ACCELERATION OPTIONS FOR AI AND HPEC

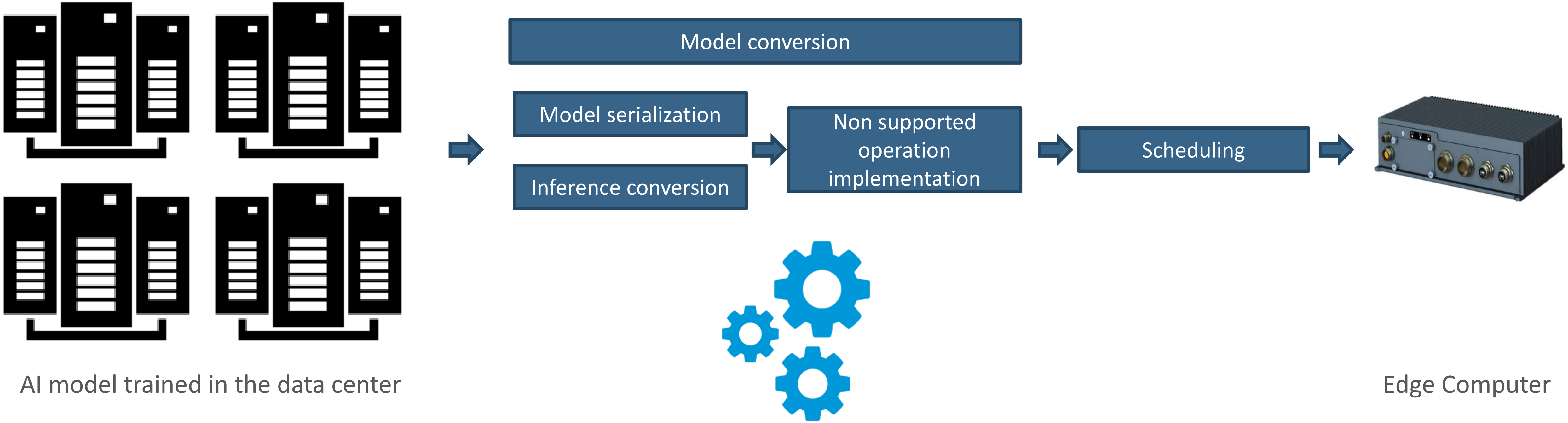
Embedded Tech Trends, 2020
Atlanta, GA

Dr. Mohamed BERGACH, Sr. Systems Architect – January 27, 2020

AGENDA

- ❑ AI from datacenter to the edge
- ❑ Software acceleration
- ❑ Hardware acceleration
- ❑ Conclusion and takeaway

