What we need to Transform Lives and Industries with On-Device AI, Cloud and 5G

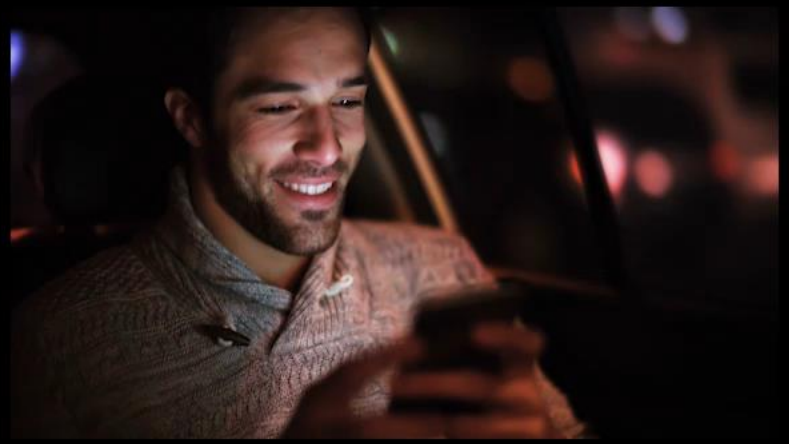
Ziad Asghar, Vice President, Product Management at Qualcomm Technologies Inc.

