

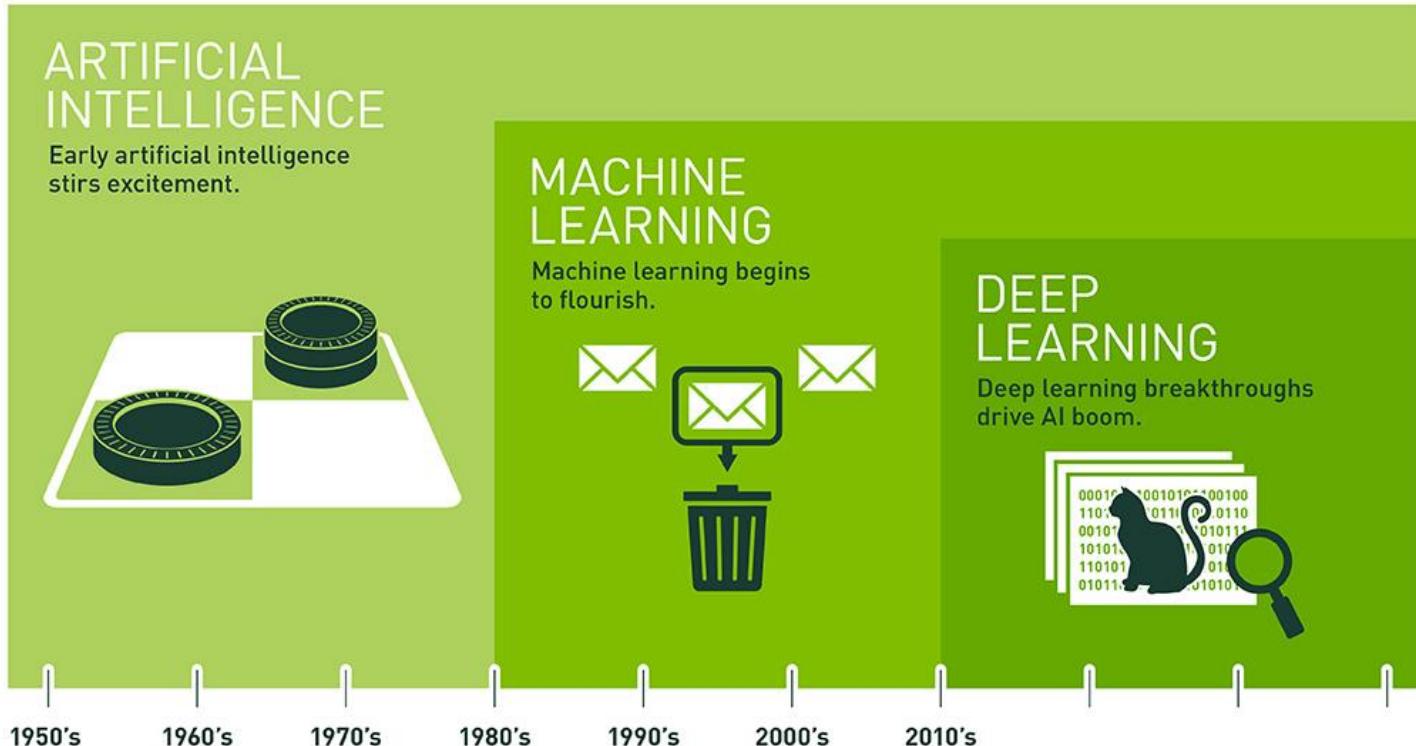
Computation Power Enabling A.I.

Tang Shan

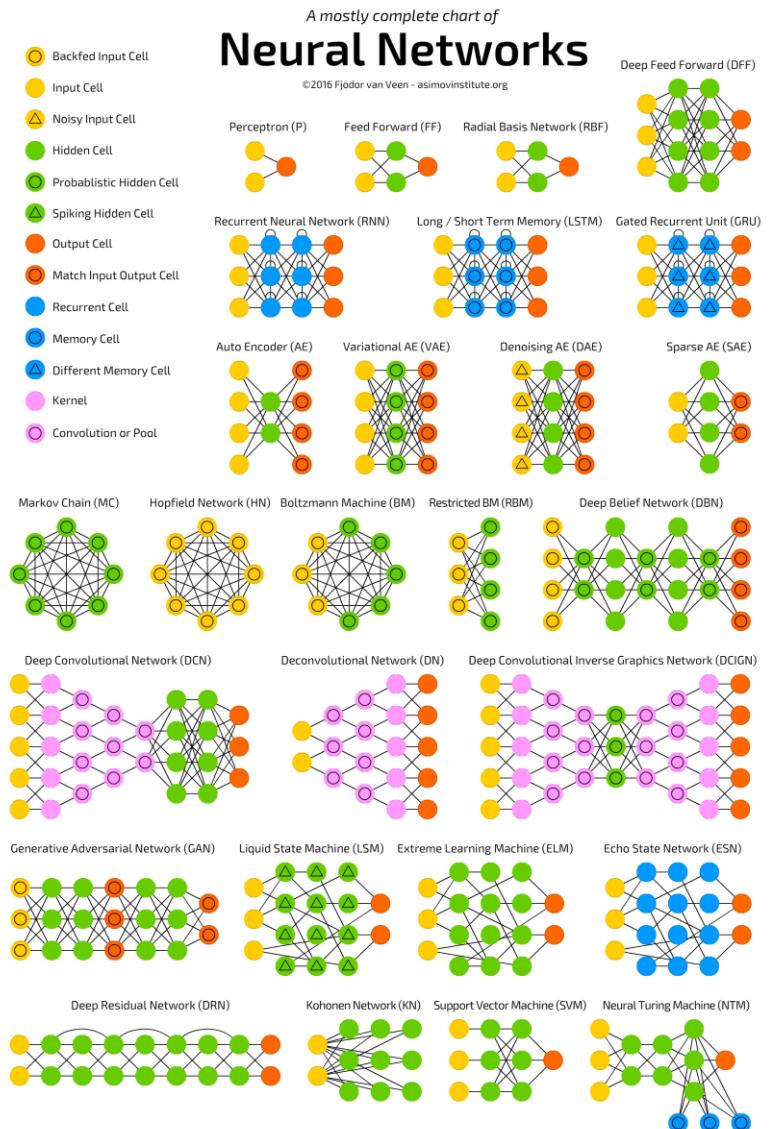
The Big Bang of Modern AI



Deep Learning and Neural Network



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

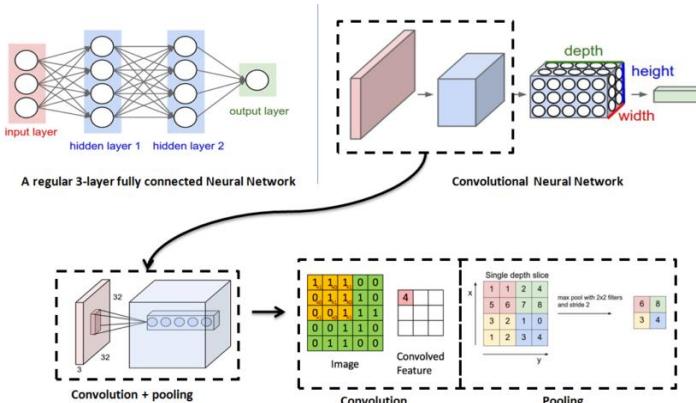


Three Keys to the Rise of Deep Learning

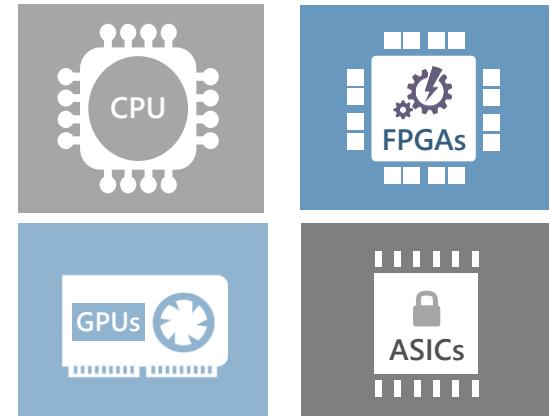
Data



Algorithm



Computation Power (Hardware)