FROM TRAINED AI MODEL TO THE EDGE



Model conversion

Model serialization

Inference conversion

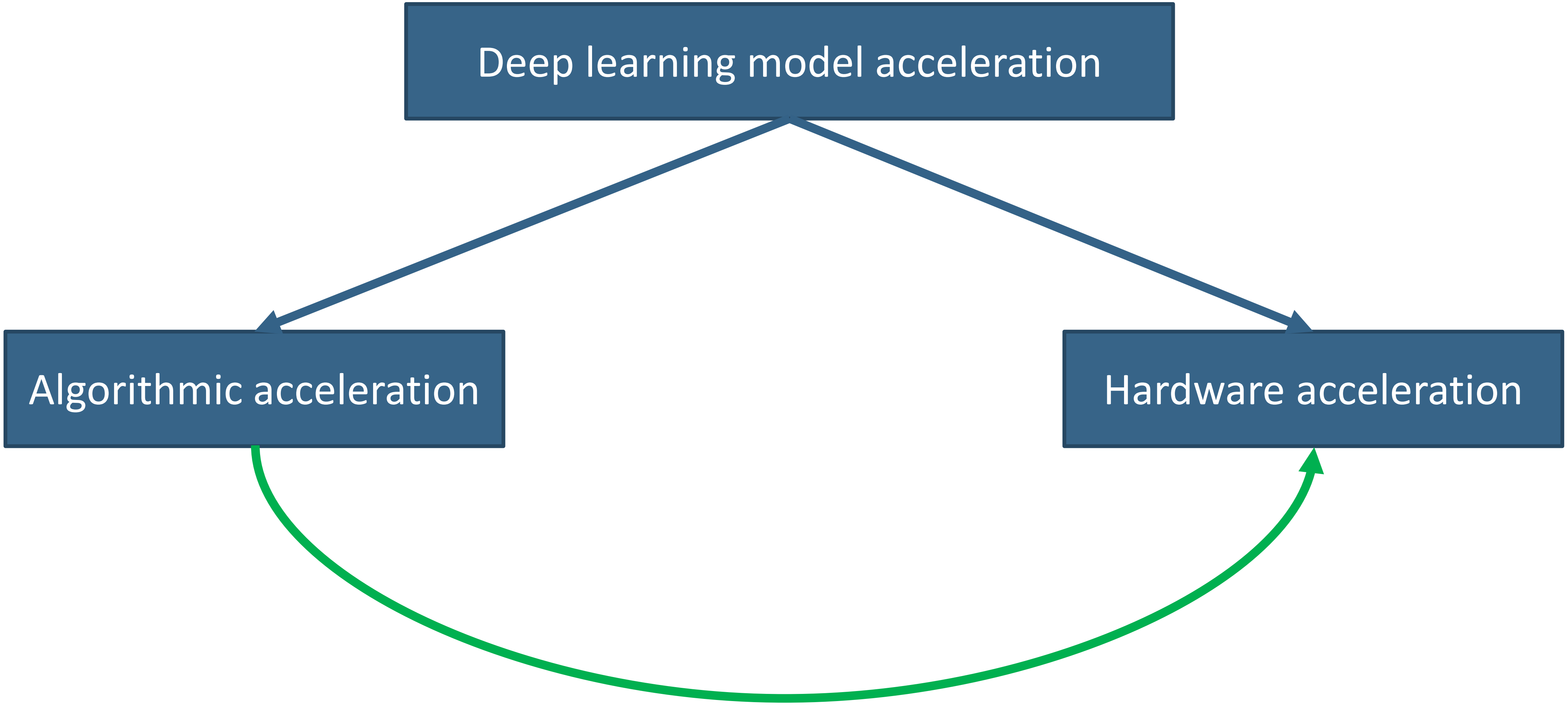
Non supported operation implementation

Scheduling

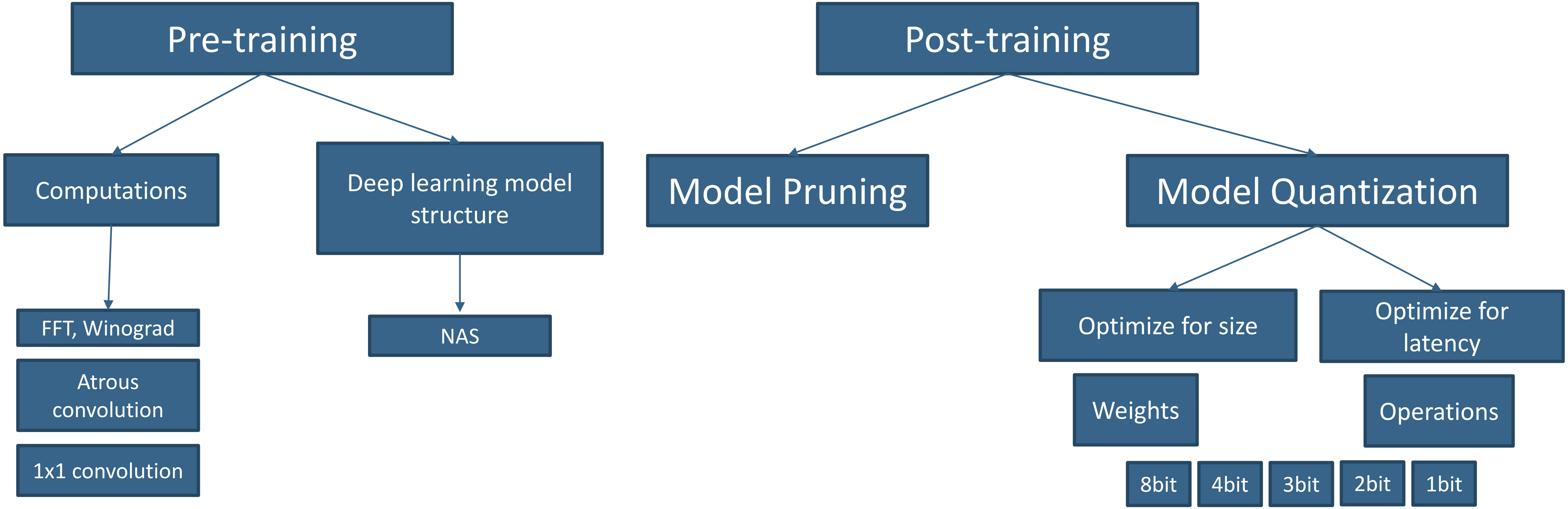
AI model trained in the data center

Edge Computer

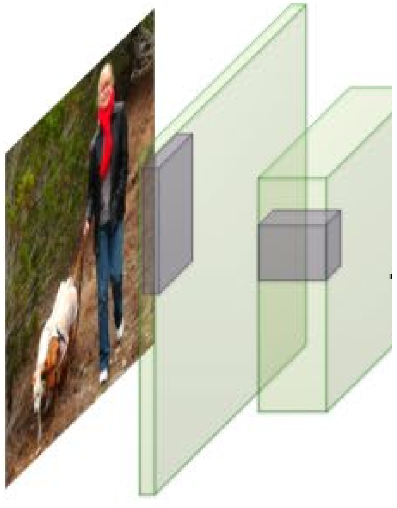
