



Master Informatics Eng.

2019/20

A.J.Proença

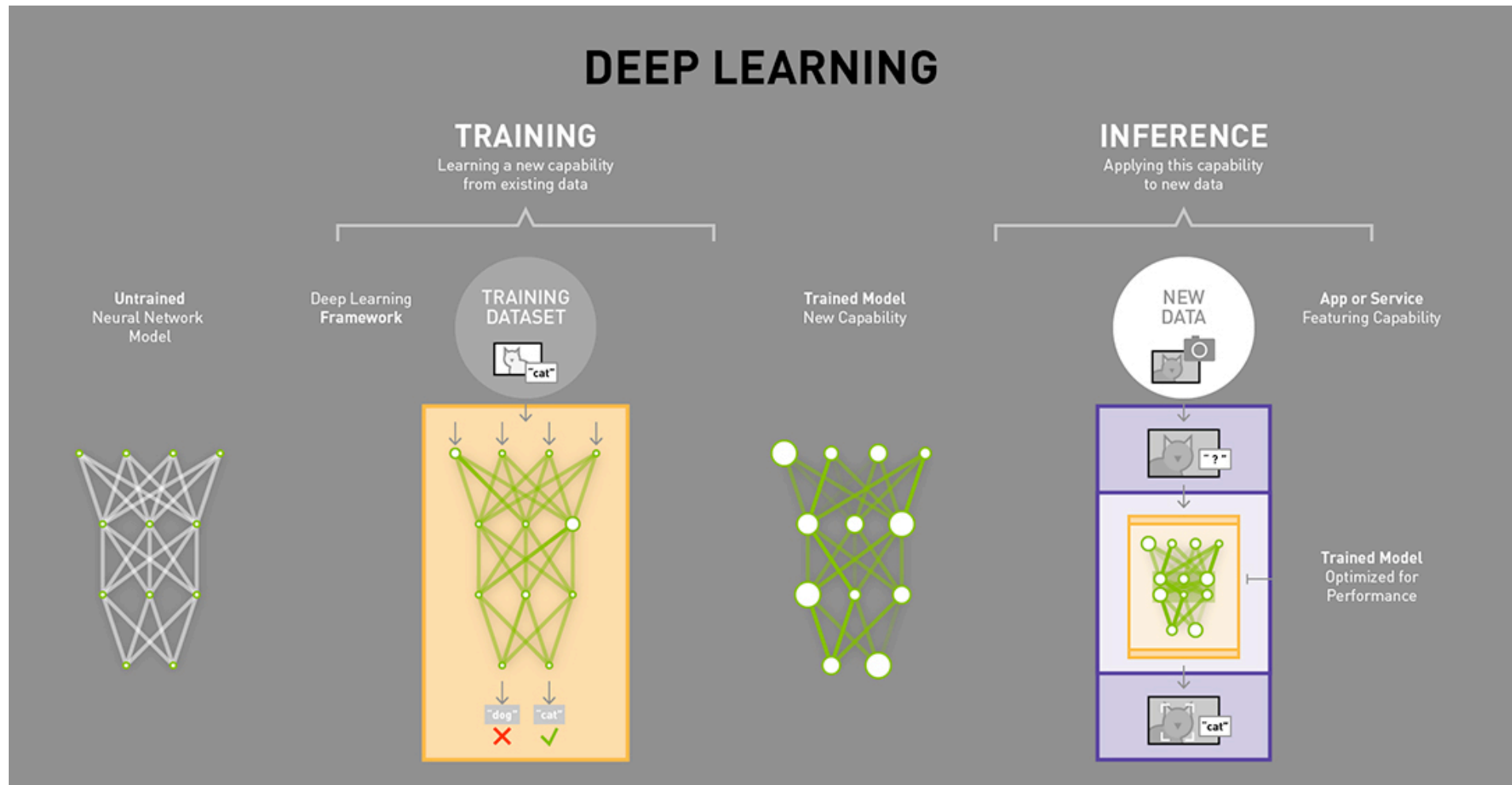
Data Parallelism 6 (*neural nets, neuromorphic computing, ...*)
(*most slides are borrowed*)

Beyond Vector/SIMD architectures



- Vector/SIMD-extended architectures are hybrid approaches
 - mix (super)scalar + vector op capabilities on a single device
 - highly pipelined approach to reduce memory access penalty
 - tightly-closed access to shared memory: lower latency
- Evolution of Vector/SIMD-extended architectures
 - **PU (Processing Unit) cores with wider vector units**
 - x86 many-core: Intel MIC / Xeon KNL
 - others: IBM Power BlueGene/Q Compute, ShenWay 260, Matrix-2000, A64FX Arm
 - **coprocessors (require a host scalar processor): accelerator devices**
 - on disjoint physical memories (e.g., Xeon KNC with PCI-Expr, PEZY-SC)
 - ISA-free architectures, code compiled to silica: FPGA
 - focus on SMT/SIMD to hide memory latency: GPU-type approach
 - tensor/neural net cores: NVidia Volta, Google TPU, Intel NNP
 - ...
 - **heterogeneous PUs in a SoC: multicore PUs with GPU-cores**
 - ...

Machine learning w/ neural nets & deep learning...

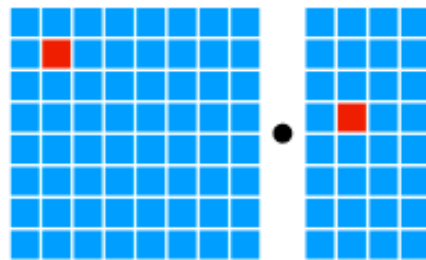


Key algorithms to train & classify use matrix dot products,
but require lower precision numbers!

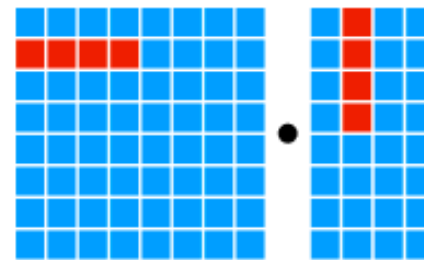
Required hardware operations & data types to train & classify neural nets



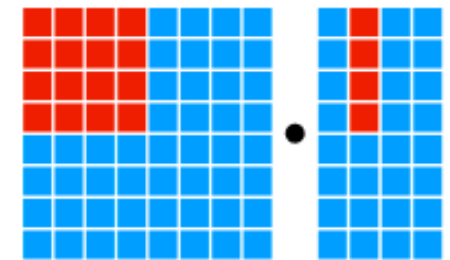
Compute primitives



scalar



vector



tensor

Data type



fp32



fp16



int8

NEW!

NVidia Volta Architecture: the new Tensor Cores



$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 FP16 or FP32

Figure 8. Tensor Core 4x4 Matrix Multiply and Accumulate



For each SM:
8x 64 FMA ops/cycle
1k FLOPS/cycle!

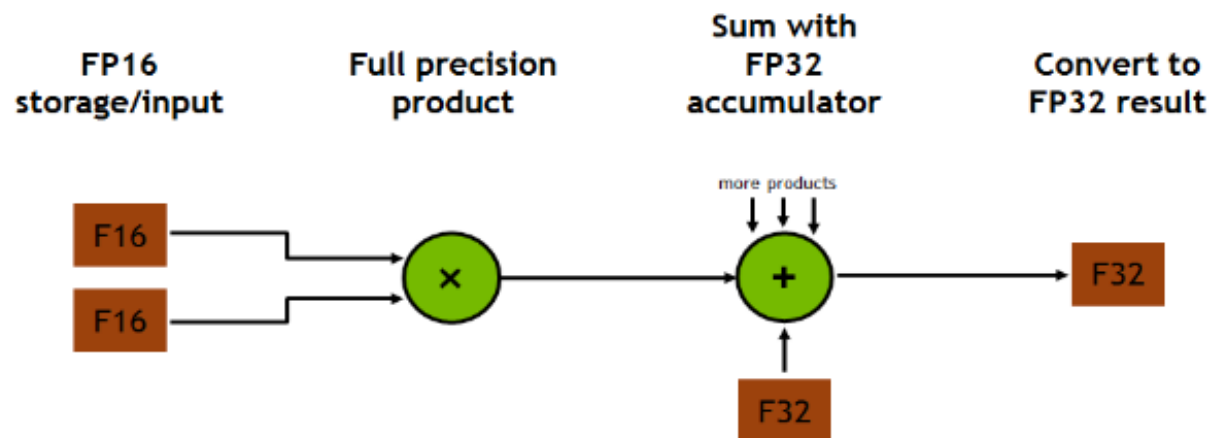


Figure 9. Mixed Precision Multiply and Accumulate in Tensor Core

NVidia competitors with neural net features: Intel Nervana Neural Network Processor, NNP



History

- Nervana Engine announced in May 2016
- Key features:
 - ASIC chip, focused on matrix multiplication, convolutions,... (for neural nets)
 - HBM2: 4x 8GiB in-package storage & 1TB/sec memory access b/w
 - no h/w managed cache hierarchy (saves die area, higher compute density)
 - built-in networking (6 bi-directional high-b/w links)
 - separate pipelines for computation and data management
 - proprietary numeric format Flexpoint
 - in-between floating point and fixed point precision
- Nervana acquired by Intel in August 2016
 - renamed the project to “Lake Crest” (prototype only)
 - later to Nervana NNP, launched in October’17 (next slides)

<https://www.top500.org/news/intel-will-ship-first-neural-network-chip-this-year/>



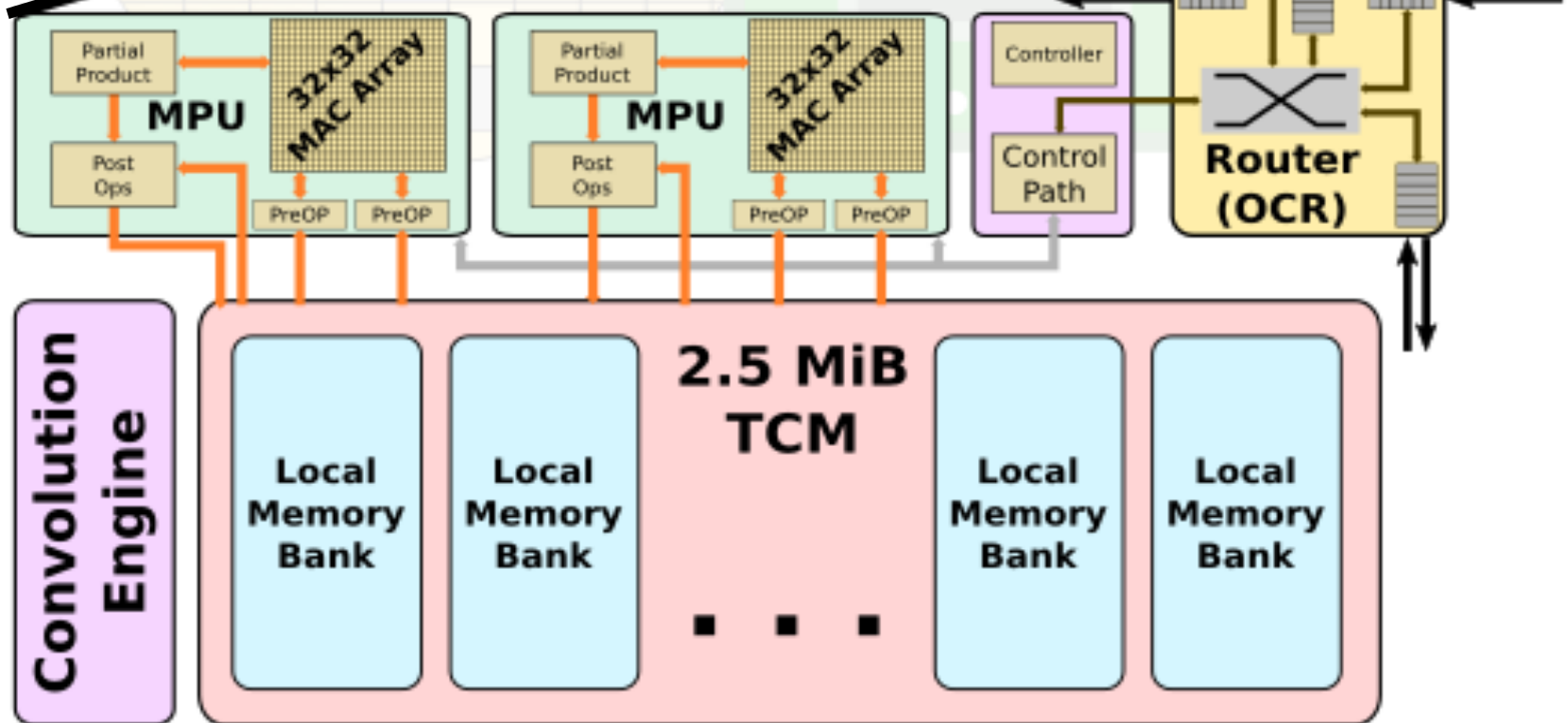
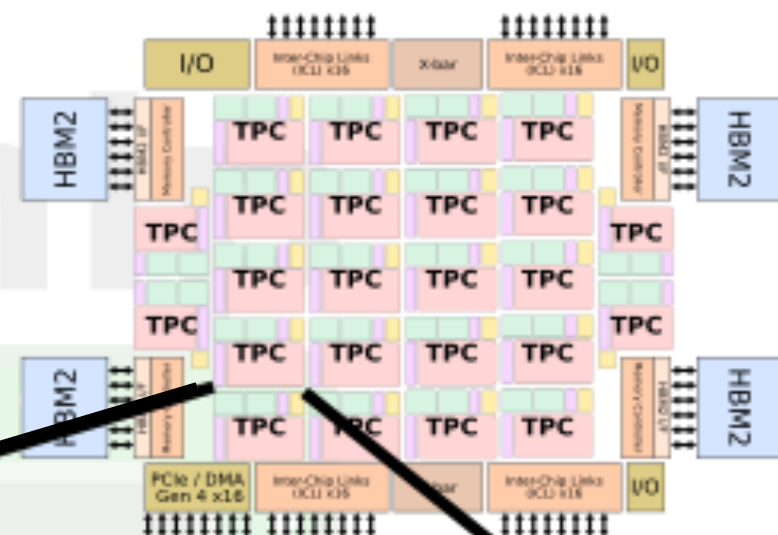
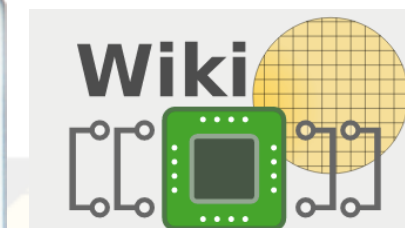
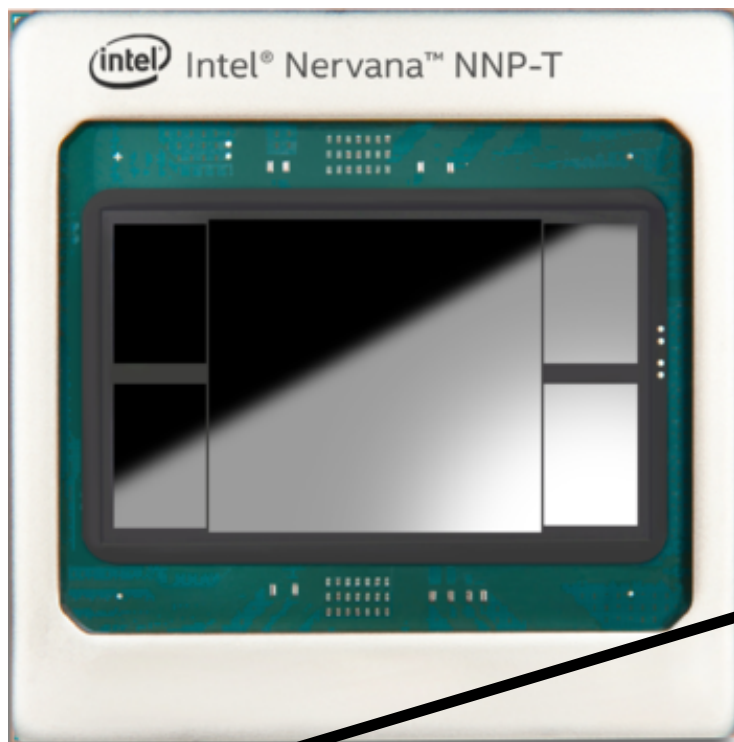
Intel NNPs : Training (NNP-T) and Inference (NNP-I)

SPRING CREST (NNP-T) SOC

- PCIe Gen 4 x16 EP
- 4x HBM2
- 64 lanes SerDes
- 24 Tensor Processors
- Up to 119 TOPS
- 60 MB on-chip distributed memory
- Management CPU and Interfaces
- 2.5D packaging

Spring Crest (NNP-T) SoC





Intel Inference neural processor (NNP-I)