AI Hardware Target Domains

Training

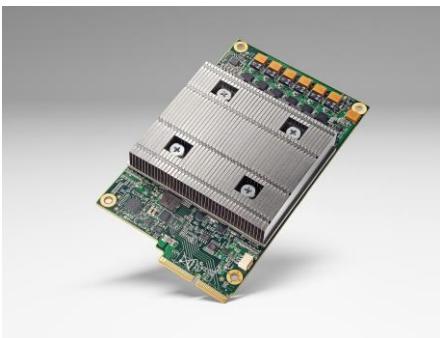
Cloud / Data Center



GPU/ASIC

Inference

Cloud / Data Center



GPU/ASIC/FPGA

Edge / Embedded



Training

Inference

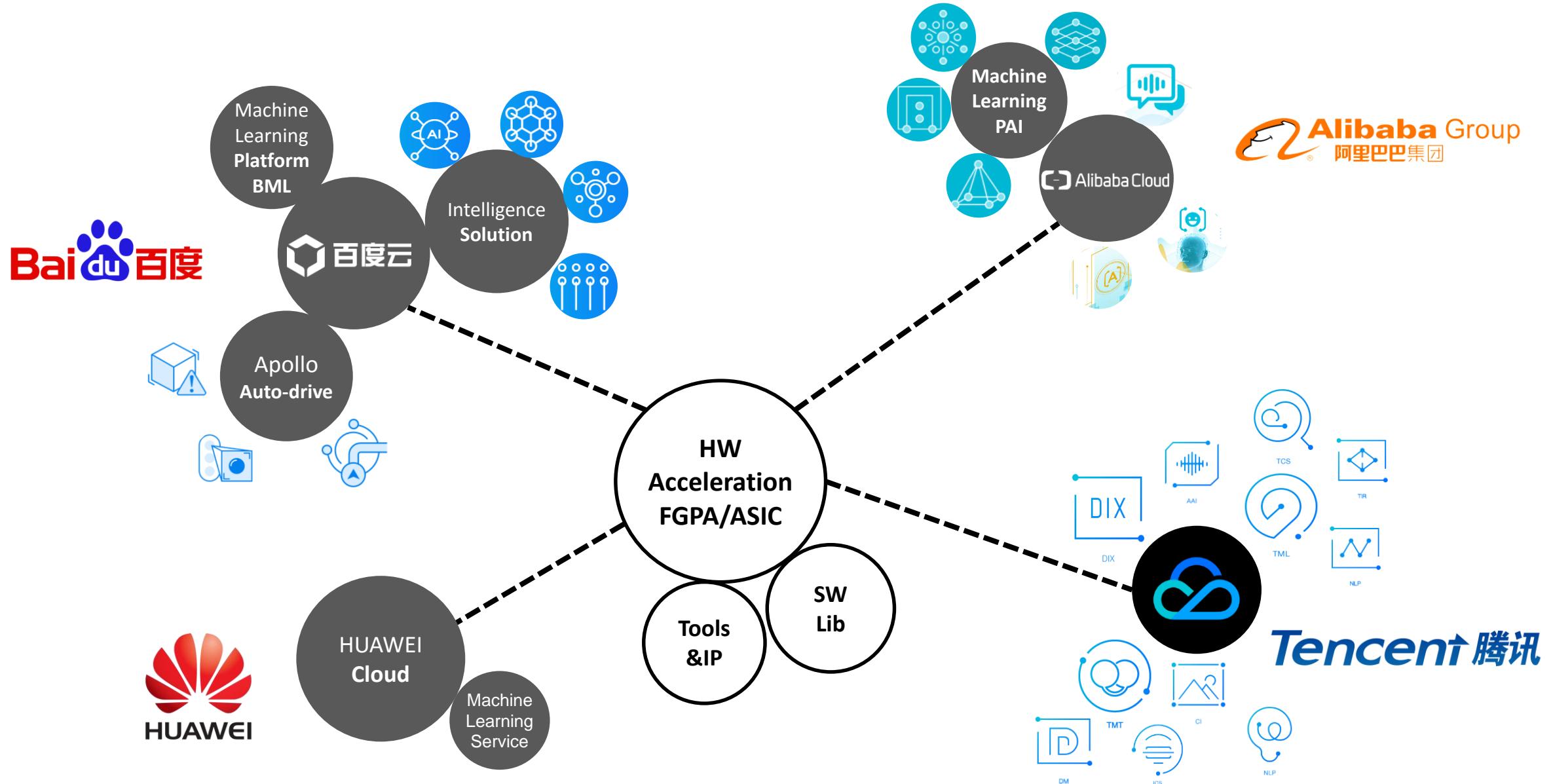
- Different Requirements & Constraints (from ADAS to wearable)
- Low ~ moderate Throughput
- Low Latency
- Power Efficiency (~10 TOPS/W)
- Low Cost



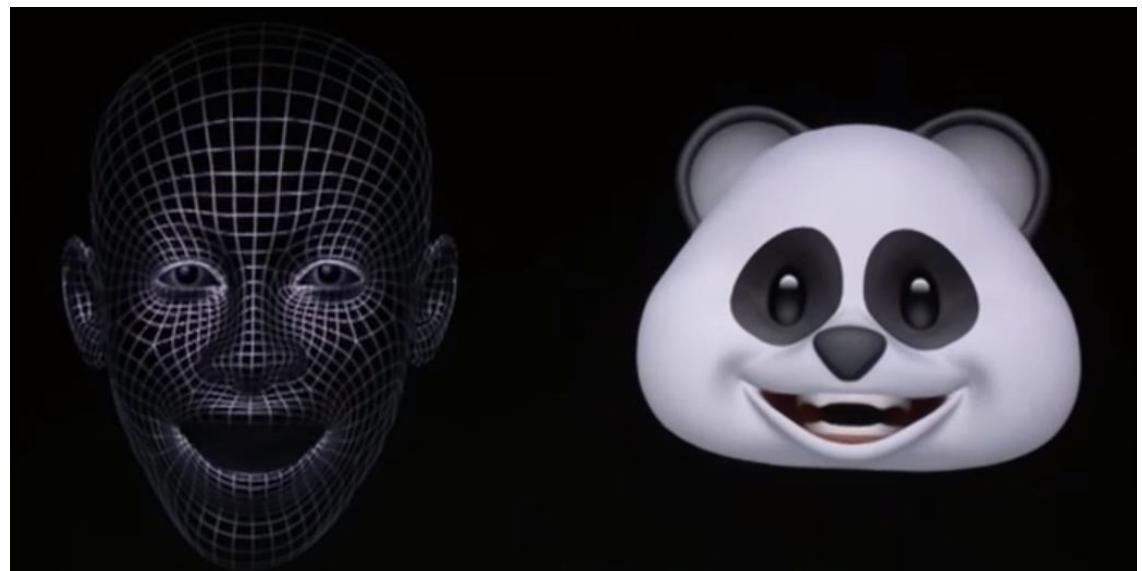
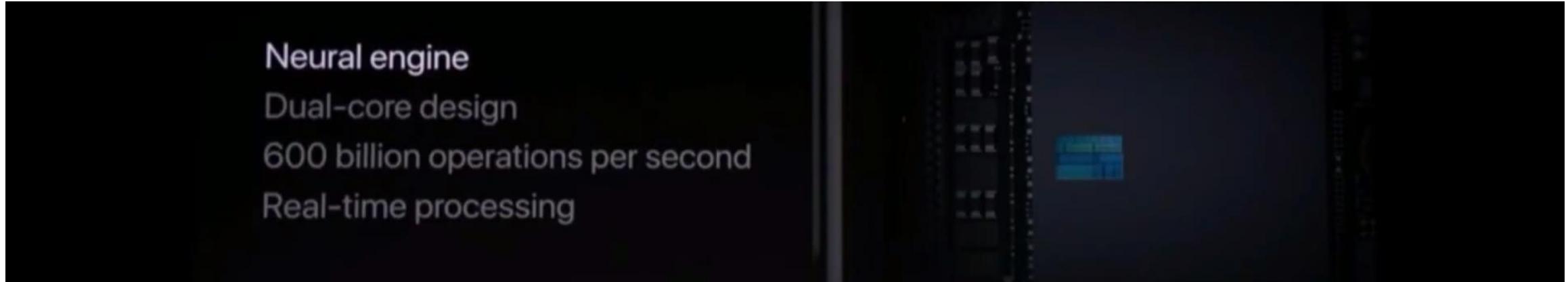
ASIC/FPGA

Edge / Embedded

Tech Giants' Cloud AI



Apple A11 Bionic



Hisilicon Kirin970

HUAWEI Kirin 970

The World's First Smartphone AI Computing Platform with a Dedicated **NPU**

Leading Process Technology
10nm Process Technology

Mobile AI Computing NPU
Up to 25x performance
Up to 50x power efficiency

High Performance 8-Core CPU
4xA73 @2.4GHz
4xA53 @1.8GHz

High Efficiency 12-Core GPU
First-to-Market Mali G72MP12

Advanced Dual ISP
4-Hybrid Focus
Low-light & Motion Shooting

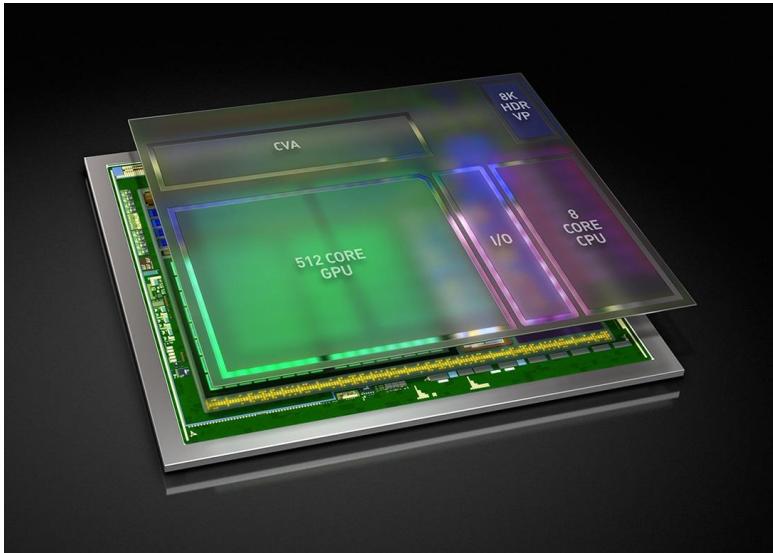
Ultra-Fast 4.5G LTE Modem
4.5G LTE Cat.18 up to 1.2Gbps Download speeds

NPU: Neural Processing Unit

© 2017 Huawei Technologies Co., Ltd.