ACCELERATION

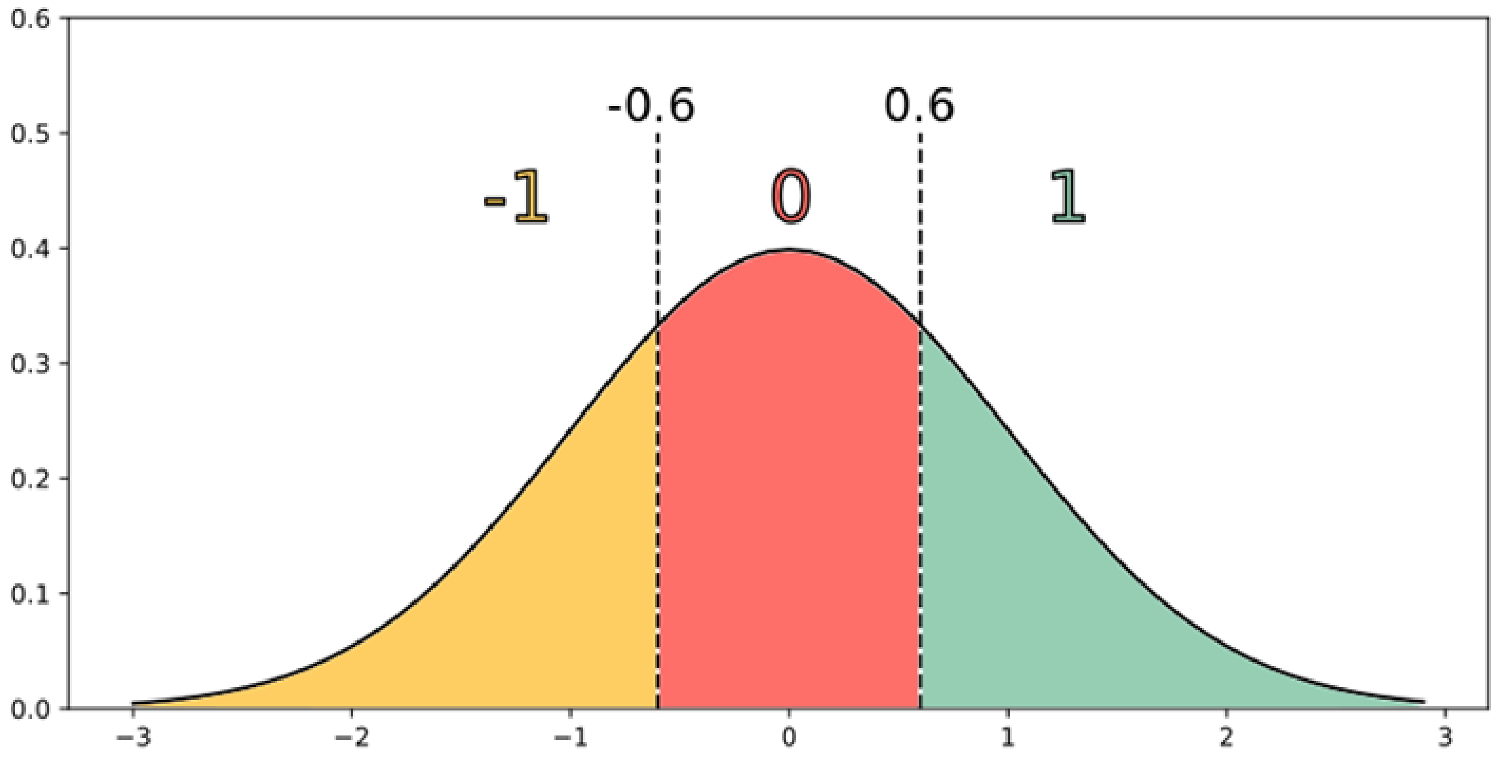


ALGORITHMIC ACCELERATION:



ALGORITHMIC ACCELERATION: QUANTIZATION

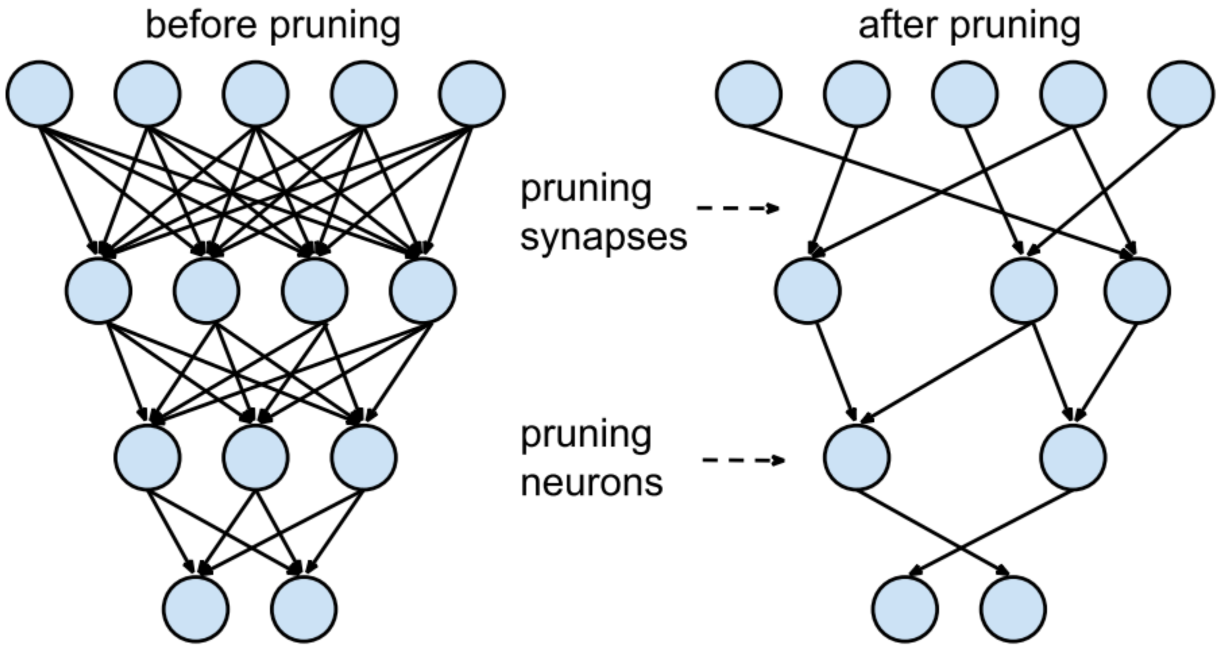




Computation Saving (Inference)	Accuracy on ImageNet (AlexNet)
1x	%56.7
~2x	%56.8
~58x	%44.2

Source: [Rastegari et al. ECCV'16]

