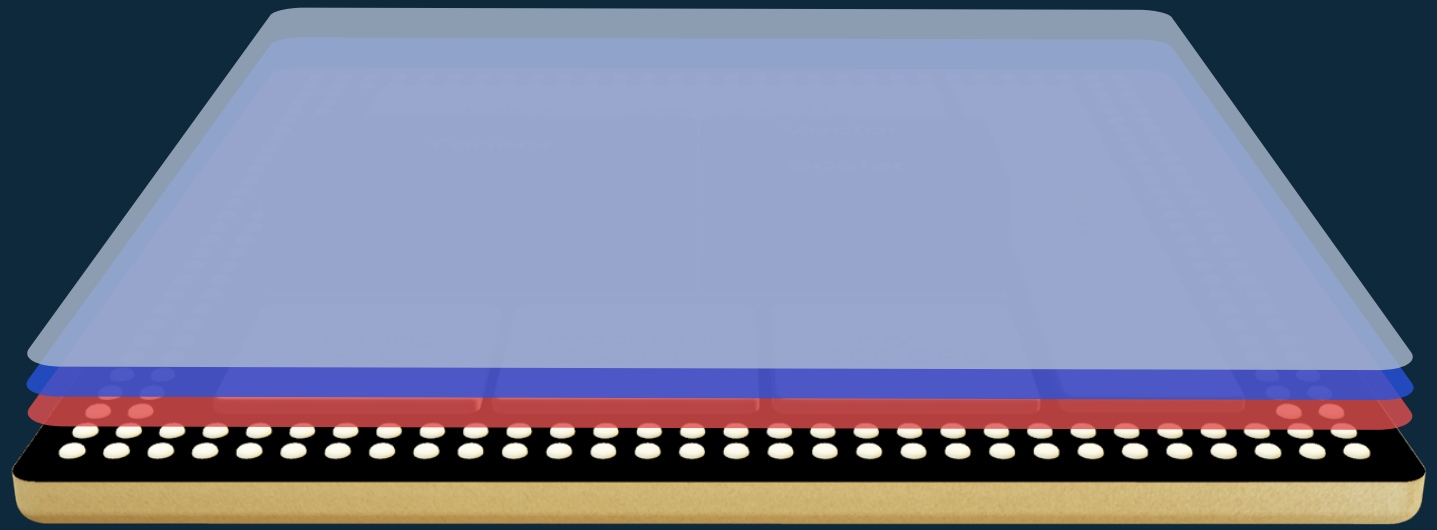
Snapdragon 865 is a leading mobile AI platform



