# Hardware acceleration of Neural Networks on Edge Devices

## A look into accelerators like the Google Edge TPU and Intel's Neural Compute Stick

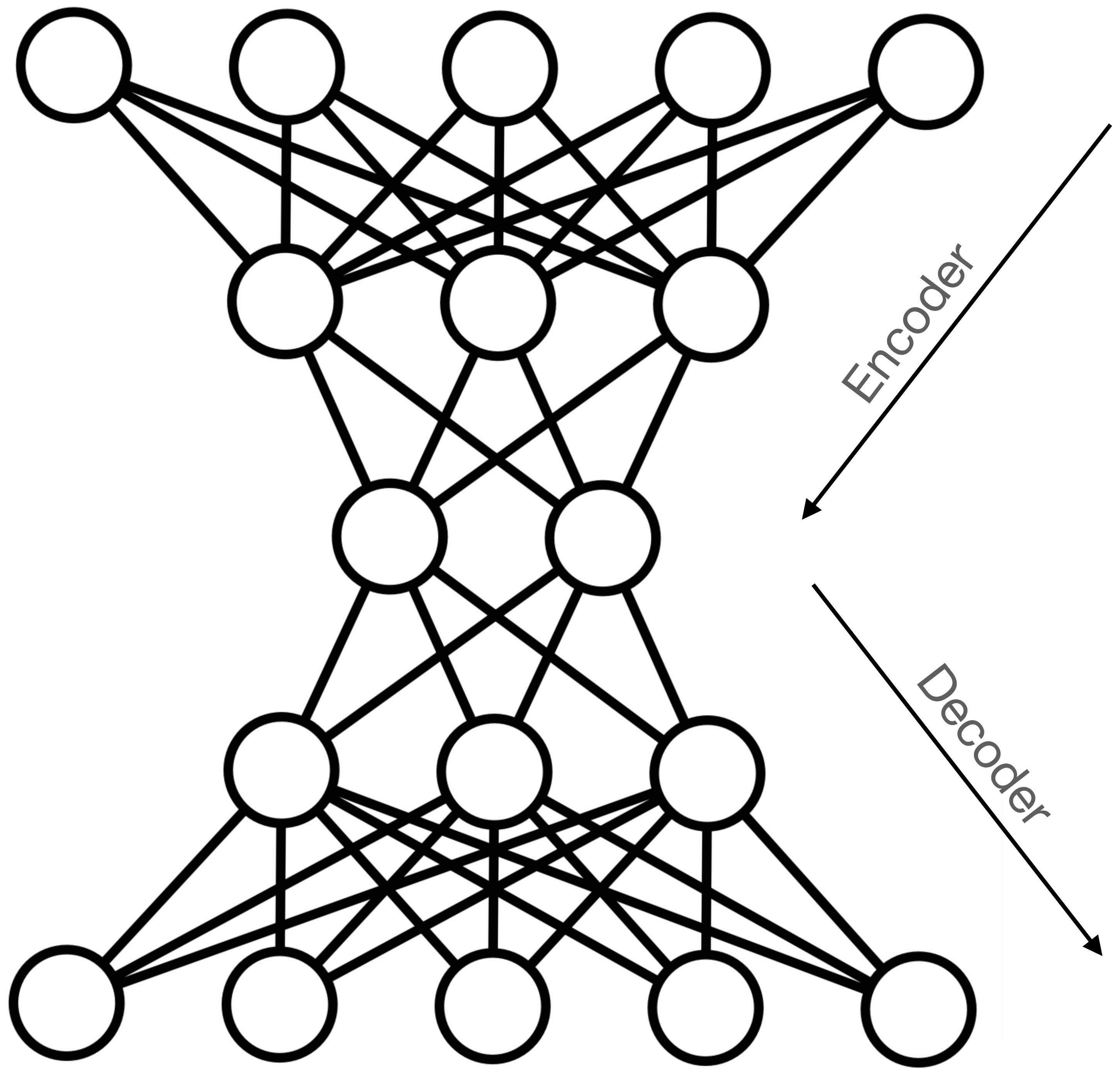
**Jannis Wolf, B.Sc., 20.5.2021**

# What makes DL computations so heavy?

## The example of an under complete autoencoder

- Autoencoder map input vector back on the input vector
- Encoder/Decoder characteristic
- Dimensionality reduction also called embeddings
- Common use cases: Denoising, Generating through Variational AE