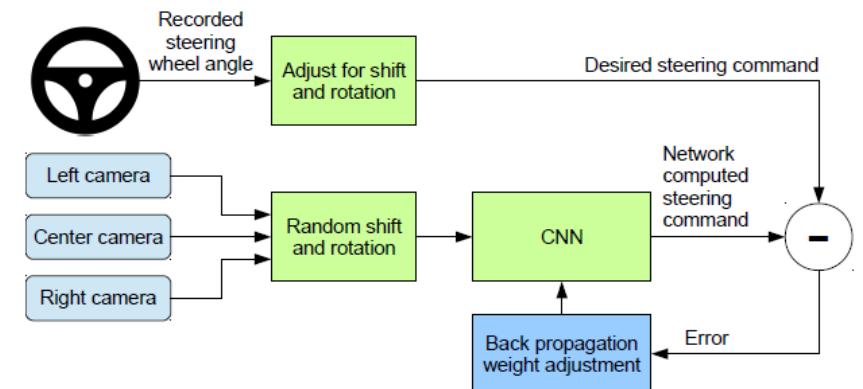
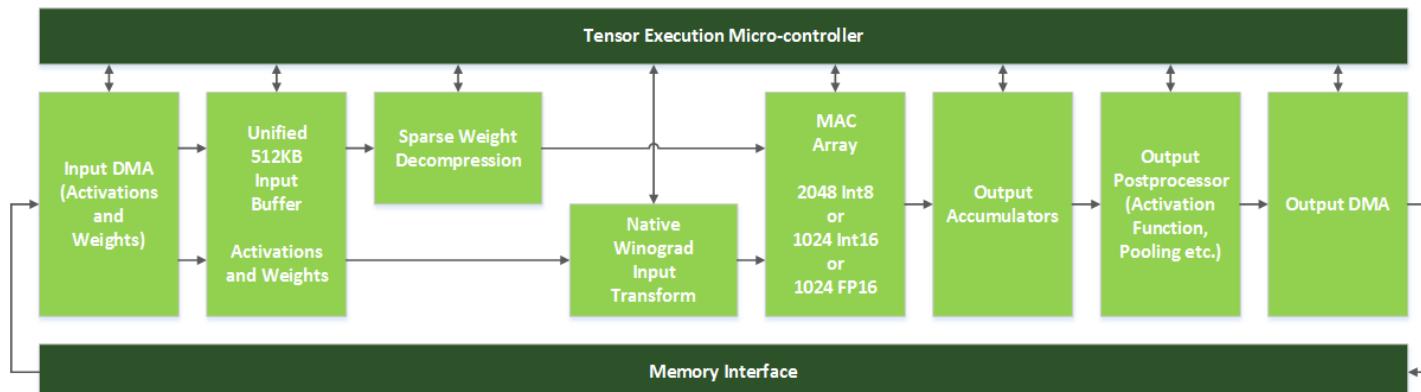
8

Nvidia SoC for Self driving cars

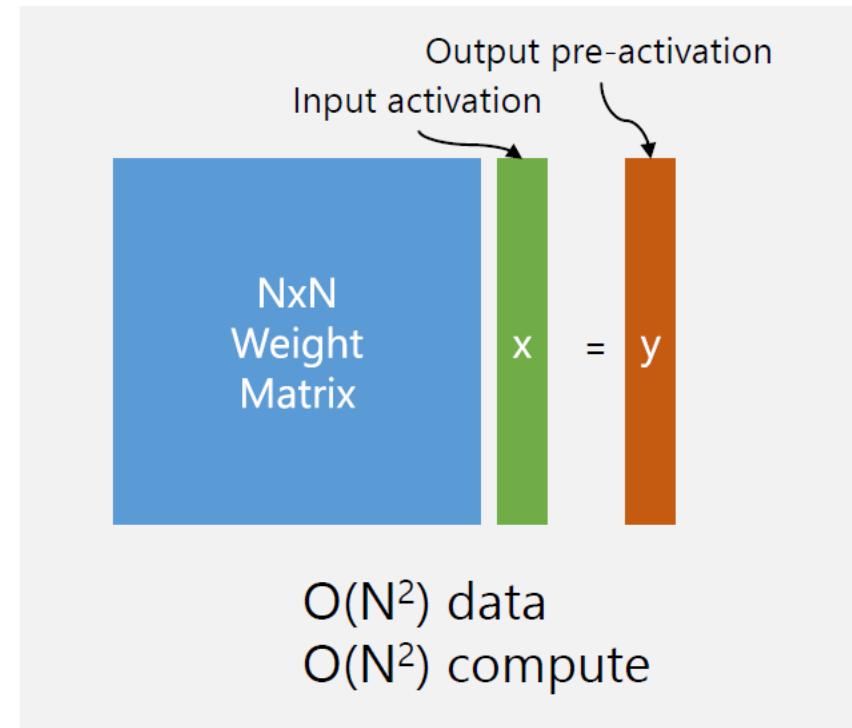
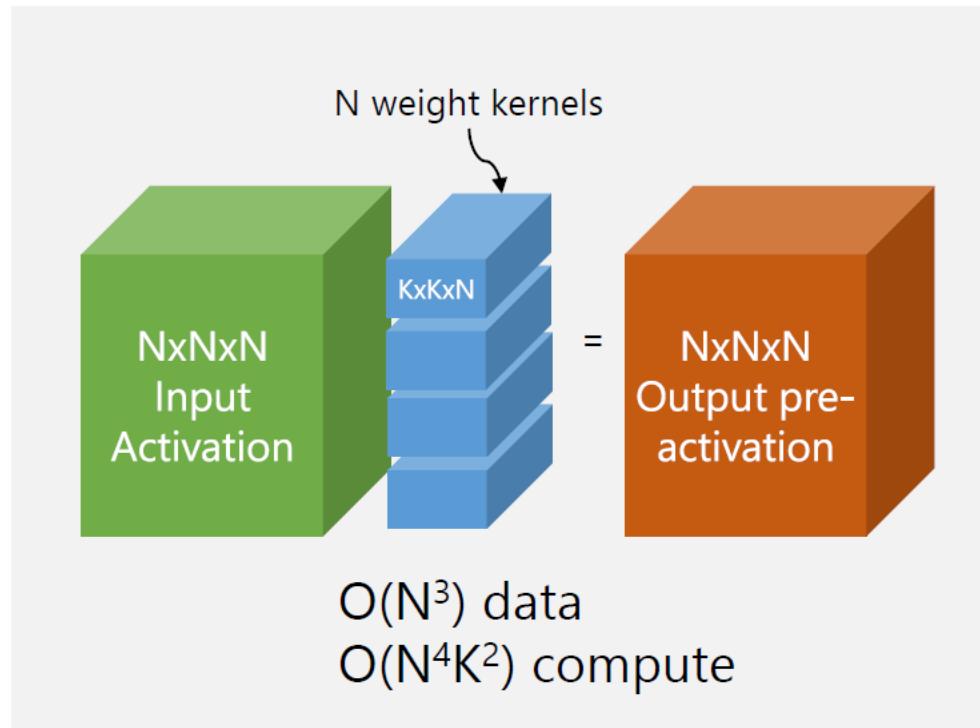


**20 TOPS DL
160 SPECINT
20W**

“In the case of Xavier, we’ve created a processor out of a CPU, GPU with CUDA and DLA. This lets you combine general purpose architecture, and domain specific accelerators. What’s great is we have an architecture that is both programmable, robust, super energy efficient and can run the entire software stack of self driving cars. We understand the entire pipeline and we have the software stack across the board.”