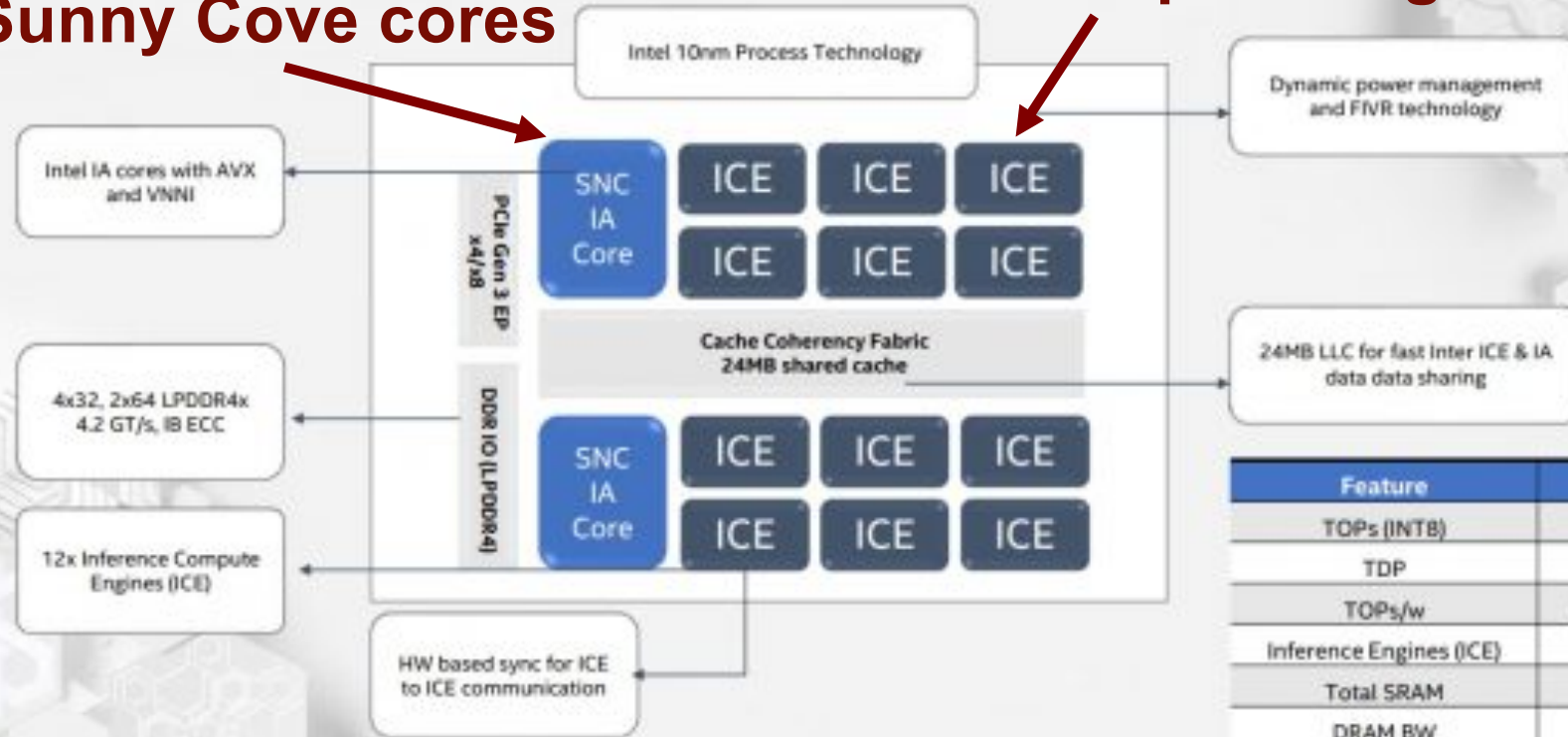


INTRODUCING - SPRING HILL (NNP-I)



Sunny Cove cores

Inference Compute Engines



Feature	SoC
TOPs (INT8)	48 - 92
TDP	10-50w
TOPs/w	2.0-4.8 TOPs/w
Inference Engines (ICE)	10-12 ICEs
Total SRAM	75MB
DRAM BW	68 GB/s

Intel doesn't expect the NNP-I to come to the retail market



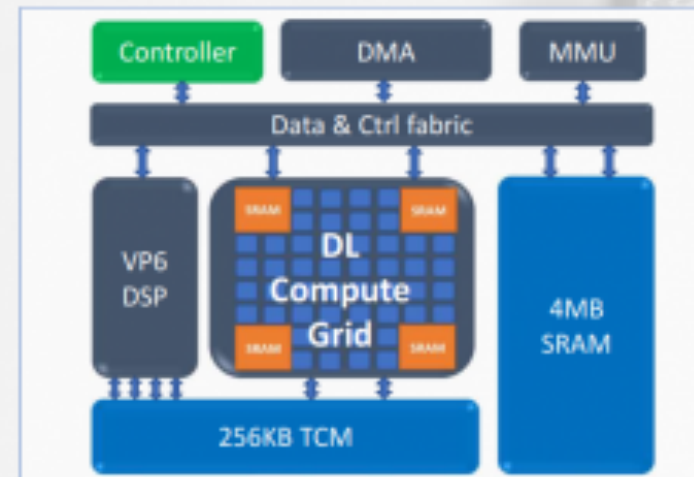
NNP-I: the ICE



ICE - INFERENCE COMPUTE ENGINE

- Deep Learning compute grid
 - 4K MAC (int8) per cycle
 - Scalable support: FP16, INT8, INT 4/2/1
 - Large internal SRAMs for power efficiency
 - Non-linear ops & Pooling
- Programmable vector processor
 - High throughput: 5 VLIW 512b
 - Extended NN support (FP16/16b/8b)
- High BW data memory access
 - Dedicated DMA – optimized for DL
 - Compression/decompression unit- support for sparse weights
- Large Local SRAM
 - Persistent data and/or storage of temporal data

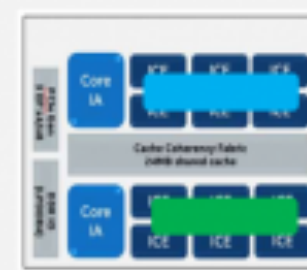
The tensor engine

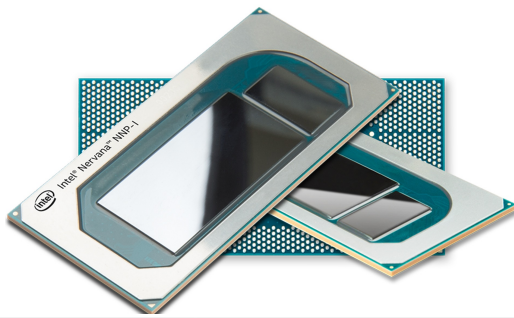


Optimized for throughput
batch: 4x6 (or 1x12)

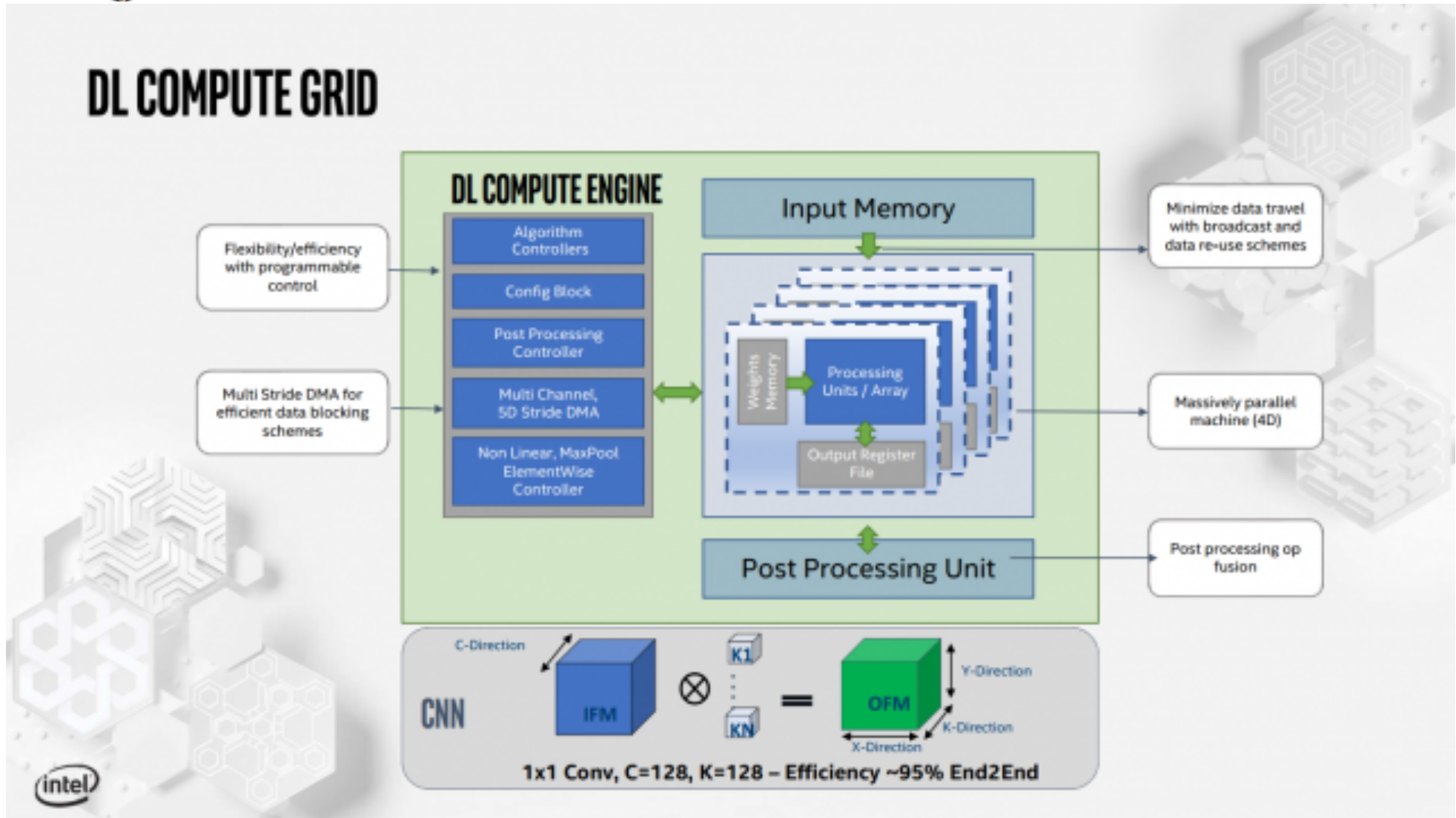


Optimized for latency
batch: 1x2





NNP-I: the Deep Learning Compute Grid

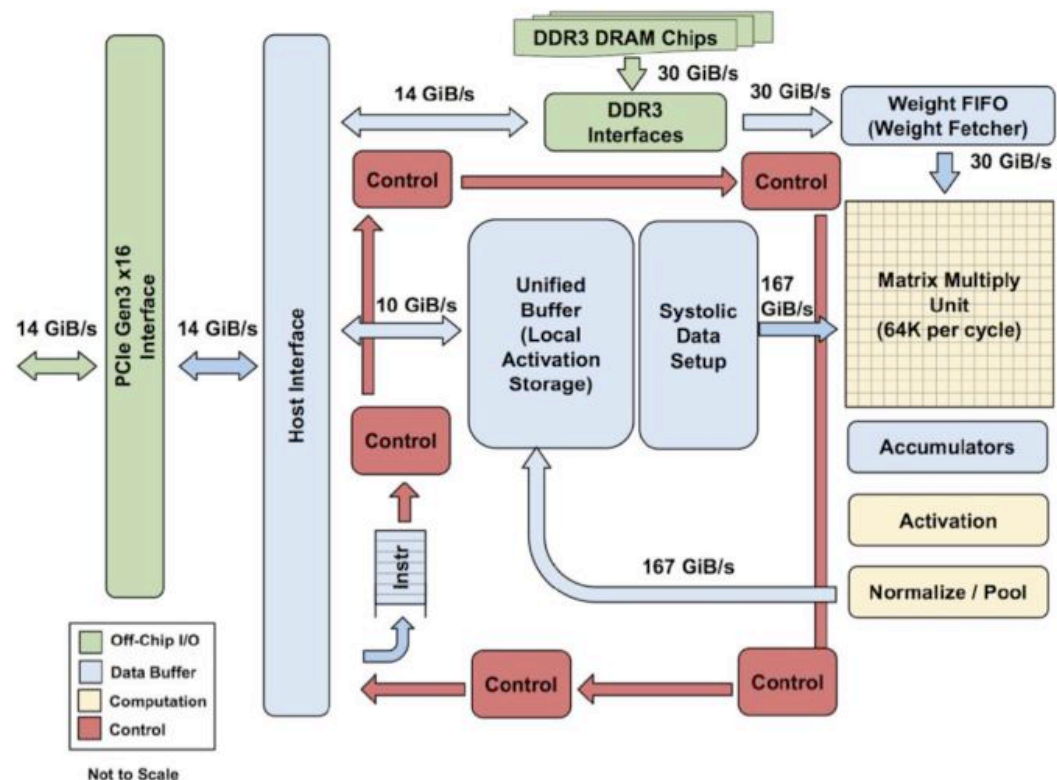


Google Tensor Processing Unit, TPU (May'16)



- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
 - $65,536 * 2 * 700M$
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU
- 4 MiB of on-chip Accumulator memory
- 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much on-chip memory vs GPU
- Two 2133MHz DDR3 DRAM channels
- 8 GiB of off-chip weight DRAM memory

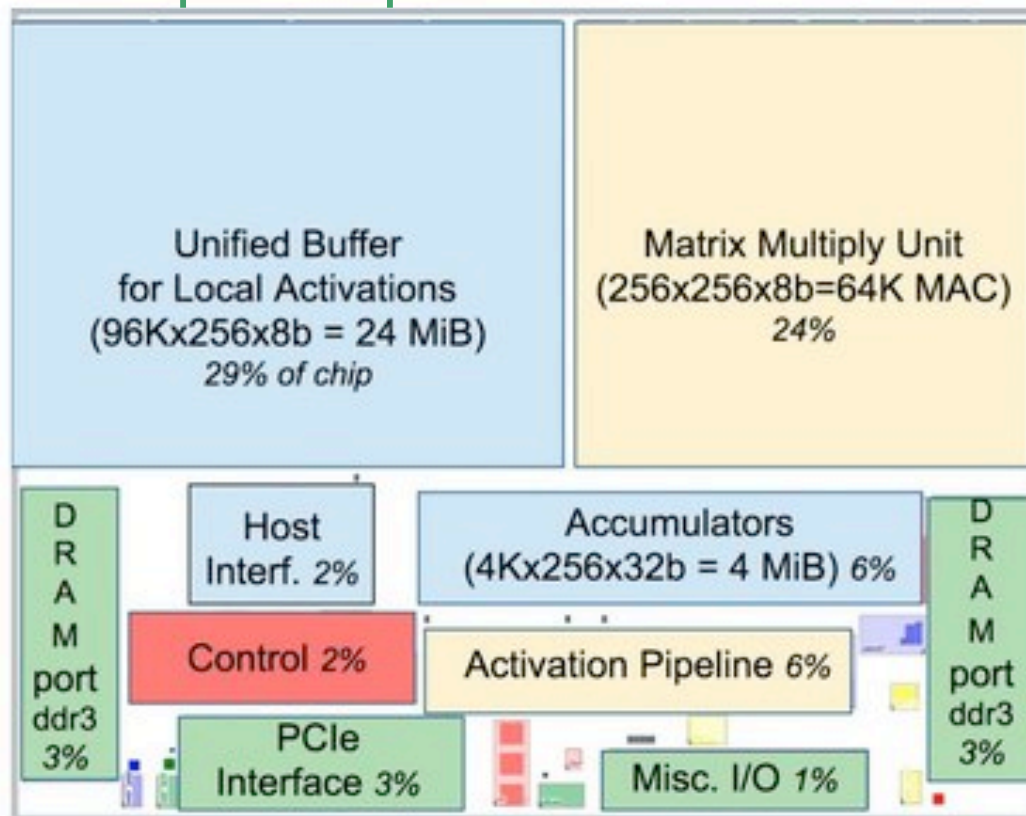
TPU: High-level Chip Architecture



Google Tensor Processing Unit, TPU (May'16)



Chip floor plan



TPU: a Neural Network Accelerator Chip



TPUs are intensively used by Google, namely in Google Photos, RankBrain, StreetView & Google Translate