Qualcomm


eXtended Reality


5G
生意正在好轉
MAN
ENG
Qualcomm
snapdragon


Smart Cropping



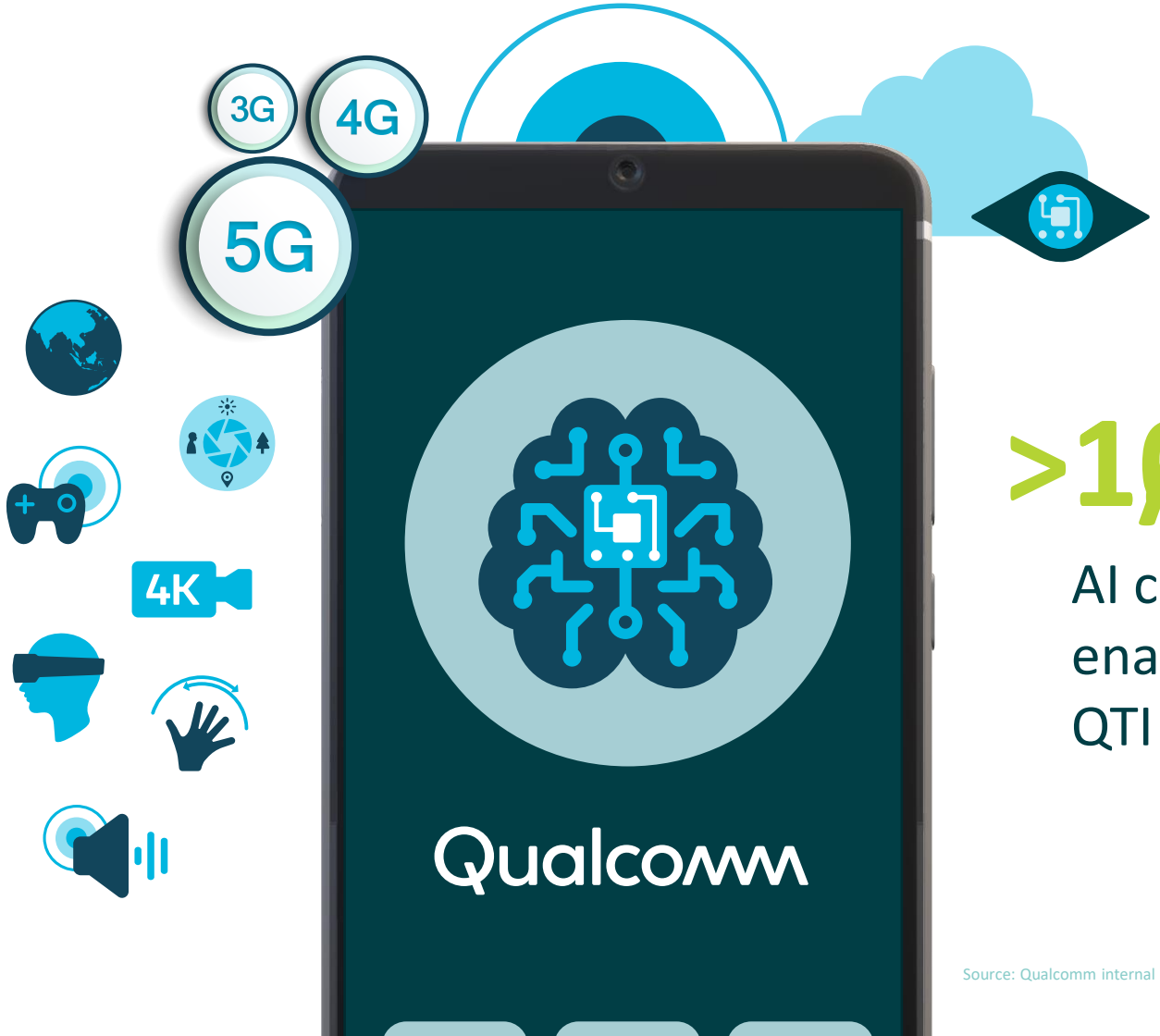

20.000
30°
On-device
Private – Personal – Immediate


5G
Public network
Private network
20.000
30°


5G
Public network
Private network
Edge cloud
20.000

Qualcomm

Mobile — THE most pervasive AI platform

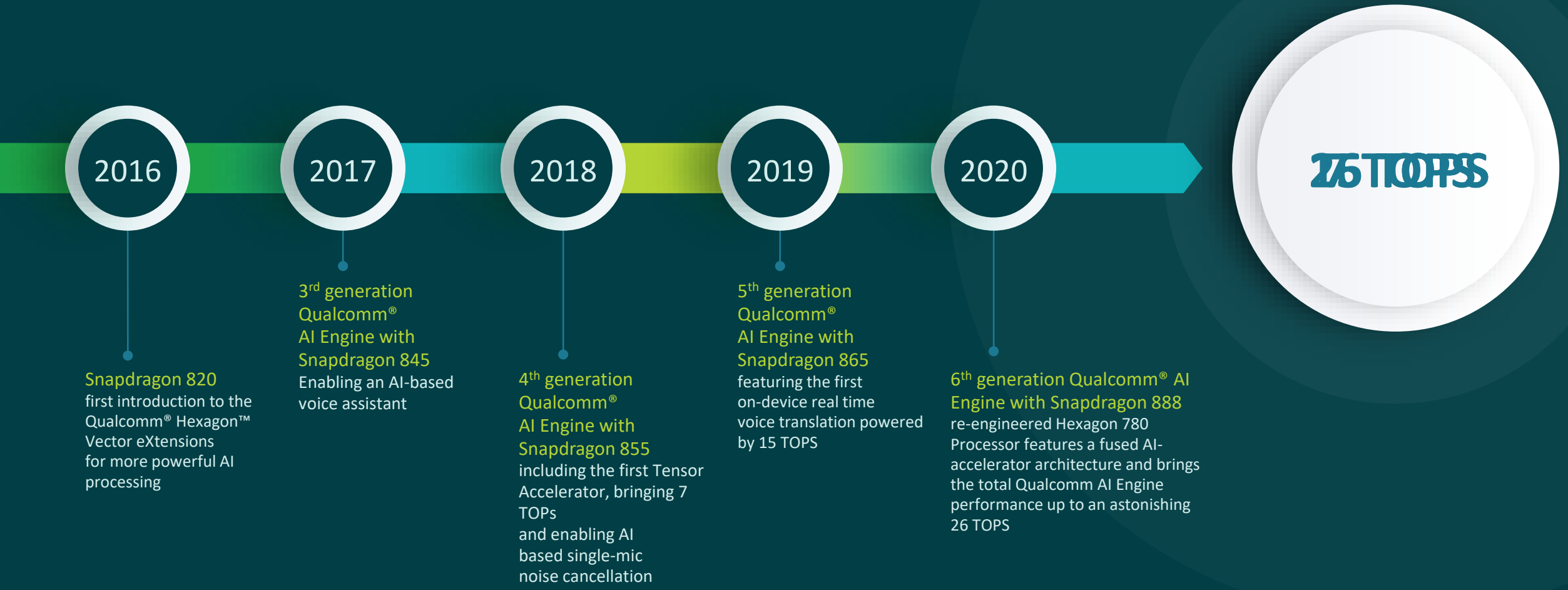


>1B

AI capable devices
enabled with
QTI technology

Source: Qualcomm internal

Snapdragon AI Milestones



2016

Snapdragon 820
first introduction to the Qualcomm® Hexagon™ Vector eXtensions for more powerful AI processing

2017

3rd generation Qualcomm® AI Engine with Snapdragon 845
Enabling an AI-based voice assistant

2018

4th generation Qualcomm® AI Engine with Snapdragon 855
including the first Tensor Accelerator, bringing 7 TOPs and enabling AI based single-mic noise cancellation

2019

5th generation Qualcomm® AI Engine with Snapdragon 865
featuring the first on-device real time voice translation powered by 15 TOPS

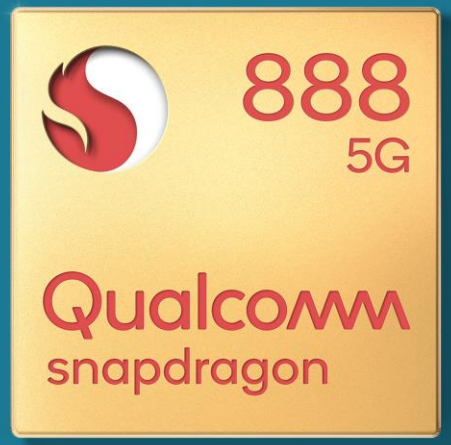
2020

6th generation Qualcomm® AI Engine with Snapdragon 888
re-engineered Hexagon 780 Processor features a fused AI-accelerator architecture and brings the total Qualcomm AI Engine performance up to an astonishing 26 TOPS

Qualcomm snapdragon

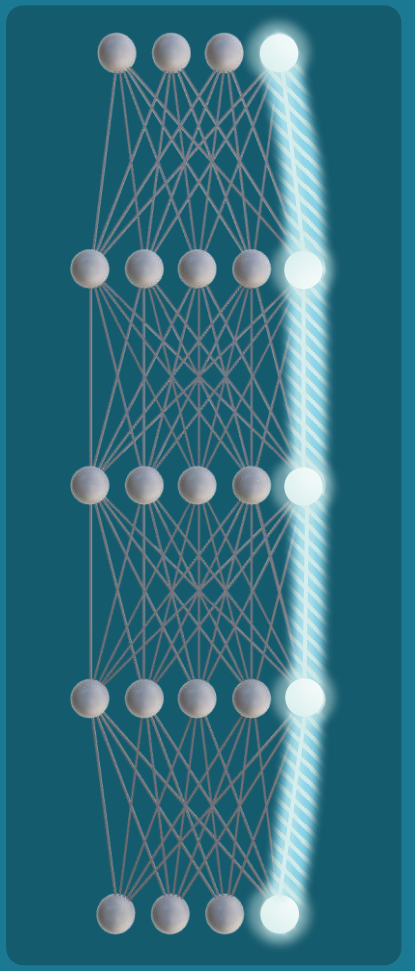
Key Ingredients for AI Leadership

Hardware



+ Software

+ Tools



Connectivity

3rd Gen Snapdragon
X60 Modem-RF System

3rd Gen mmWave

5G Carrier Aggregation*

Qualcomm® FastConnect™ 6900
6GHz, 4K QAM, BT 5.2

Performance

Kryo 680
X1 Architecture, 25% faster

Adreno 660
35% faster graphics rendering

5nm
Process Technology

Security

Hypervisor

CAI Compliant camera

AI

6th Gen
AI Engine

Hexagon 680
Fused AI Accelerators

3x Performance per watt

26 TOPS

2nd Gen
Qualcomm
Sensing Hub

1st

888 5G

Qualcomm snapdragon

Gaming

3rd Gen Qualcomm®
Snapdragon Elite Gaming™

Ultra smooth gaming:
Qualcomm® Game Quick Touch,

Desktop level features:
Qualcomm® Variable Rate Shading

Camera

2.7 Gigapixels Per Second:
35% Faster

Triple ISP:

Triple Concurrency

Triple Parallel Processing

Triple Capture:

Triple 4K HDR Video Capture

Triple 28MP Photo Capture

Burst Capture
120 photos in
1 second at 12MP

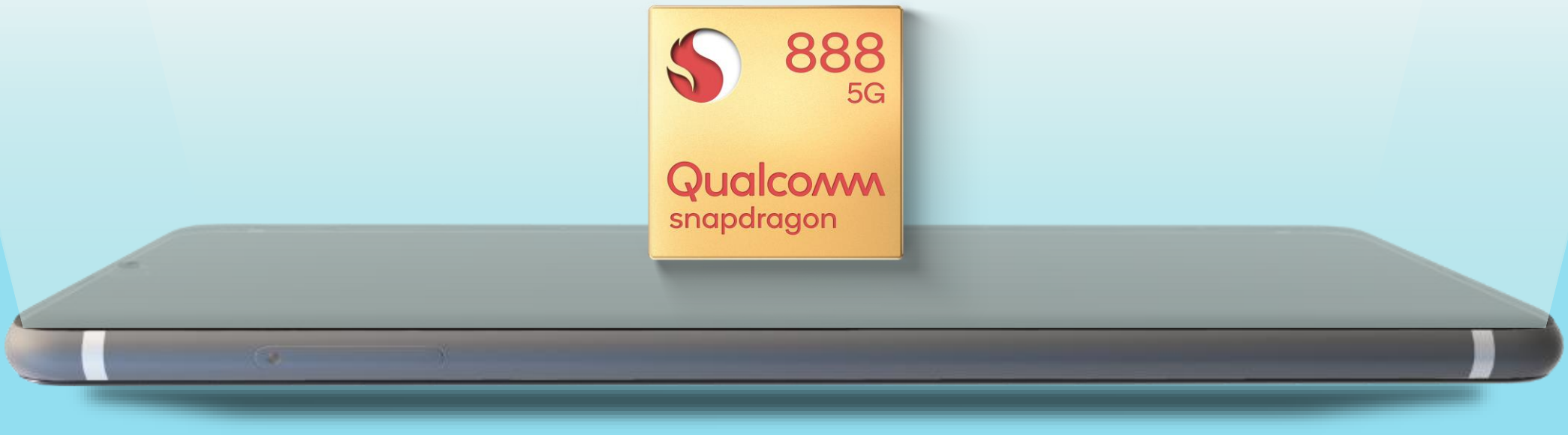
4K HDR
video capture with
computational HDR