ALGORITHMIC ACCELERATION: PRUNING



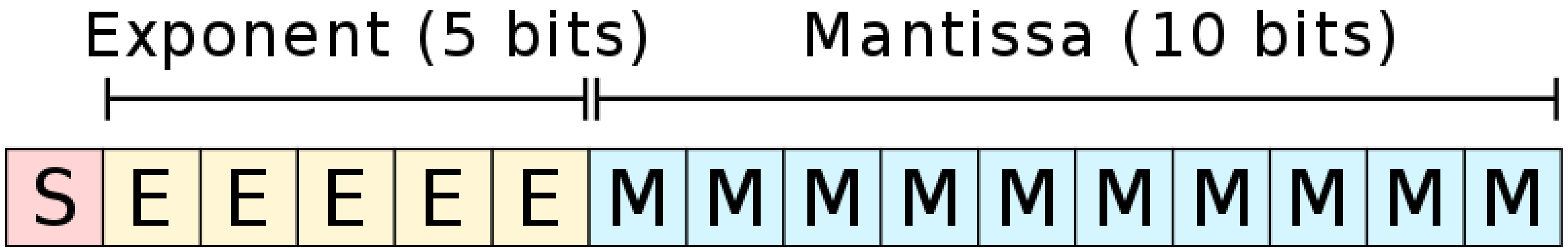
10X speedup

Source: [Han et al. NIPS’15]

FLOAT VS BFLOAT

Float16

range (~5.96e-8 to 65,504)



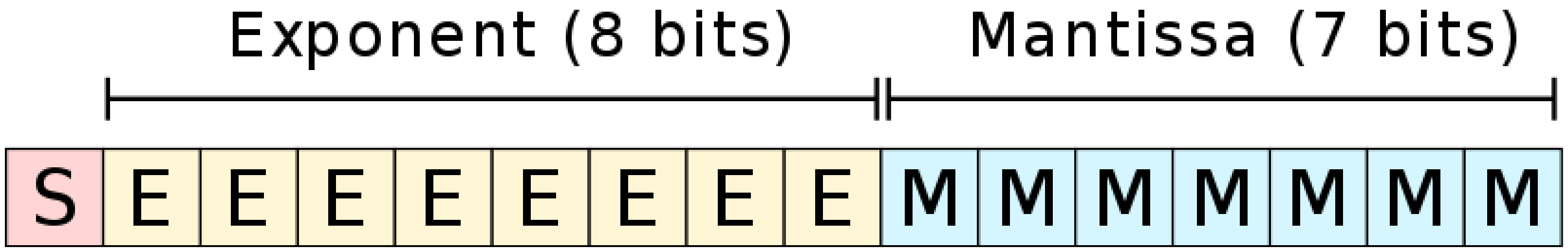
Exponent (5 bits)

Mantissa (10 bits)

S E E E E E M M M M M M M M M M

Bfloat16

Range (~1e-38 to ~3e38)



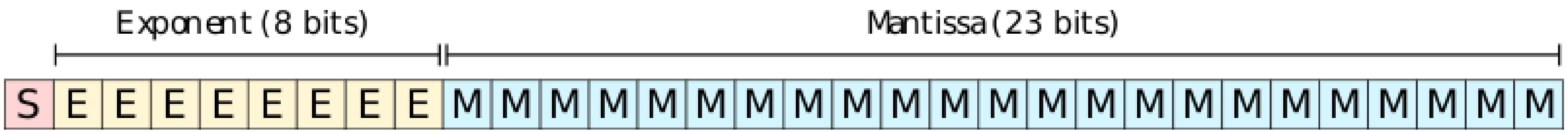
Exponent (8 bits)

Mantissa (7 bits)

S E E E E E E E E M M M M M M M

Float32

Range (~1e-38 to ~3e38)