Input feature tensor



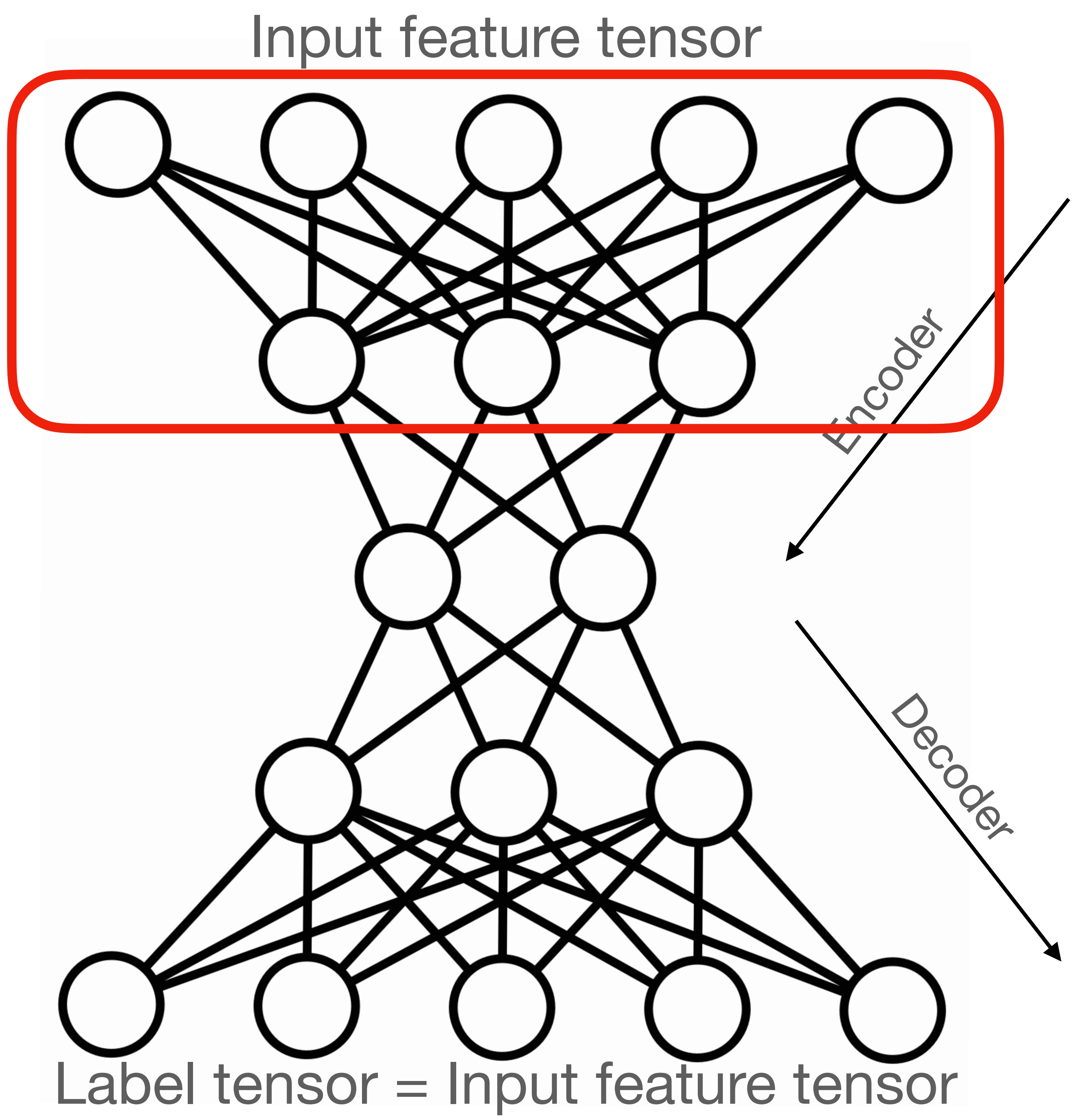
Label tensor = Input feature tensor

Undercomplete Autoencoder[1]

# What makes DL computations so heavy?

## The example of an under complete autoencoder

$$\underbrace{\begin{pmatrix} w_{1,1} & \cdots & w_{n,1} & w_{n+1,1} \\ \vdots & \ddots & \vdots & \vdots \\ w_{1,m} & \cdots & w_{n,m} & w_{n+1,m} \end{pmatrix}}_{\text{Weight tensor}} \overset{\text{Input tensor}}{\begin{pmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix}} = \underbrace{\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_m \end{pmatrix}}_{\text{Activation}}$$



Undercomplete Autoencoder [1]