low light architecture
photo capture

10-bit HDR
HEIF photo
capture

* across TDD and FDD

6th Generation
Qualcomm AI Engine



Hexagon 780 Processor

Fused AI accelerator

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

Shared Memory:

16X
dedicated memory

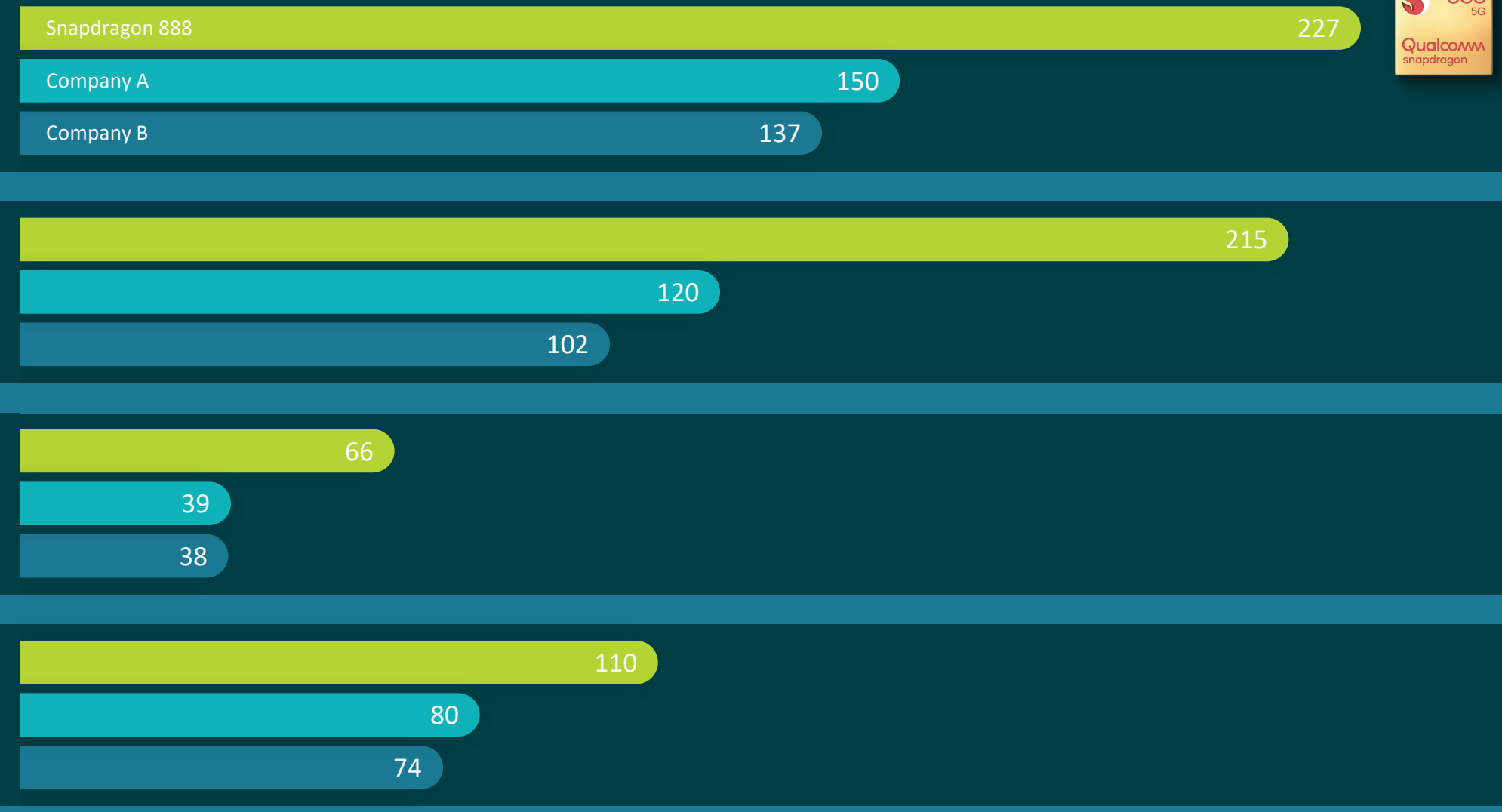
Up to
1000x
hand off time improvement in certain use cases