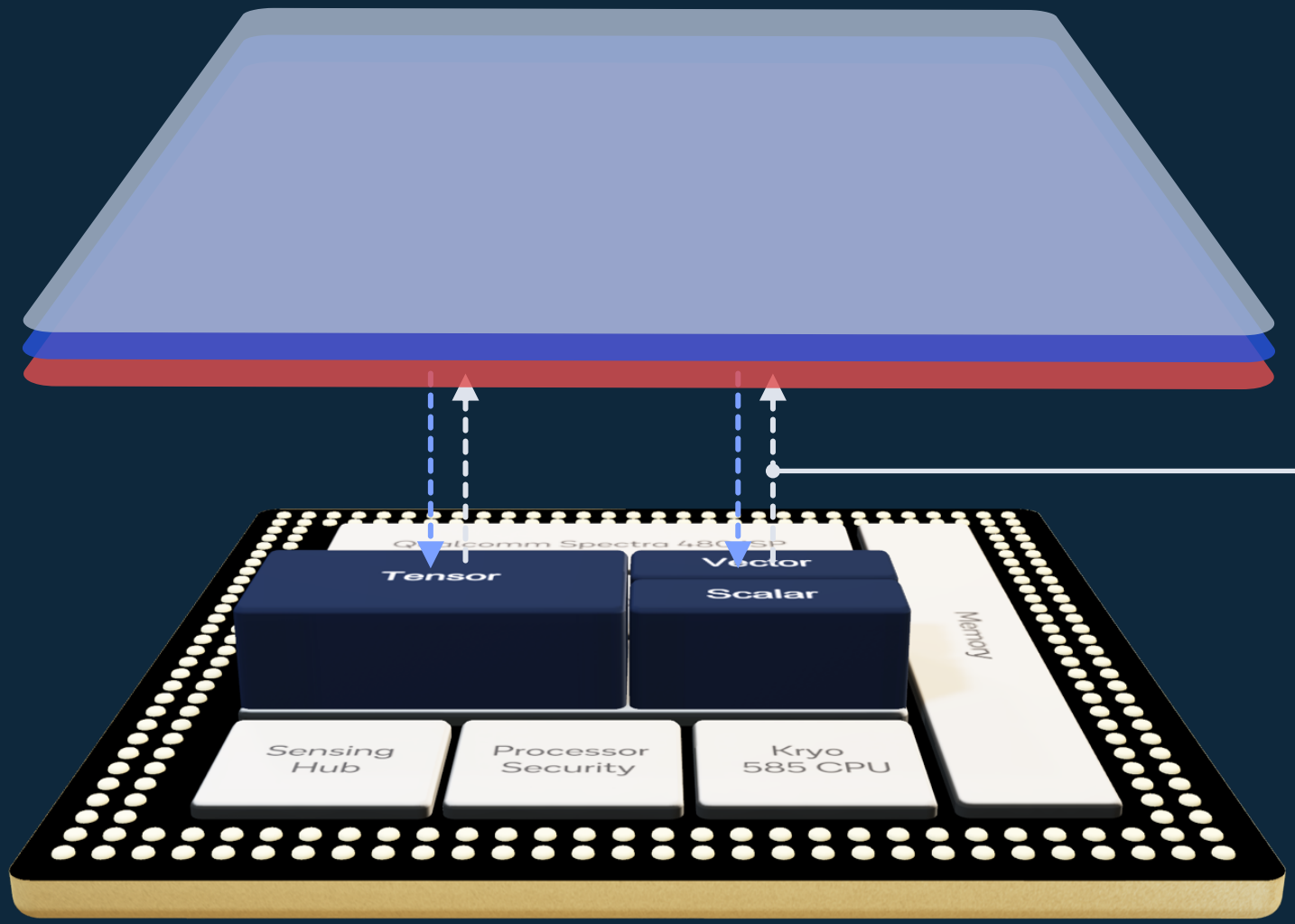

5G


Source sample text

ByteDance



ByteDance



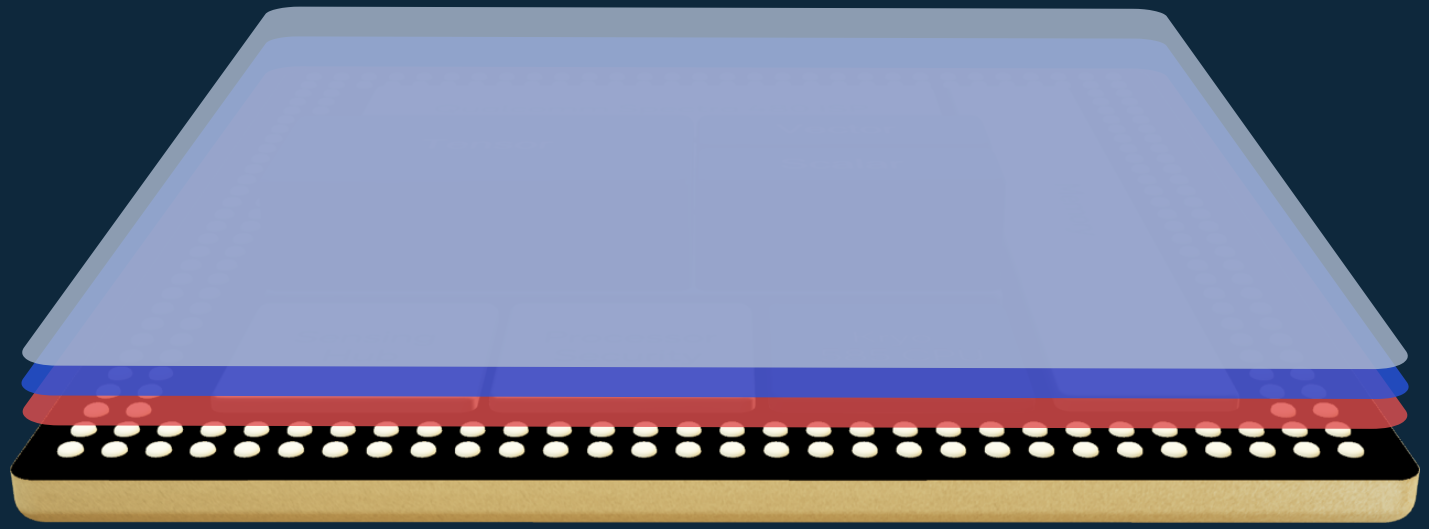
Qualcomm snapdragon

Qualcomm
Neural Processing SDK

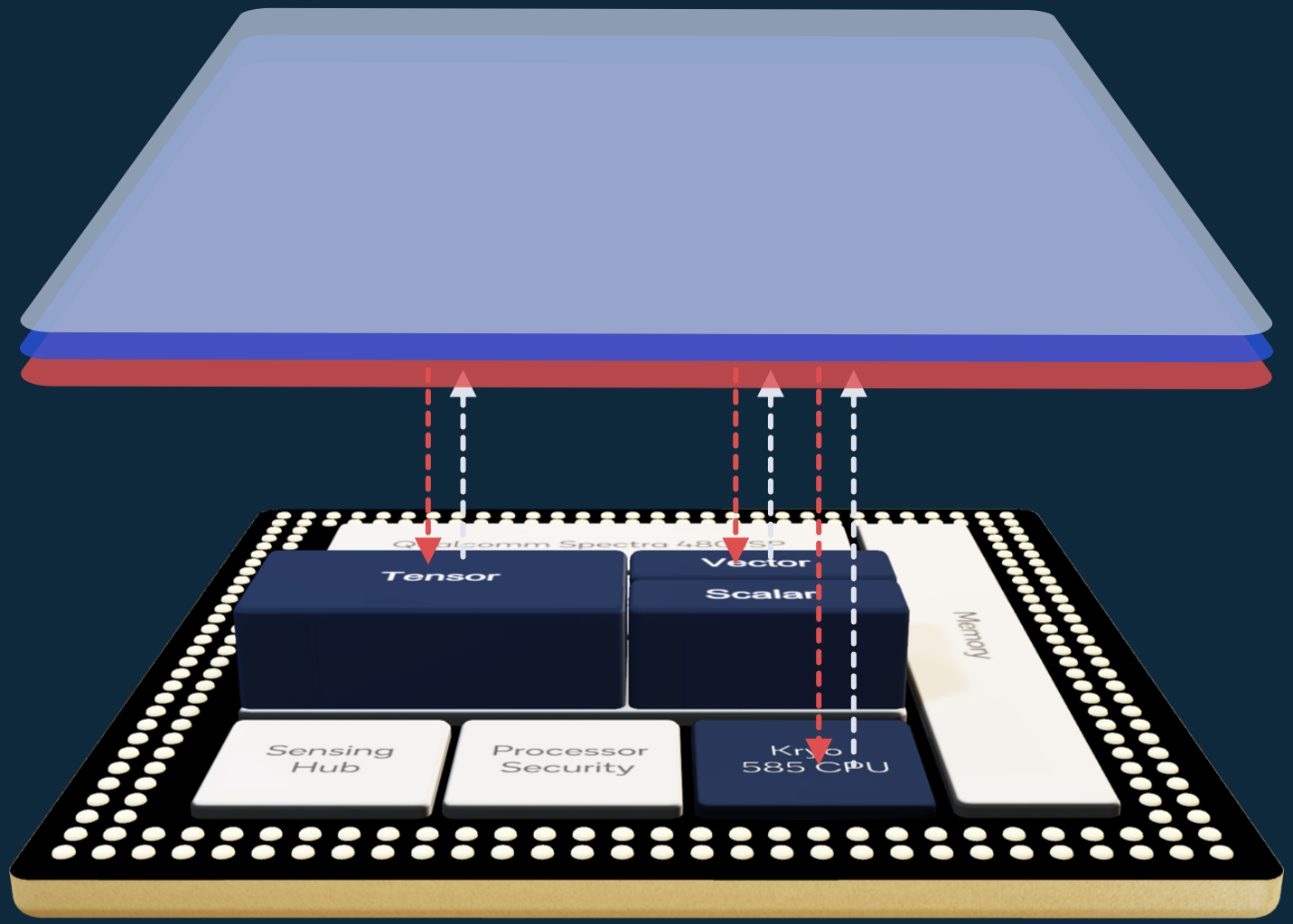
Hexagon
NN Direct

Super resolution–
Convolutional Neural Network (CNN)

loom.ai



loom.ai



Qualcomm
snapdragon

Qualcomm
Neural Processing SDK

Face detection –
Convolutional Neural Network (CNN)

Landmark detection –
Convolutional Neural Network (CNN)

Expression detection –