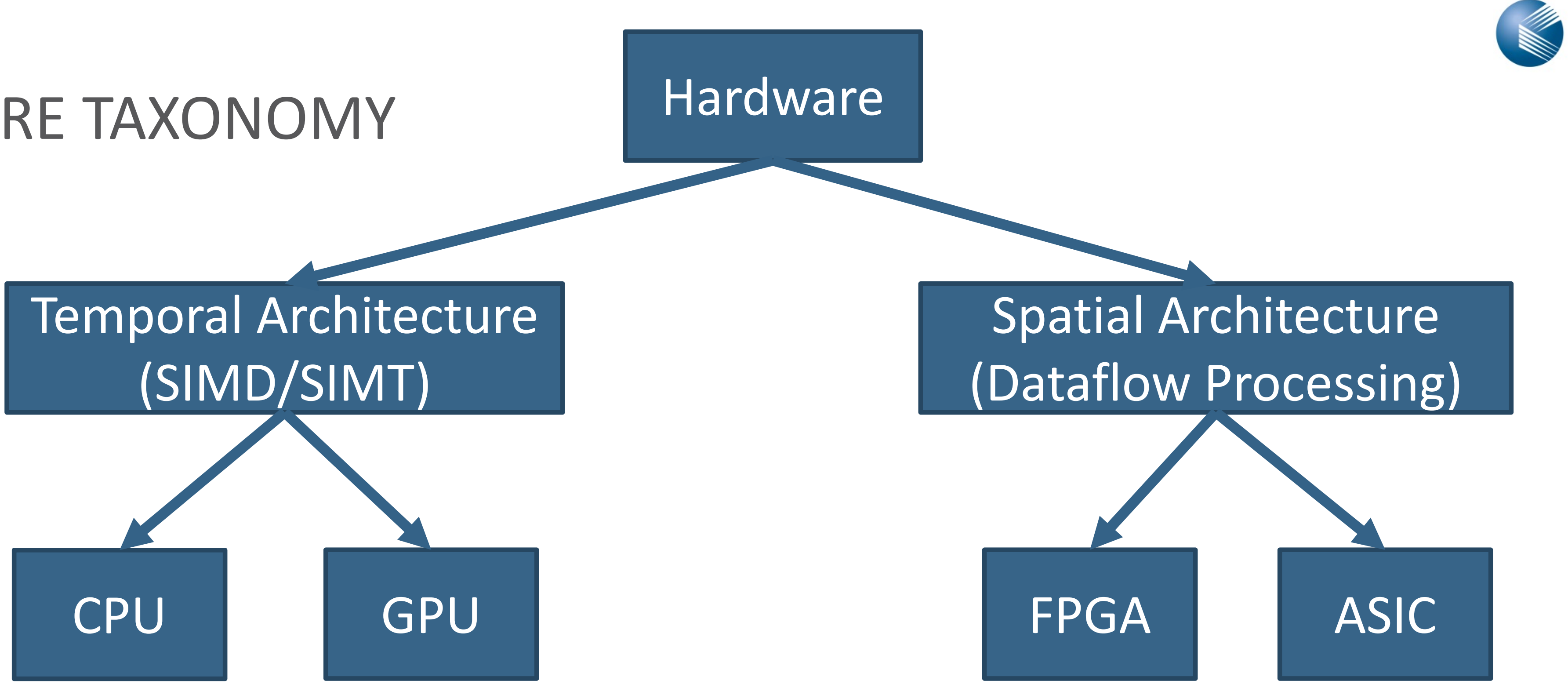


Exponent (8 bits)

Mantissa (23 bits)

S E E E E E E E E M M M M M M M M M M M M M M M M M M M M M M M

HARDWARE TAXONOMY



Hardware

Temporal Architecture (SIMD/SIMT)

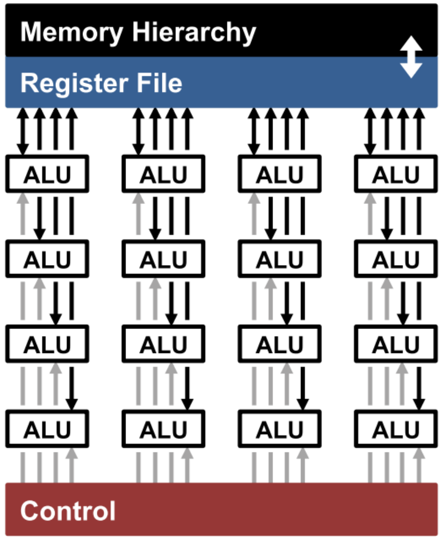
Spatial Architecture (Dataflow Processing)