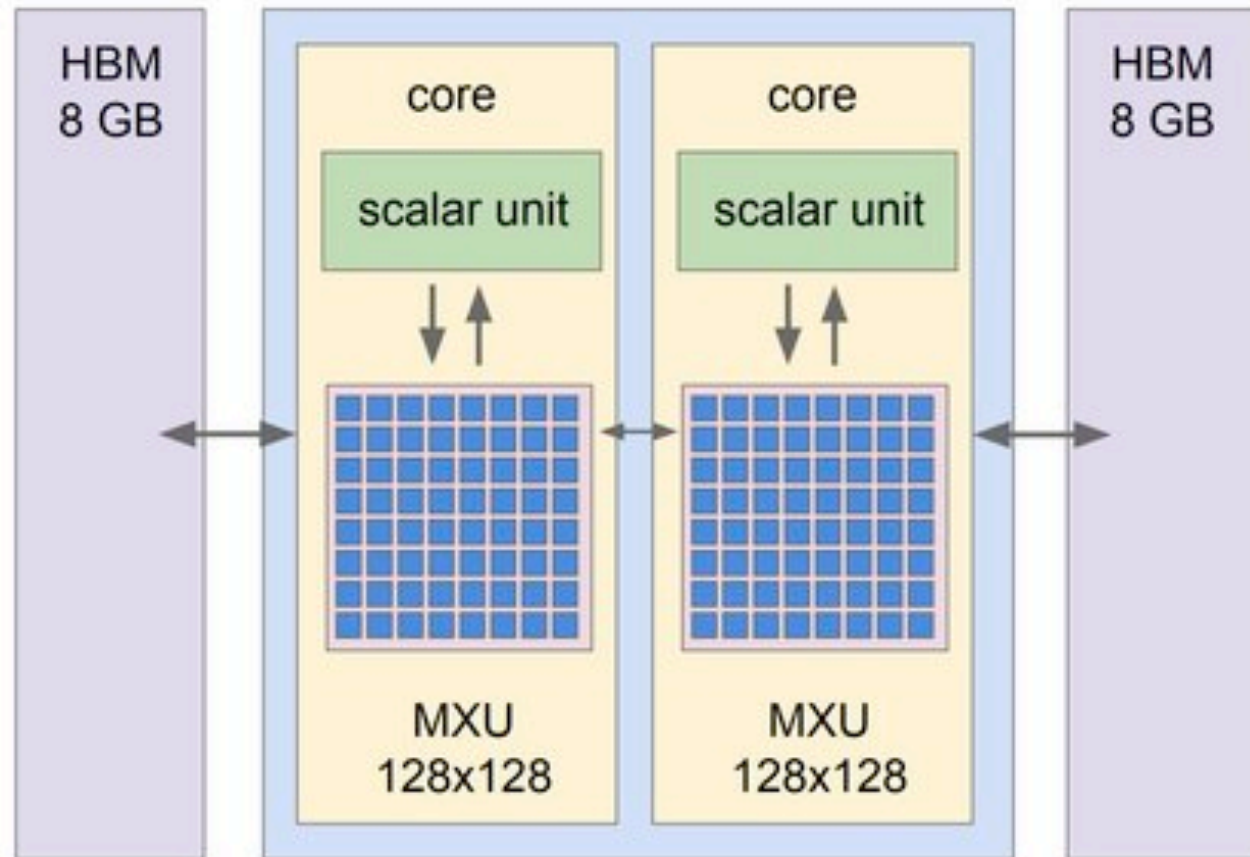
Google TPUv2 (May'17)



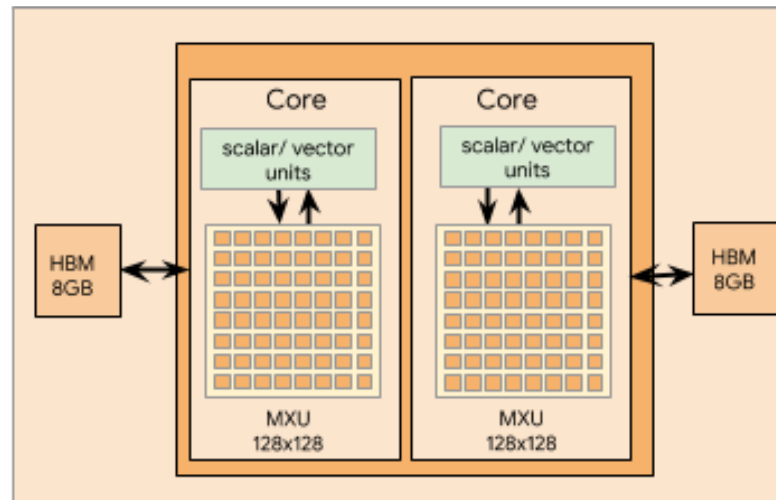
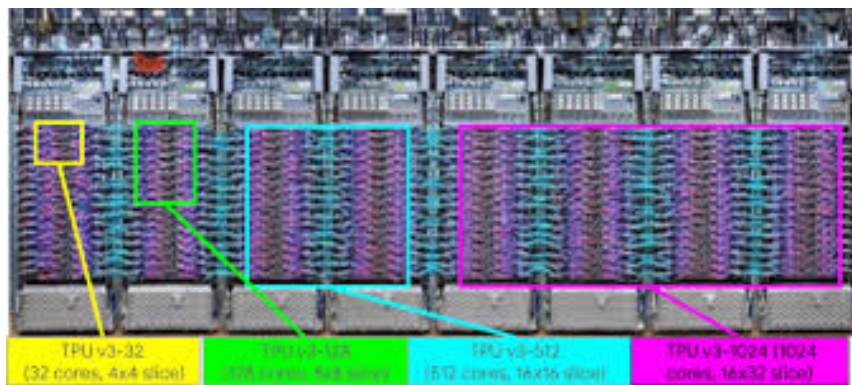
TPUv2 Chip



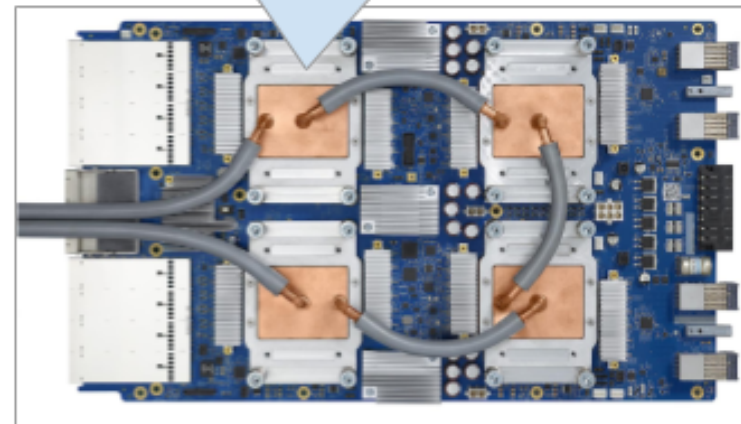
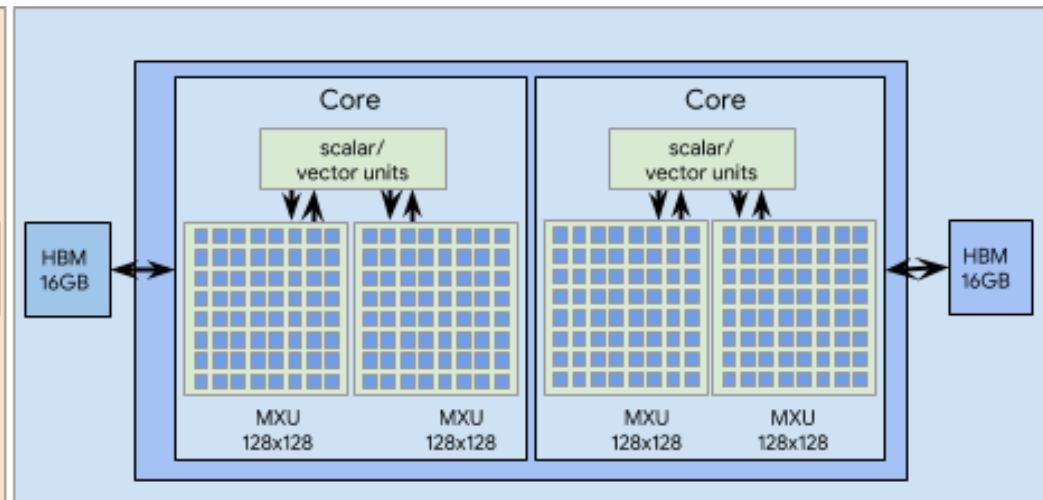
- 16 GB of HBM
- 600 GB/s mem BW
- Scalar unit: 32b float
- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS



Google TPUv3 (May'18)

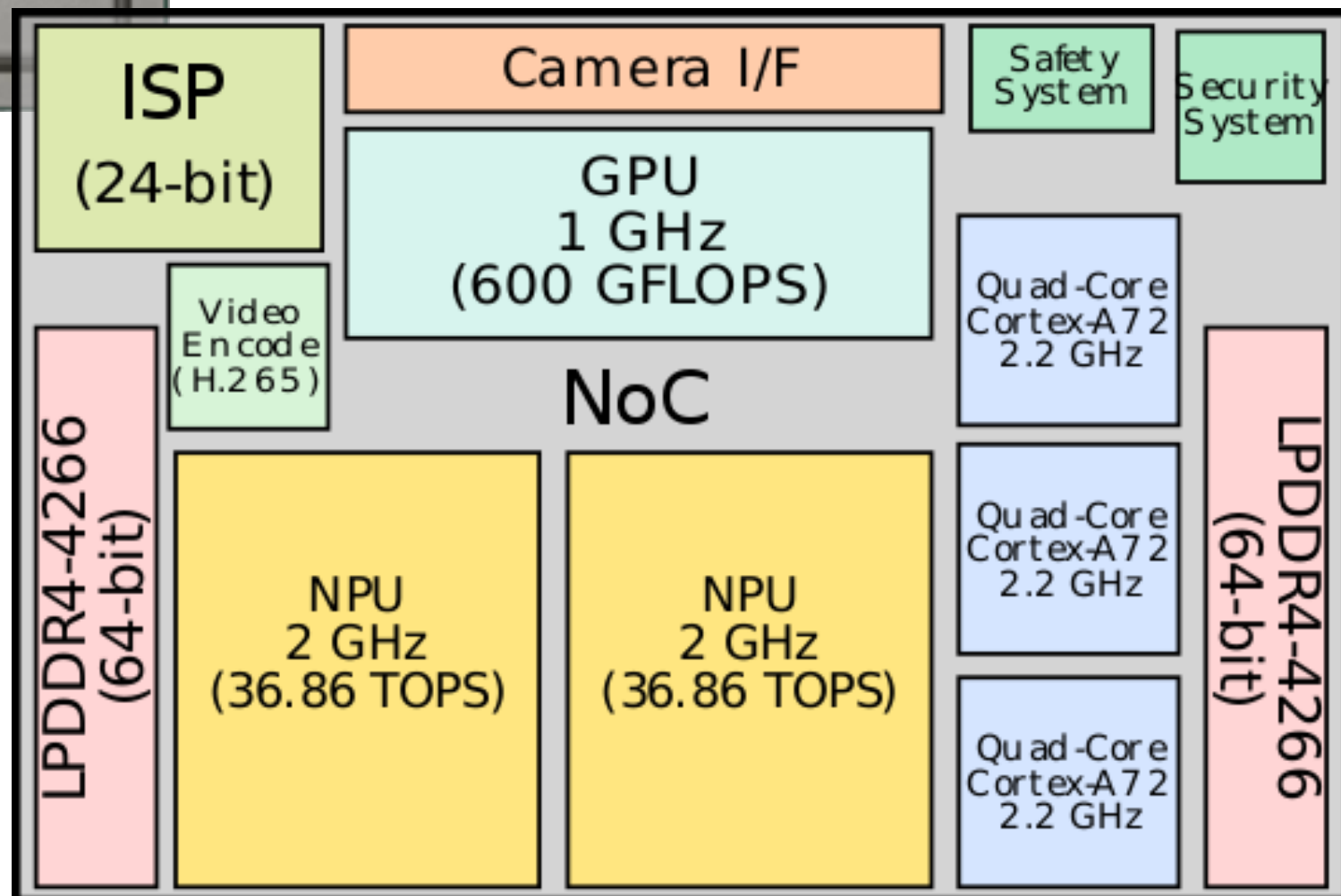


TPU v2 - 4 chips, 2 cores per chip



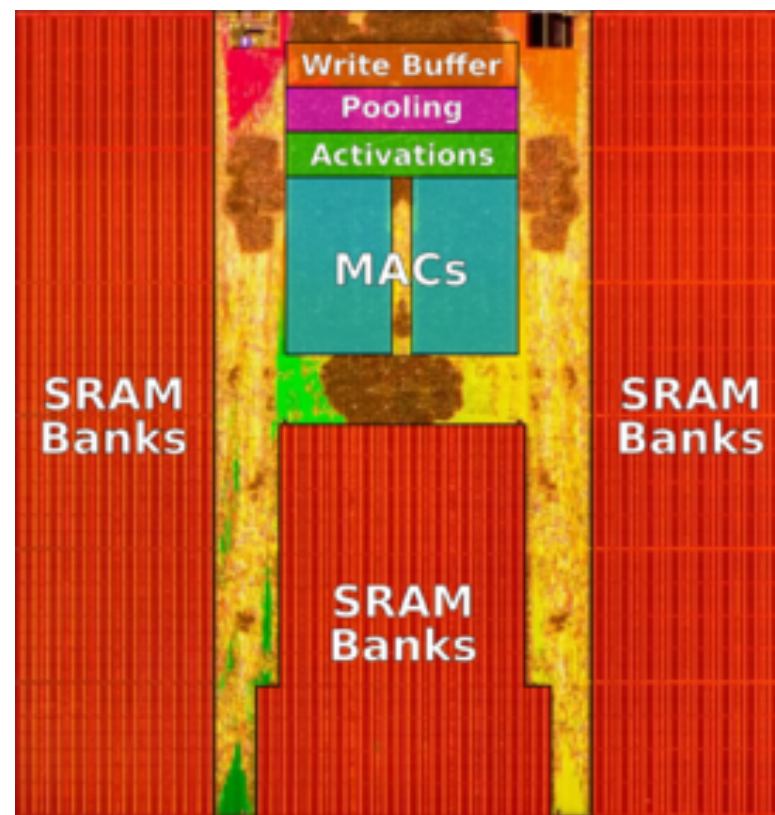
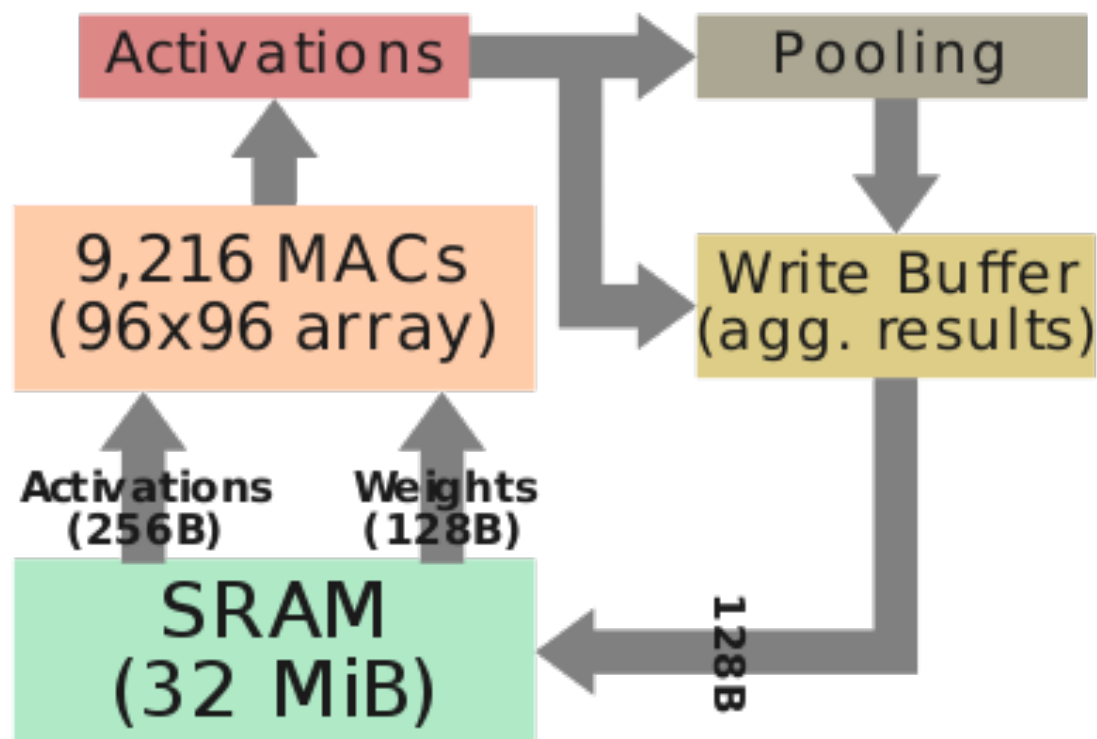
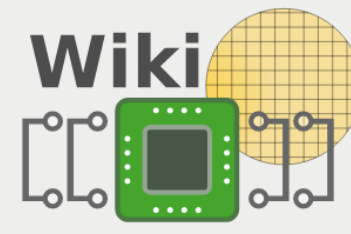
TPU v3 - 4 chips, 2 cores per chip

Tesla Full Self-Driving chip (FSD)





