Qualcomm® Cloud AI 100

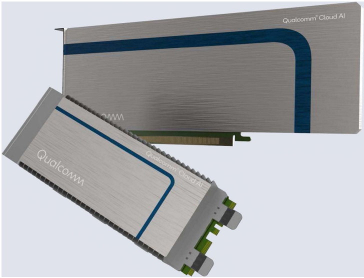
SoC Product Manual Review Presentation

Yihua Liu

UM-SJTU Joint Institute

December 9, 2021

Product Overview



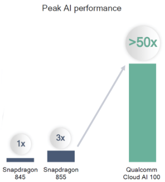
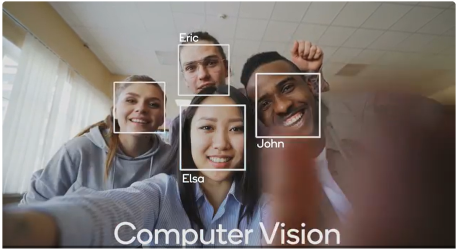


Figure. Peak AI performance [2].

- Research starts since 2016
- Qualcomm's most advanced low-power and high performance AI processing
- Powerful and efficient processing speeds: More than 10x performance per watt over the industry's most advanced AI inference solutions deployed today
- Specifically designed for processing AI inference workloads

Application Overview



Application support: Cloud AI

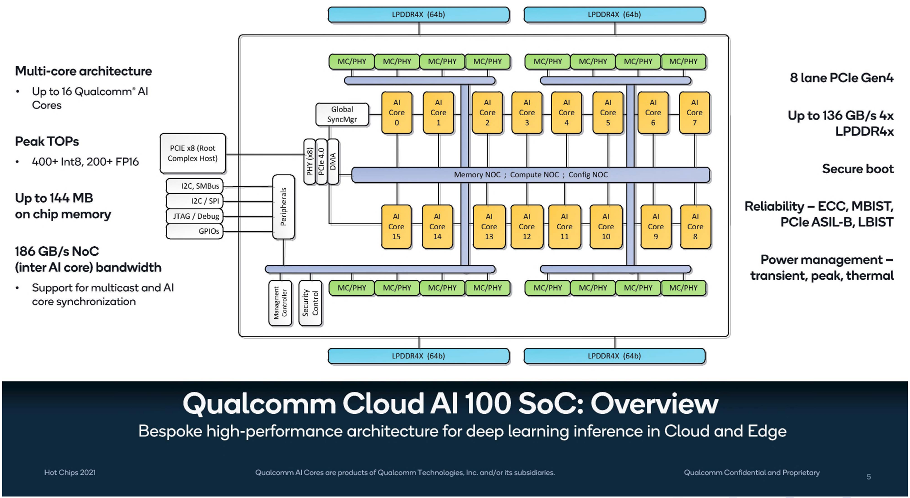
- Industry-leading 5G connectivity by Qualcomm Snapdragon X55 Modem-RF System
- Application and video processing on Qualcomm Snapdragon 865 Modular Platform
- Development kit supports leading software stacks including Pytorch, Glow, Tensorflow, Keras, and ONNX [3]

Application targets:

- Natural Language Processing
- eXtended Reality
- Translations
- Computer Vision

Product Architecture

High performance, low latency, low power, datacenter to edge



Multi-core architecture

- Up to 16 Qualcomm® AI Cores

Peak TOPs

- 400+ Int8, 200+ FP16

Up to 144 MB on chip memory

186 GB/s NoC (inter AI core) bandwidth

- Support for multicast and AI core synchronization

8 lane PCIe Gen4

Up to 136 GB/s 4x LPDDR4x

Secure boot

Reliability – ECC, MBIST, PCIe ASIL-B, LBIST

Power management – transient, peak, thermal

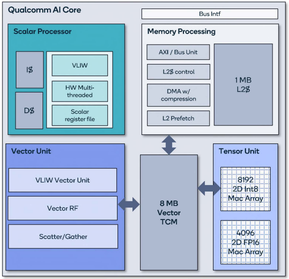
Qualcomm Cloud AI 100 SoC: Overview

Bespoke high-performance architecture for deep learning inference in Cloud and Edge