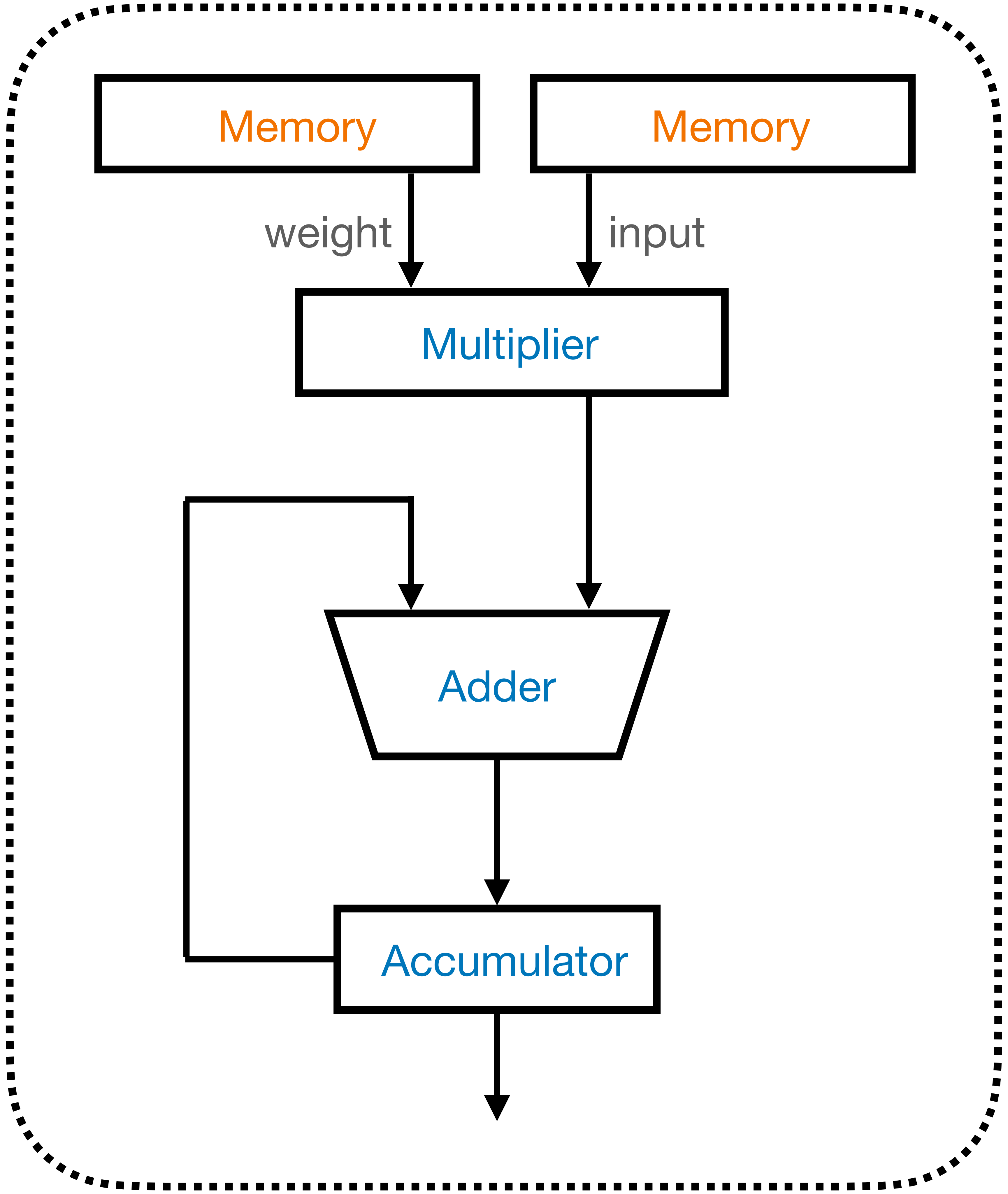
- Summation of an element wise tensor multiplication

—> Multiply Accumulate Operation

# From Software to Hardware level

## Multiply Accumulate operation (MAC)

- Two critical performance bottlenecks:
  - Enough MAC units for the computation of the tensor products
  - Intelligent memory design for loading the weights and activations



# Hardware Accelerators

Google Edge TPU [2]

## Google Edge TPU (now called Coral)

- Neural Network-focused ASIC
- 4 TOPS / 2 Watt
- Tensorflow Lite
- 8-Bit fixed point format

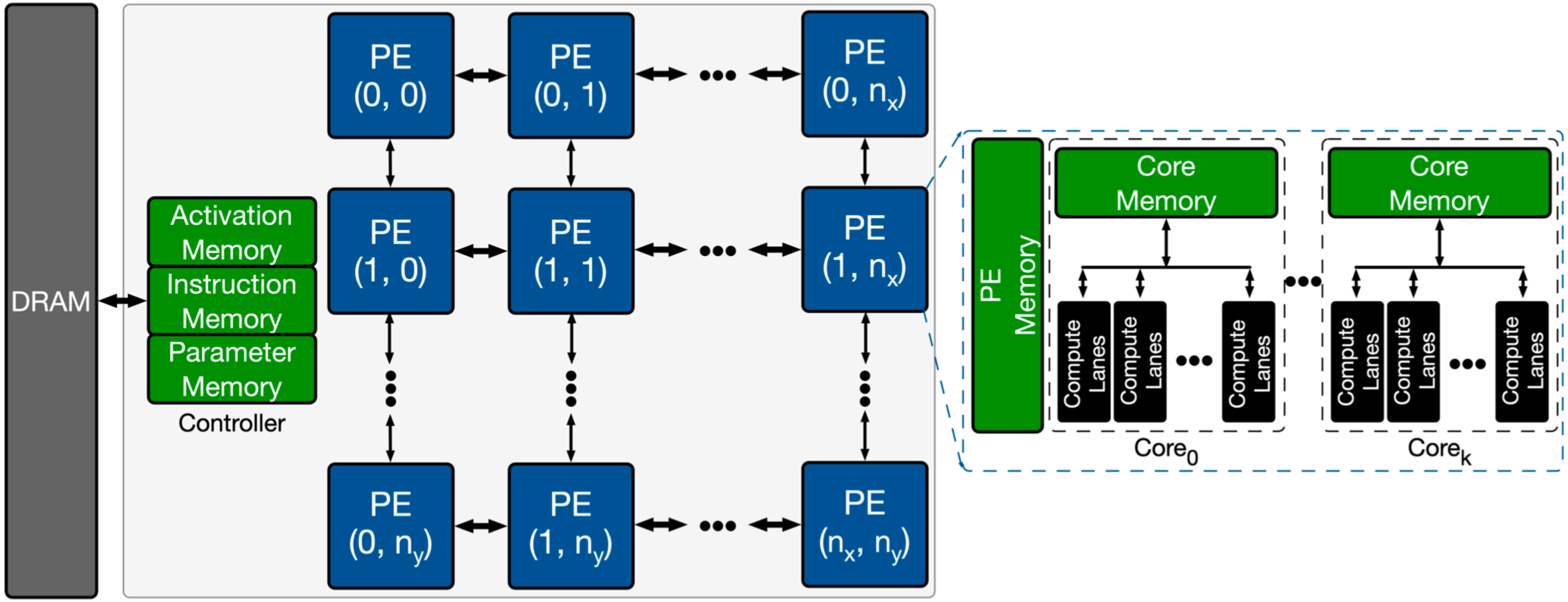
Intel NCS 2 [3]