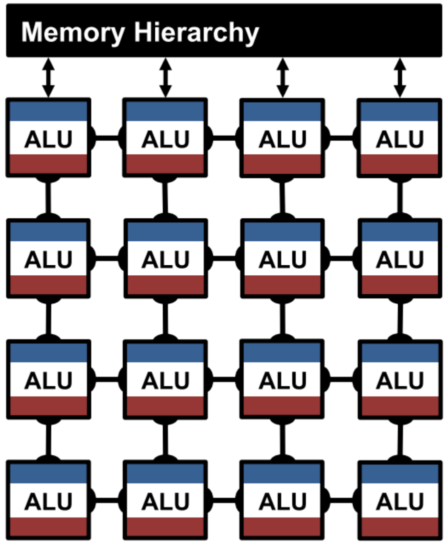
CPU

GPU

FPGA

ASIC

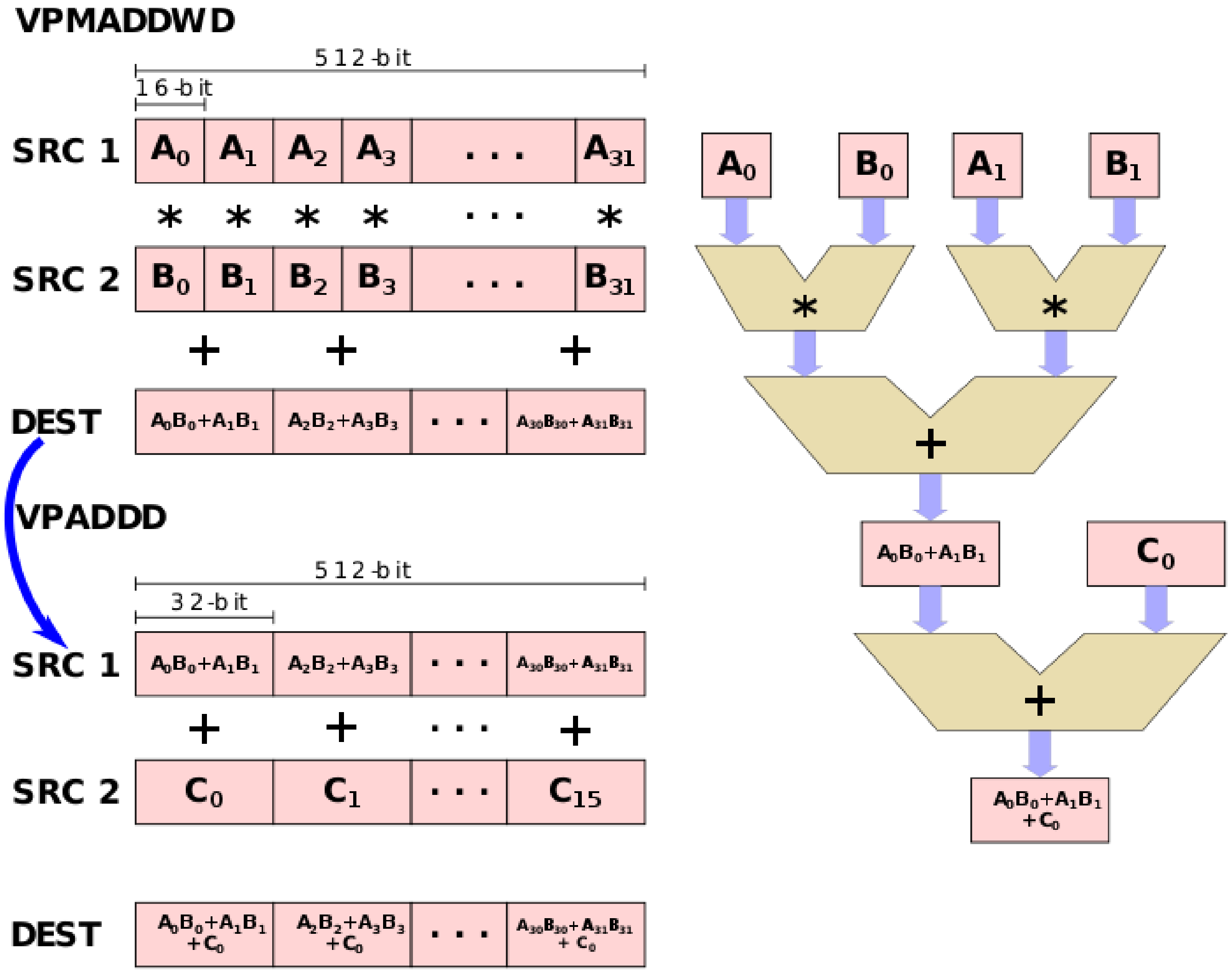




Source: [Sze et al. Proceedings of the IEEE 105(12): 2017]

CPU (AVX512, VNNI)

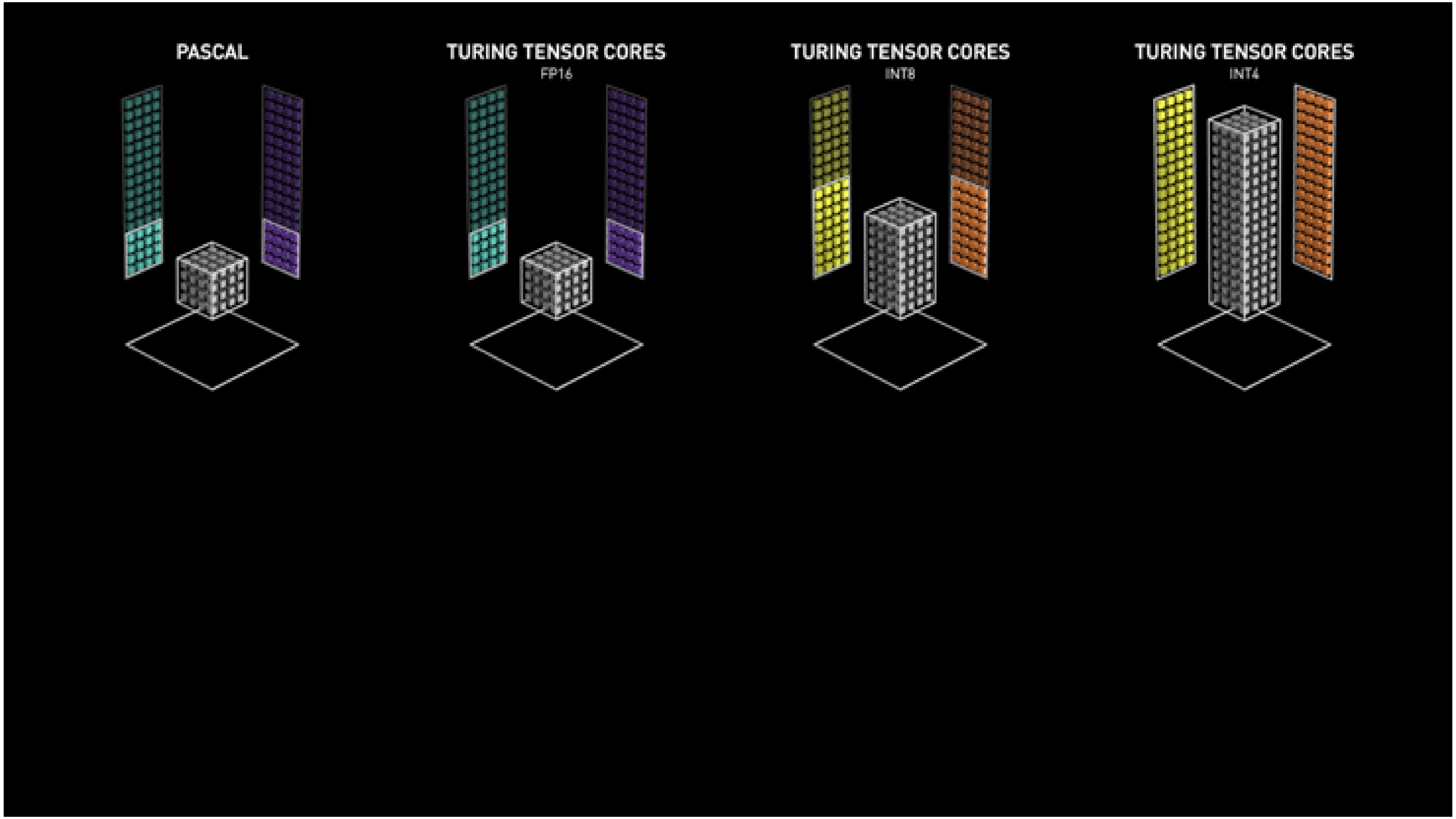
2x64x2x2 = 0.5TOPS/core
4 core Intel Tiger lake CPU at 25W TDP
2TOPS => 80GOPS/W