Qualcomm

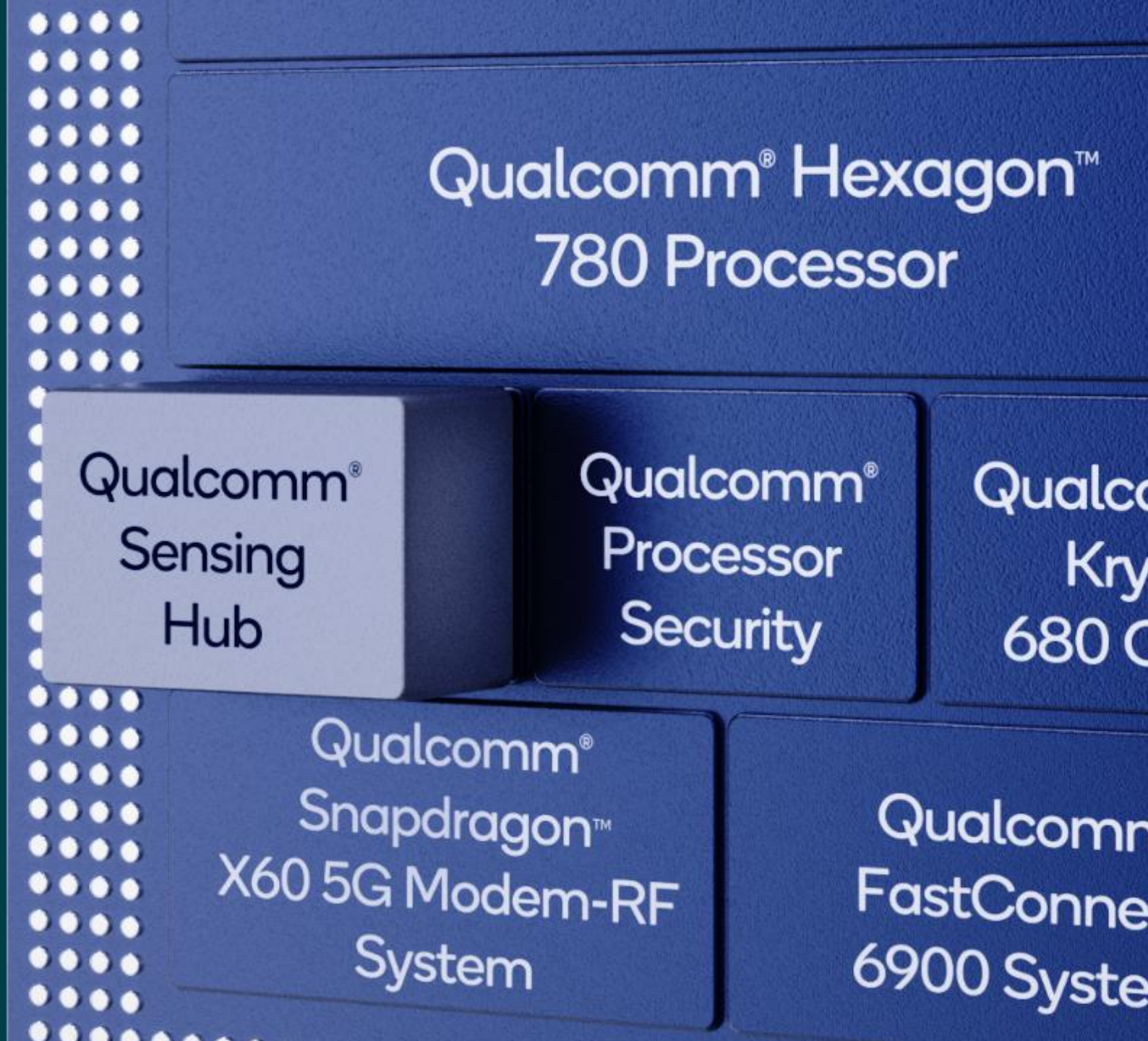
Compared to previous generation

Power Efficiency

Efficiency: (inf/s/W)

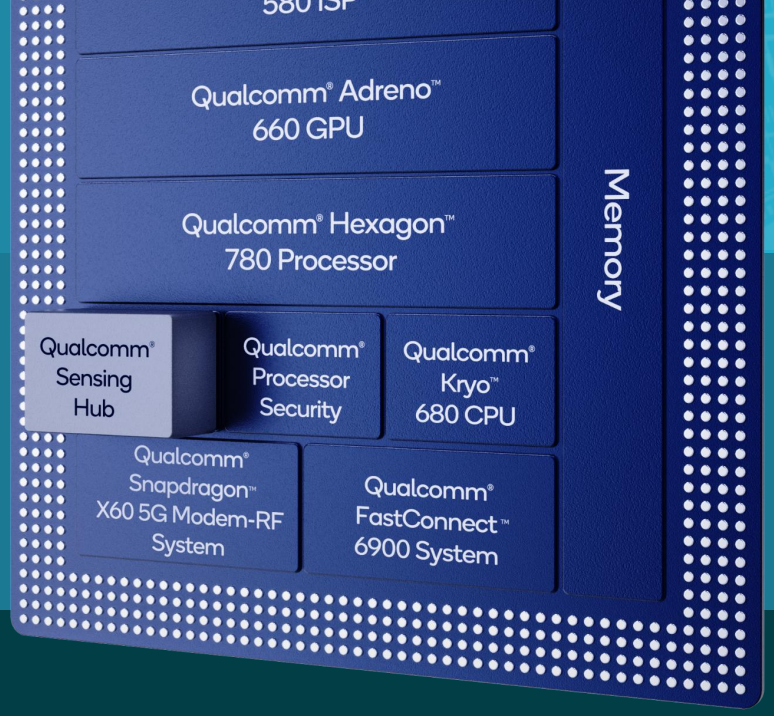


2nd Gen
Qualcomm Sensing Hub <1mA



Qualcomm

AI performance



With 2nd Gen Qualcomm Sensing Hub

1st Gen Qualcomm Sensing Hub

5X

80%

Task offload from Hexagon Processor

Sensing Hub – Use case

Qualcomm AI software stack

Supporting every AI software layer from applications to the metal

Tools + Compilers

AIMET

TVM

Models

ResNet | SSD | MobileNet | Mobile BERT | VDSR | DeepLab

Applications

Qualcomm® Neural Processing SDK

TensorFlow Lite

Android Neural Networks API

Frameworks

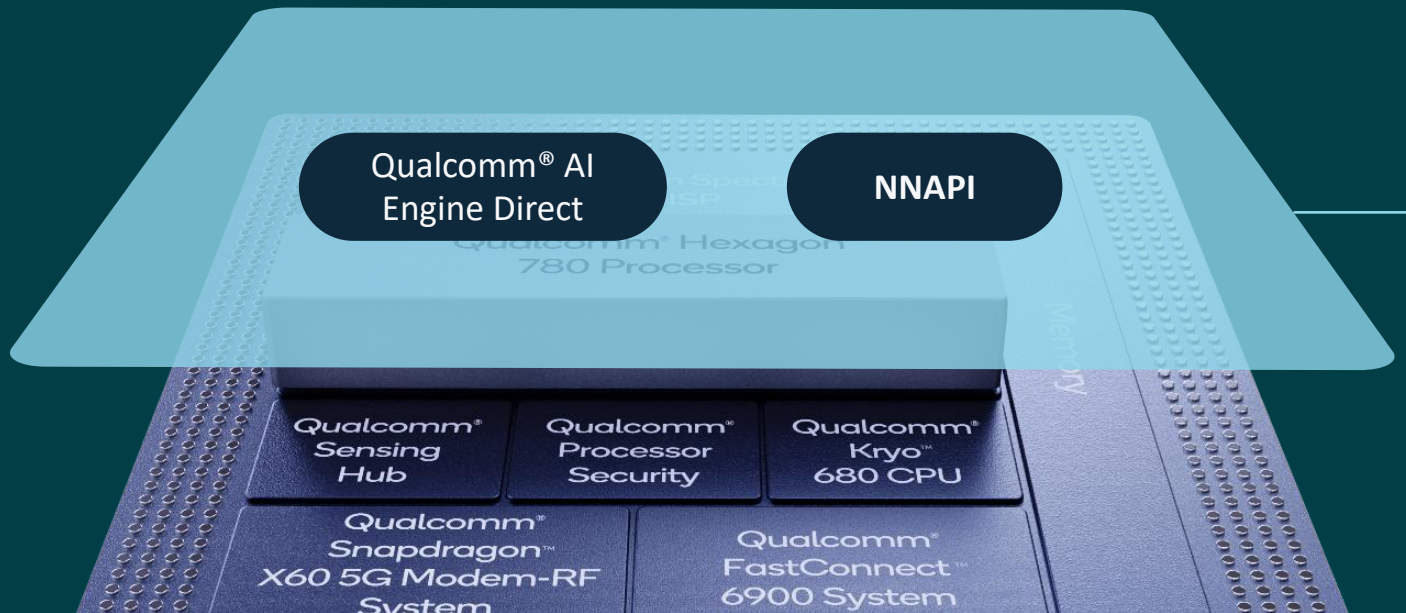
ONNX

Runtime

Qualcomm® AI Engine Direct

NNAPI

SDKs



AIMET + Tools

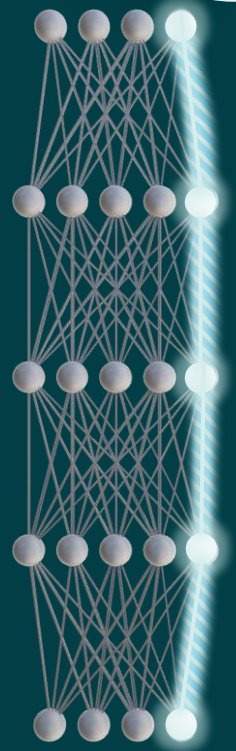
2021 embedded VISION summit

>4x Increase in performance per watt with quantization

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

AI Model Efficiency Toolkit

- Improved/ Robust quantization for INT16,8,4
- Quantization aware training with range learning
- Mix precision support
- Opensource



AI Model Efficiency Toolkit is a product of Qualcomm Innovation Center, Inc.

AIMET Model Zoo

- Image classification
- Semantic segmentation
- Super resolution
- Object detection
- Pose estimation
- Speech recognition

```python
def quantized_add_generic(size, a, b, a_offset, b_offset, a_mult, b_mult, output, target, ctx):
    # Construct the TVM computation.
    A_offset = tvm.var('A_offset', dtype='uint8')
    B_offset = tvm.var('B_offset', dtype='uint8')
    A_mult = tvm.var('A_mult', dtype='uint16')
    B_mult = tvm.var('B_mult', dtype='uint16')

    N = tvm.var("N")
    A = tvm.placeholder((N,), name='A', dtype='uint8')
    B = tvm.placeholder((N,), name='B', dtype='uint8')

    C = tvm.compute((A.shape),
          lambda i: ((A[i].astype('int32') - A_offset.astype('int32')) * A_mult.astype('int32')) +
                    ((B[i].astype('int32') - B_offset.astype('int32')) * B_mult.astype('int32')), name='C')

    # Create the schedule.
    s = tvm.create_schedule(C.op);
    px, x = s[C].split(s[C].op.axis[0], nparts=1)
    s[C].bind(px, tvm.thread_axis("pipeline"))

    # Construct the callable object "func" corresponding to the computation.
    func = tvm.build(s, [A, B, C, N, A_offset, B_offset, A_mult, B_mult], target, name='qadd_tvm')

    func(tvm.ndarray.array(a, ctx=ctx), tvm.ndarray.array(b, ctx=ctx), output,
         size, a_offset, b_offset, a_mult, b_mult)
```

tvm

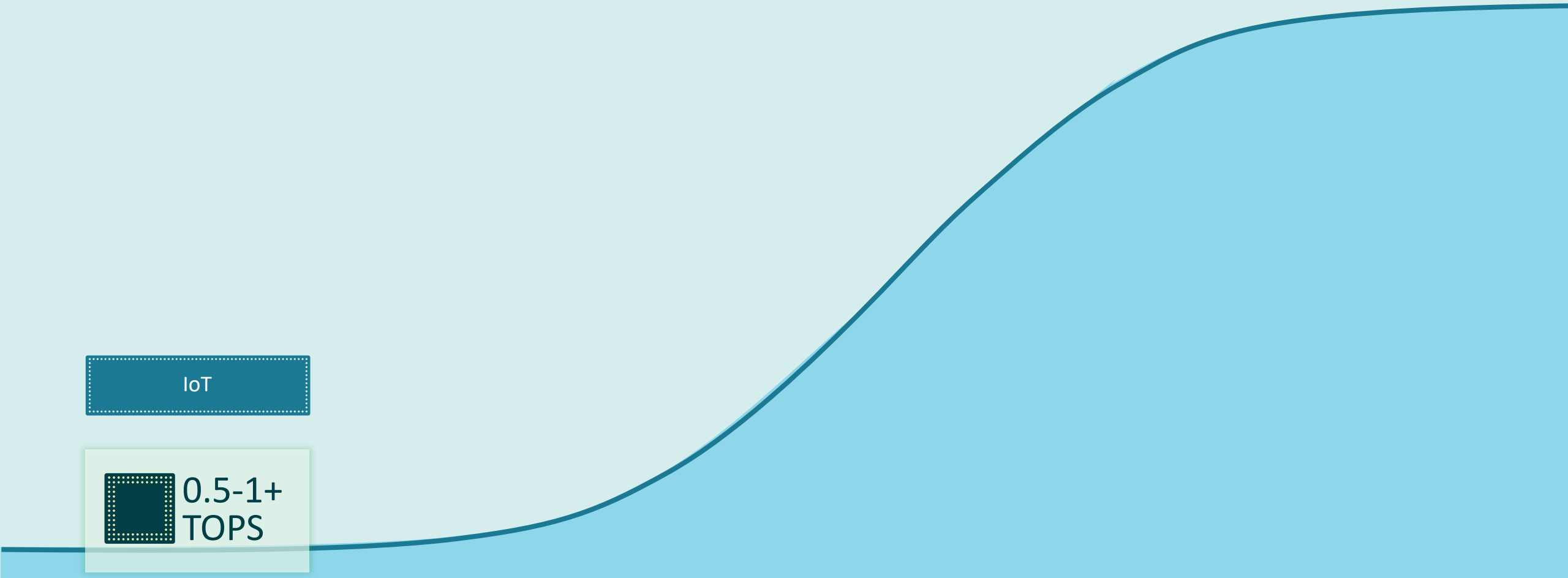
Qualcomm AI Engine direct

Custom operators efficiently written in

python™

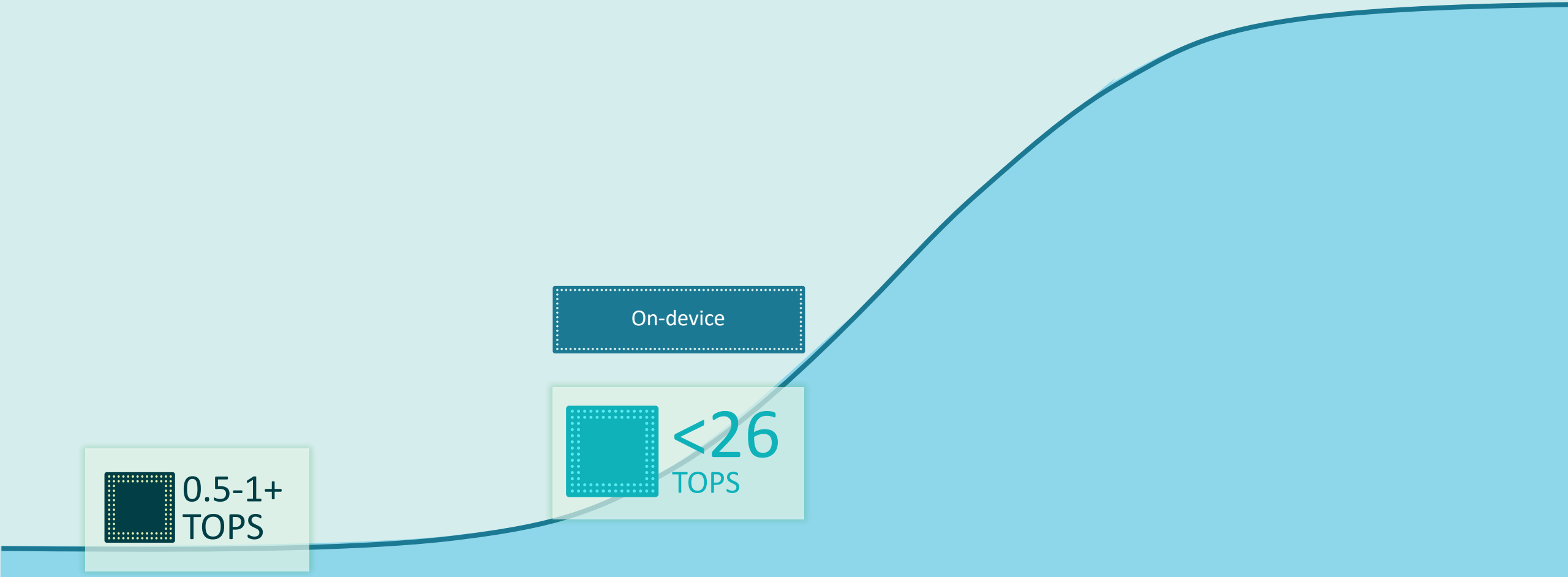
Qualcomm

AI Performance Continuum



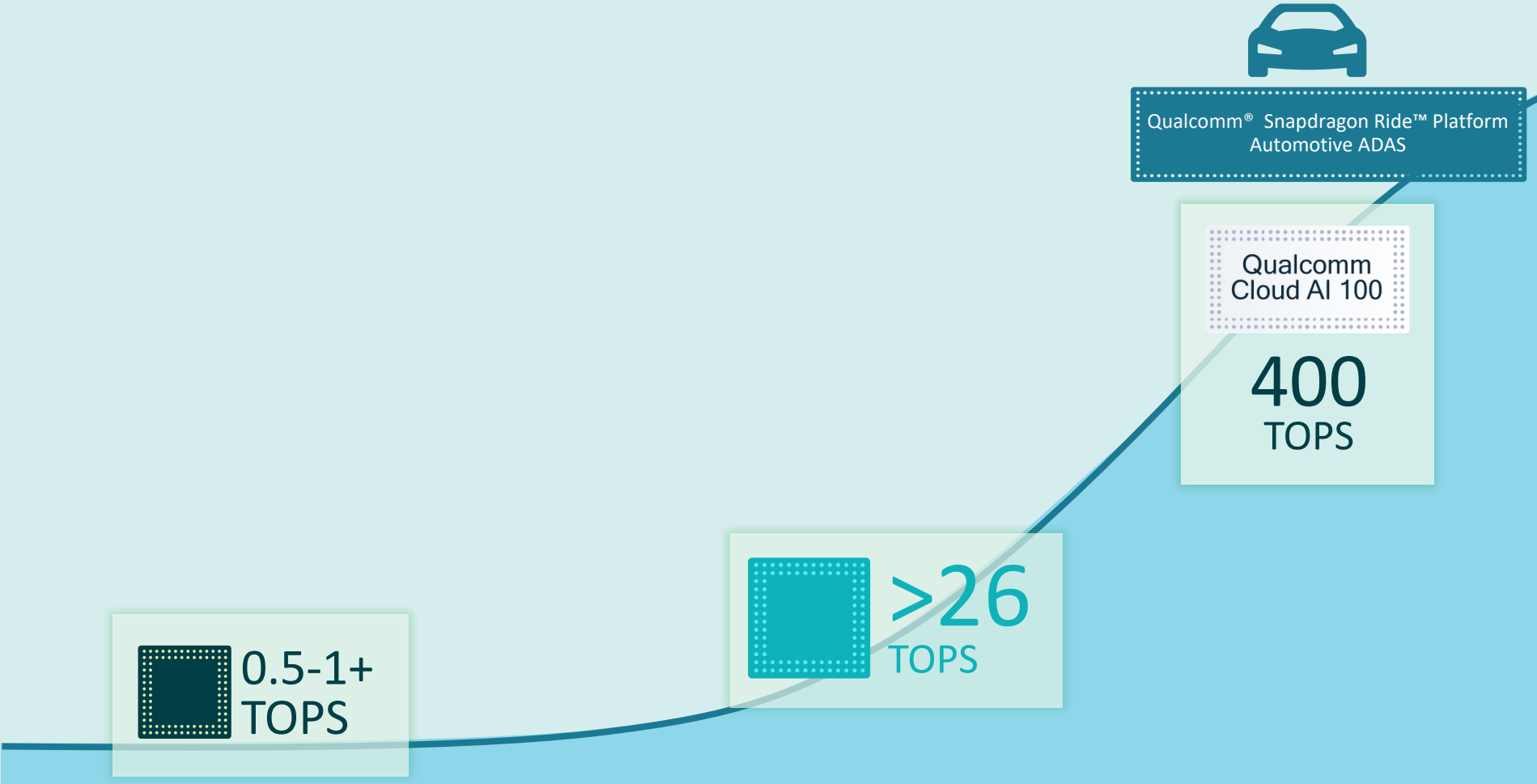

0.5-1+
TOPS
AI for IoT
Home
Industrial/Enterprise
Smart cities
Object Detection
Access Control
Semantic Filtering
Classification
Picking and Sorting
Super Resolution
<10 Concurrencies

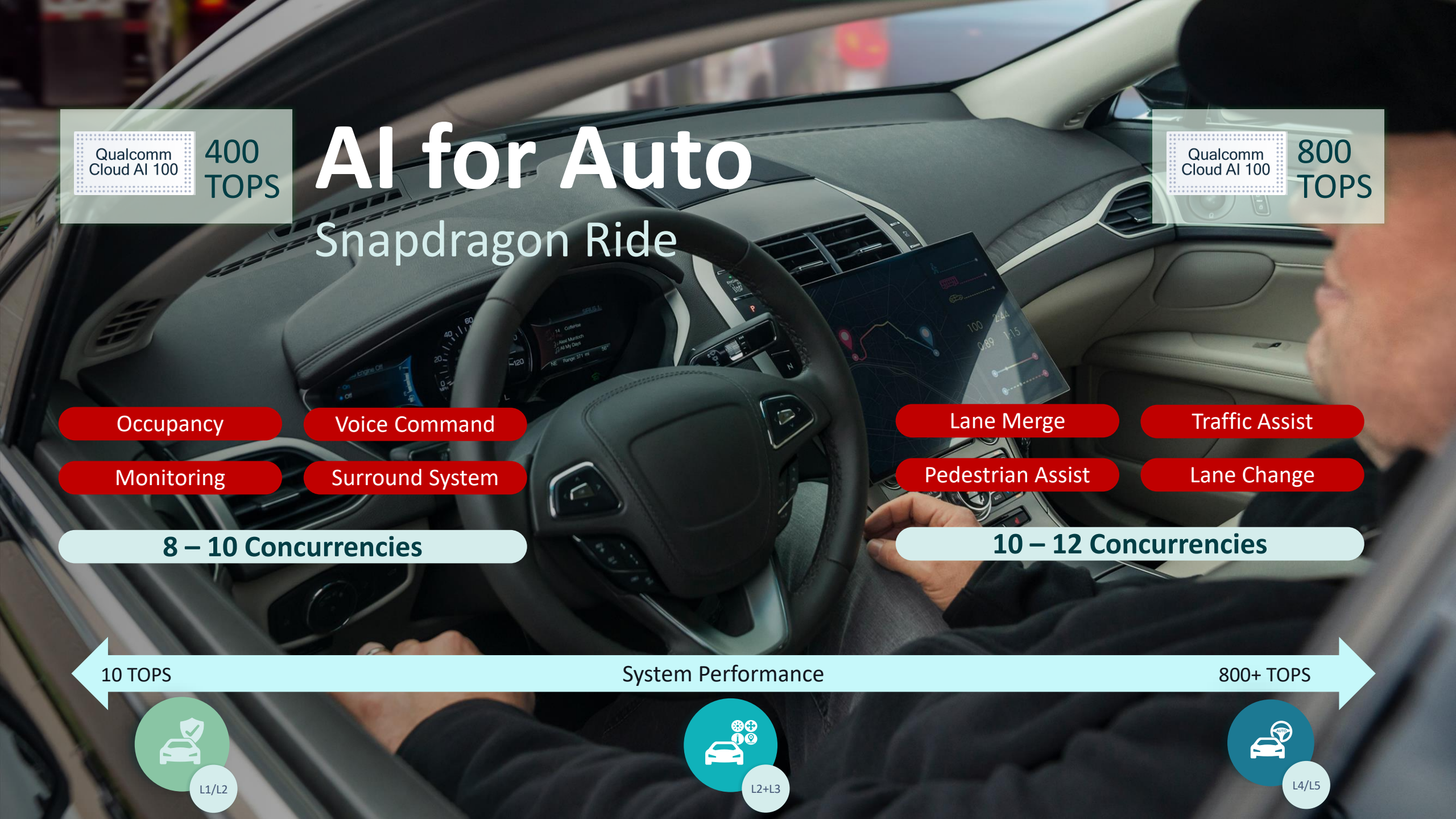
AI Performance Continuum

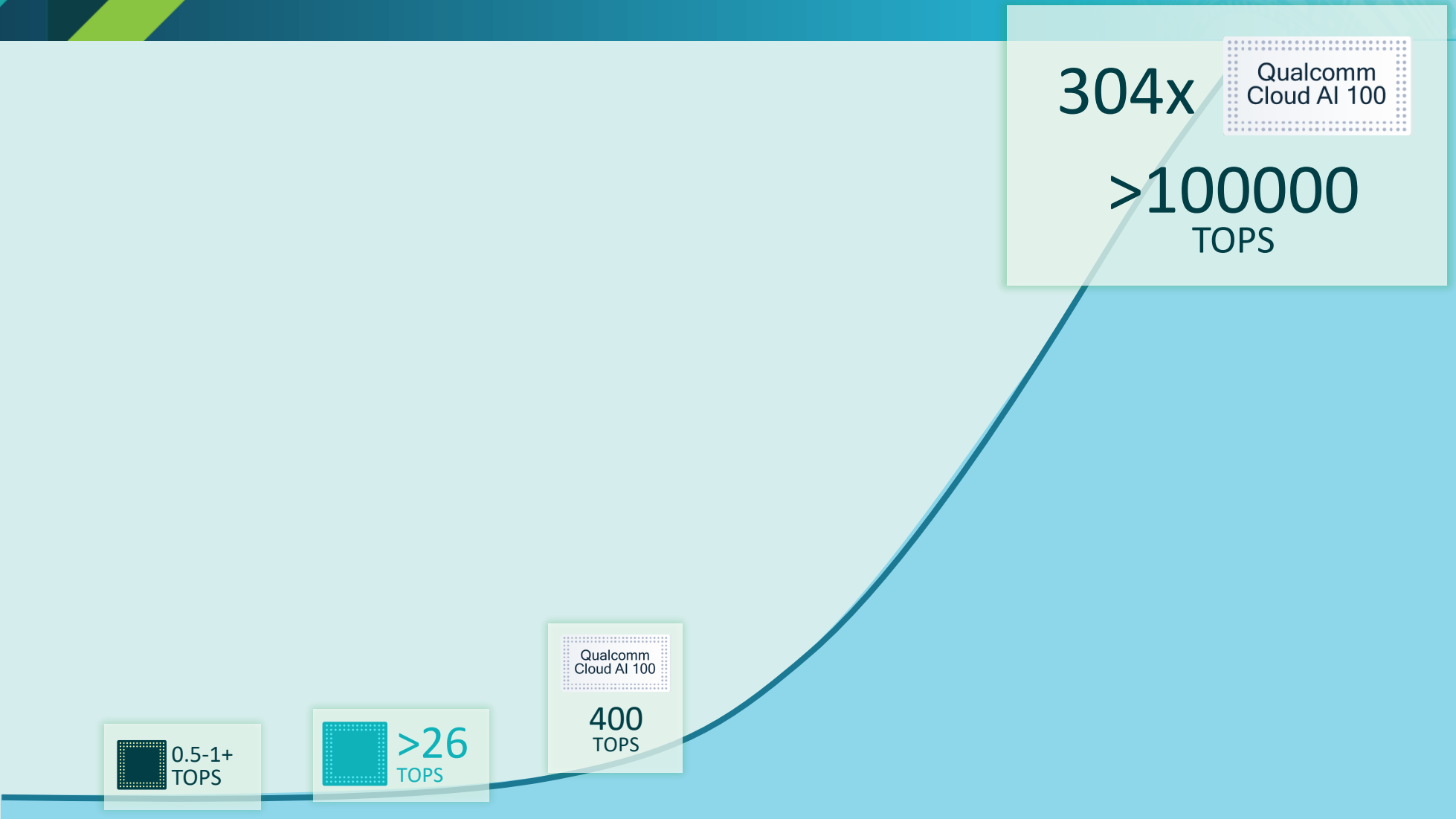



<26
TOPS
On-Device AI
Mobile
Always connected PCs
XR
Segmentation
De-noise
Semantic Filtering
Classification
Detection
Super Resolution
10 – 15 Concurrencies
Qualcomm

AI Performance Continuum




Qualcomm Cloud AI 100
400 TOPS
AI for Auto
Snapdragon Ride
Qualcomm Cloud AI 100
800 TOPS
Occupancy
Voice Command
Monitoring
Surround System
8 – 10 Concurrencies
Lane Merge
Traffic Assist
Pedestrian Assist
Lane Change
10 – 12 Concurrencies
10 TOPS
System Performance
800+ TOPS
L1/L2
L2+L3
L4/L5