The Neural Processing Unit in FSD



Competitors with neural accelerators: Amazon Web Services (AWS)



— AWS Inferentia Machine Learning Processor —

**Announced
Nov'18**

**No more
details...**



Competitors with neural accelerators: Wujian for Alibaba Aliyun cloud

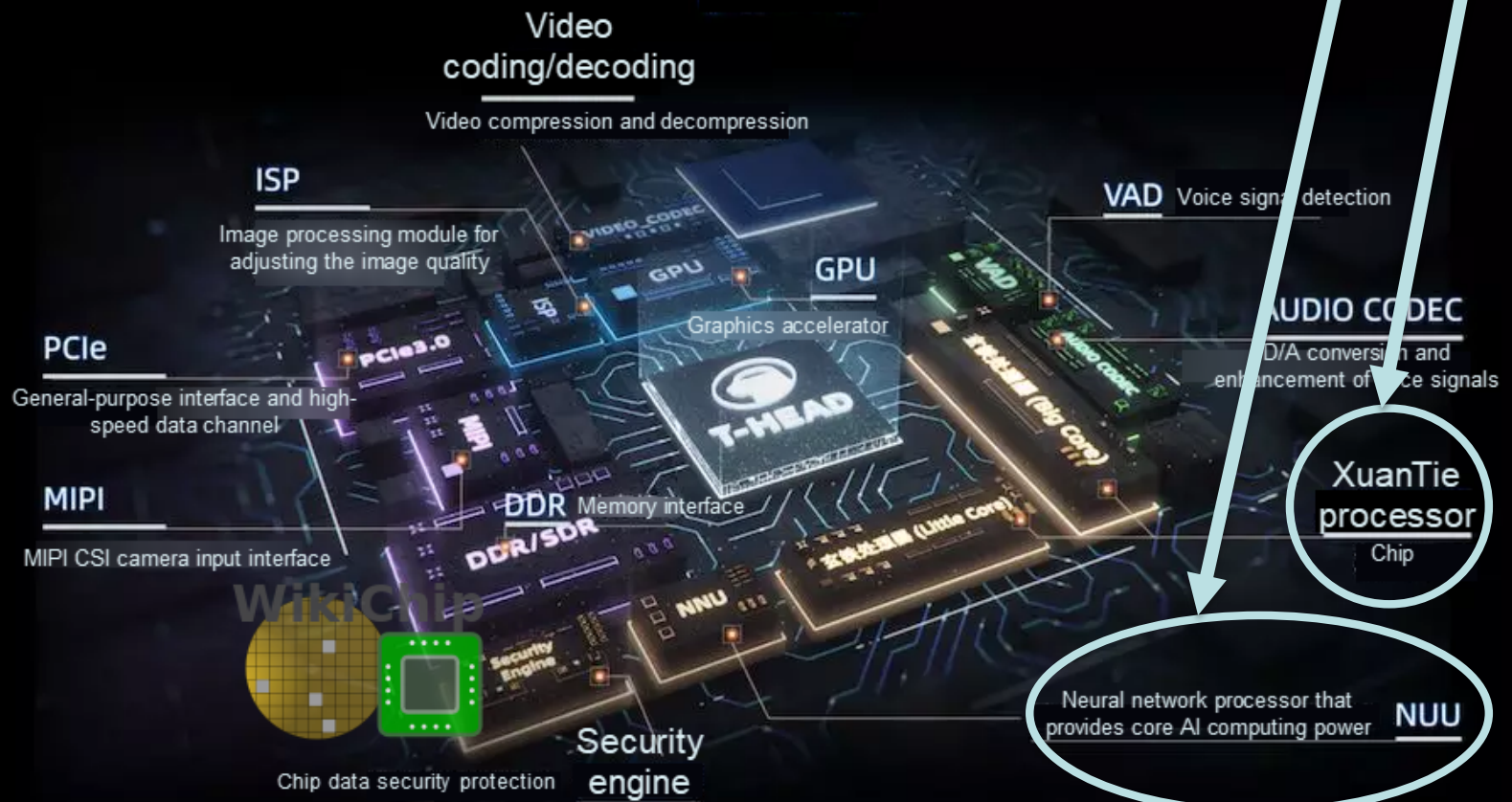


Pingtougou Pingtougou Releases Its One-stop Chip Platform

Wujian

**Wujian
SoC
platform**

**Announced
March'19**



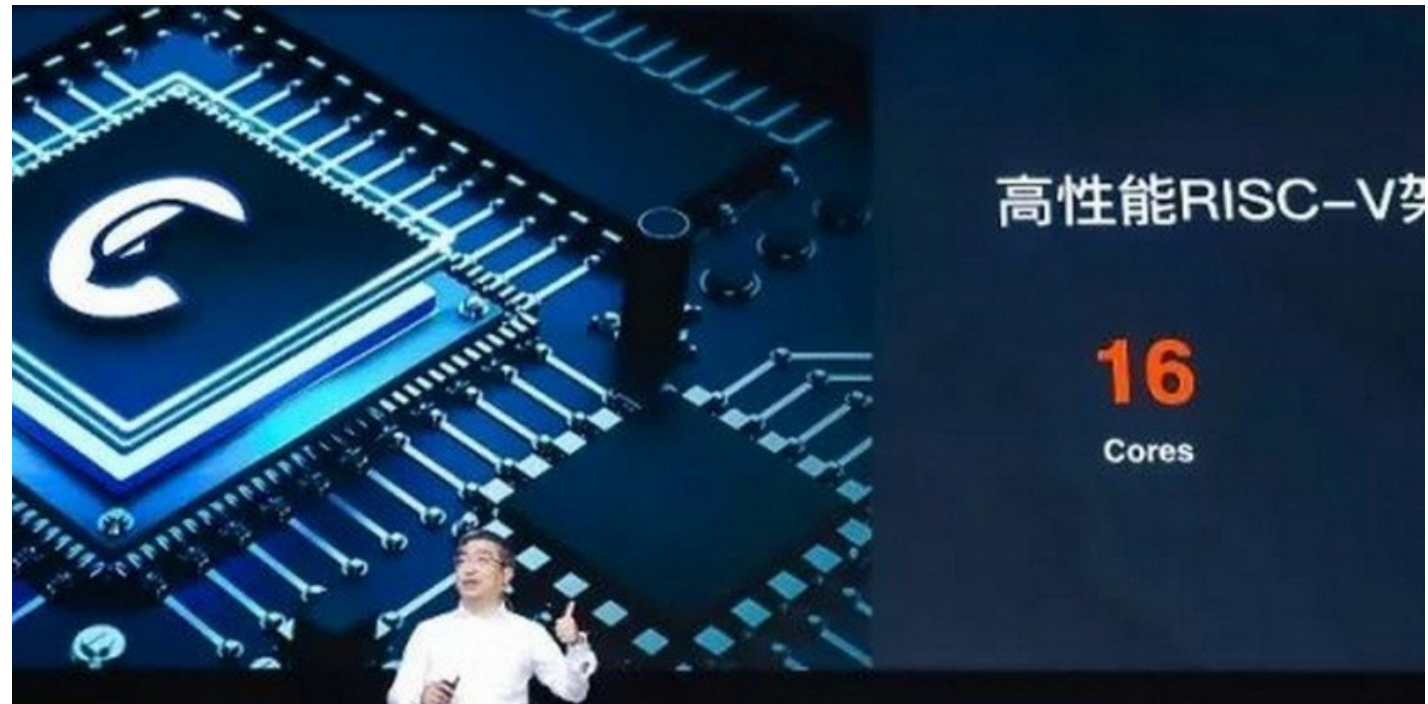
A Quality Chip

The chip architecture shown in the picture is a full-stack integration of the XuanTie CPU, storage, high-speed interface, voice, visual and some other features.

The platform allows chip enterprises to design customized chips and reduce the chip design cycle and cost by 50%.

Wujian from Alibaba : the 16-core XuanTie & the NUU

**16-core
RISC-V
XuanTie**



**NPU
Hanguang 800**





Cerebras Wafer Scale Engine (WSE): *the largest chip ever built*



46,225 mm² chip

56x larger than the biggest GPU ever made

400,000 core

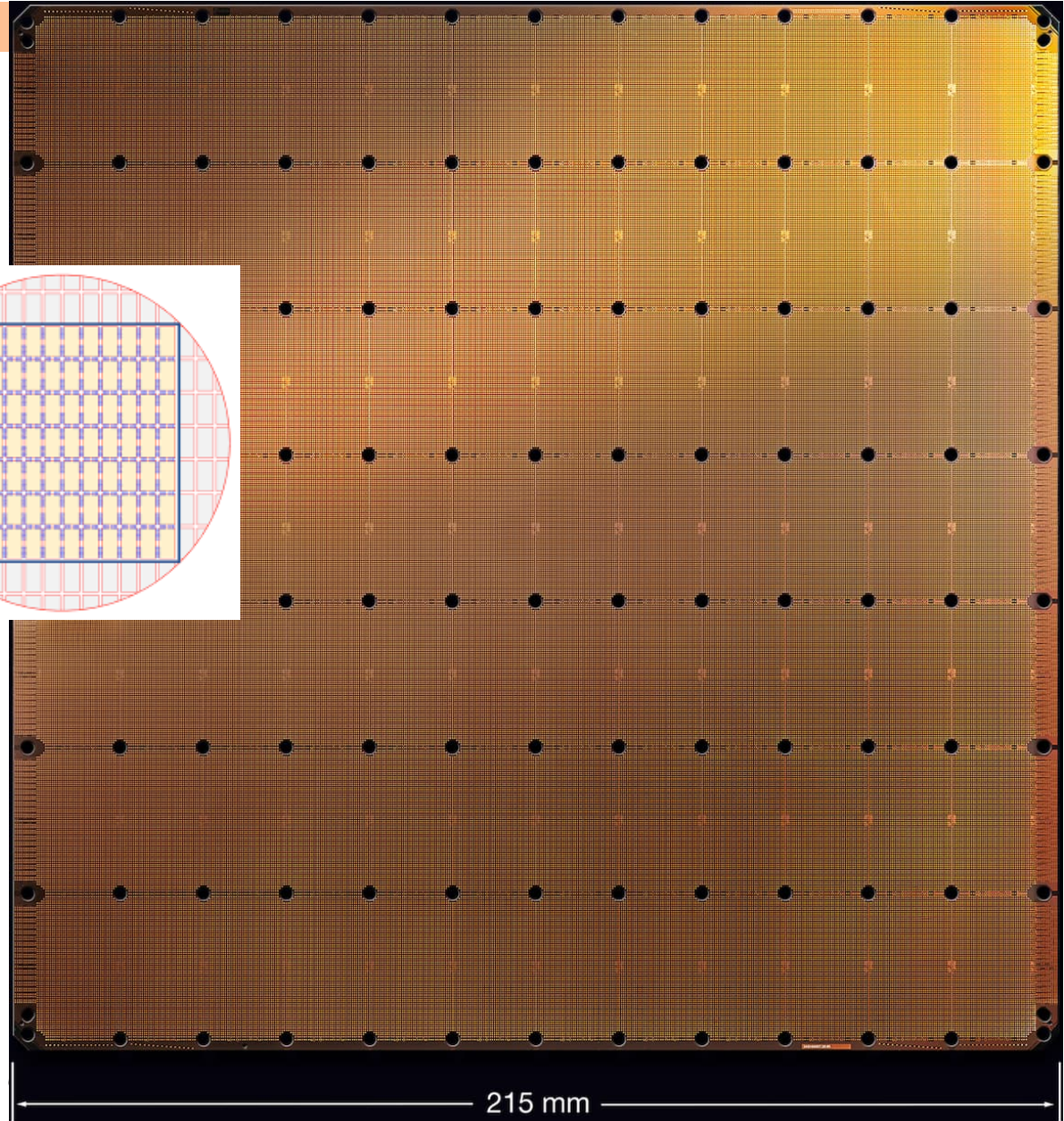
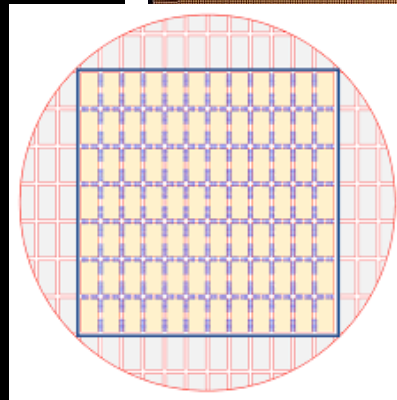
78x more cores

18 GB on-chip SRAM

3000x more on-chip memory

100 Pb/s interconnect

33,000x more bandwidth



**400,000
Sparse Linear Algebra
(SLA) cores
(AI-optimized)**

AJProença, Advanced Architectures, MiEI,



Packing the Cerebras WSE: clusters of CS-1s (Argonne National Lab)



SOFTWARE PLATFORM

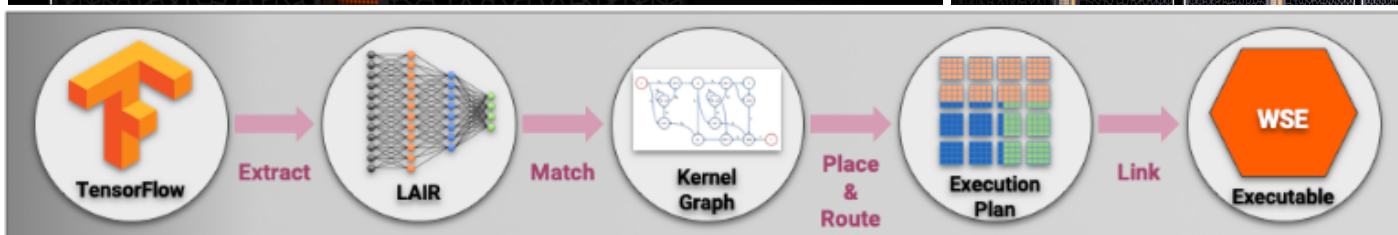
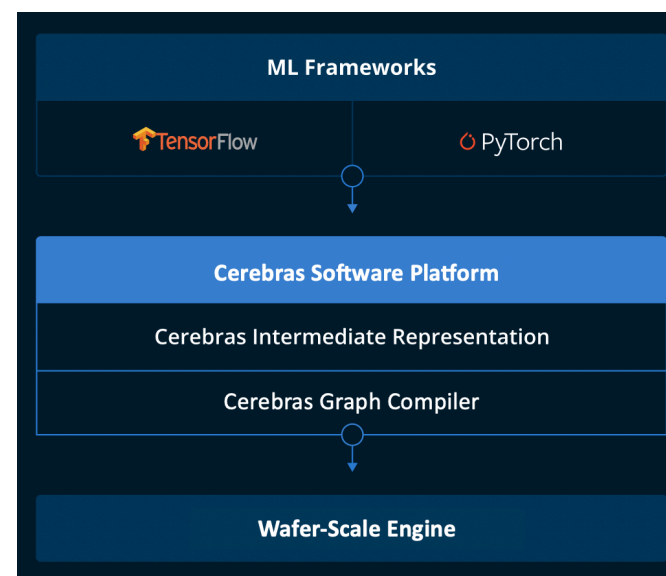
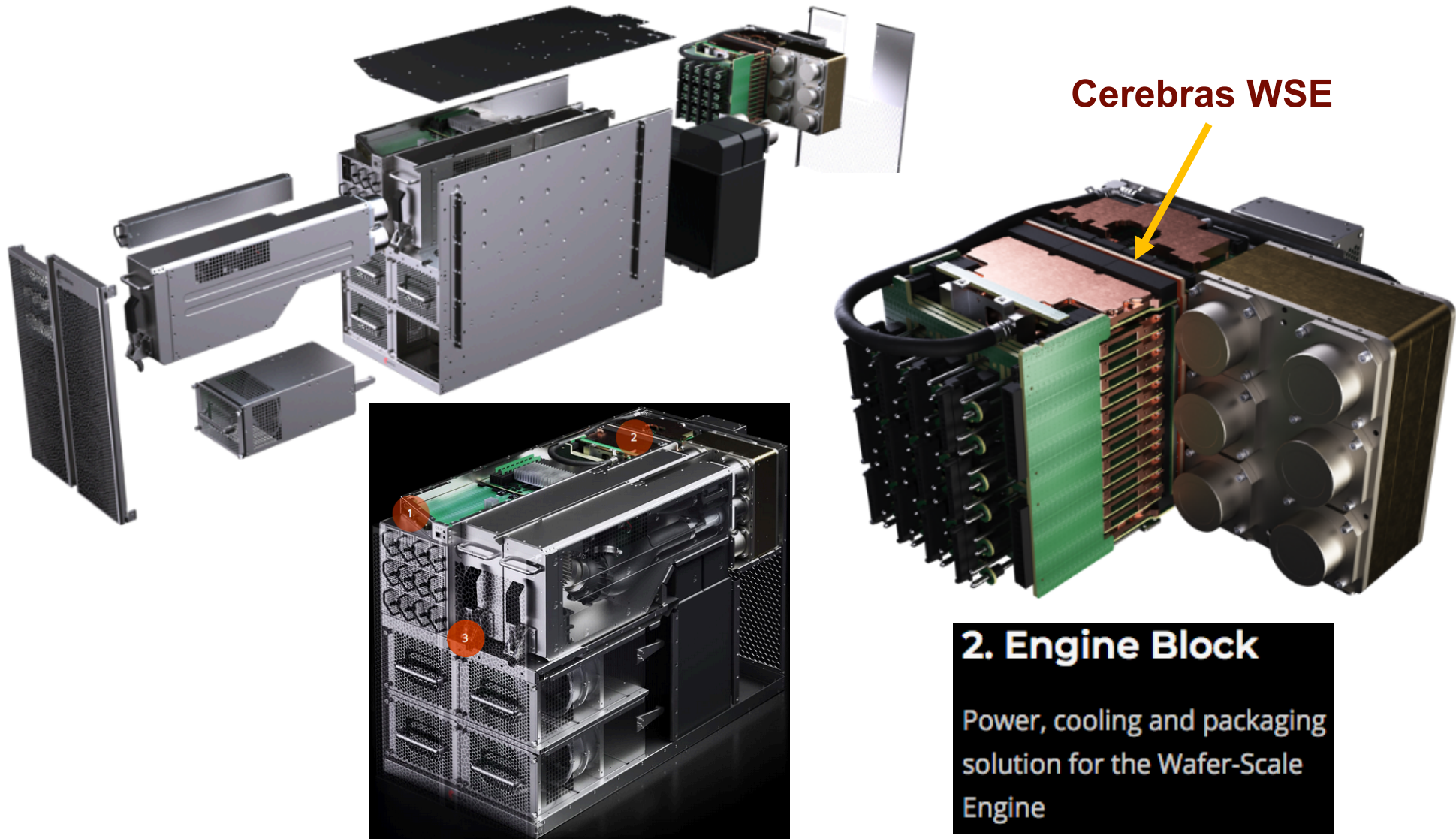