VPMADDWD
512-bit
16-bit
SRC 1
A_0 A_1 A_2 A_3 . . . A_{31}
* * * * . . . *
SRC 2
B_0 B_1 B_2 B_3 . . . B_{31}
+ + +
DEST
$A_0B_0+A_1B_1$ $A_2B_2+A_3B_3$ . . . $A_{30}B_{30}+A_{31}B_{31}$
VPADDD
512-bit
32-bit
SRC 1
$A_0B_0+A_1B_1$ $A_2B_2+A_3B_3$ . . . $A_{30}B_{30}+A_{31}B_{31}$
+ + . . . +
SRC 2
C_0 C_1 . . . C_{15}
DEST
$A_0B_0+A_1B_1+C_0$ $A_2B_2+A_3B_3+C_0$ . . . $A_{30}B_{30}+A_{31}B_{31}+C_0$
A_0 B_0 A_1 B_1
* *
+
$A_0B_0+A_1B_1$
C_0
+
$A_0B_0+A_1B_1+C_0$

Source: https://en.wikichip.org/wiki/x86/avx512vnni

GPU (TENSOR CORES)

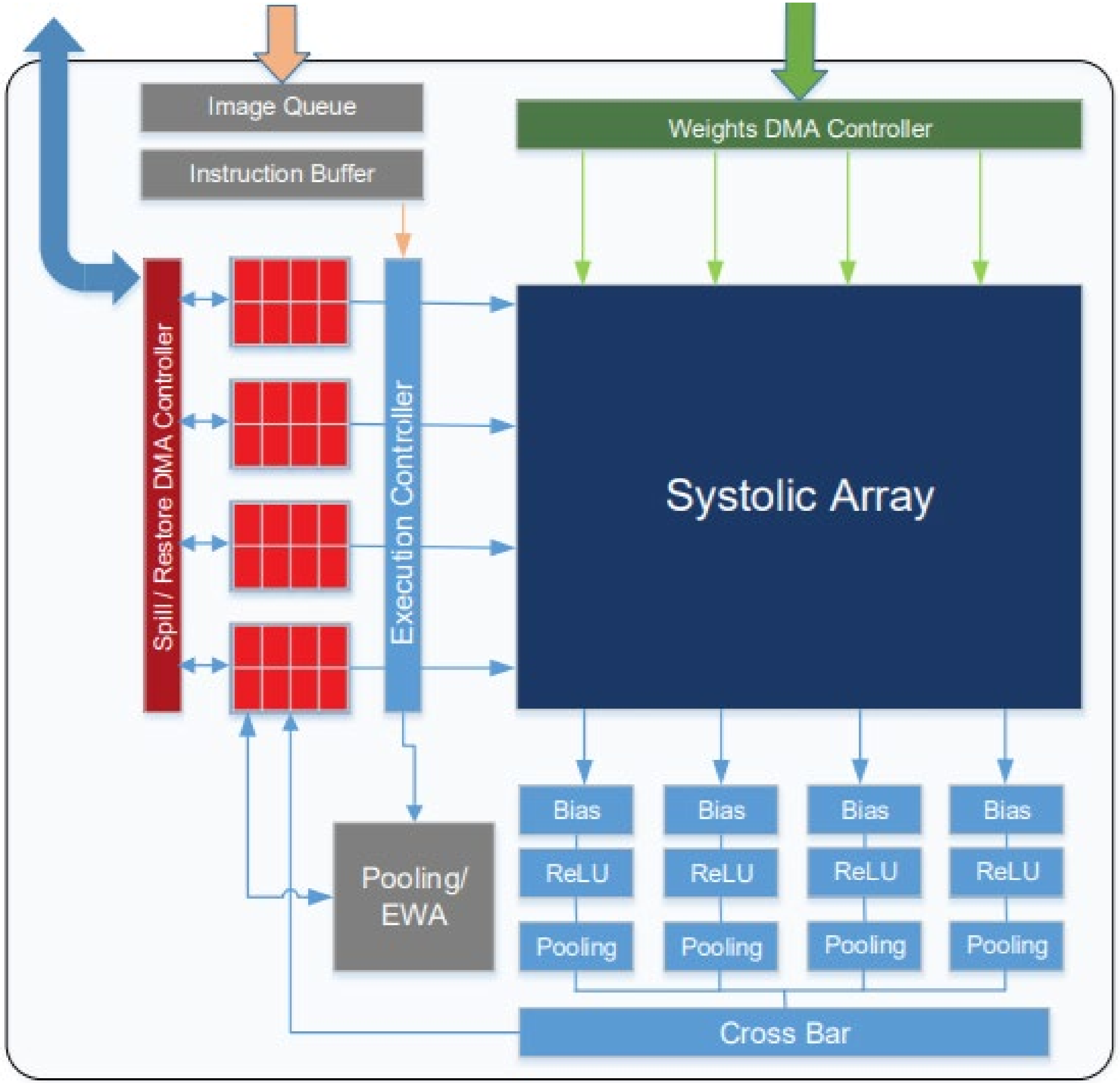
Nvidia Tesla T4
130TOPs for 75W TDP
.5TOPs/W



