

August 2021



Qualcomm[®] Cloud AI 100

12 TOPS/W Scalable, High Performance and
Low Latency Deep Learning Inference Accelerator

Karam Chatha

Senior Director, Engineering
Qualcomm Technologies, Inc.

Future of AI in Data Center Demands Breakthrough Technology

Compute power not keeping up with business needs to deliver best in class services

Social

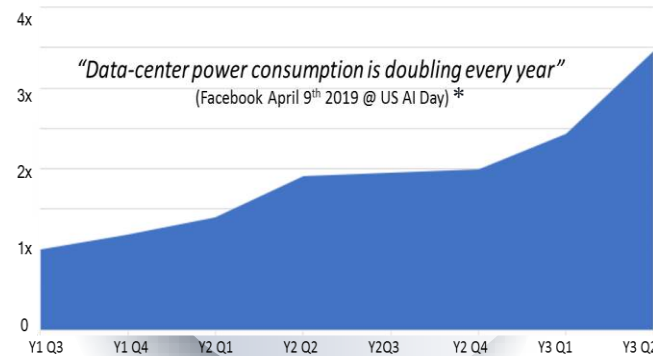
- Social media graphs
- Language translation
- Content identification
- AR/VR content generation
- Bots/assistants

Commerce

- Recommendation systems
- Business intelligence & insights
- Fraud detection
- Retail loss prevention
- Conversational interfaces
- HPC

Industrial

- Manufacturing defect and loss detection
- Factory safety
- Health and sciences
- Public safety



AI Ubiquitous in Data Center

- AI fundamental for next gen business analytics for best customer experience and insights
- Velocity of insight key to business leadership

Infrastructure Under Pressure

- Staying ahead of the AI curve is increasing demand on infrastructure cost and power

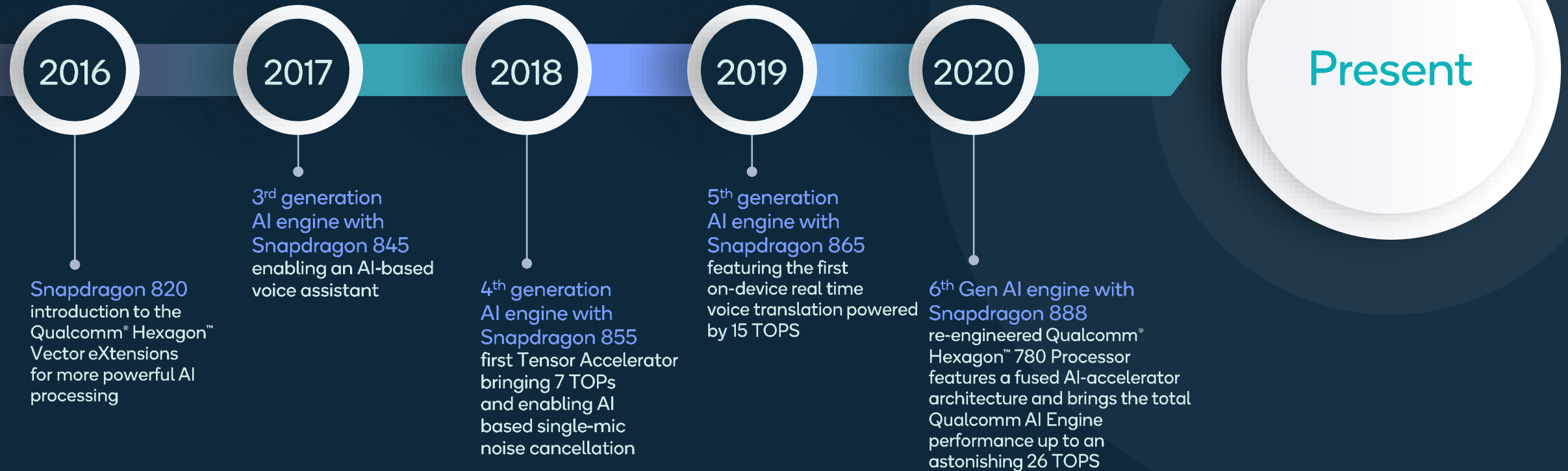
Heterogeneous Approach

- Energy-efficient inference acceleration enables AI to scale beyond constraints of general-purpose servers
- AI computation on optimized hardware reduces data center Opex and frees up server resources to drive business value application software

Need for high performance and power efficient inference accelerator in Data Centers