Computations for Neural Networks

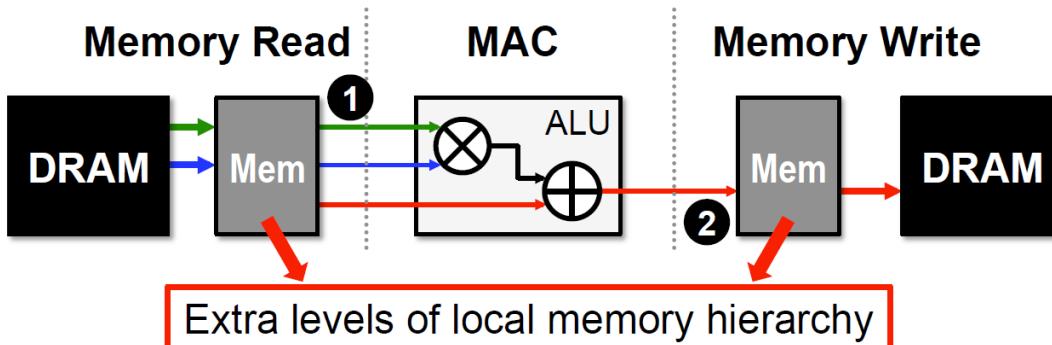


Convolutional Neural Network (CNN)
High Compute-to-Data Ratio

MLPs, LSTMs, GRUs
Low compute-to-data ratio

Optimizations – Two Directions

Reduce Memory Access



Opportunities: ① data reuse ② local accumulation

- ① Can reduce DRAM reads of **filter/fmap** by up to **500x**
 - ② **Partial sum** accumulation does **NOT** have to access DRAM
- Example: DRAM access in AlexNet can be reduced from **2896M** to **61M** (best case)

Reduce Storage/Computation

- Reduce size of operands for storage/compute
 - Floating point → Fixed point
 - Bit-width reduction
 - Non-linear quantization
- Reduce number of operations for storage/compute
 - Exploit Activation Statistics (Compression)
 - Network Pruning
 - Compact Network Architectures

NVIDIA Volta

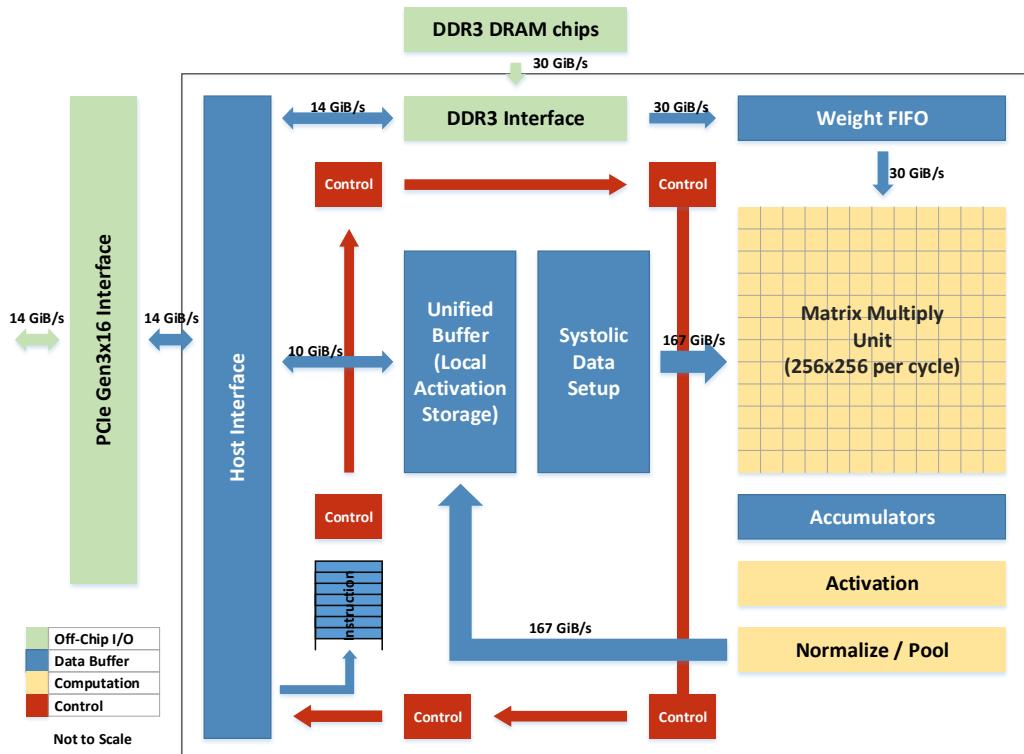


NVIDIA TESLA V100 SPECIFICATIONS

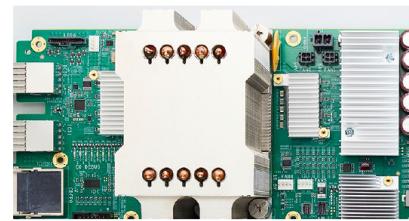
| | Tesla V100 for NVLink | Tesla V100 for PCIe |
|---|--|---|
| PERFORMANCE with NVIDIA GPU Boost™ | DOUBLE-PRECISION 7.8 TeraFLOPS | DOUBLE-PRECISION 7 TeraFLOPS |
| SINGLE-PRECISION | 15.7 TeraFLOPS | SINGLE-PRECISION 14 TeraFLOPS |
| DEEP LEARNING | 125 TeraFLOPS | DEEP LEARNING 112 TeraFLOPS |
| INTERCONNECT BANDWIDTH Bi-Directional | NVLINK 300 GB/s | PCIE 32 GB/s |
| MEMORY CoWoS Stacked HBM2 | CAPACITY 16 GB HBM2 | BANDWIDTH 900 GB/s |
| POWER Max Consumption | 300 WATTS | 250 WATTS |

Source: NVIDIA

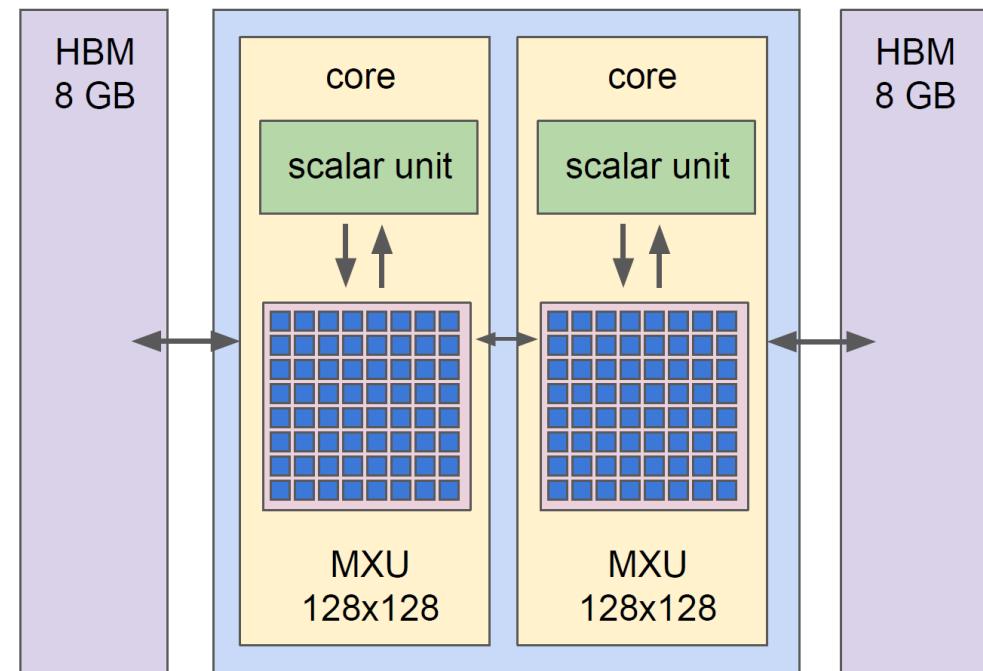
Google TPU and TPU2



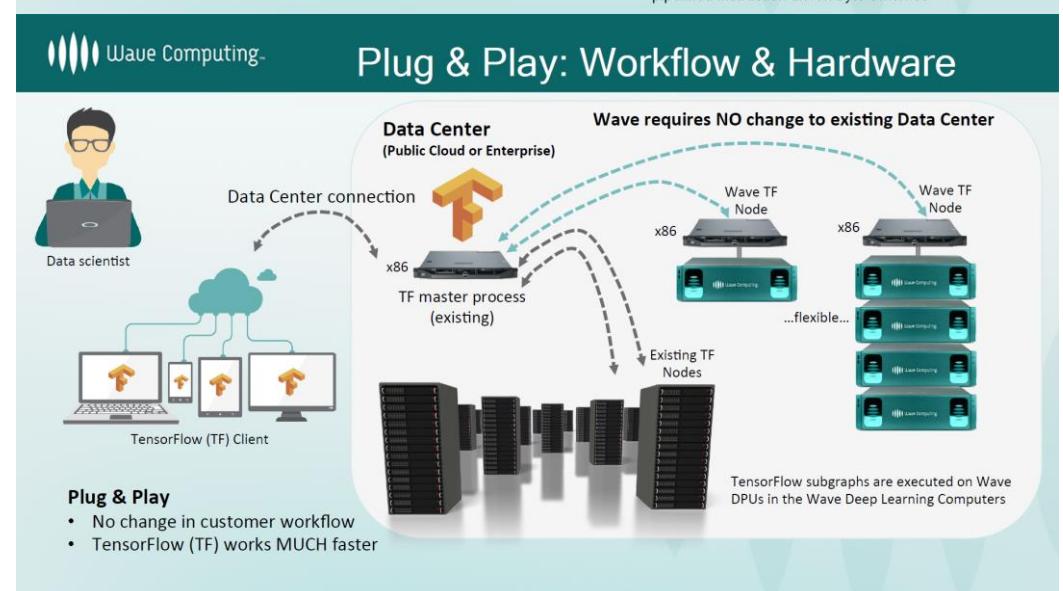
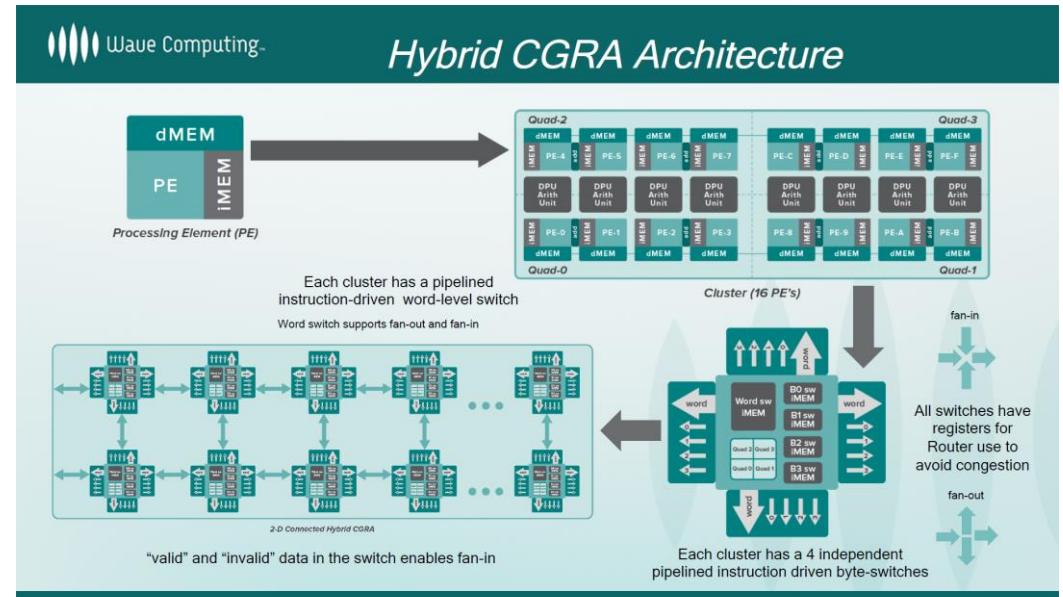
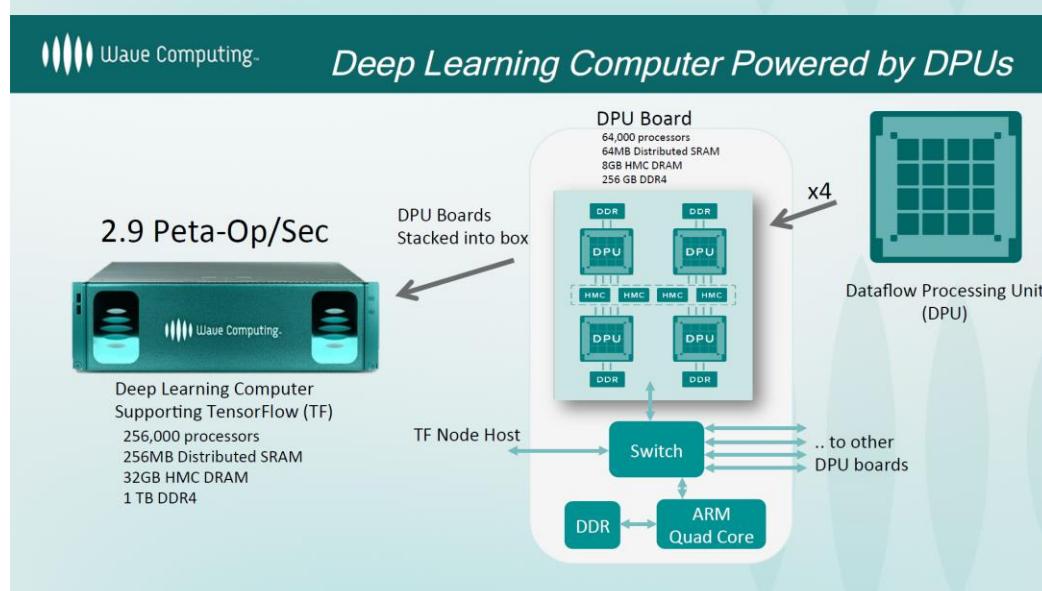
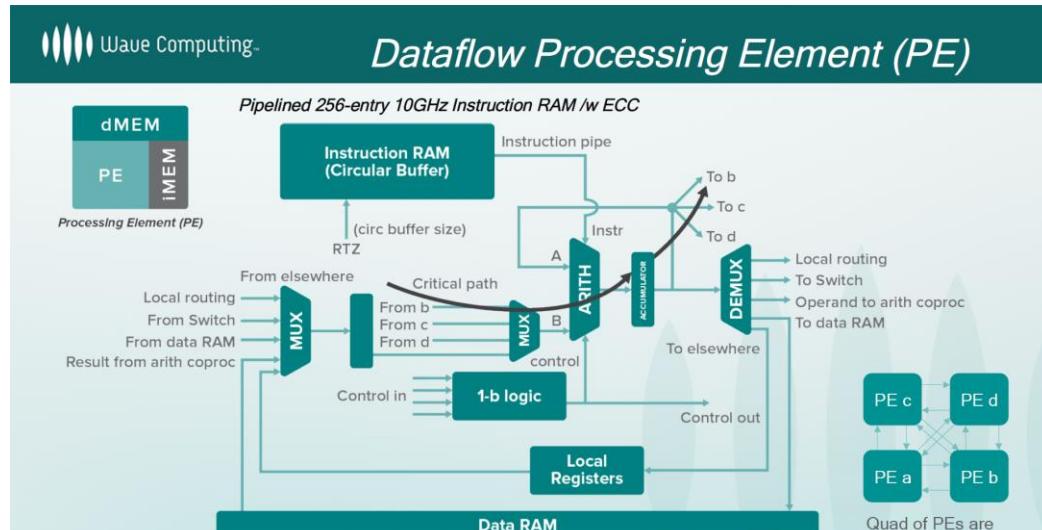
TPUv2 Chip



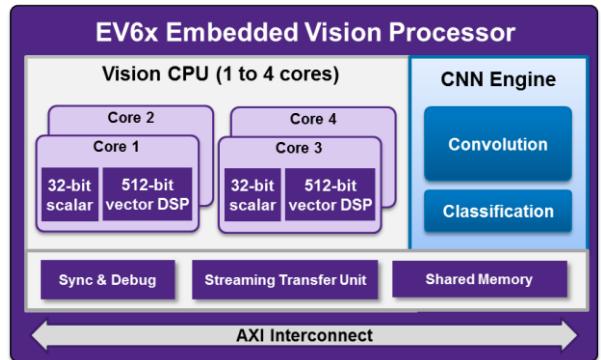
- 16 GB of HBM
- 600 GB/s mem BW
- Scalar unit: 32b float
- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS



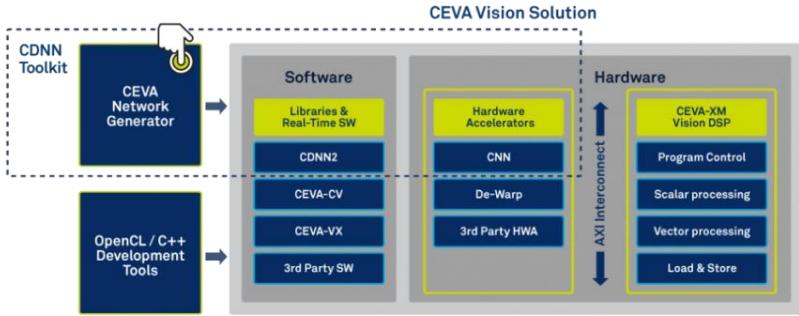
Wave Computing



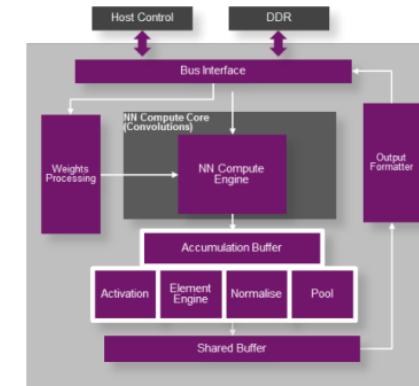
IPs for Embedded Application



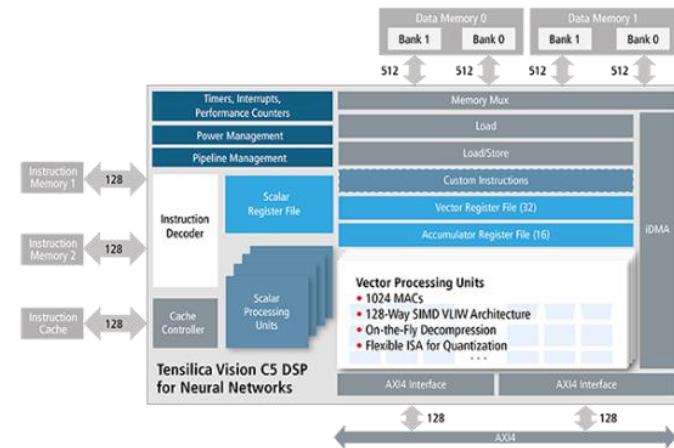
Synopsys EV



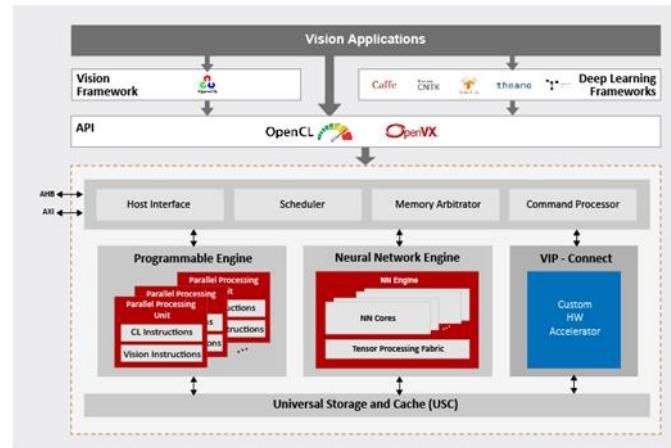
CEVA XM



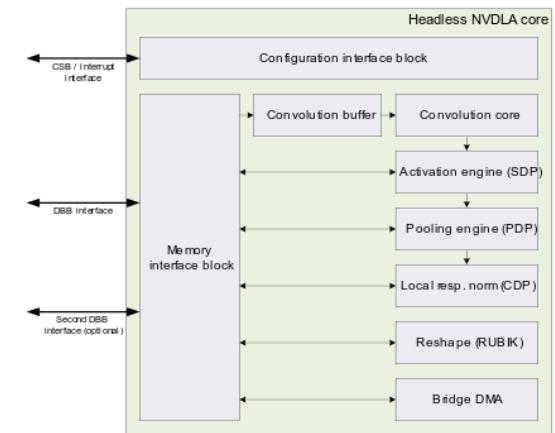
Imagination Neural Network Accelerator (NNA)



Cadence Tensilica C5: Dedicated NN DSP

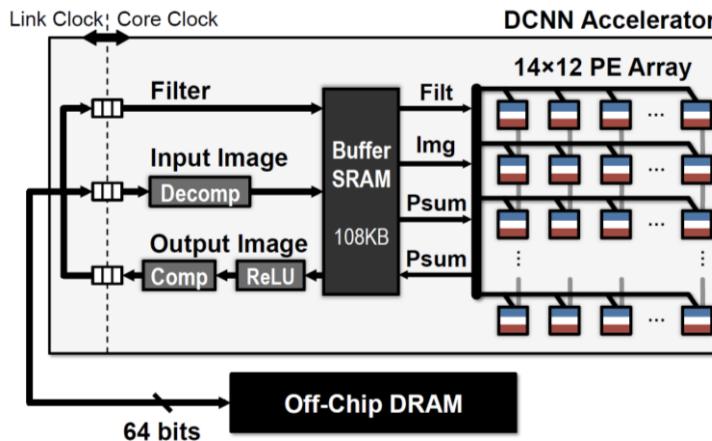


Vivante VIP8000



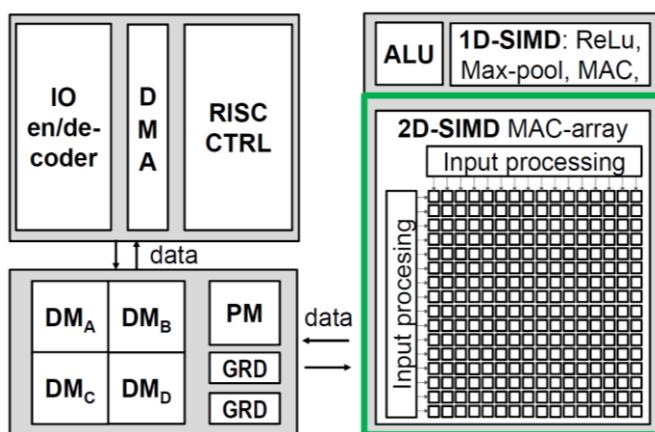
Nvidia NVDLA

Deep Learning Processor from Academia

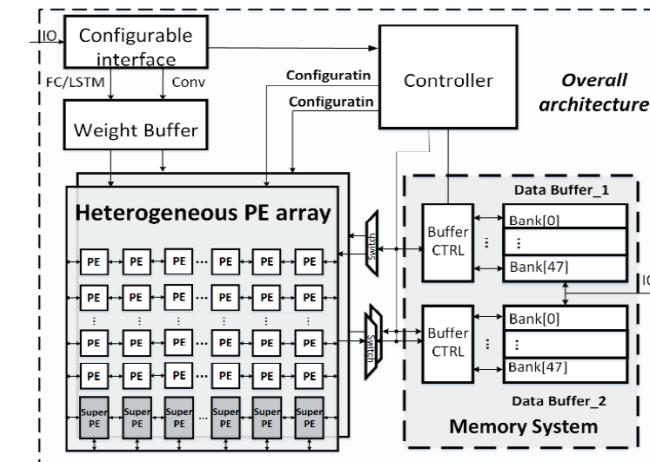


Eyeriss MIT

A 2D-SIMD DVAFS Architecture

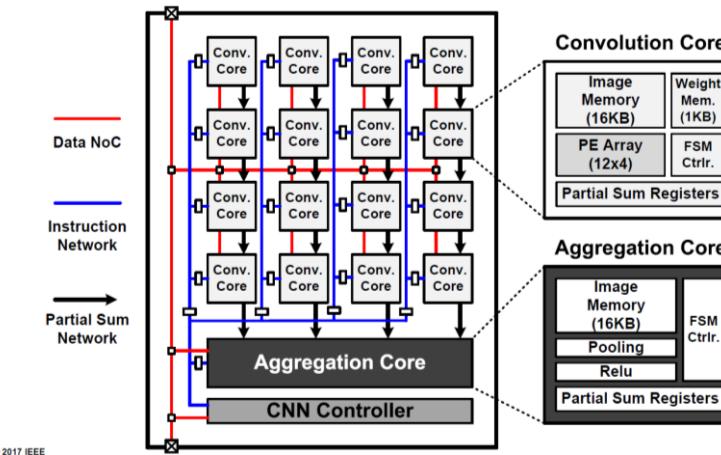


ENVSION ESAT/MICAS - KU Leuven



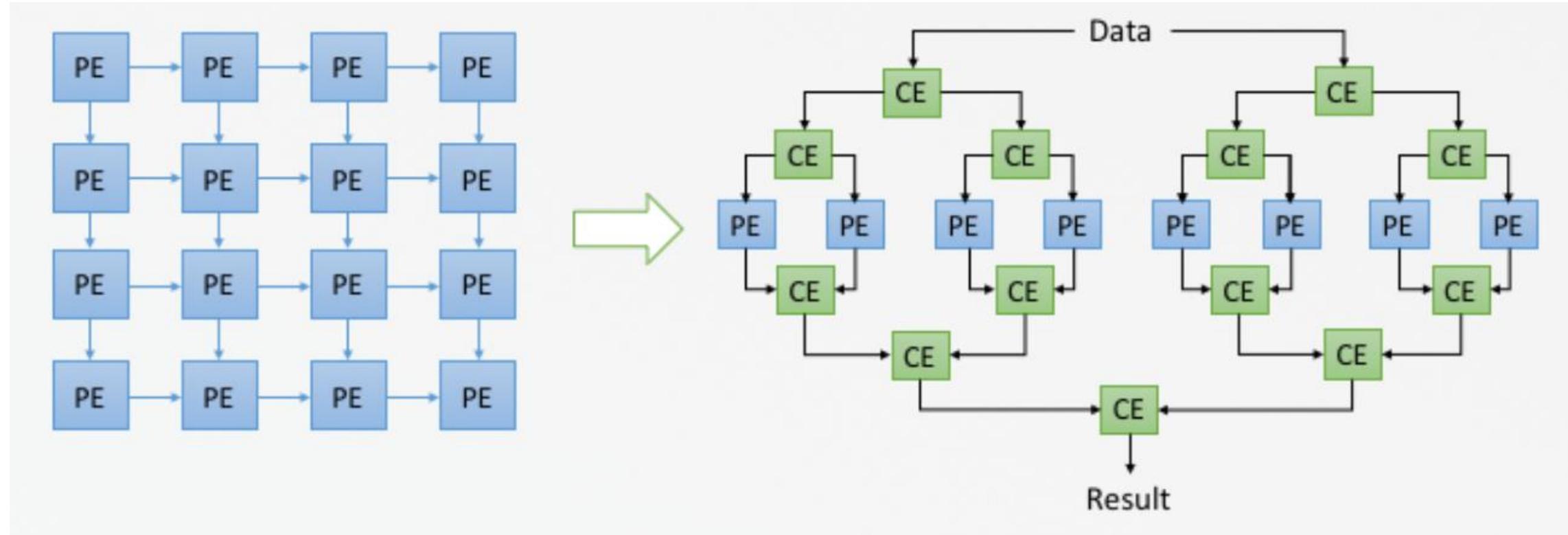
Tsinghua Thinker

- **Distributed memory-based architecture**

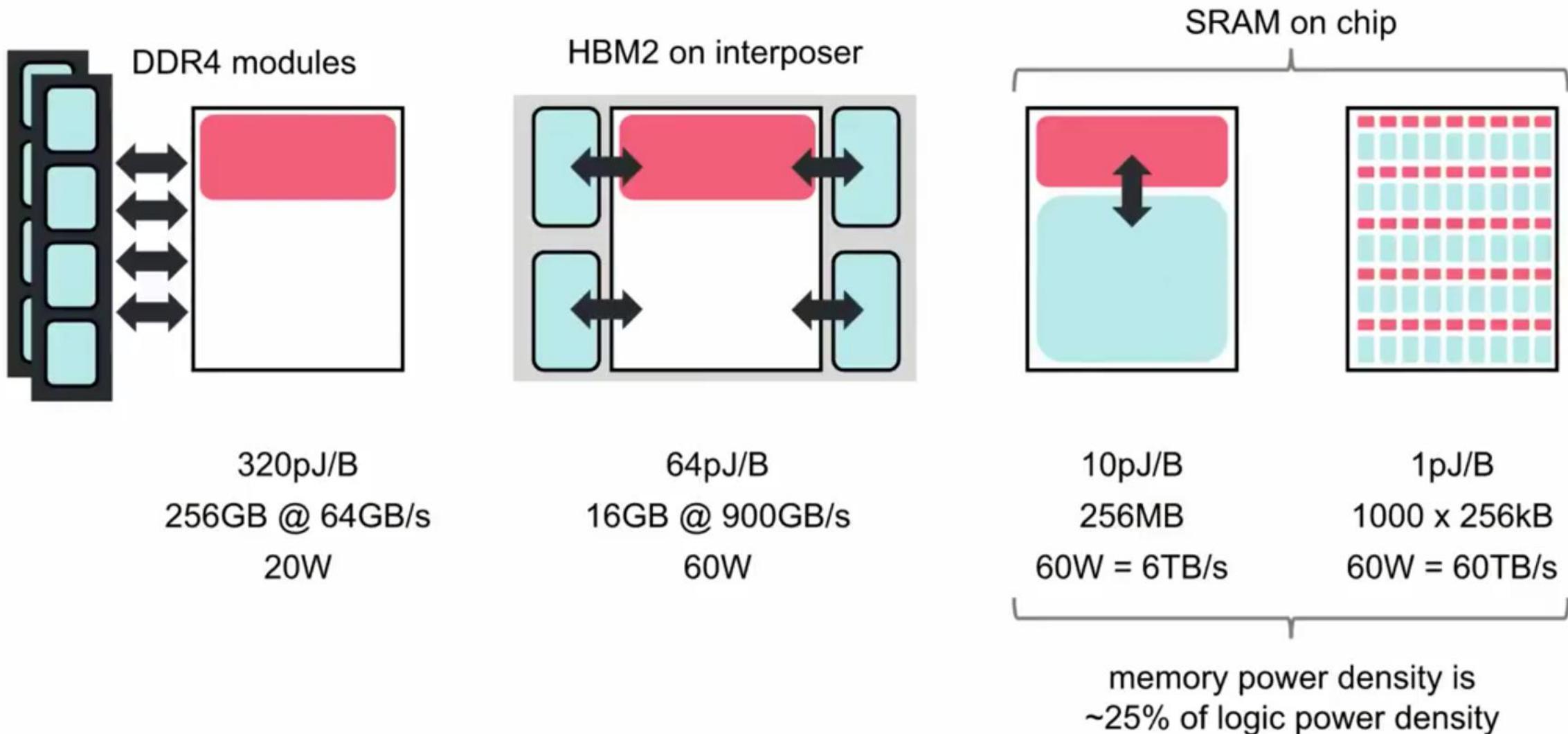


KAIST (ISSCC 2017)

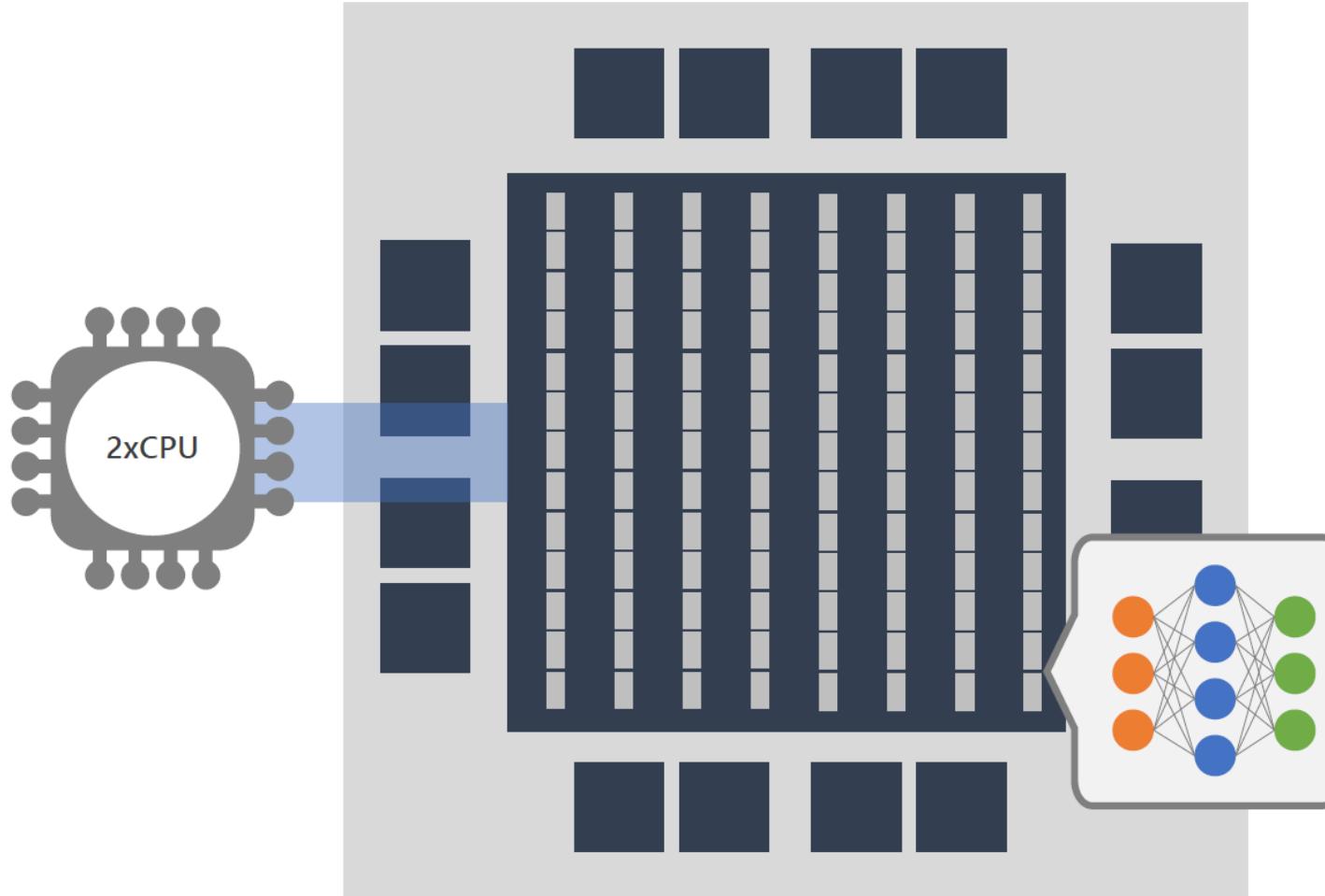
Architecture Discussions: Systolic Array ! or Not?



Architecture Discussions: “Dark Silicon” challenge, All in SRAM?



Architecture Discussions: Persistent



Observations

State-of-art FPGAs have $O(10K)$ distributed Block RAMs $O(10MB)$
→ Tens of TB/sec of memory BW

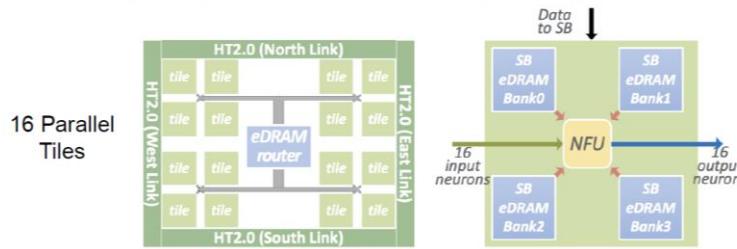
Large-scale cloud services and
DNN models run persistently

*Solution: persist all model
parameters in FPGA on-chip
memory during service lifetime*

Architecture Discussions: Advanced

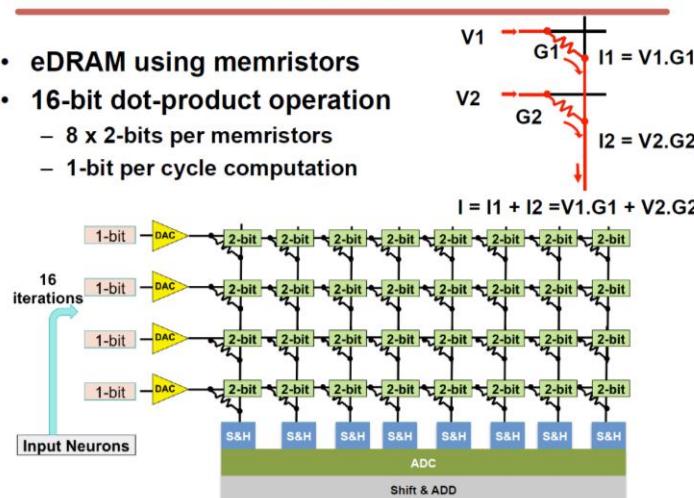
eDRAM (DaDianNao)