Hot Chips 2021 Qualcomm AI Cores are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm Confidential and Proprietary 5

Main Features

Qualcomm AI Core



- Scalar - VLIW architecture
- Vector tightly couple memory (VTCM)
- Vector unit
- Tensor unit

SoC Power	12.05 W	19.74 W	69.26W
TOPs	149.01	196.64	363.02
SoC TOPs/W	12.37	9.98	5.24

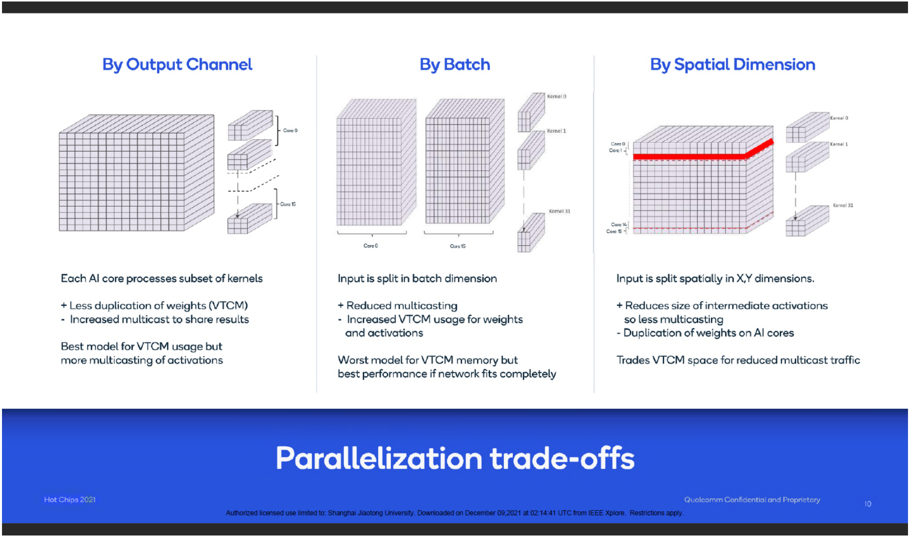
Table. Performance and power measured [1].

Specifications

- Data Types: FP16, INT8, INT16, FP32
- On Die SRAM: 144MB (9MB Each AI Core)
- AI Cores: Up to 16
- Process Node and Technology: 7 nm
- Card: Dual M.2 (edge): 70 TOPS15W TDP, Dual M.2: 200 TOPS 25W TDP, PCIe: 400 TOPS 75W TDP
  On Card DRAM: Up to 32GB w/ 4x64 LPDDR4x 2.1GHz [4]

Major Uniqueness

Parallelization trade-offs



By Output Channel

Each AI core processes subset of kernels

+ Less duplication of weights (VTCM)
- Increased multicast to share results

Best model for VTCM usage but more multicasting of activations

By Batch

Input is split in batch dimension

+ Reduced multicasting
- Increased VTCM usage for weights and activations

Worst model for VTCM memory but best performance if network fits completely

By Spatial Dimension

Input is split spatially in X,Y dimensions.

+ Reduces size of intermediate activations so less multicasting
- Duplication of weights on AI cores

Trades VTCM space for reduced multicast traffic

Parallelization trade-offs

Reference

[1] Karam Chatha. "Qualcomm® Cloud AI 100: 12TOPS/W Scalable, High Performance and Low Latency Deep Learning Inference Accelerator". In: *2021 IEEE Hot Chips 33 Symposium (HCS)*. IEEE. 2021, pp. 1–19.

[2] Dylan McGrath. "Qualcomm Targets AI Inferencing in the Cloud". In: *EE Times* (Apr. 10, 2019). URL: https://www.eetimes.com/qualcomm-targets-ai-inferencing-in-the-cloud/.

[3] Qualcomm. *CLOUD AI. The future of AI edge-to-cloud computing is here*. 2021. URL: https://www.qualcomm.com/products/cloud-artificial-intelligence.

[4] Qualcomm. *CLOUD AI 100*. 2021. URL: https://www.qualcomm.com/products/cloud-ai-100.

Thanks!