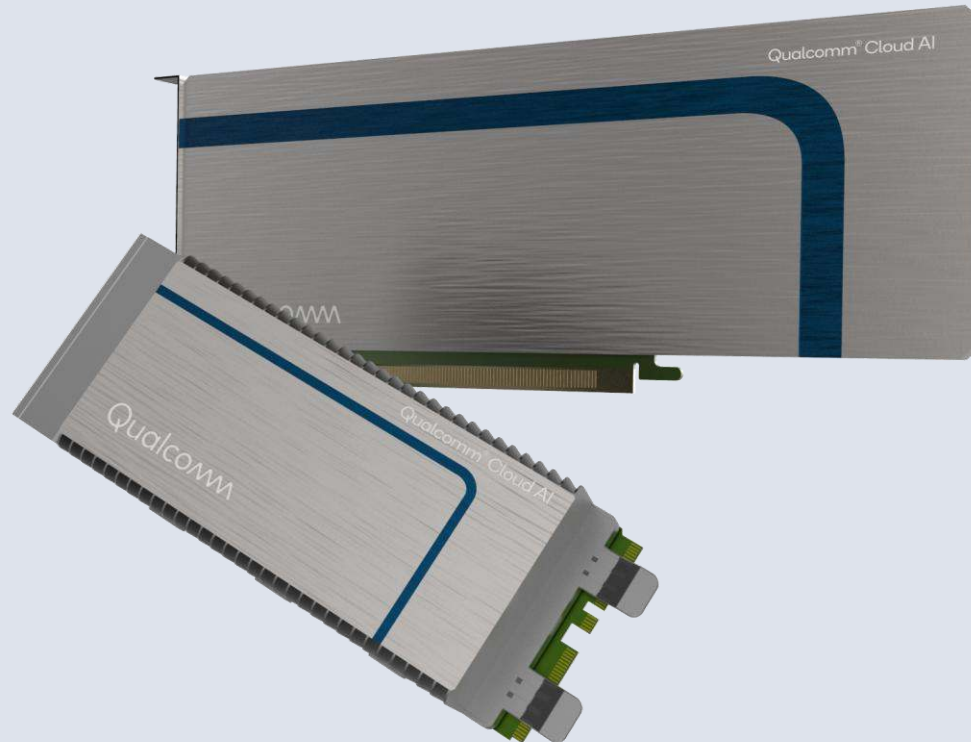
13 Years of AI research

6 Generations of Inference Engines



Qualcomm Cloud AI 100 scalable across Cloud to Edge

High performance, low power architecture for Datacenter to Edge



Utilizes over a decade of research and development delivering high-performance, low power deep learning inference acceleration technology