Figure 8: A high-level overview of the compilation process for the WSE



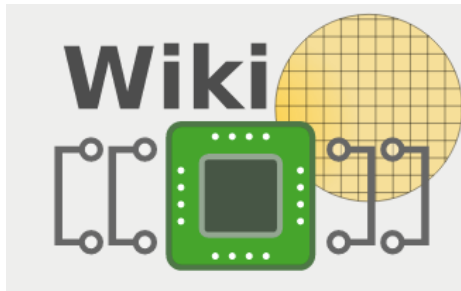
The CS-1 internal structure



Beyond Vector/SIMD architectures



- Vector/SIMD-extended architectures are hybrid approaches
 - mix (super)scalar + vector op capabilities on a single device
 - highly pipelined approach to reduce memory access penalty
 - tightly-closed access to shared memory: lower latency
- Evolution of Vector/SIMD-extended architectures
 - **PU (Processing Unit) cores with wider vector units**
 - x86 many-core: Intel MIC / Xeon KNL
 - others: IBM Power BlueGene/Q Compute, ShenWay 260, Matrix-2000, A64FX Arm
 - **coprocessors (require a host scalar processor): accelerator devices**
 - on disjoint physical memories (e.g., Xeon KNC with PCI-Expr, PEZY-SC)
 - ISA-free architectures, code compiled to silica: FPGA
 - focus on SIMT/SIMD to hide memory latency: GPU-type approach
 - tensor/neural net cores: NVidia Volta, Google TPU, Intel NNP
 - neuromorphic computing: IBM TrueNorth, Intel Loihi
 - **heterogeneous PUs in a SoC: multicore PUs with GPU-cores**
 - ...



Neuromorphic landscape: IBM TrueNorth, Intel Loihi,

Neuromorphic Chip

A **neuromorphic chip** is any **integrated circuit** that attempts to imitate the function of a biological brain by mimicking things such as cranial nerves, neurons, and synapses.

RECENT NEUROMORPHIC CHIPS

System	TrueNorth	Loihi	SpiNNaker 2
Process	28 nm	14 nm	22 nm (FDX)
Transistors	5.4 Billion	2.07 Billion	
Neuromorphic Cores	4,096	128	144
Neurons	1 Million	131 Thousand	
Synapses	256 Million	130 Million	

The IBM TrueNorth chip array (August'2014)

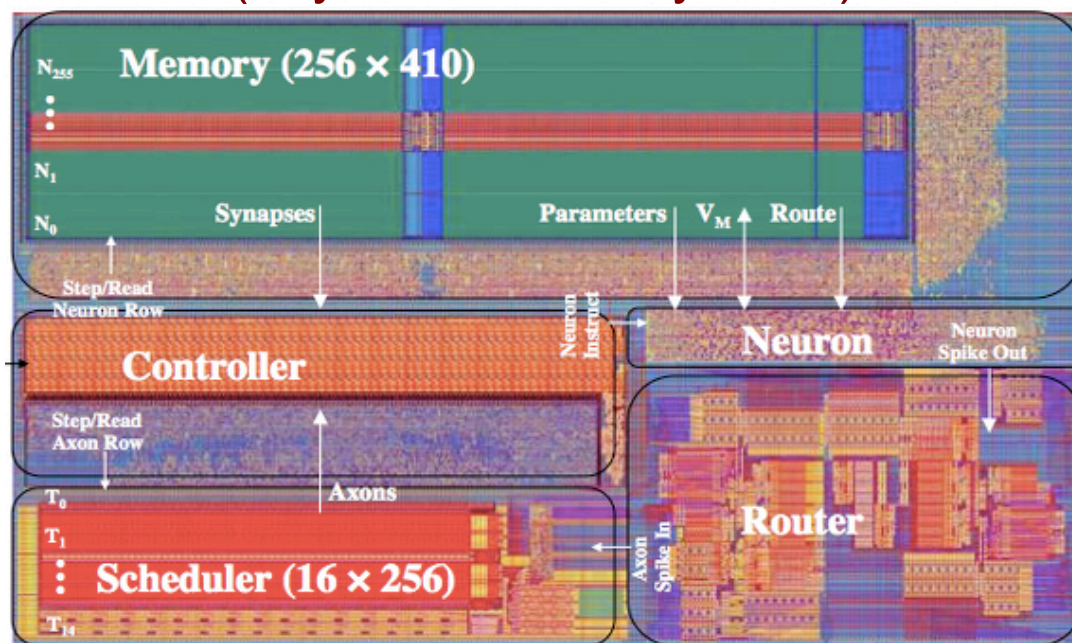
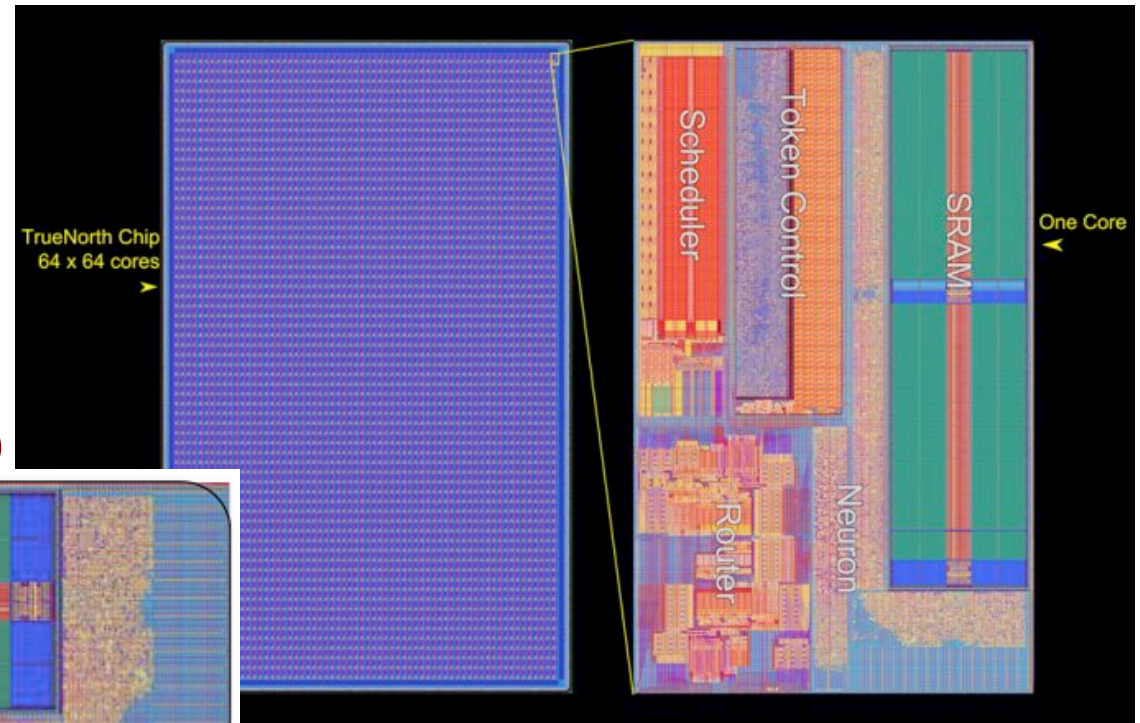


TrueNorth Chip:

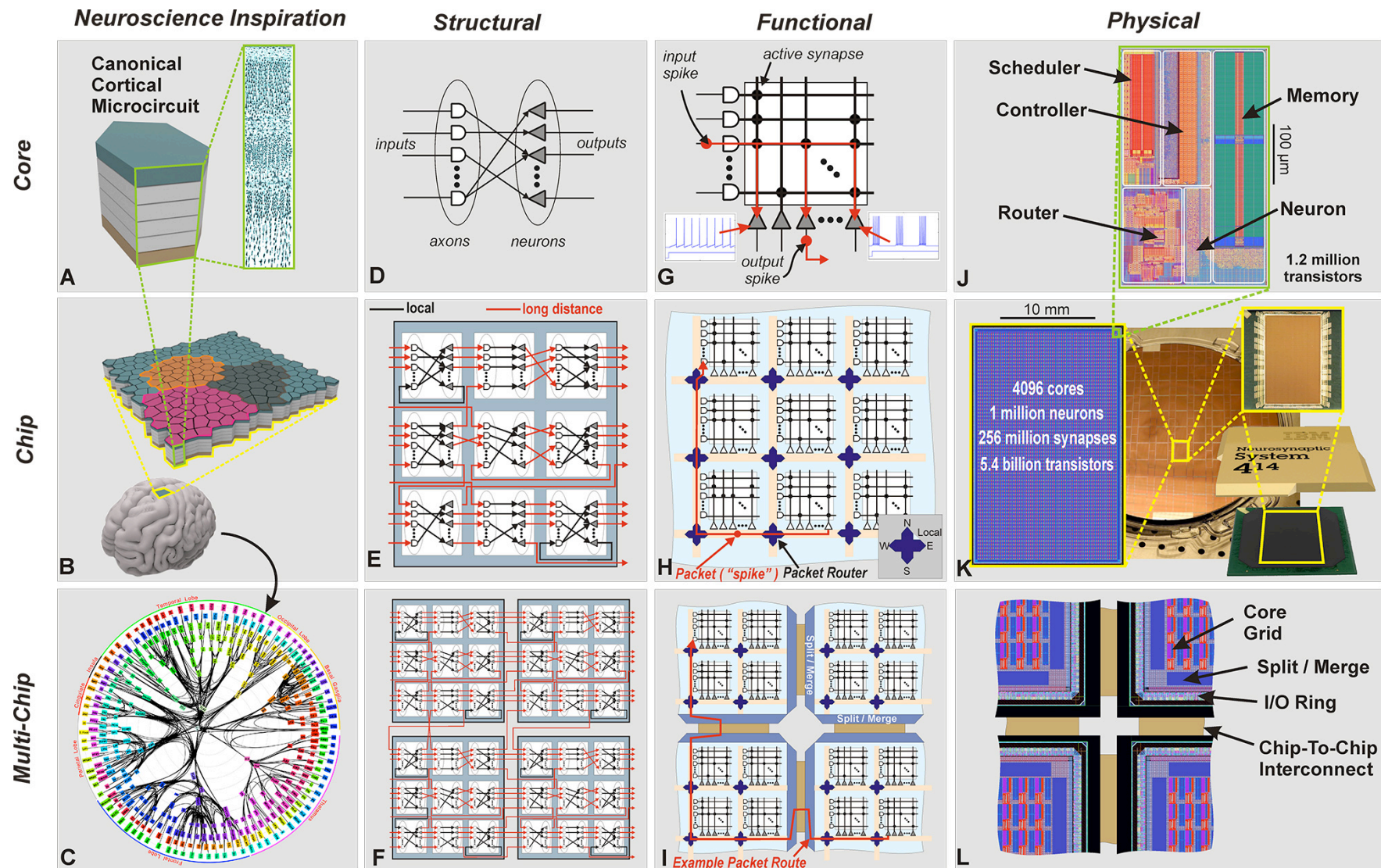
- 4096 neurosynaptic cores

Each core:

- 256 inputs (axons)
- 256 outputs (neurons)
- RAM w/ data for each neuron
- router (any neuron to any axon)

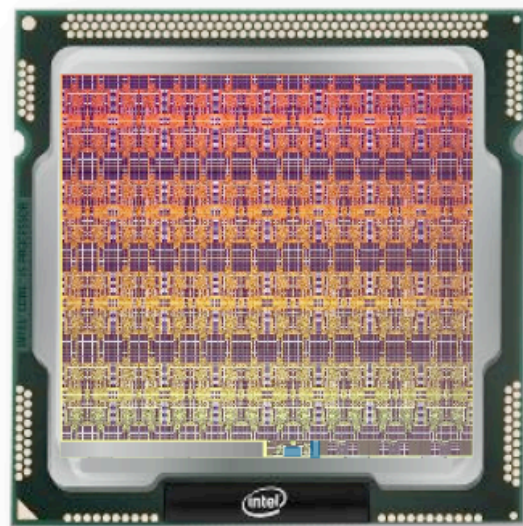


The IBM TrueNorth architecture





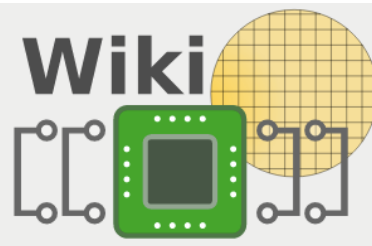
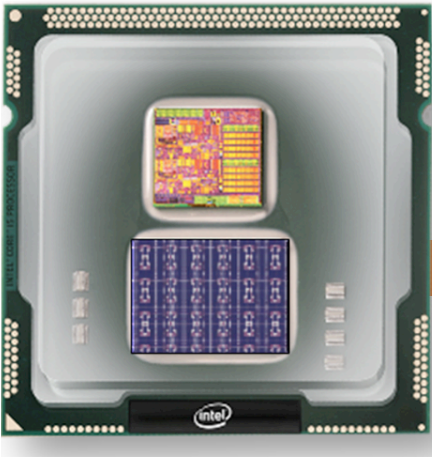
Loihi at a Glance



**Integrated
Memory + Compute
Neuromorphic Architecture**

Key Properties

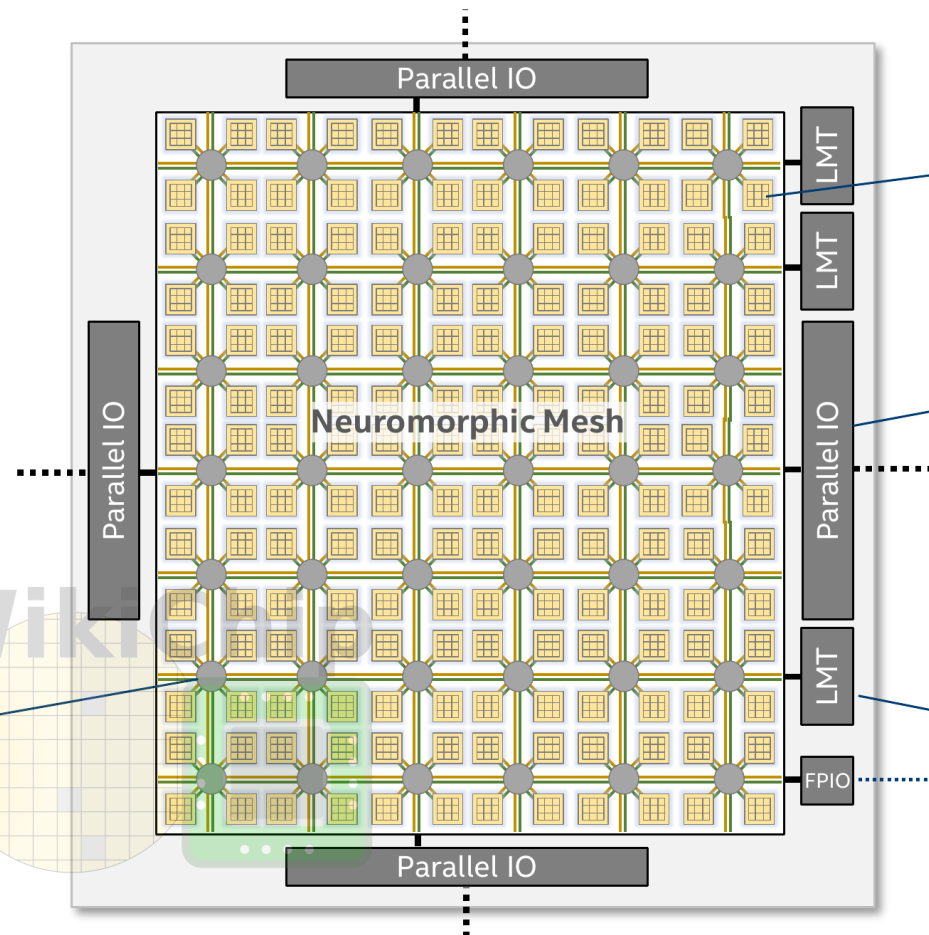
- 128 neuromorphic cores supporting up to 128k neurons and 128M synapses with an **advanced SNN feature set**.
- **Scalable on-chip learning** capabilities to support a range of learning paradigms (unsupervised, supervised, reinforcement-based, and others)
- Supports **highly complex neural network topologies** (up to 2000-way fan-out between neurons)
- Fully digital **asynchronous** implementation
- Fabricated in Intel's **14nm FinFET process** technology



Intel Loihi chip

Chip Architecture

Technology:	14nm
Die Area:	60 mm ²
Core area:	0.41 mm ²
NmC cores:	128 cores
x86 cores:	3 LMT cores
Max # neurons:	128K neurons
Max # synapses:	128M synapses
Transistors:	2.07 billion



Neuromorphic core

- LIF neuron model
- Programmable learning
- 128 KB synaptic memory
- Up to 1,024 neurons
- Asynchronous design

Parallel off-chip interfaces

- Two-phase asynchronous
- Single-ended signaling
- 100-200 MB/s BW

Embedded x86 processors

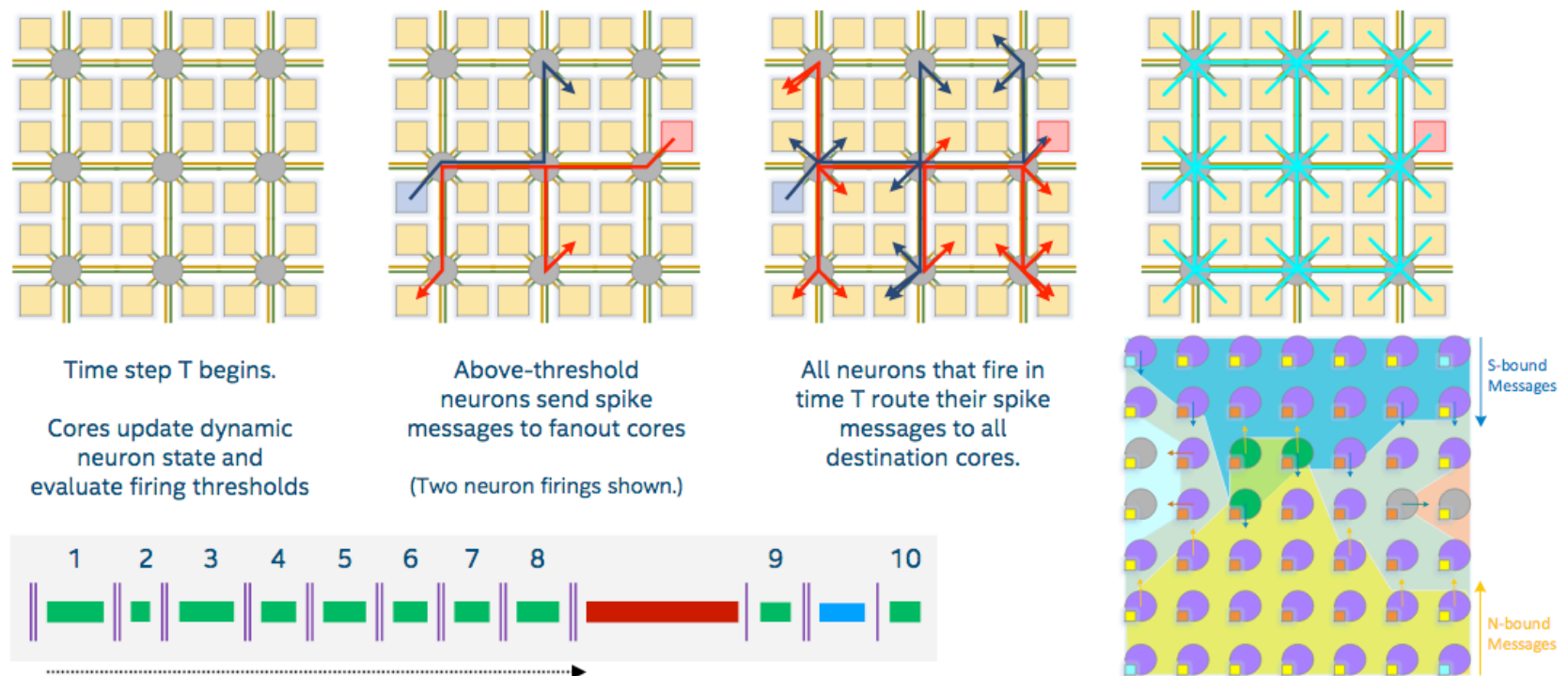
- Efficient spike-based communication with neuromorphic cores
- Data encoding/decoding
- Network configuration
- Synchronous design

Low-overhead NoC fabric

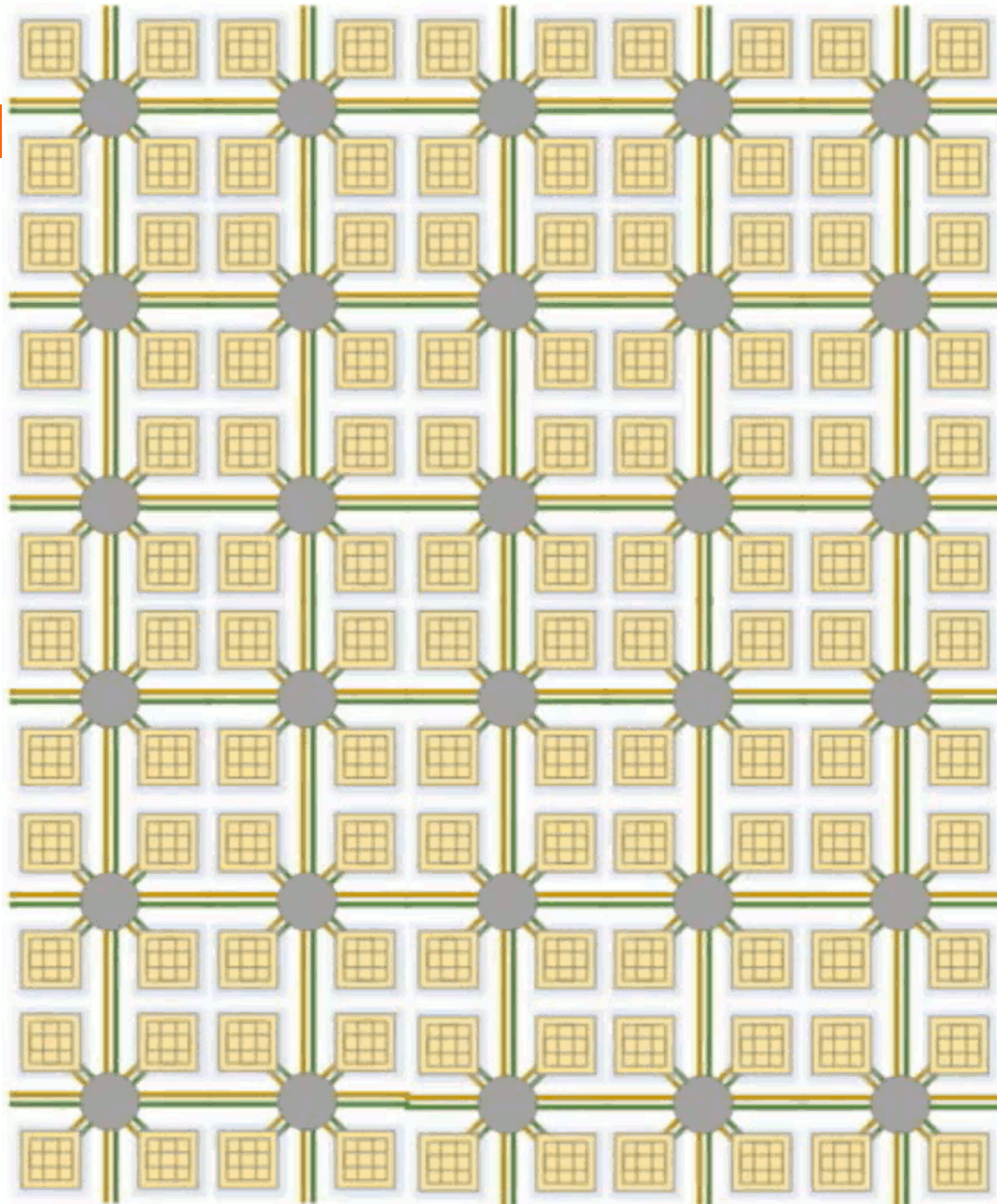
- 8x16-core 2D mesh
- Scalable to 1000's cores
- Dimension order routed
- Two physical fabrics
- 8 GB/s per hop



Mesh Operation



$t = 1$

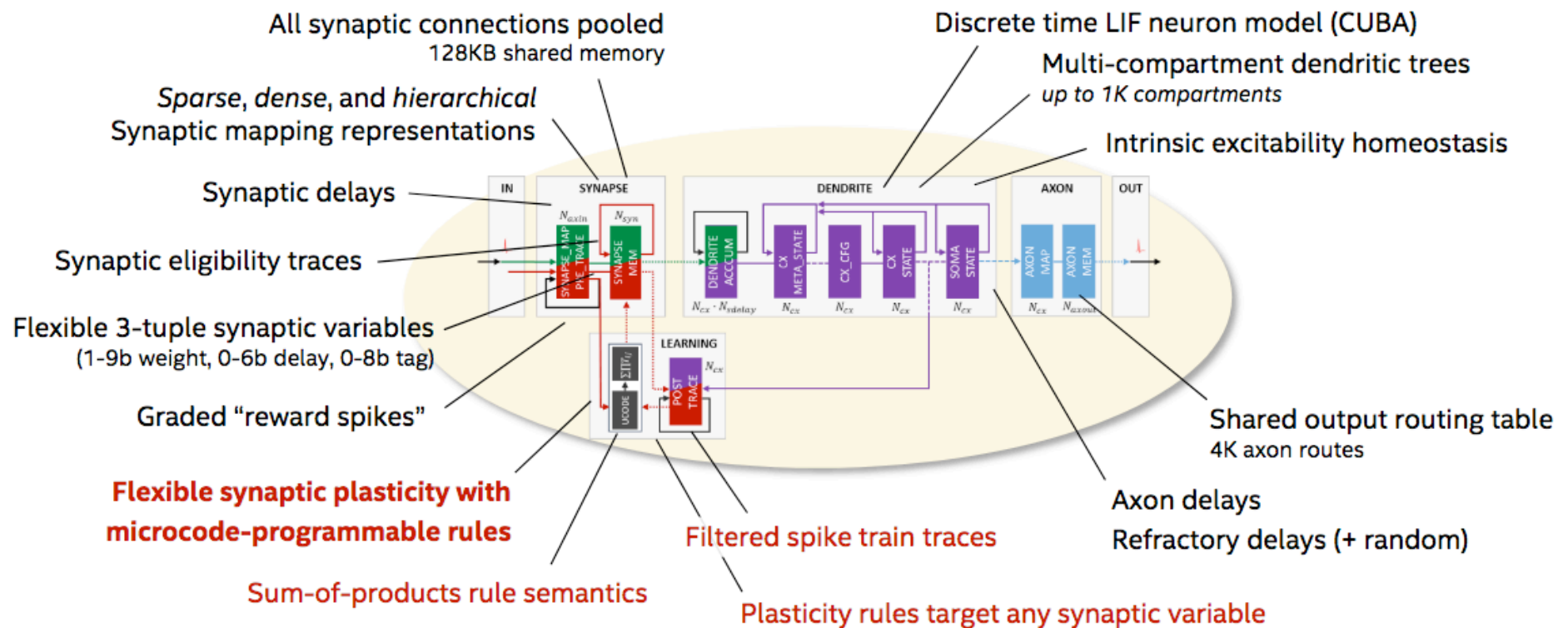


The Intel Loihi

Loihi (pronounced *low-ee-hee*) is a **neuromorphic** research test chip designed by **Intel** Labs that uses an asynchronous **spiking neural network** (SNN) to implement adaptive self-modifying event-driven fine-grained parallel computations used to implement learning and inference with high efficiency. The chip is a 128-neuromorphic cores **many-core** IC fabricated on Intel's **14 nm process** and features a unique programmable microcode learning engine for on-chip **SNN** training. The chip was formally presented at the 2018 Neuro Inspired Computational Elements (NICE) workshop in Oregon.



Neuromorphic Core Architecture

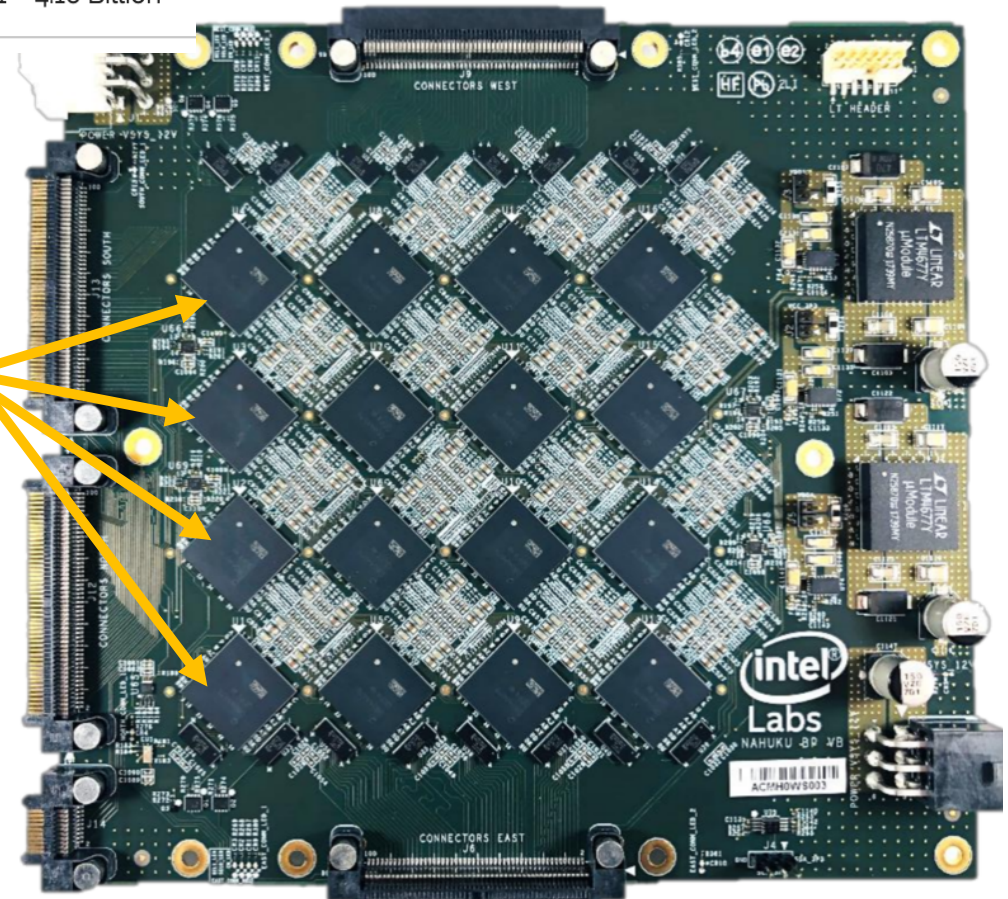


Systems with Intel Loihi

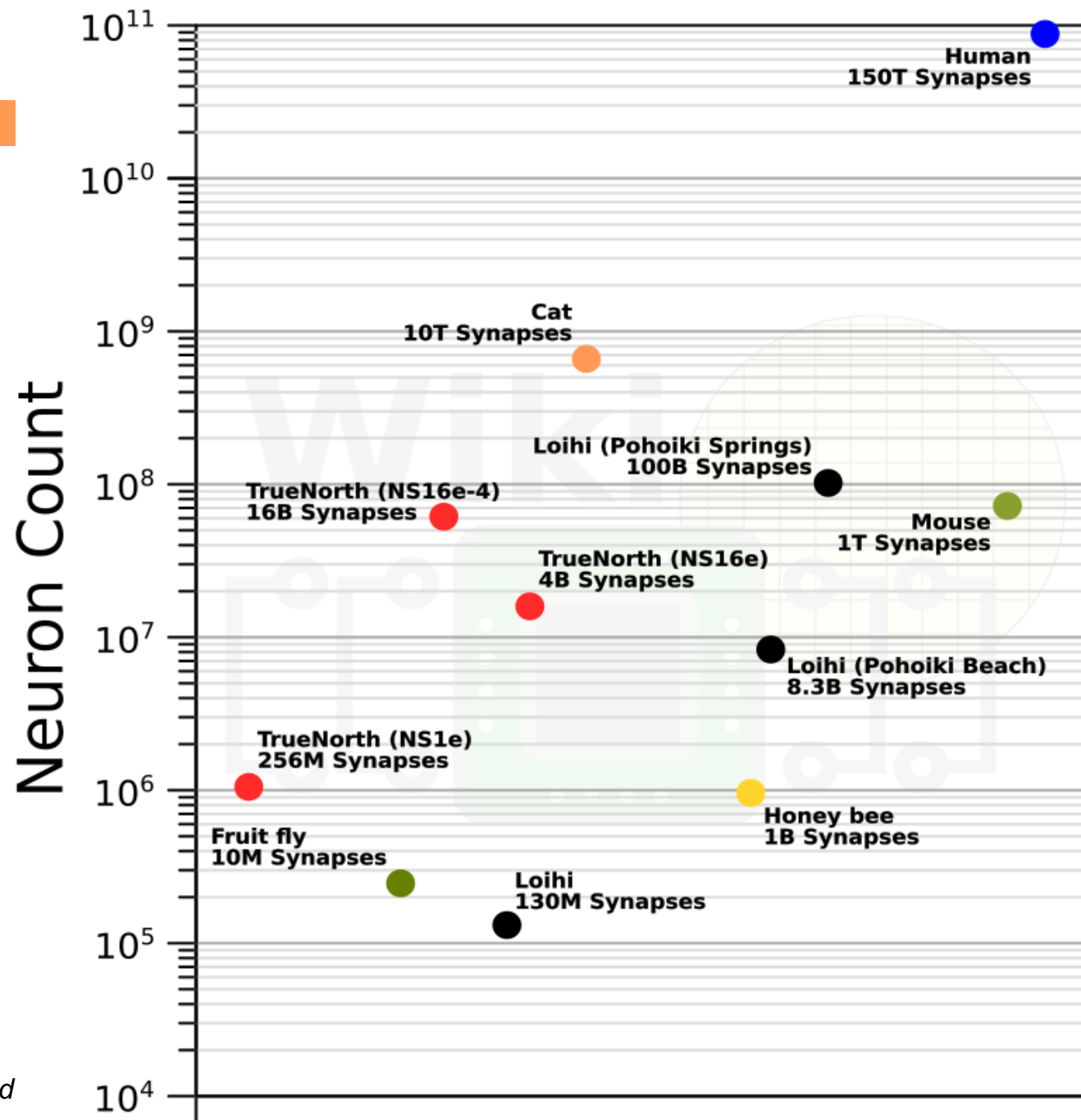
LOIHI-BOARDS

Board	Kapoho Bay	Wolf Mountain	Nahuku
Configuration	2 x Loihi	4 x Loihi	8-32 x Loihi
Neuromorphic Cores	256	512	1,024 – 4,096
Neurons	262K	524K	1 – 4.2 Million
Synapses	260 Million	520 Million	1 – 4.16 Billion

**Nahuku board:
16 Loihi chips
at each side
in a 4x4 mesh**



Neuron and Synapse Counts



Beyond Vector/SIMD architectures

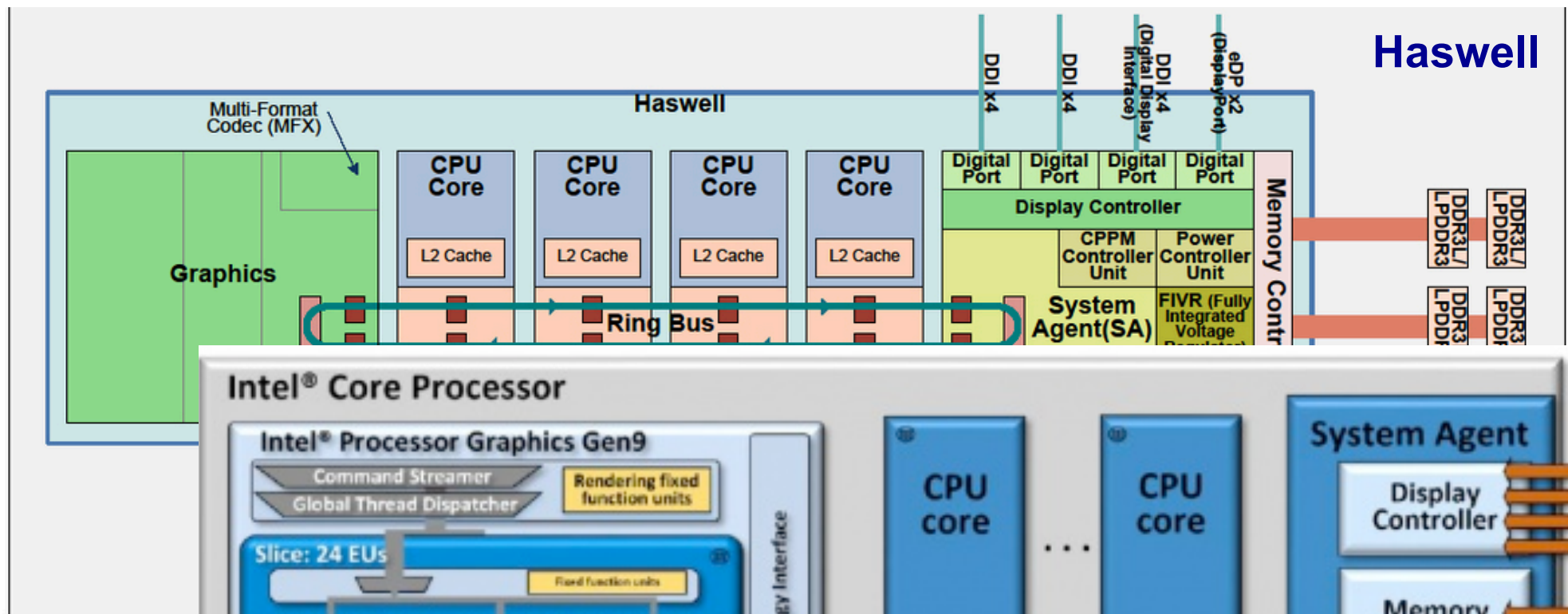


- Vector/SIMD-extended architectures are hybrid approaches
 - mix (super)scalar + vector op capabilities on a single device
 - highly pipelined approach to reduce memory access penalty
 - tightly-closed access to shared memory: lower latency
- Evolution of Vector/SIMD-extended architectures
 - **PU (Processing Unit) cores with wider vector units**
 - x86 many-core: Intel MIC / Xeon KNL
 - others: IBM Power BlueGene/Q Compute, ShenWay 260, Matrix-2000, A64FX Arm
 - **coprocessors (require a host scalar processor): accelerator devices**
 - on disjoint physical memories (e.g., Xeon KNC with PCI-Expr, PEZY-SC)
 - ISA-free architectures, code compiled to silica: FPGA
 - focus on SIMT/SIMD to hide memory latency: GPU-type approach
 - tensor/neural net cores: NVidia Volta, Google TPU, Intel NNP
 - neuromorphic computing: IBM TrueNorth, Intel Loihi
 - **heterogeneous PUs in a SoC: multicore PUs with GPU-cores**
 - x86 multicore coupled with GPU cores: Intel i5/i7
 - ARMv8 cores coupled with GPU cores: NVidia Tegra... Xavier

Intel multicore coupled with GPU-cores

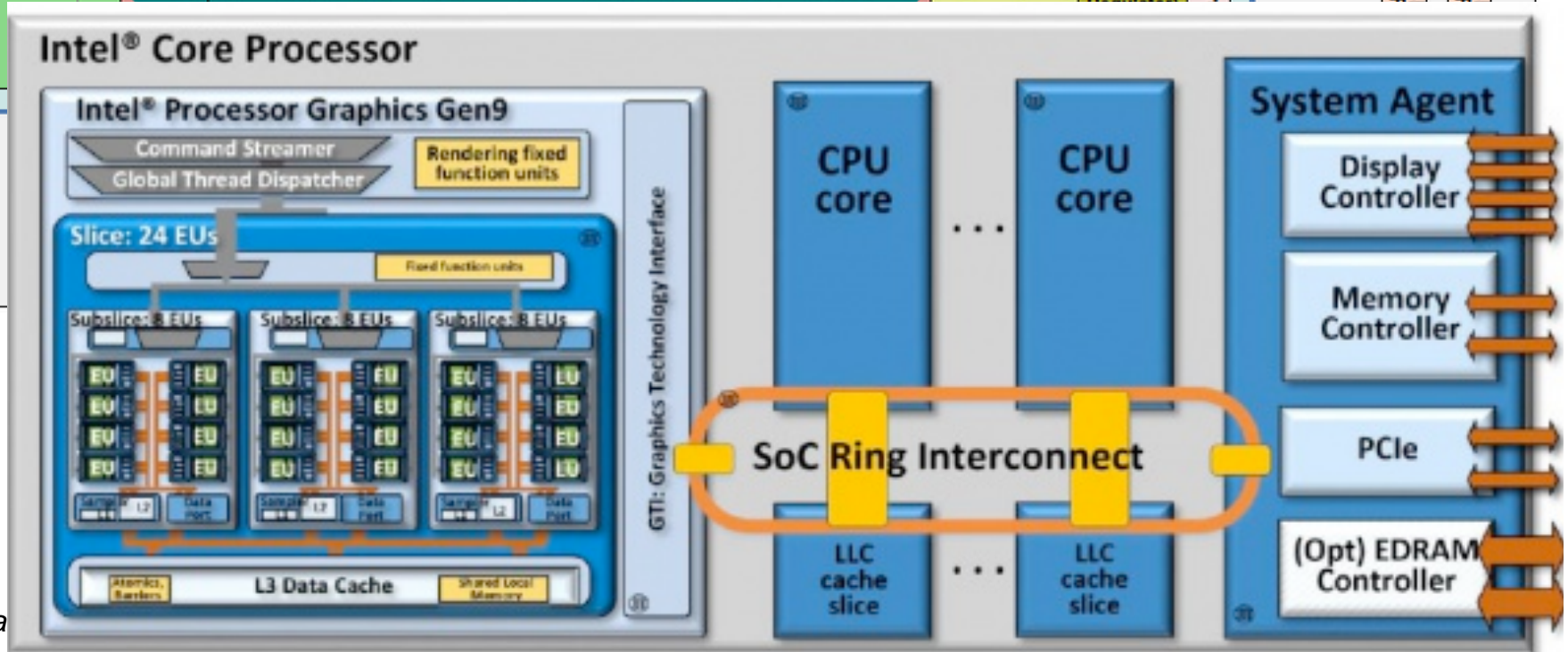


Haswell



Skylake

AJProença



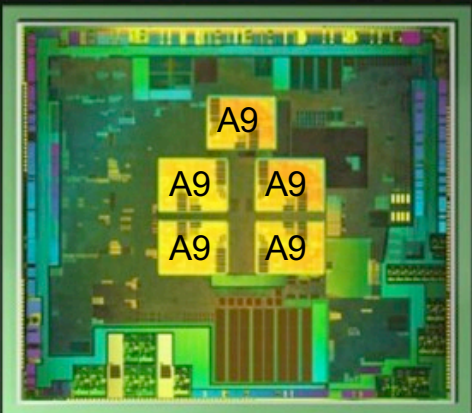
NVidia Tegra: SoC partnership with ARM (1)



- Tegra 2 in Android (2010) ...
- **Tegra 3** in Audi infotainment (2012) ...

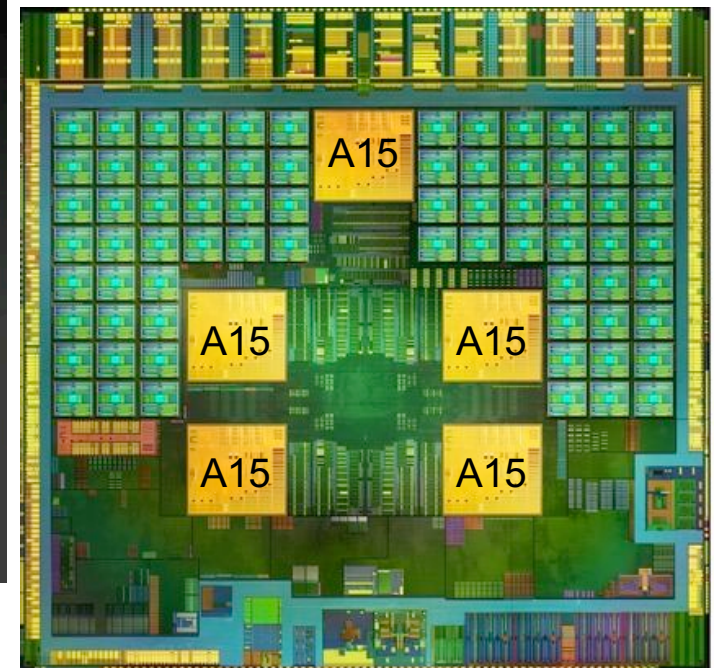
Tegra 3 The World's First Mobile Quad Core, with 5th Companion Core for Low Power

CPU	Quad Core, with 5 th Companion Core – Up to 1.4GHz Single Core, 1.3GHz Quad Core
GPU	Up to 3x Higher GPU Performance – 12 Core GeForce GPU
VIDEO	Blu-Ray Quality Video – 1080p High Profile @ 40Mbps
POWER	Lower Power than Tegra 2 – Variable Symmetric Multiprocessing (vSMP)
MEMORY	Up to 3x Higher Memory Bandwidth – DDR3L-1500, LPDDR2-1066
IMAGING	Up to 2x Faster ISP (Image Signal Processor)
AUDIO	HD Audio, 7.1 channel surround
STORAGE	2-6x Faster – e.MMC 4.41, SD3.0, SATA-II



Tegra 3 Nov'2011

Tegra 4:
replace the 32-bit ARM
Cortex A9 by
Cortex A15, and add
72 CUDA-cores



Tegra 4 May'2013

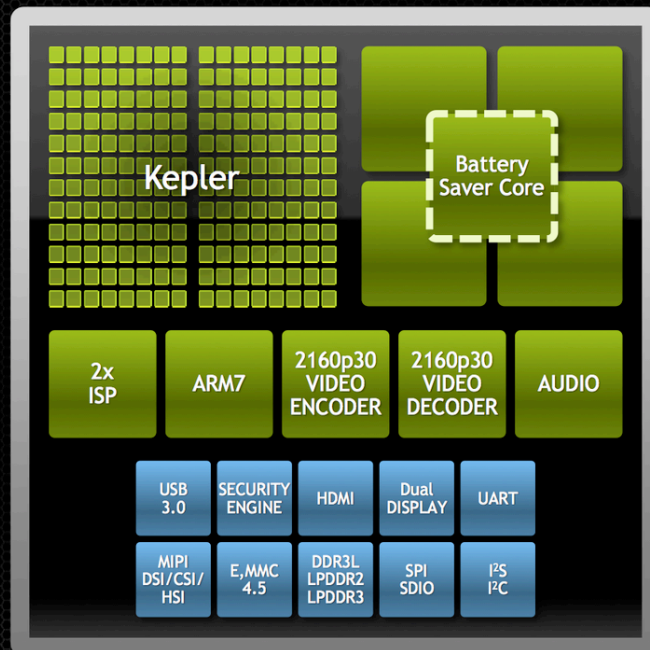
NVidia Tegra: SoC partnership with ARM (2)



Replace the GPU block by 192 GPU-cores (*from Kepler*) and offer either 32/64-bit CPU cores => **Tegra K1**

Tegra K1

Apr'2014



GPU	Kepler GPU (192 CUDA Cores) Open GL 4.4, OpenGL ES3.0, DX11, CUDA 6
CPU	Quad Core Cortex A15 "r3" With 5 th Battery-Saver Core; 2MB L2 cache
CAMERA	Dual High Performance ISP 1.2 Gigapixel throughput, 100MP sensor
POWER	Lower Power 28HPM, Battery Saver Core
DISPLAY	4K panel, 4K HDMI DSI, eDP, LVDS, High Speed HDMI 1.4a