## Intel Neural Compute Stick 2

- 16 128-Bit VLIW Vector Processors
- 2 TOPS / Watt
- Multiple frontends incl. Tensorflow
- Floating and fixed point

# Google Edge TPU

## Technical Overview

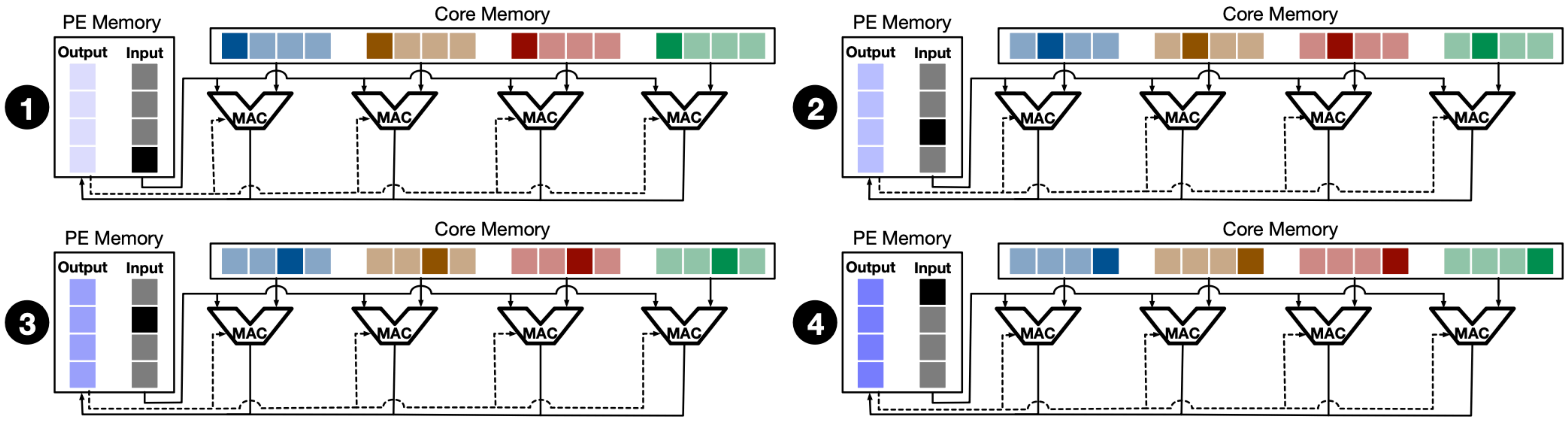


Technical overview of the Google Edge TPU [4]

- Cluster of processing elements (PE)
- Arbitrary number lets space for perfect coordination with software (= definition of ASIC)

# Google Edge TPU

## Inside a processing unit (PE)



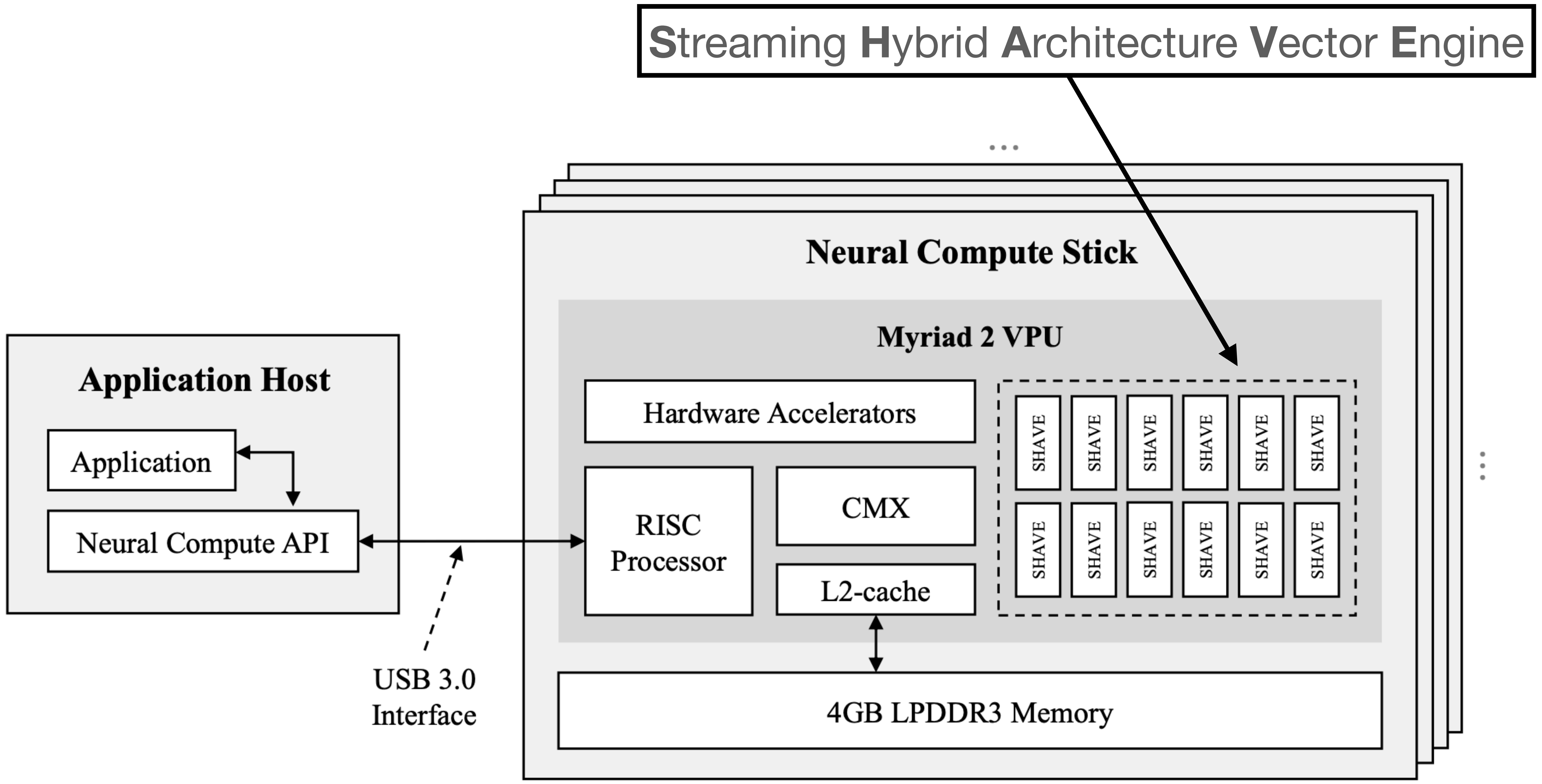
Inside a processing unit of the Google Edge TPU [4]

- Parallel MAC computing lanes inside parallel compute cores
- Shared PE memory enables parallel calculations in a SIMD fashion

# Intel Neural Compute Stick 2

## Technical Overview

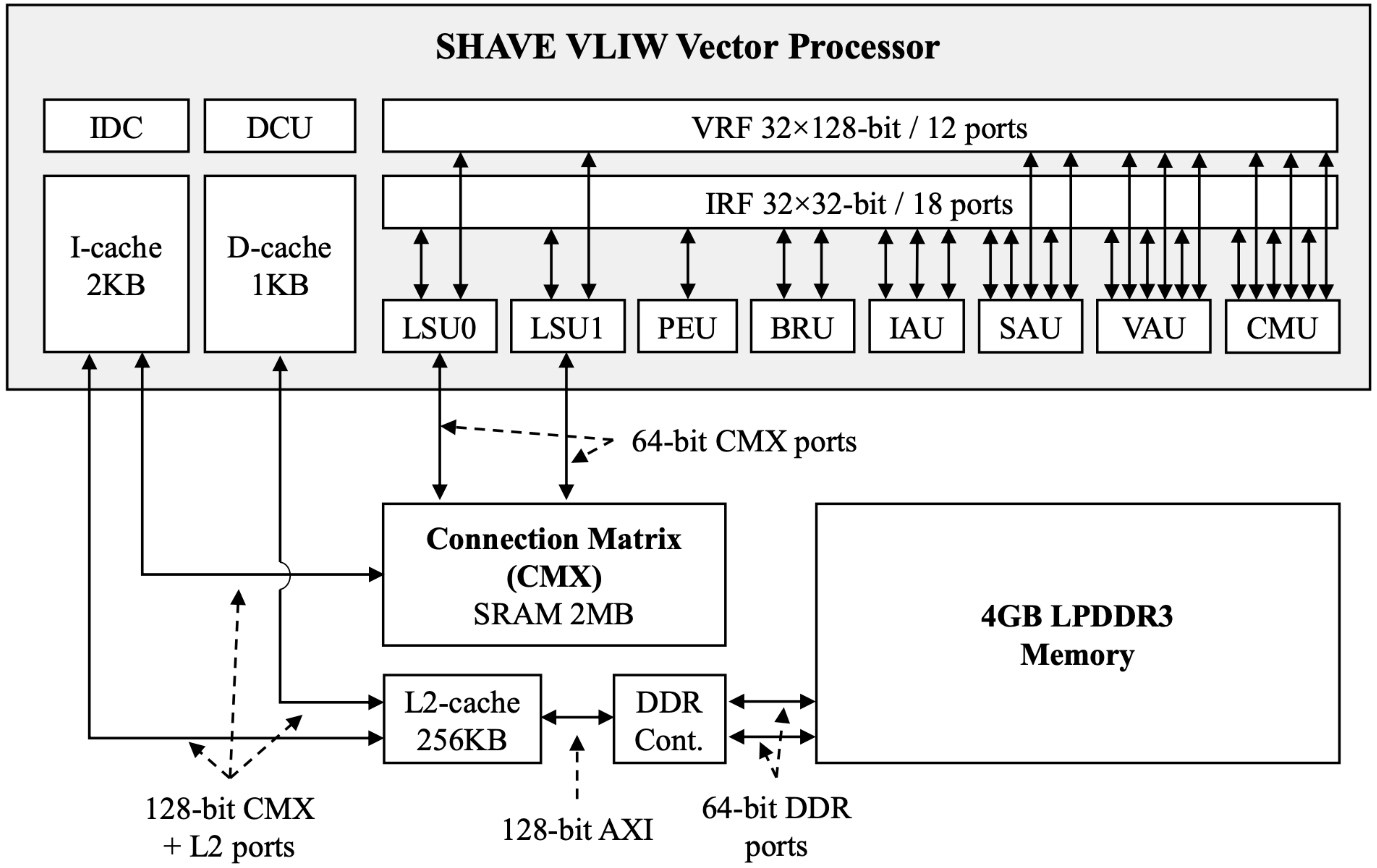
- Application host connected via USB
- Neural Compute API offers frontend
- RISC Processor handles communication and runs realtime OS
- 16 SHAVE Cores for parallel SIMD computations
- Stacking of multiple NCS possible



Neural Compute Stick [5]

# Intel Neural Compute Stick 2

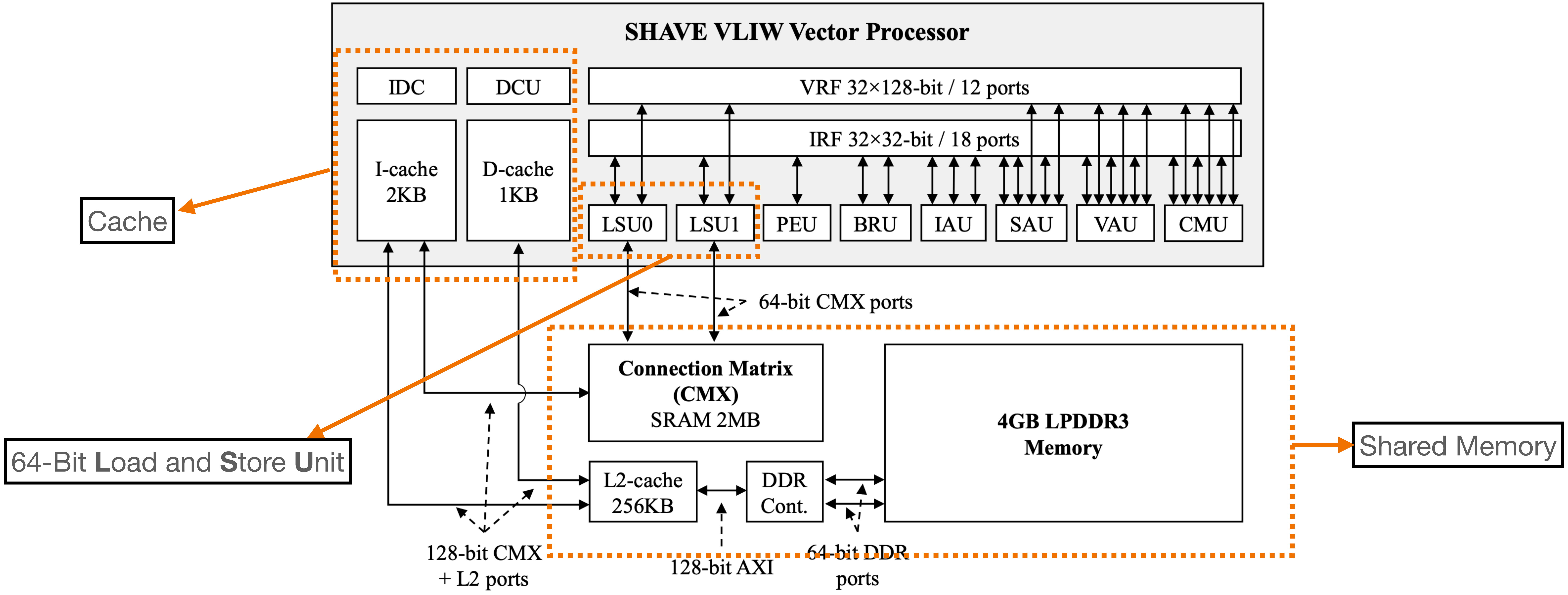
## Inside a SHAVE Core



SHAVE Processor [5]

# Intel Neural Compute Stick 2

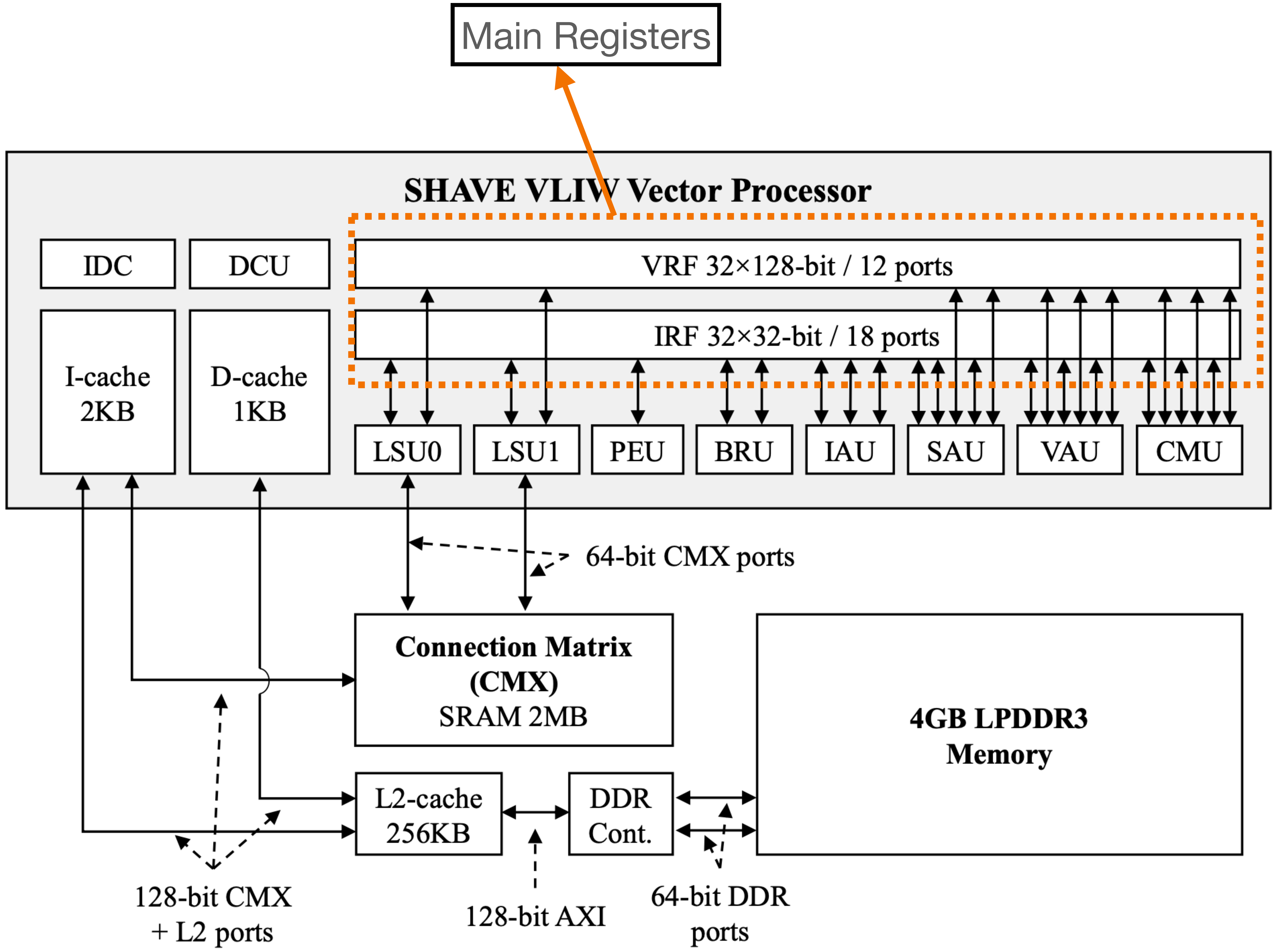
## Inside a SHAVE Core



SHAVE Processor [5]