Convolutional Neural Network (CNN)

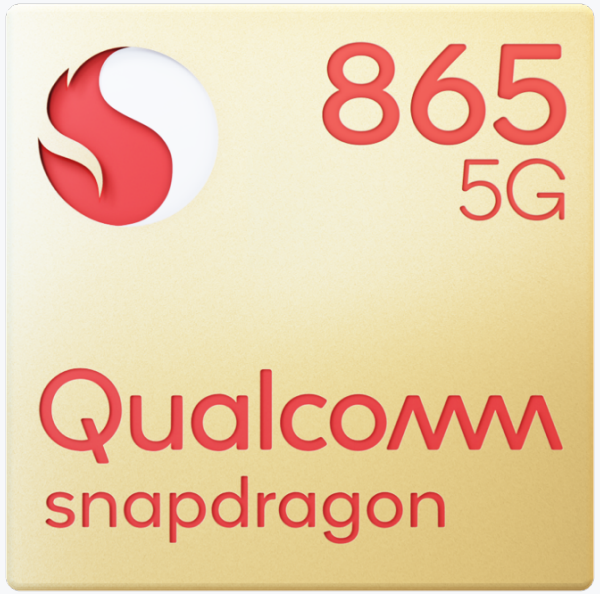
Replace face with avatar –
Convolutional Neural Network (CNN)


Gaming
Game Center
X Mode
CPU
GPU
Game Profiles
Fan speed
Memory Used
24 hours 13 minut


XR
CPU
GPU
Memory
Sensor processing
ISP
DSP
Display processing
BT FM
Wi-Fi
GNSS
USB
PCI
LTE modem
Multimedia processing


Contextual Awareness

AI highlights



Adreno 650

New AI mixed precision instructions

2x higher TOPS

16-bit and 32-bit FP

Hexagon 698

4x higher TOPS

Up to 35% power savings

Deep learning bandwidth compression

LP-DDR5 Memory

30% more bandwidth

5th gen Qualcomm AI engine

15 TOPS

Qualcomm Neural Processing SDK

New features and improvements

NNAPI support

3x number of accelated operators

Optimized for Google ASR and Google Lens

Hexagon NN Direct

Provide developers direct access to Hexagon

AI model efficiency toolkit

Data free quantization

Quantization aware training

Model compression

Qualcomm

Thank you

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog