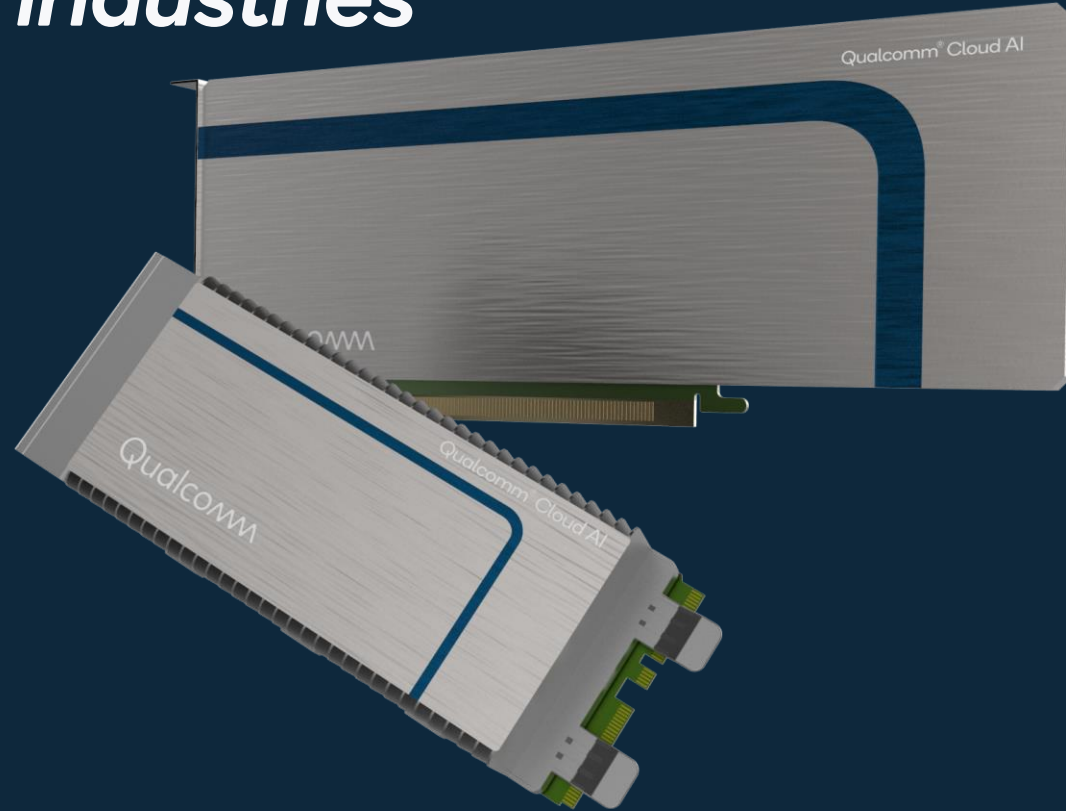


# Qualcomm<sup>®</sup> Cloud AI 100 Announcement

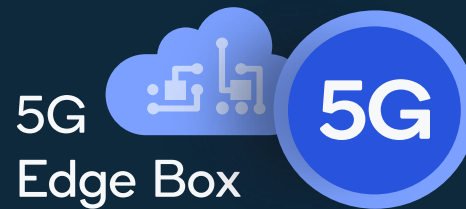
**Under Embargo until Sept 16 at 6:30am PT**

Qualcomm Cloud AI is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

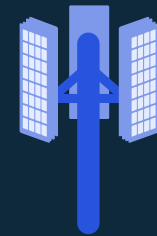
# Qualcomm Cloud AI 100 addressing edge-to-cloud industries



Data Center/  
Cloud Edge



5G  
Edge Box



5G  
Infrastructure



# Datacenter



Personalized  
Purchase recommendations



Personalized  
Purchase advertisements



# Edge Box

Pedestrian alert  
Crossing & blind spot assist

Road safety  
Intersection management  
assist



# 5G Infra



*Powering the shopping  
of the future*

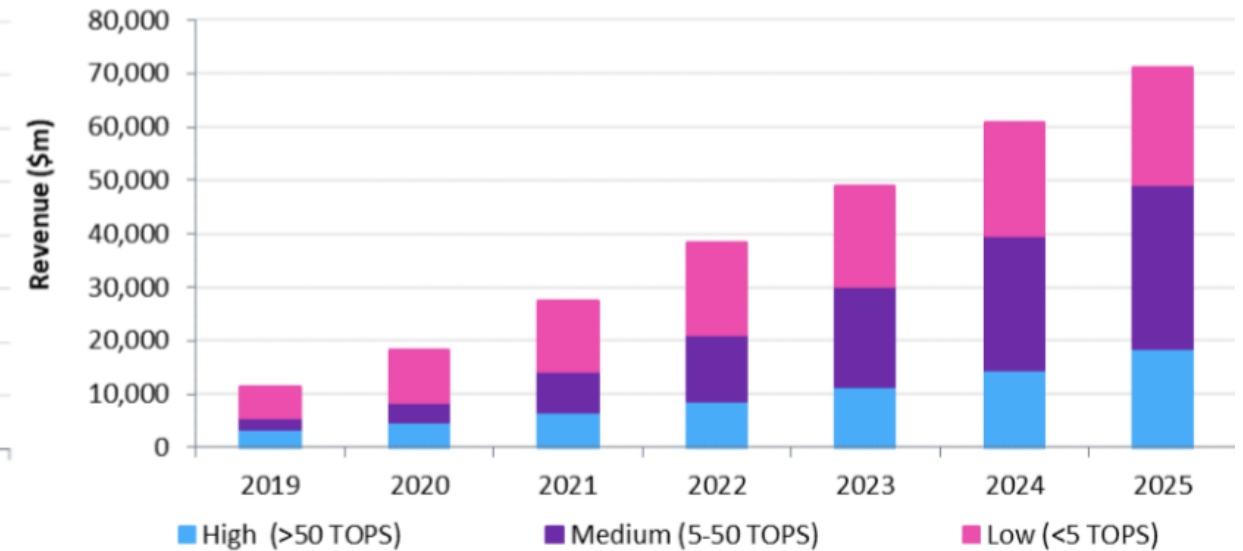
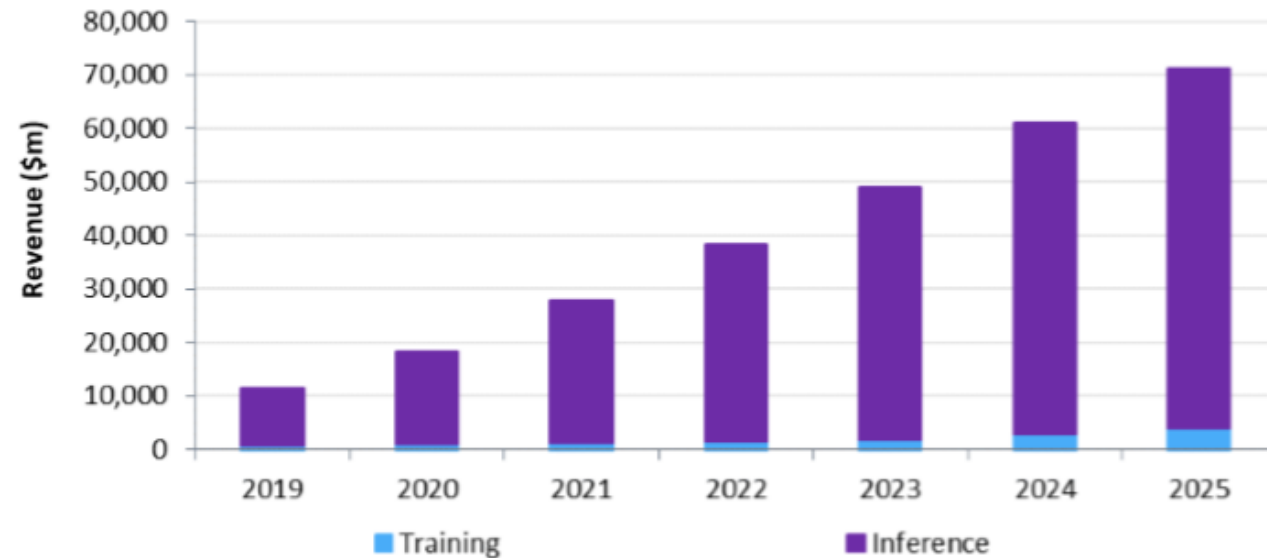
*Pioneering  
safety*

*Reinventing the  
communication  
experience*

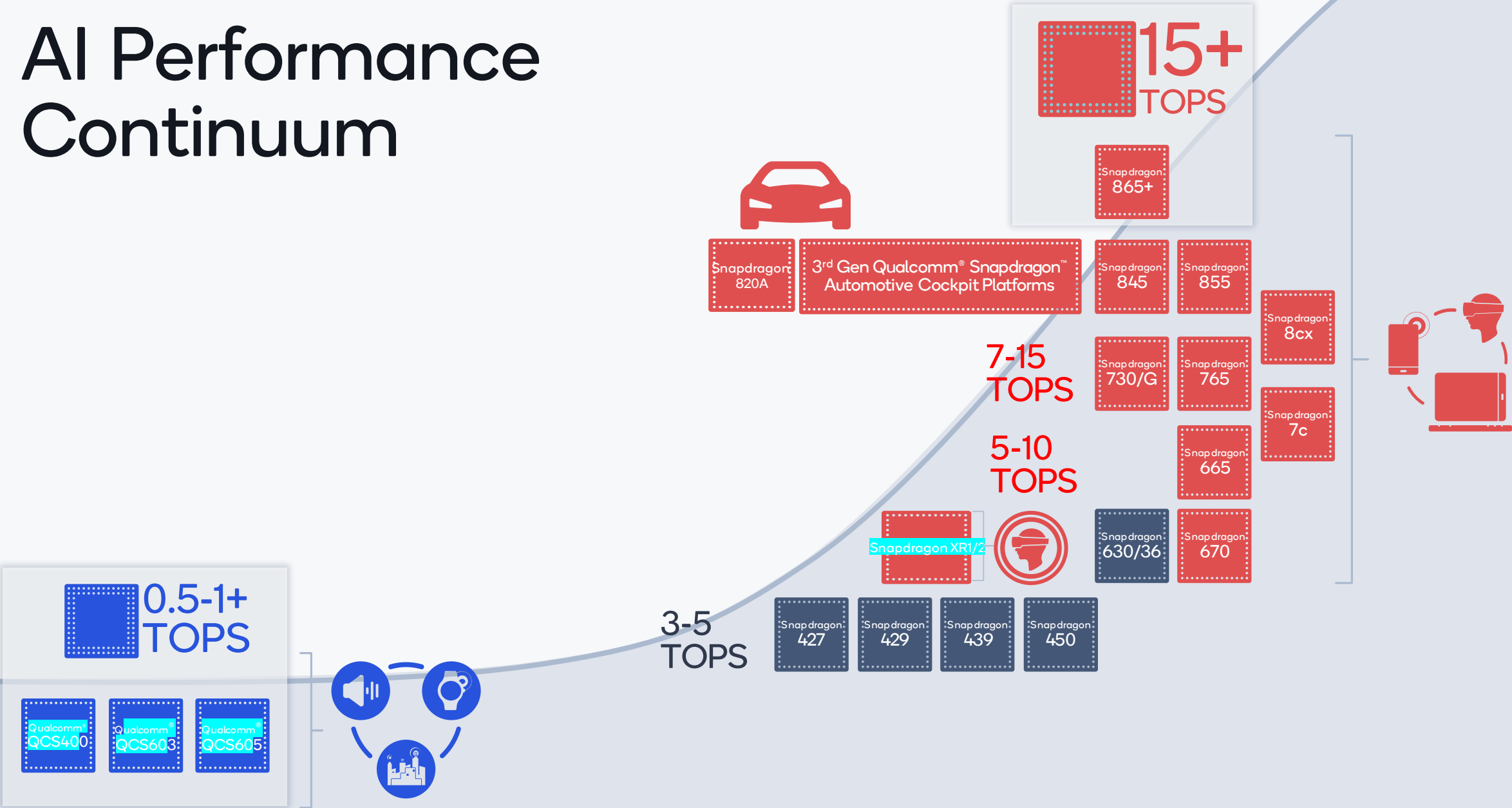
# A reminder why we are doing this...

COMPUTING IS MOVING TOWARD THE EDGE; SIGNIFICANT OPPORTUNITY FOR ENTRY

Explosive growth in Inference; High and Medium performance addressed well by Qualcomm Cloud AI 100




# AI Performance Continuum





# Full Production Cycle


Final Silicon hitting peak TOPS and Performance per watt



0.5-1+  
TOPS



>15  
TOPS



Qualcomm  
Cloud AI 100

>50  
TOPS

DM.2e



Qualcomm  
Cloud AI 100

200  
TOPS

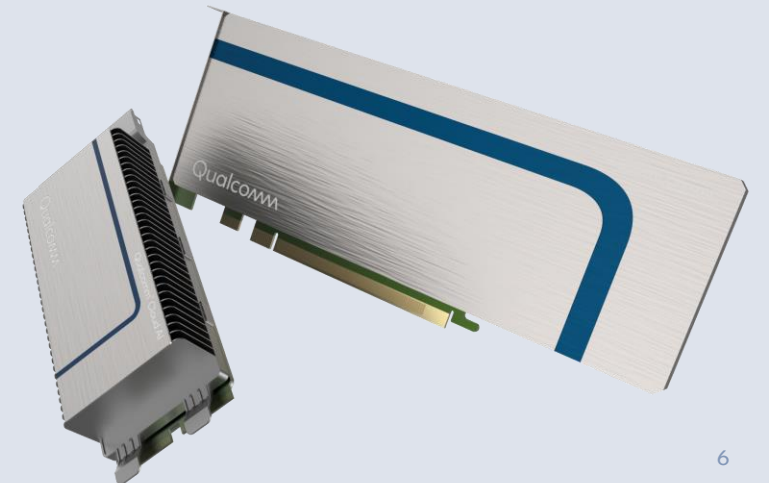
DM.2



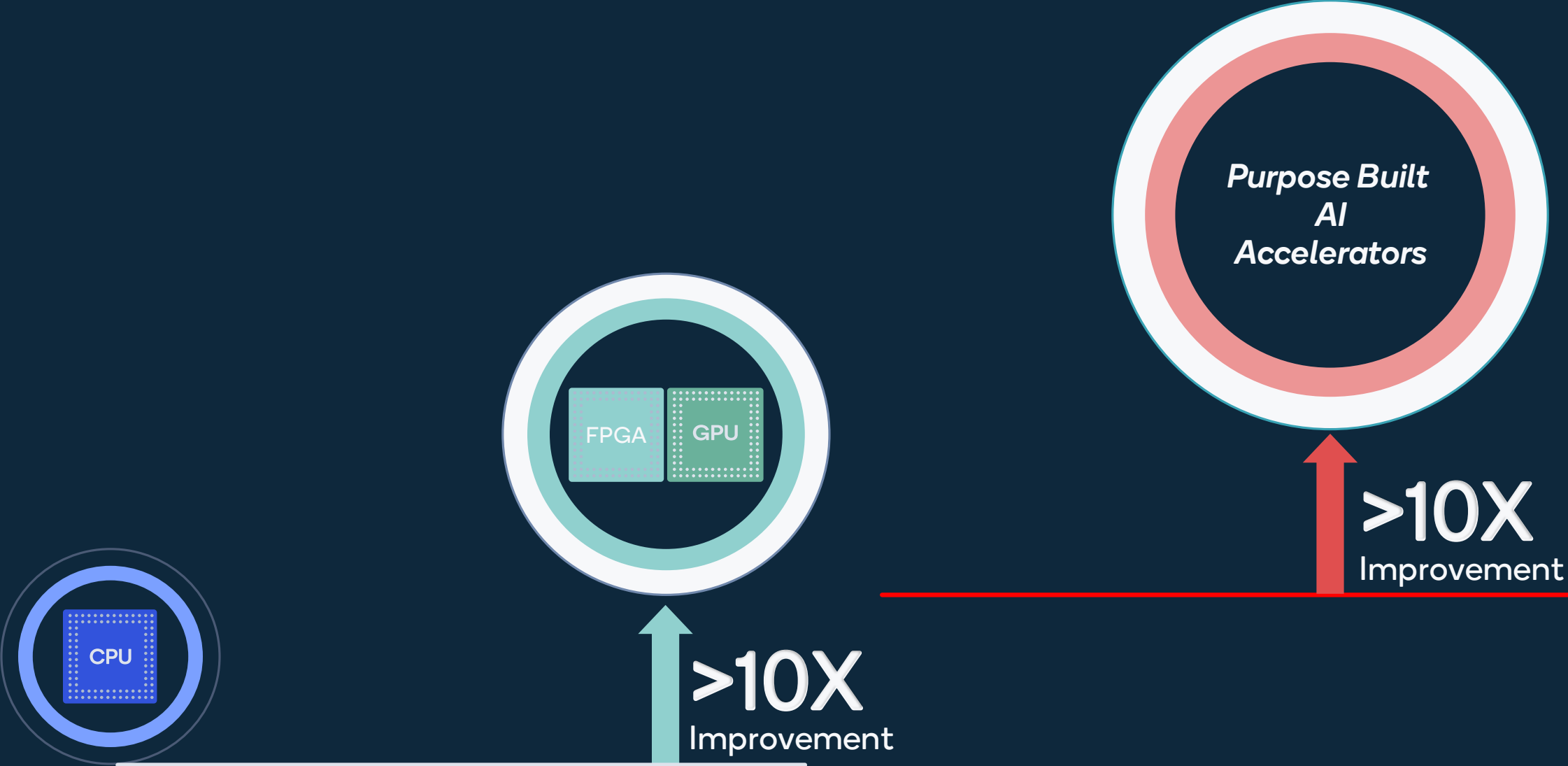
Qualcomm  
Cloud AI 100

400  
TOPS

PCIe

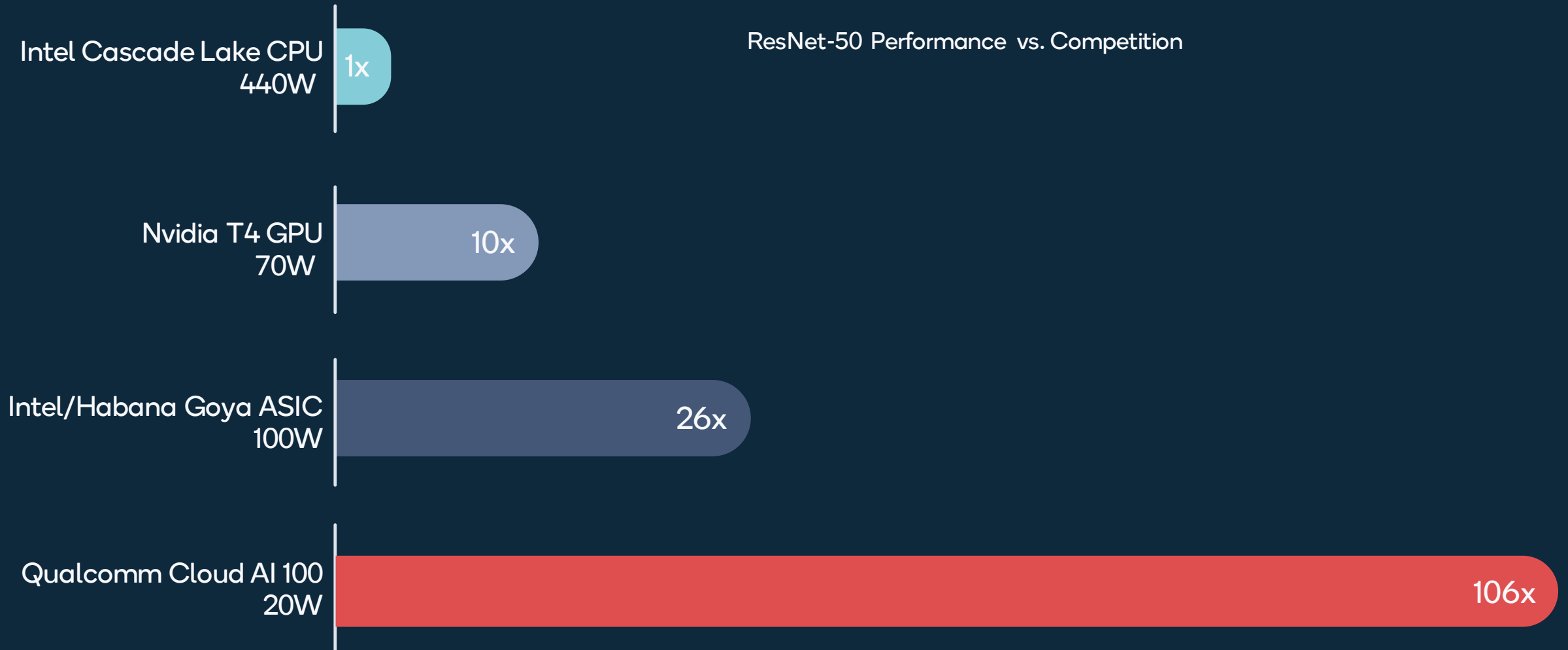


# An architecture shift in AI cloud inferencing



# Industry Leading Inference Performance

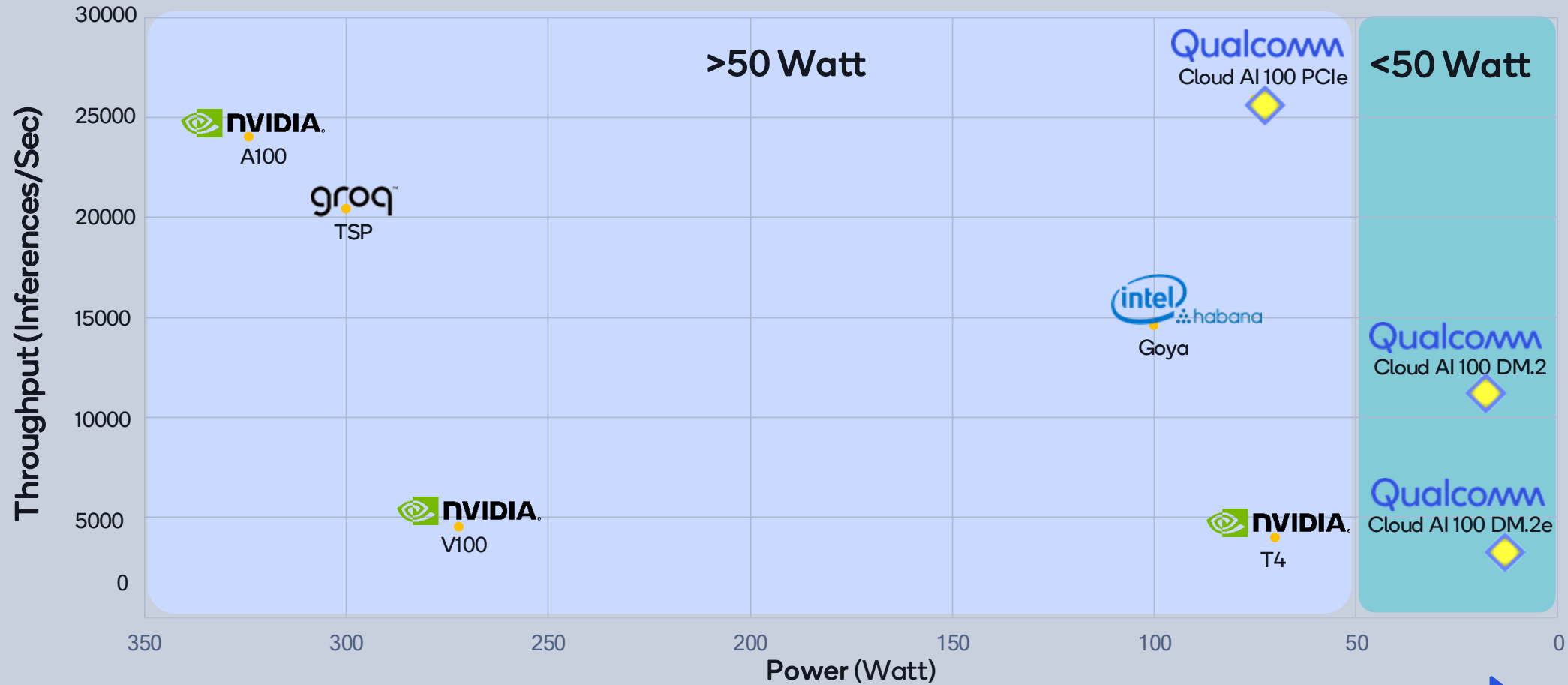
Qualcomm Cloud AI 100 outperforms Competition by 4~10x (Inference/Sec/W)





# Performance & Efficiency Leadership

ResNet-50 Benchmark



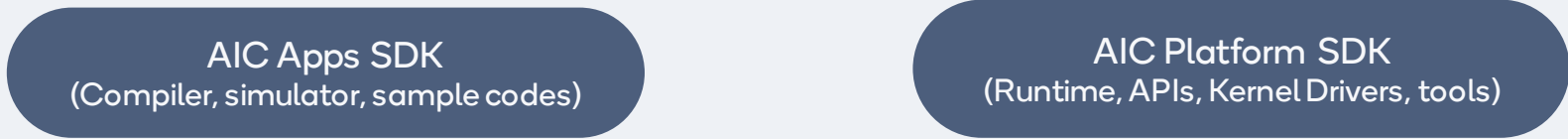
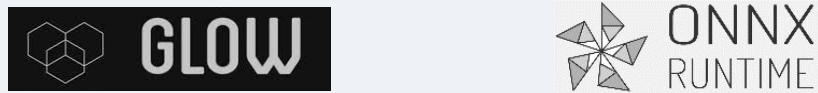
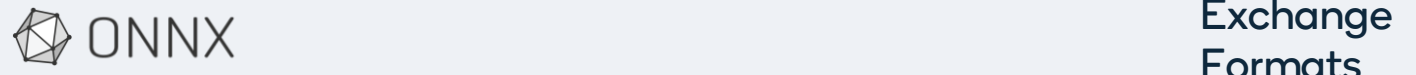
**Power (Watt) - Lower is Better**

Source: Habana and Nvidia websites, Linley report for . Batch size = 8 for Nvidia T4, V100, Habana Goya, Groq TSP and Cloud AI100. Batch size unknown for Nvidia A100



## 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



Applications

Models

Frameworks

Exchange Formats

Runtimes

SDKs

Hardware

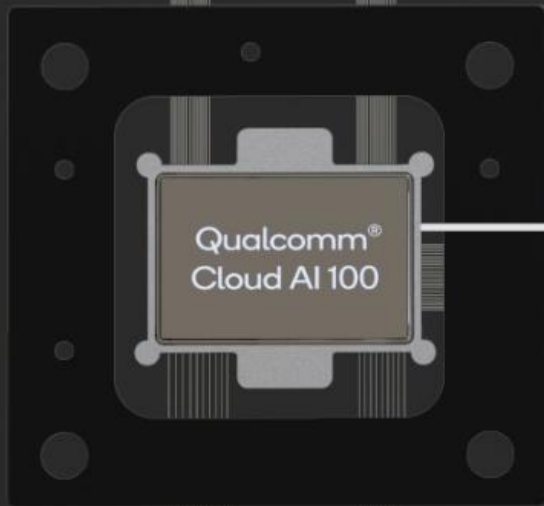
Introducing...  
The Qualcomm Cloud AI 100  
Edge Development Kit



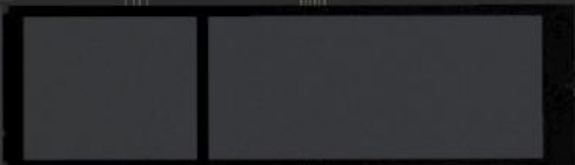
**Snapdragon 865 Modular Platform**  
Application & Video Processing



**Snapdragon X55 Modem-RF System**  
Industry Leading 5G Connectivity



**Cloud AI 100**  
Advanced Low-Power and  
High Performance AI





## Qualcomm Cloud AI 100 Summary

### Full Production Cycle/Final Silicon

- Sampling now to multiple customers
- Shipping 1<sup>st</sup> Half 2021

### Edge Development Kit

- Sampling October 2020
- Single Solution: one-stop-shop with AI inferencing, host processor and 5G connectivity
- Pre-certified modem module with industry leading 5G
- Cutting-edge performance per Watt



# Product Brief

## Qualcomm® Cloud AI 100





Schedule	Sampling Now Commercial launch 1H 21
Process Node	7nm
Card Types	Dual M.2 (edge): 15W TDP Dual M.2: 25W TDP PCIe (HHHL): 75W TDP
Card Performance (Raw TOPS)	Dual M.2 (edge): 70 TOPS Dual M.2: 200 TOPS PCIe: 400 TOPS
AI Cores	Up to 16 cores
On Die SRAM	144MB (9MB Each AI Core)
On Card DRAM	Up to 32GB w/ 4x64 LPDDR4x @ 2.1GHz
PCIe (connection to the host)	8 lane Gen3/4 (PCIe) or 4 lane Gen3/4 (Dual M.2)
Data Types	INT8, INT16, FP16, FP32

## Qualcomm® Cloud AI 100 Edge Development Kit

Schedule	Shipping October 2020
Targeted Application	5G Intelligent Edge Compute Appliance
Dimensions (mm)	231 (w) x 250 (h) x 84 (d) without antenna and stand 231 (w) x 504 (h) x 120 (d) with antenna and stand
Operating Temp Range (Ambient)	0C – 50C
Host SOC	Qualcomm® Snapdragon™ 865 Mobile Platform Qualcomm® Kryo™ 585 CPU cores up to 2.84GHz Integrated video pipeline to process 24 streams of FHD @ 30fps simultaneously 12GB LPDDR5 memory
Operating System	CentOS 8.0
AI Accelerator	Qualcomm Cloud AI 100 (DM.2 edge variant)
Connectivity	Snapdragon X55 5G modem on M.2 module (global sub6 variant) Ethernet (LAN and WAN)
Storage	NVMe SSD



# 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 is a trademark 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.