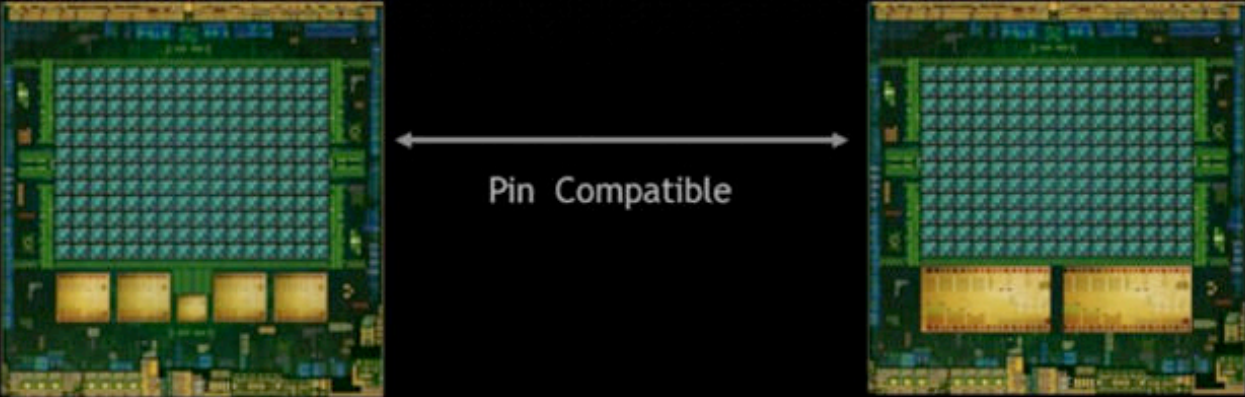
NVidia Tegra: SoC partnership with ARM (2)



Replace the GPU block by 192 GPU-cores (*from Kepler*) and offer either 32/64-bit CPU cores => **Tegra K1**

TEGRA K1 Apr'2014

One Chip – Two Versions



Pin Compatible

Quad Core +1 (<i>battery saver</i>)	Dual Super Core
32-bit	64-bit
3-way Superscalar	7-way Superscalar
Up to 2.3GHz	Up to 2.5GHz
32K+32K L1\$	128K+64K L1\$

NVidia Tegra: SoC partnership with ARM (3)



- Replace the 5x 32-bit ARM by 2x4 32-bit Cortex (A57 & A53) and the 192 Kepler CUDA cores by 256 Maxwell => **Tegra X1**
May'2015



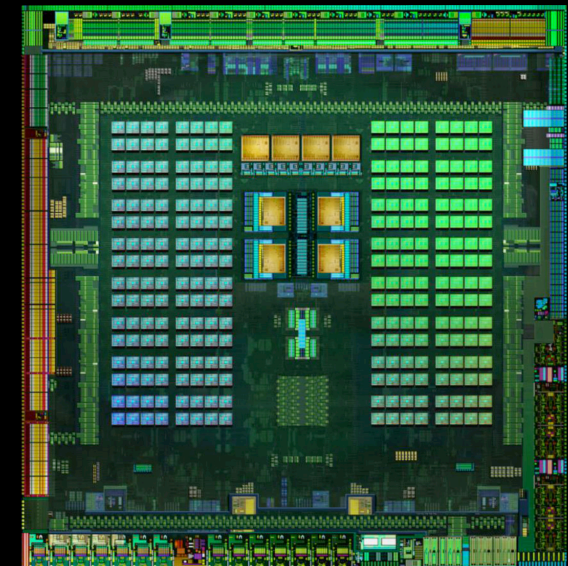
TEGRA X1 CPU CONFIGURATION

4 HIGH PERFORMANCE A57 BIG CORES

- 2MB L2 cache
- 48KB L1 instruction cache
- 32KB L1 data cache

4 HIGH EFFICIENCY A53 LITTLE CORES

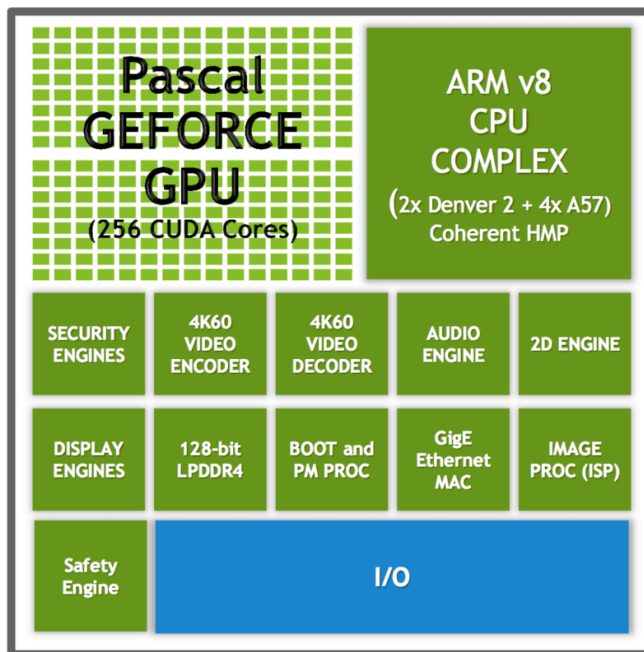
- 512KB L2 cache
- 32KB L1 instruction cache
- 32KB L1 data cache



NVidia Tegra: pathway towards ARM-64 (1)

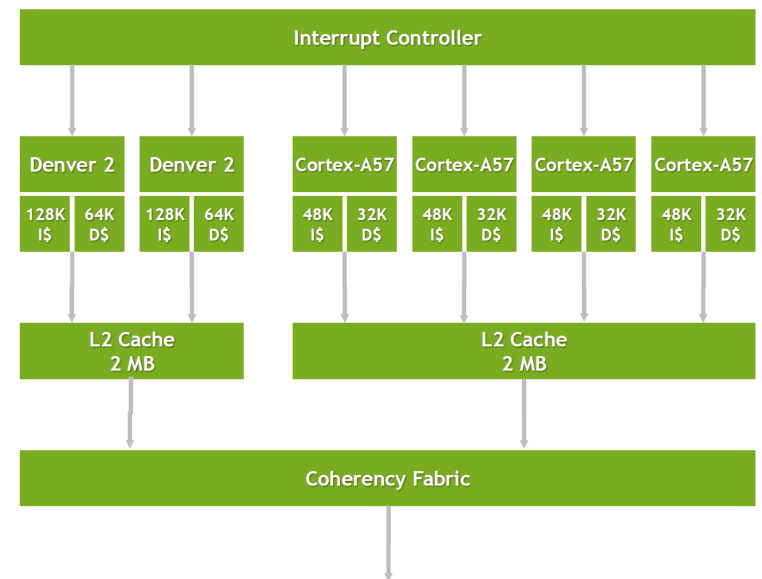


- Upgrade 32-bit ARM to 64-bit ARM (*Denver 2 & A57*) and replace Maxwell cores by Pascal ones => **Parker** Aug'2016



“PARKER” CPU COMPLEX

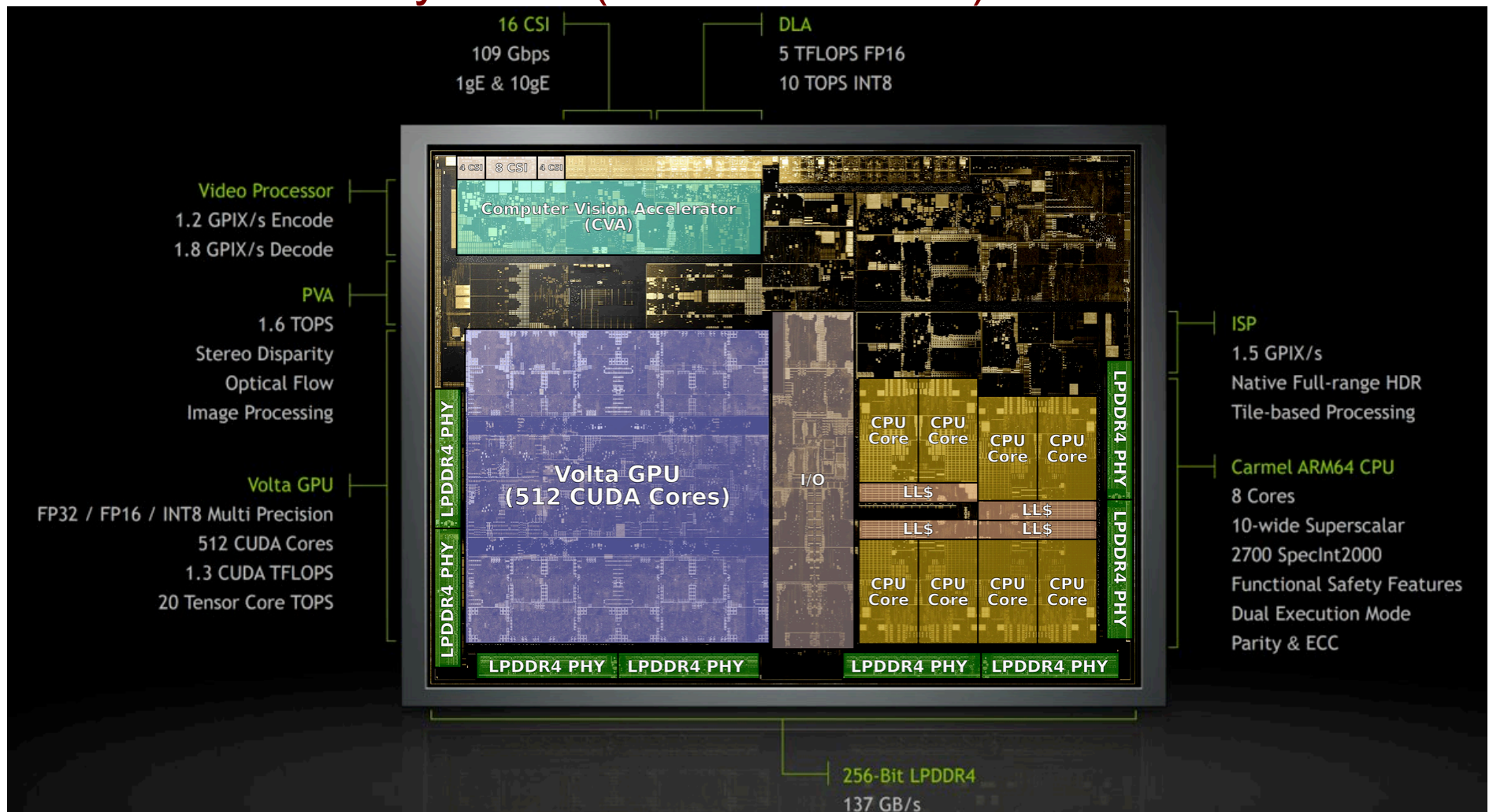
- 2x Denver2 + 4x Cortex-A57
- Fully Coherent HMP system
 - Proprietary Coherent Interconnect
- ARM V8 64-bit
- Highest performance ARM CPU
 - 2nd generation Denver core
 - Significant Perf/W improvements
- Dynamic Code Optimization
 - OoO execution without the power
 - Optimize once, use many times
- 7-wide superscalar
- Low power retention states



NVIDIA Tegra: pathway towards ARM-64 (2)

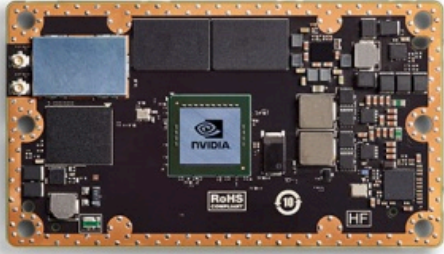
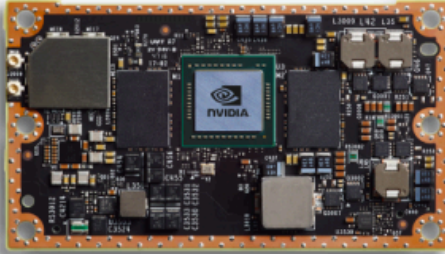
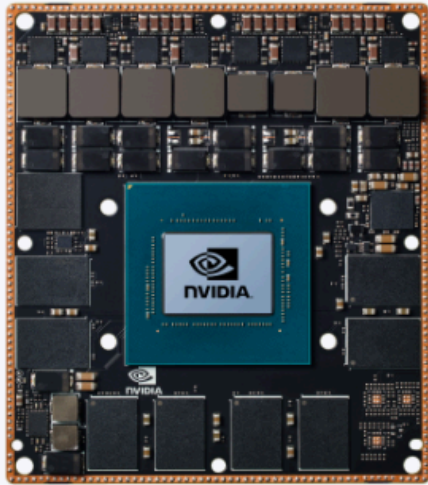


- Increment #ARMv8-cores (*custom architecture*) and replace Pascal cores by Volta (w/ tensor cores) => **Xavier** Jun'2018



The Jetson line of embedded Linux AI and computer vision compute modules

NVIDIA Jetson Modules

Features	Jetson TX1	Jetson TX2 / TX2i	Jetson AGX Xavier
			
CPU	ARM Cortex-A57 (quad-core) @ 1.73GHz (2MB L2)	ARM Cortex-A57 (quad-core) @ 2GHz (2MB L2) + NVIDIA Denver2 (dual-core) @ 2GHz (2MB L2)	NVIDIA Carmel ARMv8.2 (octal-core) @ 2.26GHz (4x2MB L2 + 4MB L3)
GPU	256-core NVIDIA Maxwell @ 998MHz	256-core NVIDIA Pascal @ 1300MHz	512-core NVIDIA Volta @ 1377 MHz + 64 Tensor Cores
DL	GPU support (CUDA, cuDNN, TensorRT)		dual NVIDIA Deep Learning Accelerators (DLAs)
Memory	4GB 64-bit LPDDR4 @ 1600MHz 25.6 GB/s	8GB 128-bit LPDDR4 @ 1866Mhz 58.3 GB/s	16GB 256-bit LPDDR4x @ 2133MHz 137GB/s
Storage	16GB eMMC 5.1	32GB eMMC 5.1	
Vision	GPU support (CUDA, VisionWorks, OpenCV)		7-way VLIW Vision Accelerator
Encoder	4Kp30, (2x) 1080p60, (4x) 1080p30	4Kp60, (3x) 4Kp30, (4x) 1080p60, (8x) 1080p30	(4x) 4Kp60, (8x) 4Kp30, (16x) 1080p60, (32x) 1080p30
Decoder	4Kp60, (4x) 1080p60, (8x) 1080p30	(2x) 4Kp60, (4x) 4Kp30, (7x) 1080p60, (14x) 1080p30	(2x) 8Kp30, (6x) 4Kp60, (12x) 4Kp30, (26x) 1080p60
Camera	12 lanes MIPI CSI-2 1.5 Gbps per lane	12 lanes MIPI CSI-2 2.5 Gbps per lane	16 lanes MIPI CSI-2 6.8125Gbps per lane (C-PHY)
Display	2x HDMI 2.0 / DP 1.2 / eDP 1.2 2x MIPI DSI		(3x) eDP 1.4 / DP 1.2 / HDMI 2.0 @ 4Kp60
Wireless	802.11a/b/g/n/ac 2x2 867Mbps Bluetooth 4.0	802.11a/b/g/n/ac 2x2 867Mbps Bluetooth 4.1	M.2 Key-E site on carrier
Ethernet	10/100/1000 BASE-T Ethernet		
USB	USB 3.0 + USB 2.0		(3x) USB 3.1 + (4x) USB 2.0
PCIe	PCIe Gen 2 x5 1x4 + 1x1	PCIe Gen 2 x5 1x4 + 1x1 or 2x1 + 1x2	PCIe Gen 4 x16 1x8 + 1x4 + 1x2 + 2x1

Competitors of the Jetson line: Google Coral Edge TPU, Raspberry Pi + Intel NCS



NVIDIA Jetson Nano



Google Coral Edge TPU



Raspberry Pi 3B +
Intel Neural Compute Stick 2

<https://augmentedstartups.com/movidius-ncs-with-raspberry-pi-vs-google-edge-tpu-coral-vs-nvidia-jetson-nano-a-quick-comparison/>

Beyond Vector/SIMD architectures



- Vector/SIMD-extended architectures are hybrid approaches
 - mix (super)**scalar + vector** op capabilities on a single device
 - **highly pipelined** approach to reduce memory access penalty
 - **tightly-closed access to shared memory**: lower latency
- Evolution of Vector/SIMD-extended architectures
 - **PU** (**P**rocessing **U**nit) **cores with wider vector units**
 - x86 many-core: **Intel MIC / Xeon KNL**
 - others: **IBM Power BlueGene/Q, ShenWay 260, Matrix-2000, A64FX Arm**
 - **coprocessors (require a host scalar processor): accelerator devices**
 - on disjoint physical memories (e.g., **Xeon KNC** with **PCI-Expr, PEZY-SC**)
 - ISA-free architectures, code compiled to silica: **FPGA**
 - focus on SIMT/SIMD to hide memory latency: **GPU-type** approach
 - tensor/neural net cores: **NVidia Volta, Google TPU, Intel NNP**
 - neuromorphic computing: **IBM TrueNorth, Intel Loihi**
 - **heterogeneous PUs in a SoC: multicore PUs with GPU-cores**
 - x86 multicore coupled with GPU cores: **Intel i5/i7**
 - ARMv8 cores coupled with GPU cores: **NVidia Tegra... Xavier**