Qualcomm

Qualcomm Cloud AI 100

400 TOPS

Qualcomm Cloud AI 100

Industrial/Enterprise Infrastructure

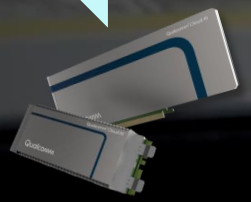
Auto

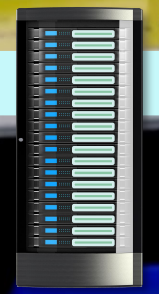
Recommendation | Linguistics | Communication

Classification | Pedestrian Assist | Intersection Assist

3 – 5 Concurrencies

400 TOPS ← Performance → 125 POPS

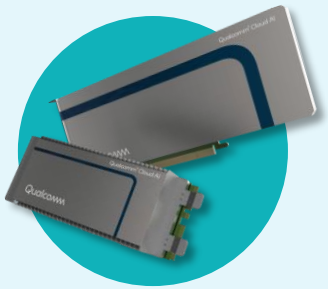




Qualcomm

Cloud AI 100

Qualcomm Cloud AI 100



400 TOPS

1 card

Gigabyte platform

6.4 POPS

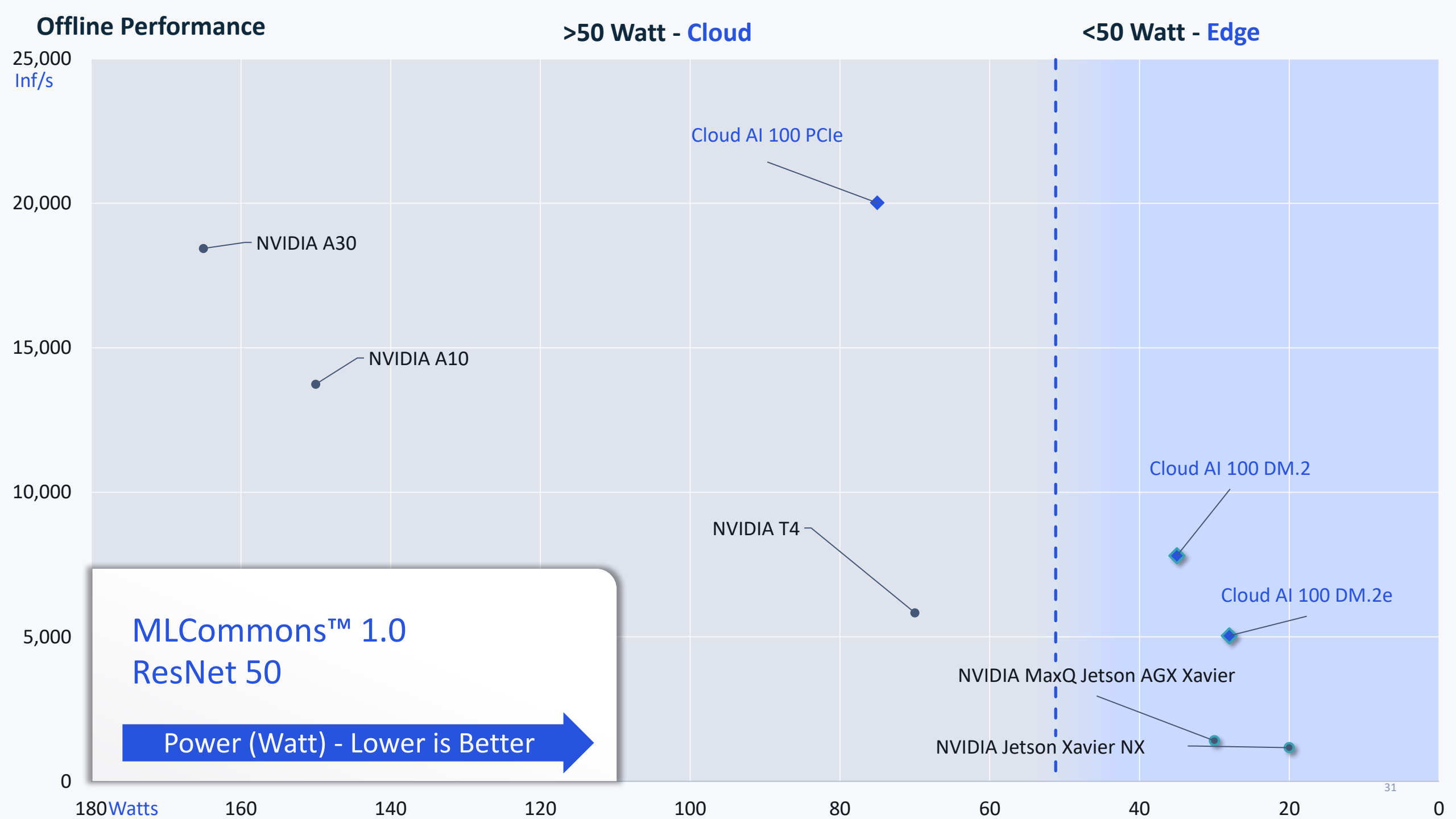
16 cards in one server

Full server rack

125+ POPS

20 server units in one server rack

1 Peta OPS = 1000 TOPS


Offline Performance
>50 Watt - Cloud
<50 Watt - Edge
25,000
Inf/s
20,000
15,000
10,000
5,000
0
Cloud AI 100 PCIe
NVIDIA A30
NVIDIA A10
NVIDIA T4
Cloud AI 100 DM.2
Cloud AI 100 DM.2e
NVIDIA MaxQ Jetson AGX Xavier
NVIDIA Jetson Xavier NX
MLCommons™ 1.0
ResNet 50
Power (Watt) - Lower is Better
180 Watts
160
140
120
100
80
60
40
20
0

The future of AI

Qualcomm

We're creating a future of distributed intelligence

Qualcomm

Thank You

Qualcomm

Resource Slide

- **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=CjwKCAjw19z6BRAYEiwAmo64LfQjU8vqH8TxqKTM2PZQp8JibXrjev85wLfKFknJnS_b494yZ7e_WhoCPQkQAvD_BwE

- **Qualcomm® Platform Solution Ecosystem:**

https://www.qualcomm.com/support/qan/platform-solutions-ecosystem

- **GitHub AI Model Efficiency Toolkit (AIMET):**

https://github.com/quic/aimet

- **Qualcomm Mobile AI page:**

https://www.qualcomm.com/products/smartphones/mobile-ai

- **Qualcomm Mobile AI blog:**

https://www.qualcomm.com/news/onq/2020/12/02/exploring-ai-capabilities-qualcomm-snapdragon-888-mobile-platform

- **Qualcomm Cloud AI 100 blog:**

https://www.qualcomm.com/news/onq/2021/03/15/qualcomm-cloud-ai-100-amd-epyc-7003-series-processor-and-gigabyte-server

Qualcomm