Focus on ML inference across Cloud and Edge applications

Collaborating with industry leaders for first-generation success

Architecture for scalable technology across generations

Multi-core architecture

- Up to 16 Qualcomm® AI Cores

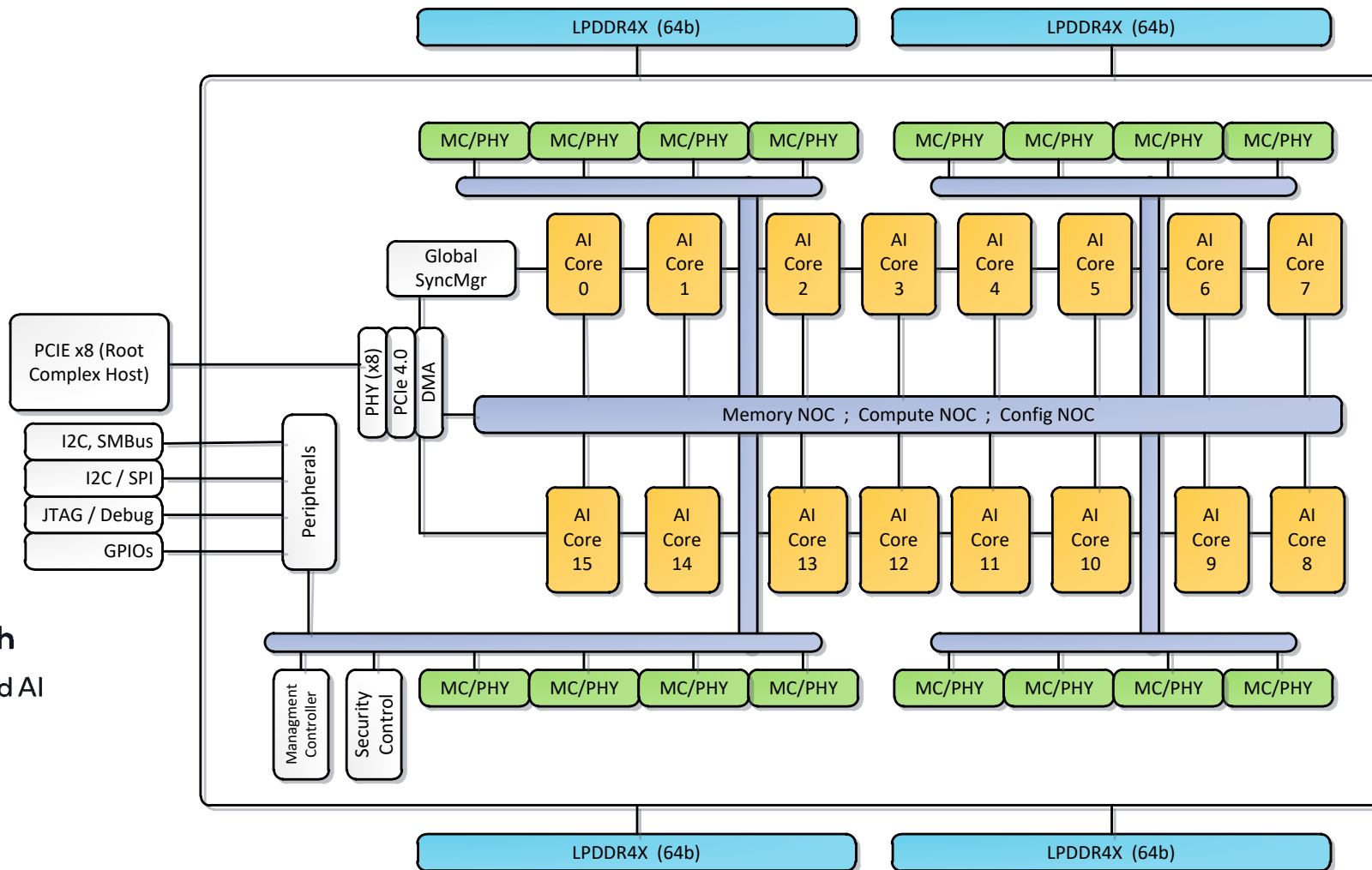
Peak TOPs

- 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

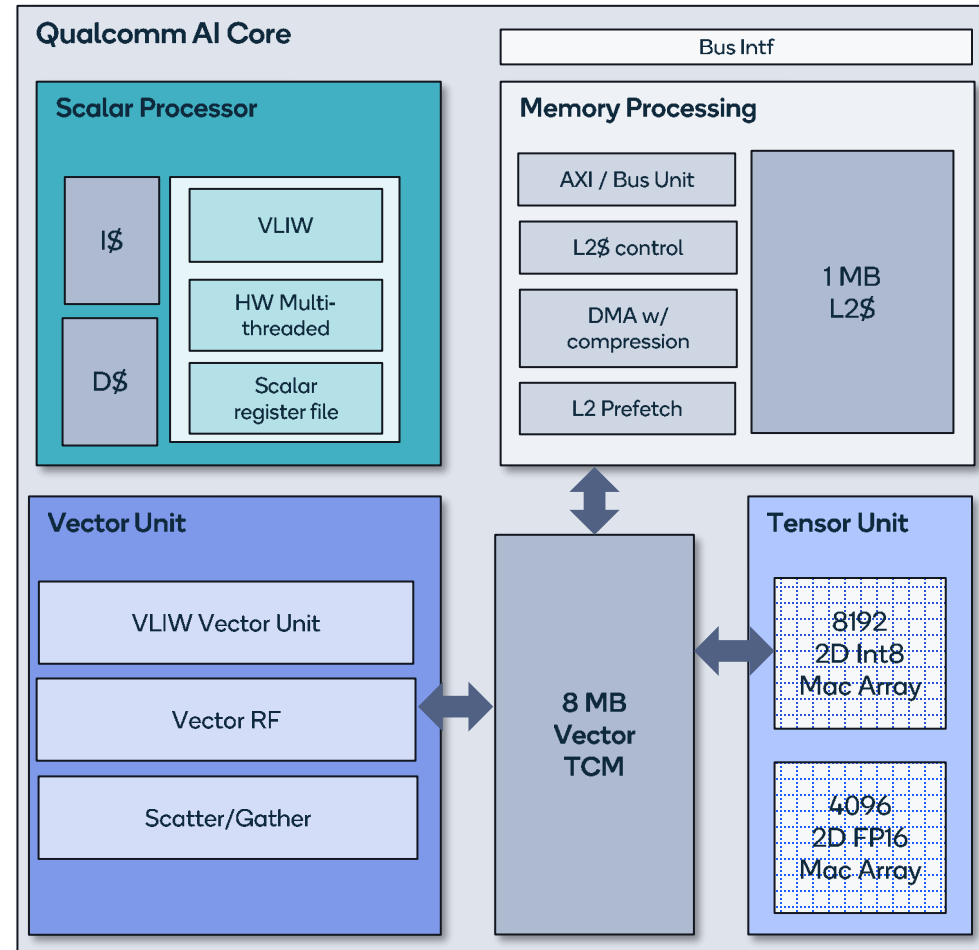
Bespoke high-performance architecture for deep learning inference in Cloud and Edge