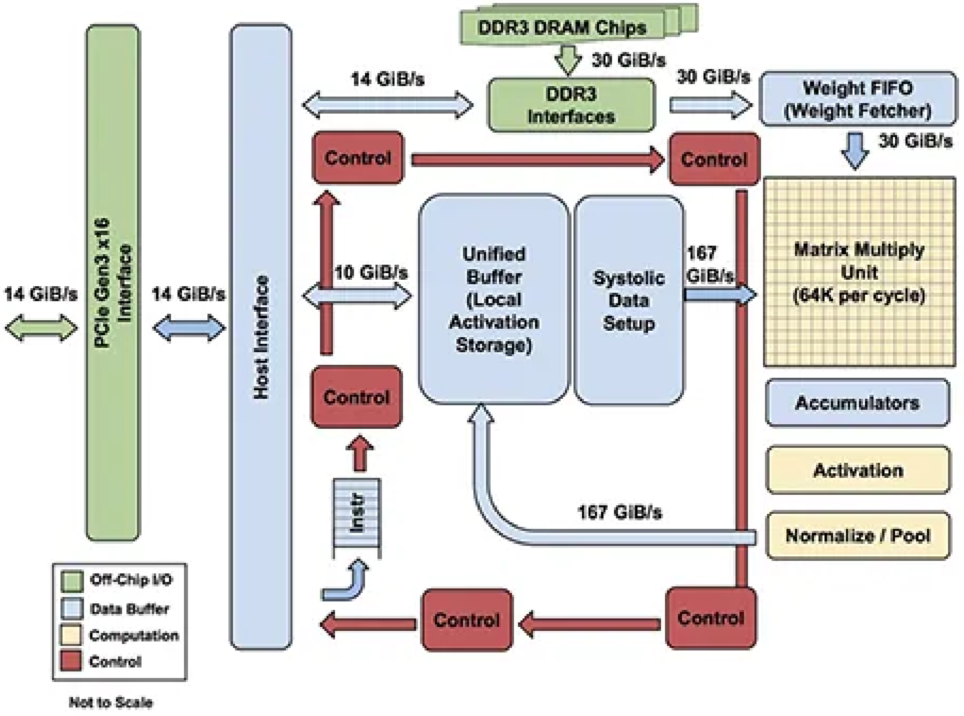
Source: https://www.nvidia.com/en-us/data-center/tensorcore/

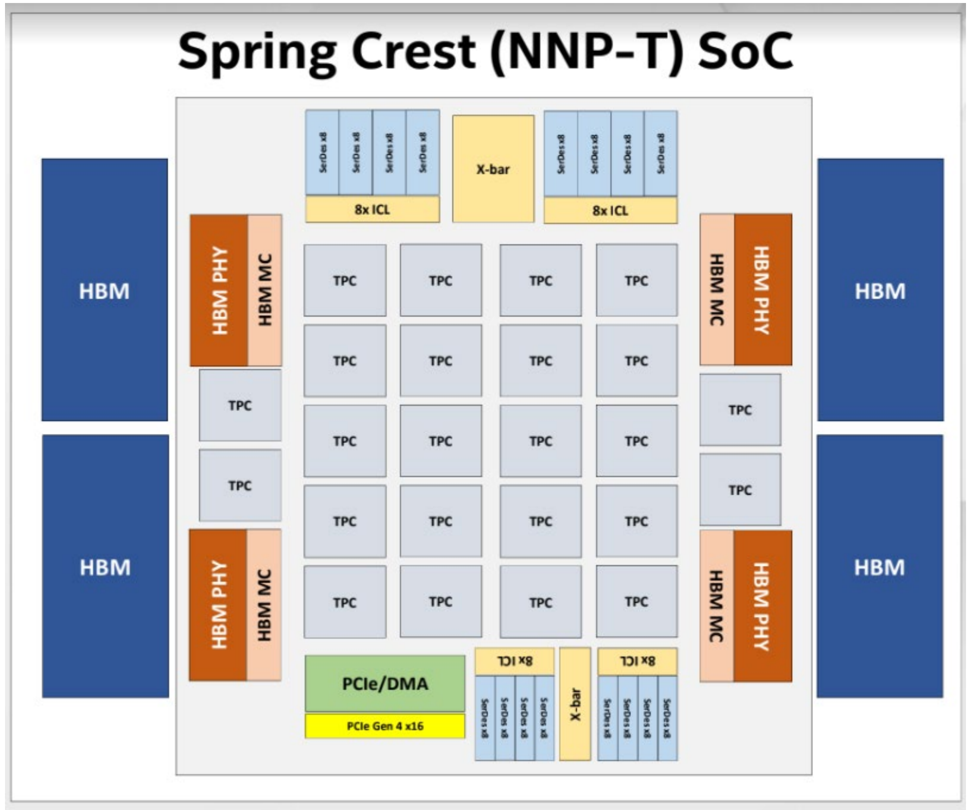
FPGA

XILINX xDNN
75W
21TOPS
0.3TOPS/W
Very low latency



ASIC





Edge TPU
4TOPS
2W
2TOPS/W

Google TPU

Intel Nervana Spring Crest
119TOPS
200W TDP
0.6TOPS/W

NNP-T

HARDWARE COMPARISON

HW Accelerator	TOPS/W	Latency	Efficiency
Intel Tiger lake CPU	0.08	★★☆☆	★★☆☆
Nvidia T4 GPU	0.5	★☆☆☆	★★☆☆
XILINX xDNN FPGA	0.3	★★★★	★★★★
Intel NNP	0.6	★★★☆	★★★☆
Google Edge TPU	2	★★☆☆	★★★☆

TAKEAWAY

- ✓ AI on the Edge is a reality now, every major smartphone has an NPU
- ✓ GPUs are dominant right now but not in the future
- ✓ Specialized hardware are much more power efficient
- ✓ The software ecosystem is key ingredient to for best performance
- ✓ Quantization is a performance booster that needs to be considered

Questions?

OUR CHANGING WORLD IS FILLED WITH

BOUNDLESS OPPORTUNITIES

Dr. Mohamed BERGACH, Sr. Systems Architect – January 27, 2020