# Intel Neural Compute Stick 2

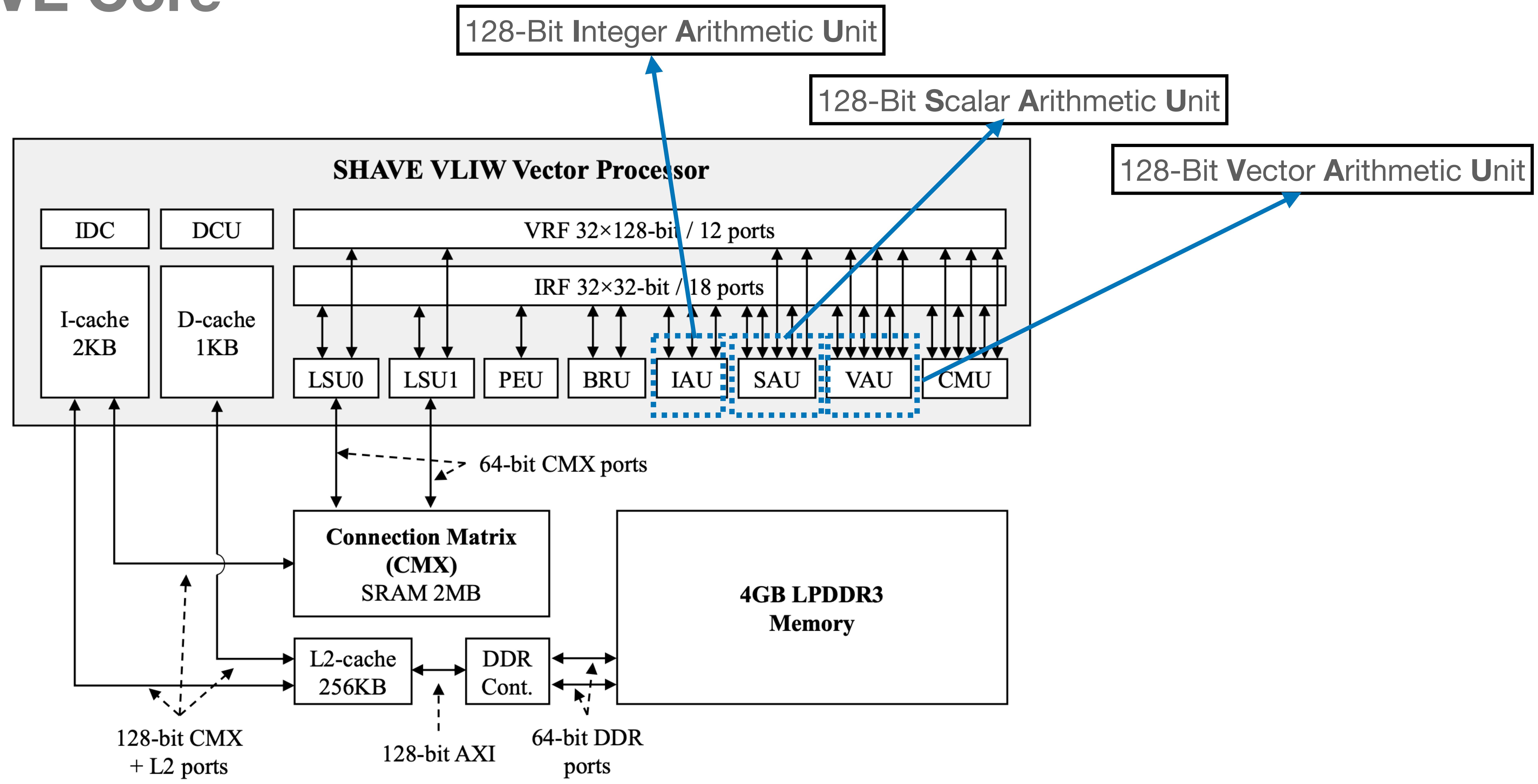
## Inside a SHAVE Core



SHAVE Processor [5]

# Intel Neural Compute Stick 2

## Inside a SHAVE Core



SHAVE Processor [5]

# Why new hardware if there are CPUs/GPUs?

## TPUs operate between traditional hardware

**Advantages:**

- Cheap (59-69 $)
- low power solution (1-2 Watt)
- Realtime inference for slow devices (Concept: Train in the cloud, inference everywhere (on the edge))
- Application specific design (No resources wasted)

*Performance comparisons in the practical presentation!*

# Coming back to the project

## Outlook

- Todo: Deploying the pretrained autoencoder on the Edge and the NCS
- Difficulties:
    - Different frontends (Tensorflow vs. Tensorflow Lite)
    - Conversion to fixed point (Quantization [6])
- Performance measure:
    - Accuracy / Reconstruction Loss (drop because of fixed point?)
    - Latency (realtime capabilities)
    - Power usage (usefulness as a low power accelerator)

# Thanks for your attention!

## Any questions?

**Jannis Wolf, B.Sc., 20.5.2021**

# Sources

[1] Charte et al., A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines, 2018

[2] https://coral.ai/static/files/Coral-USB-Accelerator-datasheet.pdf (accessed 19.5.21)

[3] https://www.intel.com/content/dam/support/us/en/documents/boardsandkits/neural-compute-sticks/NCS2_Datasheet-English.pdf (accessed 19.5.21)

[4] Yazdanbakhsh et al., An Evaluation of Edge TPU Accelerators for Convolutional Neural Networks, 2021

[5] Rivas-Gomez et al., Exploring the Vision Processing Unit as Co-processor for Inference, 2018

[6] https://heartbeat.fritz.ai/8-bit-quantization-and-tensorflow-lite-speeding-up-mobile-inference-with-low-precision-a882dfcafbbd (accessed 19.5.21)