- Advantages of eDRAM
 - 2.85x higher density than SRAM
 - 321x more energy-efficient than DRAM (DDR3)
- Store weights in eDRAM (36MB)
 - Target fully connected layers since dominated by weights



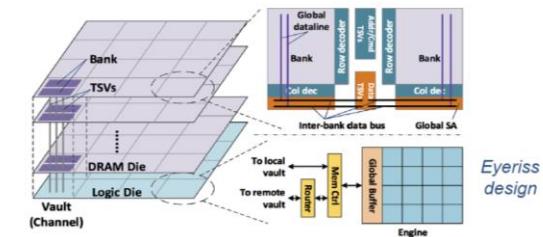
ISAAC

- eDRAM using memristors
- 16-bit dot-product operation
 - 8 x 2-bits per memristors
 - 1-bit per cycle computation



Stacked DRAM (TETRIS)

- Explores the use of HMC with the Eyeriss spatial architecture and row stationary dataflow
- Allocates more area to the computation (PE array) than on-chip memory (global buffer) to exploit the low energy and high throughput properties of the HMC
 - 1.5x energy reduction, 4.1x higher throughput vs. 2-D DRAM

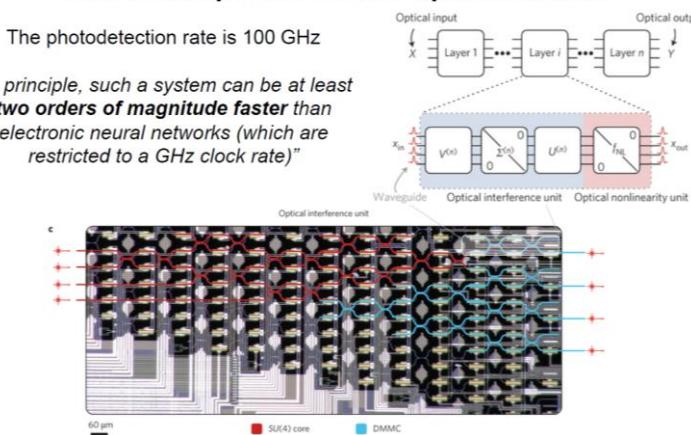


Optical Neural Network

Matrix Multiplication in the Optical Domain

The photodetection rate is 100 GHz

"In principle, such a system can be at least two orders of magnitude faster than electronic neural networks (which are restricted to a GHz clock rate)"



The Software – NVIDIA Example

Training with Popular Frameworks

Caffe



MINERVA

mxnet

K
KERAS

TensorFlow

Microsoft
CNTK

theano

DL4J
Deeplearning4j



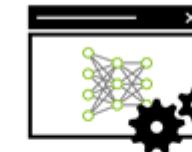
MatConvNet



Deploying with TensorRT



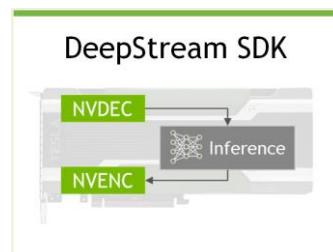
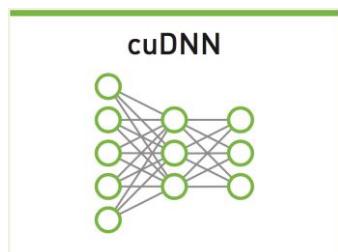
Trained
Neural
Network



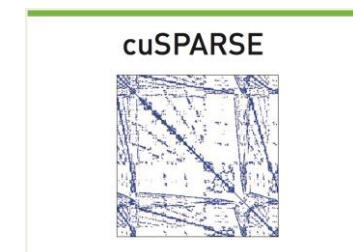
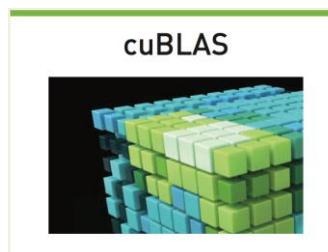
TensorRT
Optimizer



TensorRT
Runtime
Engine



Deep Learning Libraries

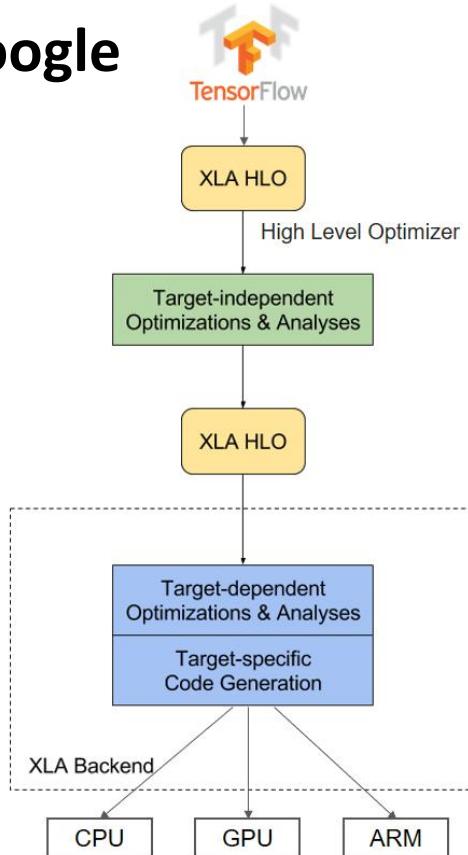


CUDA Toolkit

Source: NVIDIA

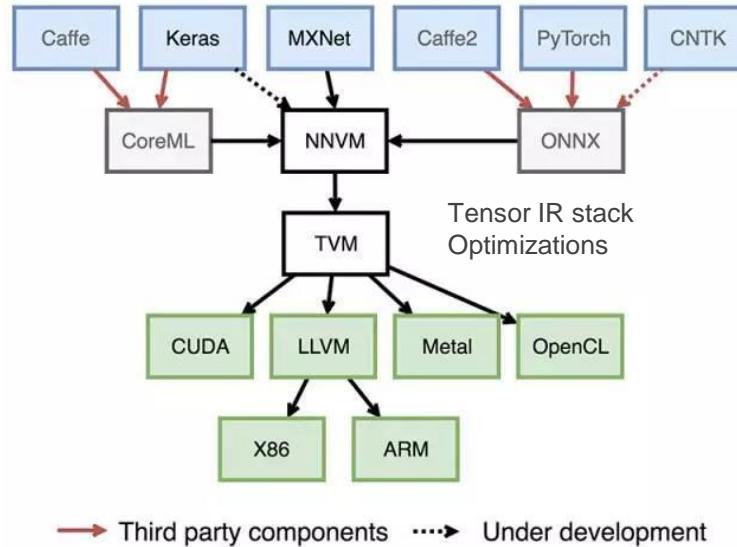
The Software – Compilers

Google



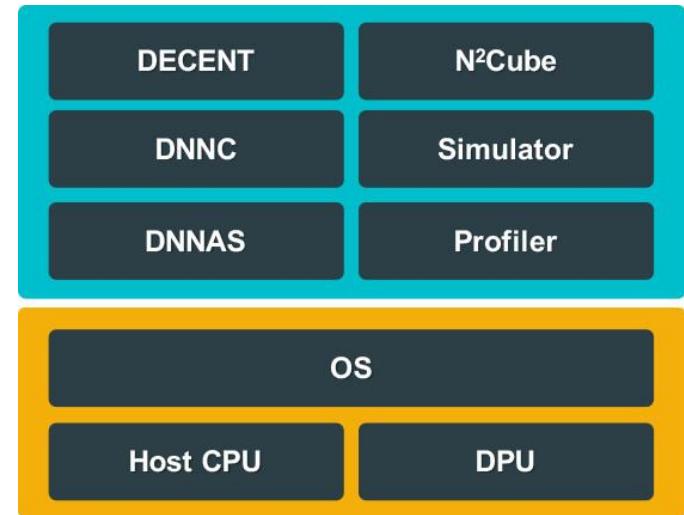
XLA:
a domain-specific compiler

AWS



NNVM Compiler:
Open End-to-End Compiler

DeePhi



DNNDK (Deep Neural Network Development Kit) Framework

Who is working on IC/IP for AI

| | | |
|---------------------------|---|----|
| IC Giants | Intel, Qualcomm, Nvidia, AMD, Apple, Xilinx, IBM, HiSilicon | 8 |
| Cloud/HPC | Google, Amazon_AWS, Microsoft, Aliyun, Tencent Cloud, Baidu, Baidu Cloud, HUAWEI Cloud, Fujitsu | 9 |
| IP Vendors | ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon | 6 |
| Startups in China | Cambricon, Horizon Robotics, DeePhi, Bitmain, Chipintelli | 5 |
| Startups Worldwide | Cerebras, Wave Computing, Graphcore, PEZY, KnuEdge, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, Almotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, TERADEEP, DEEP VISION, Groq, KAIST DNPU | 23 |



Thank you!