Scalar - VLIW architecture

- 4 Way VLIW
- Rich instruction set (over 1800+ instructions)
- Multi-threaded scalar core
- 1MB L2 cache
- Precision - FP32, FP16, Int16, Int8

Vector tightly couple memory (VTCM)

- 8MB memory that reduces DDR spillage
- Accessible by scalar, vector and tensor units



Vector unit

- Rich instruction set for AI, CV and image processing (over 700+ instructions)
- Precision - FP32, FP16, Int16, Int8
- 512 Int8 MAC/clock cycle
- 256 FP16 MAC/clock cycle

Tensor unit

- High performance and low power accelerator for linear algebra (125+ instructions)
- Precision - FP16, Int8
- 8192 Int8 MAC/clock cycle
- 4096 FP16 MAC/clock cycle

Qualcomm AI Core

Low power and high-performance deep learning inference

Qualcomm Cloud AI 100 SoC

Power efficiency

SoC Power	12.05 W	19.74 W	69.26 W
TOPs	149.01	196.94	363.02
SoC TOPs/W	12.37	9.98	5.24

Performance and power measured for typical 3x3 convolution operator (Int8) found in deep convolution neural networks (DCNN). Input activation assumes 50% zeroes which is typical for DCNN with Relu operators. Weights are uniformly distributed.

Industry leading TOPS/W for deep learning inference

SoC architecture specialized for AI inference

- Multi-core architecture with up to 16 AI cores
- Software managed 144 MB of on-chip memory
 - Reduces DDR BW and power
 - Enables on-chip storage of entire weights for many networks
- High BW on-chip NOC (with multicast support)
 - Enables splitting network operators across cores
 - Network activations are shared via multicast

Utilize industry leading low power IP from mobile

- 6th generation AI core (DSP + AI acceleration)
- Tensor unit is 5X more power efficient than a vector unit
- 8 MB memory per core maximizes data re-use and lowers power

Advanced technology node – 7nm

Compiler (software) defined multi-core schedule optimizing power and performance



Computer Vision



Speech



Autonomous



Language Translation



Recommendation System

Applications

ResNet

SSD

GNMT

BERT

DLRM

DIN

Other Models

Models



PyTorch

Caffe



PaddlePaddle

Frameworks



ONNX

Exchange Formats



ONNX
RUNTIME

Runtimes

AIC Apps SDK
(Compiler, simulator, sample codes)

AIC Platform SDK
(Runtime, APIs, Kernel Drivers, tools)

SDKs

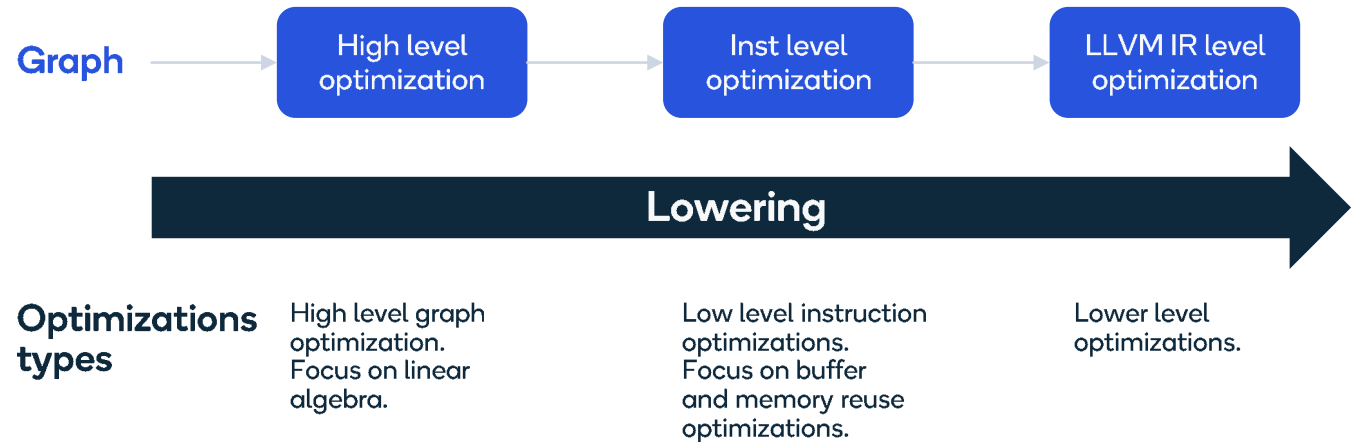
AI Inference Accelerator Cards

Hardware



Qualcomm Cloud AI 100 Parallelizing Compiler

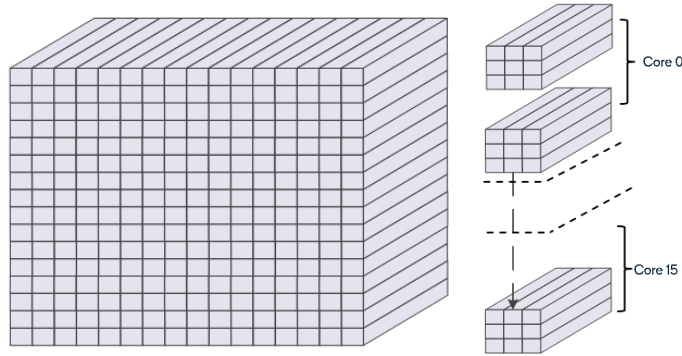
Key component for delivering power efficient performance on multi-core architecture



Translates high-level description of neural network into machine code

- Parallelizes inference computation across the multi-core SoC
- Splits data across the AI cores and synchronizes computation
- Parallelizes across tensor/vector/scalar operations within an AI core
- Supports Int8, FP16 and mixed precision operations
- Optimizes KPI – inf/s, latency, power
- Performs optimizations in 3 phases

By Output Channel

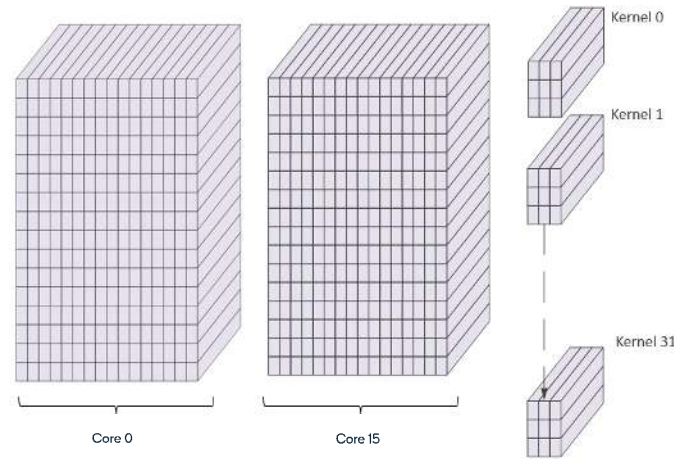


Each AI core processes subset of kernels

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

Best model for VTCM usage but more multicasting of activations

By Batch

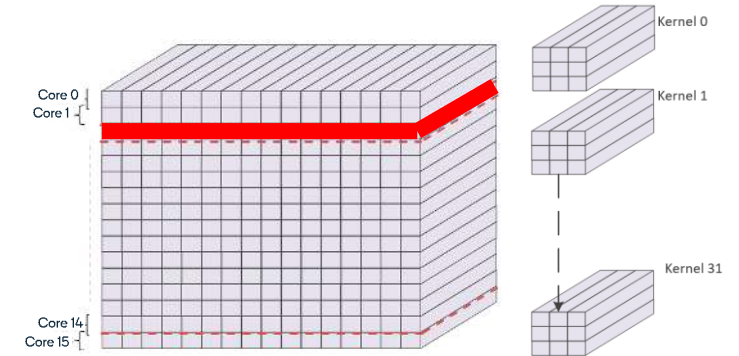


Input is split in batch dimension

- + 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.

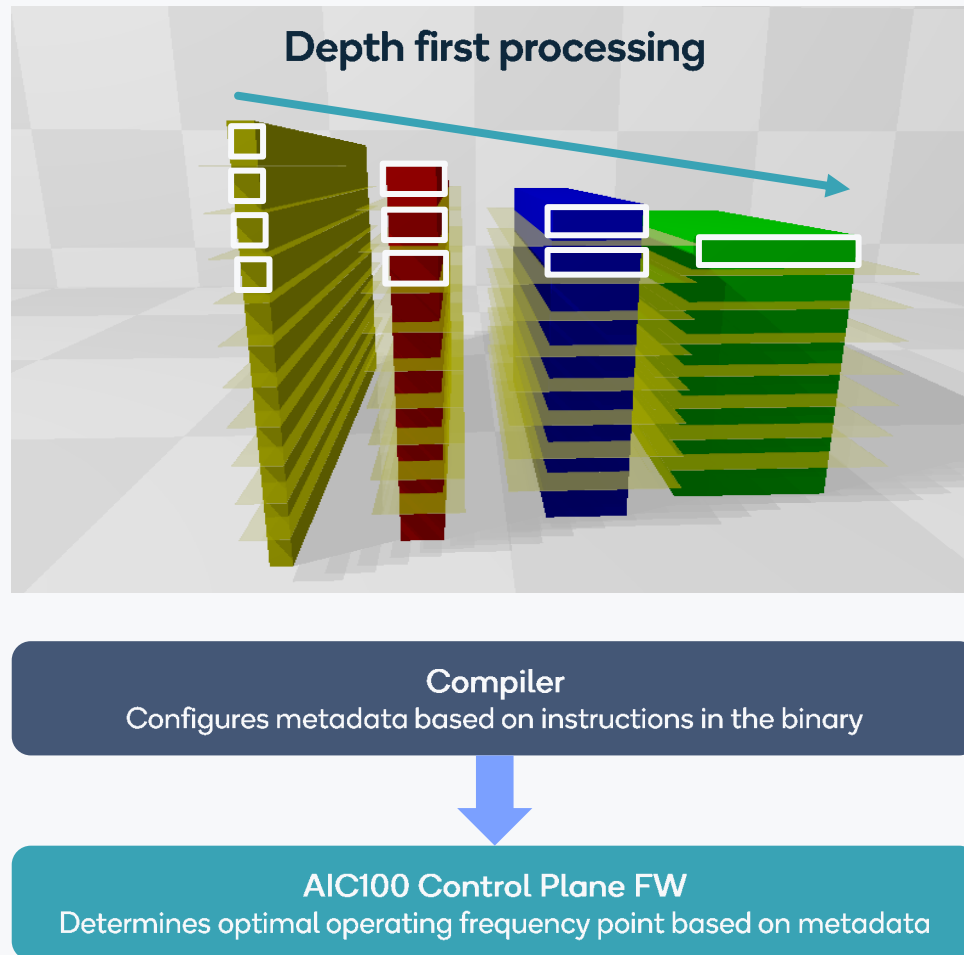
- + Reduces size of intermediate activations so less multicasting
- Duplication of weights on AI cores

Trades VTCM space for reduced multicast traffic

Parallelization trade-offs

Low power optimizations

Compiler driven optimizations for improving power efficiency



Depth first scheduling

- Minimizes spillage to DDR by processing the network graph in a depth first manner
- Reduces DDR power consumption, and improves performance
- Particularly effective for large resolution images
- RN34-SSD - 5.3X lower DDR BW, 3.5X higher inf/sec, 2X higher inf/sec/W

Scheduling to minimize peak power excursions

- Schedules the network graph across ML cores such that computation is not correlated

Instruction usage-based power management

- Compiler generates metadata for power management controller

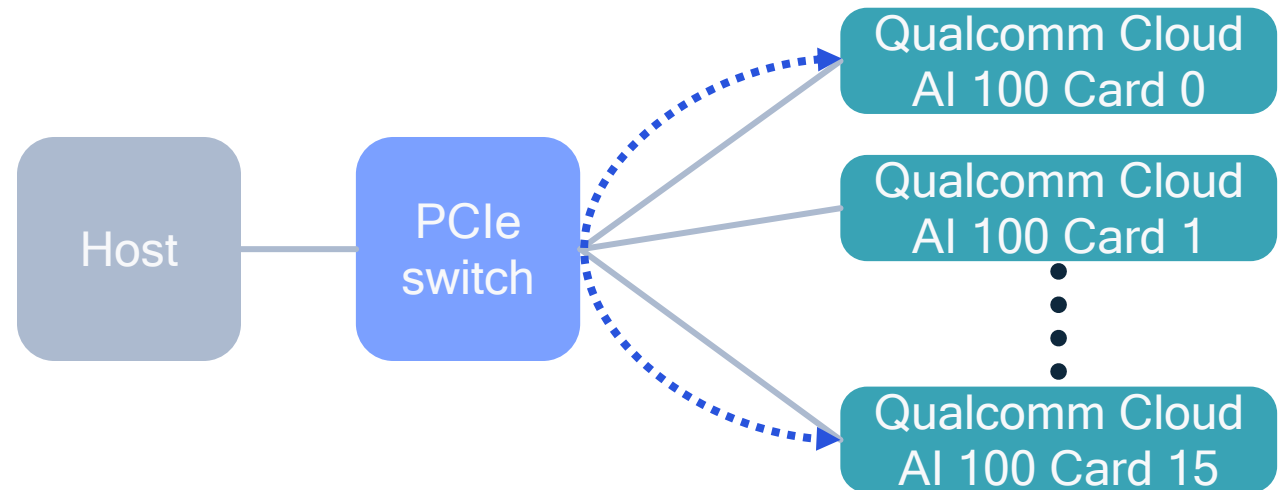
Qualcomm Cloud AI 100 peer-to-peer communication

Partition network across multiple Qualcomm Cloud AI 100 cards

- Performance boost when network fits entirely in TCM across multiple cards
- Support for networks that might exceed card DDR capacity

PCIe switch for peer-to-peer transfer between cards

- Transparent to the host
- Low latency data transfer between peers



AI Model Efficiency Toolkit

Quantization and compression for high performance inference



Improved/Robust
quantization for
INT16,8,4

Quantization
aware training
with range
learning

Model
Compression

Mix precision
support

Opensource

Benchmarks	Precision	25 W Card TDP		75W Card TDP	
		Performance (inf/s)	Efficiency (inf/s/W)	Performance (inf/s)	Efficiency (inf/s/W)
ResNet50v1.5	Int8	11118	553	22252	370
ResNet34-SSD	Int8	234	11	424	7
MobileNetv1-SSD	Int8	12499	582	23198	335
BERT Base	Mixed	1952	100	3688	54
BERT Large	Mixed	620	29	1084	15

Measurements with ML Commons™ data set

BERT benchmarks (Base and Large) at sequence lengths of 128

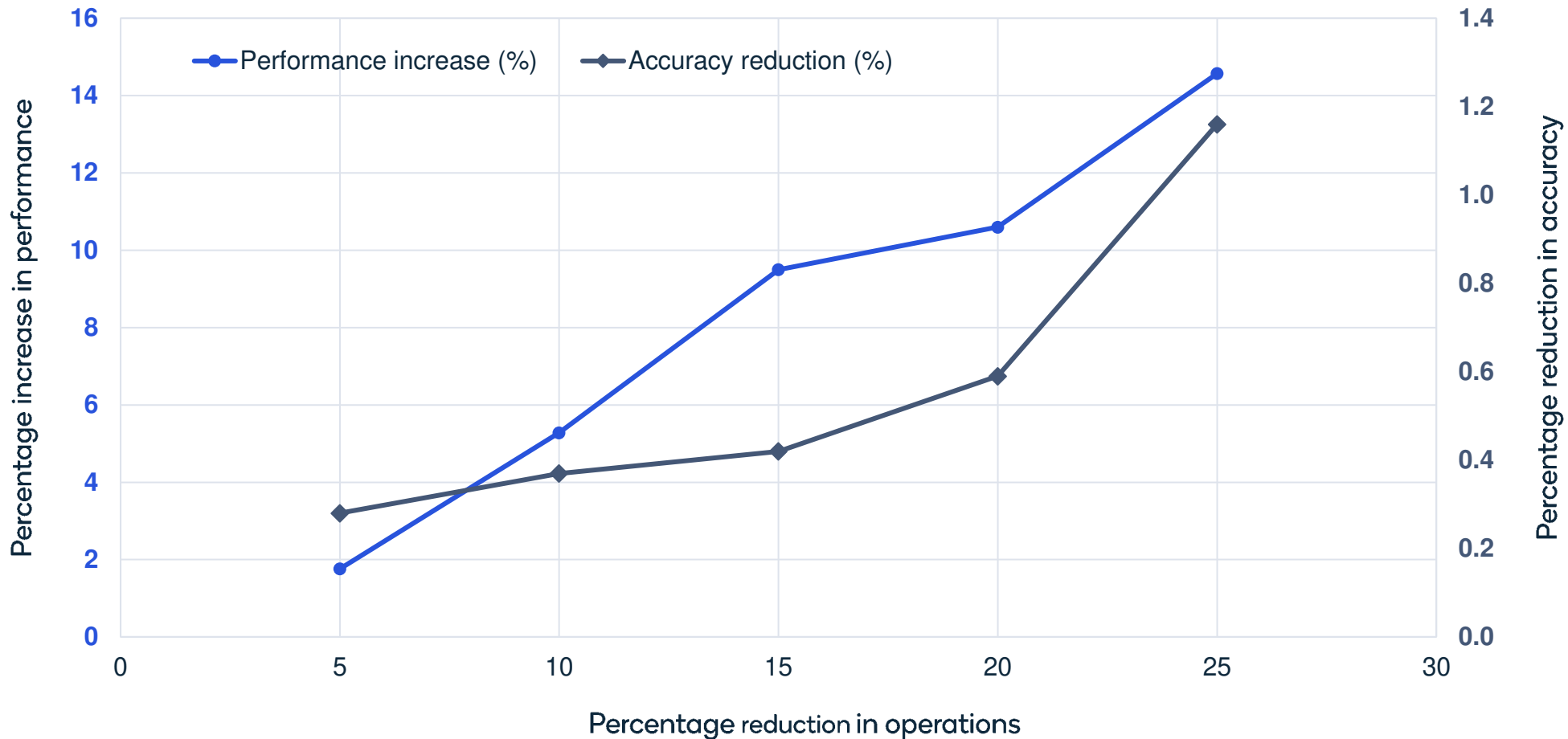
Benchmarks: Performance and Power Efficiency

Benchmarks – Performance versus batch size/latency



Peak performance with batch sizes of 4 and 8
Smaller batch sizes imply low latency for high performance

ResNet50 – Compression with AIMET



~15% increase in
ResNet50
performance for 1.1%
reduction in
accuracy (mAP
of 75.06)

Qualcomm Cloud AI 100 scalability

Performance and power scalability, across multiple platforms

Qualcomm Cloud Edge AI 100 Development Kit

Snapdragon 865

Qualcomm Cloud AI 100 (DM.2e)

Snapdragon X55 5G Modem

>50 TOPs of AI processing

15W TDP, passive cooled



Gigabyte G292-Z43 Cloud Inference Server

Supports up to 16x PCIe HHHL

6 Peta Ops of AI processing

2200W TDP, active cooled

Qualcomm Cloud AI 100

addressing edge-to-cloud industries

Data Center / Cloud Edge



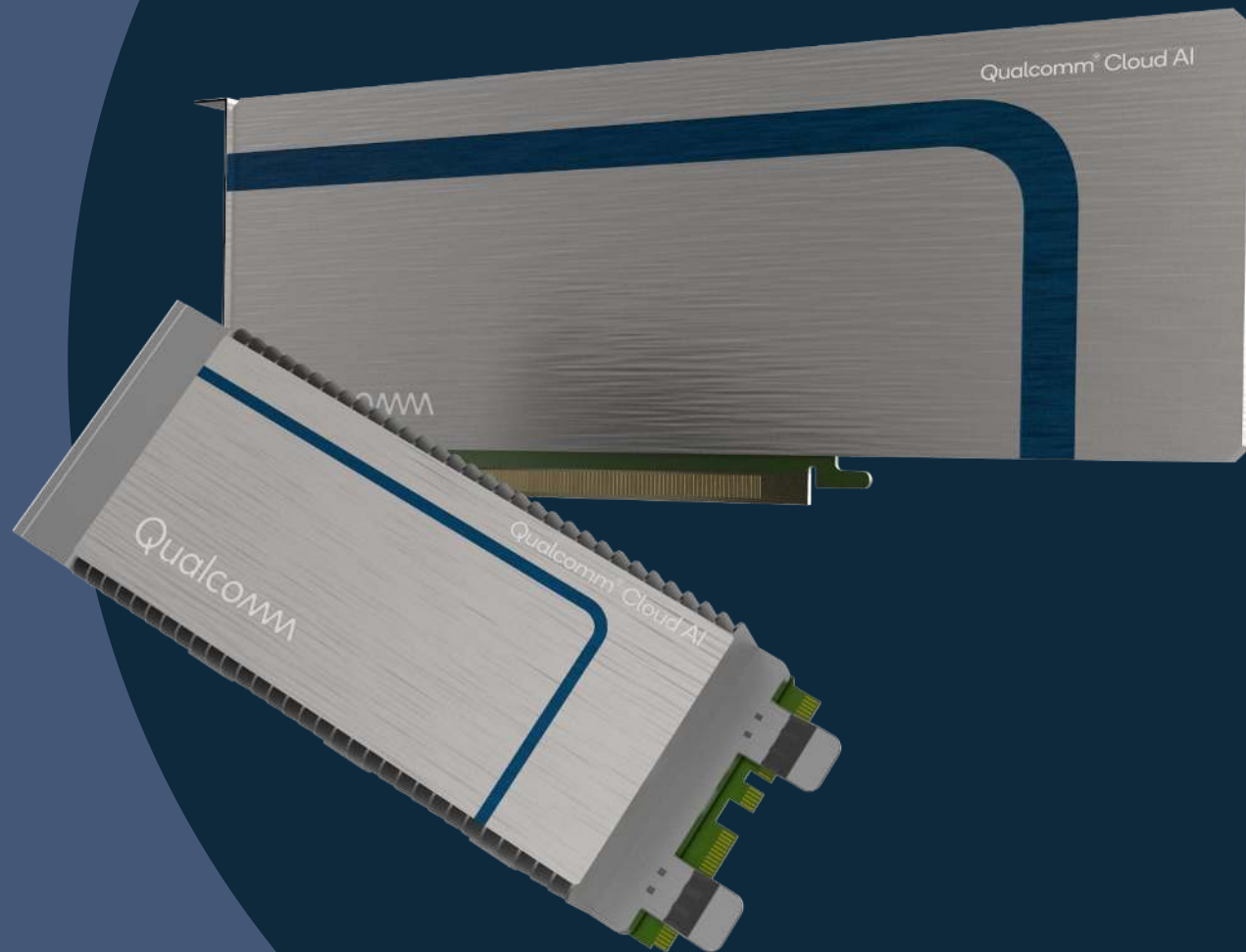
5G Edge Box



ADAS







5G Infrastructure





Thank you

Follow us on:    

For more information, visit us at:

www.qualcomm.com &

www.qualcomm.com/blog

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

©2018-2021 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. 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 our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.