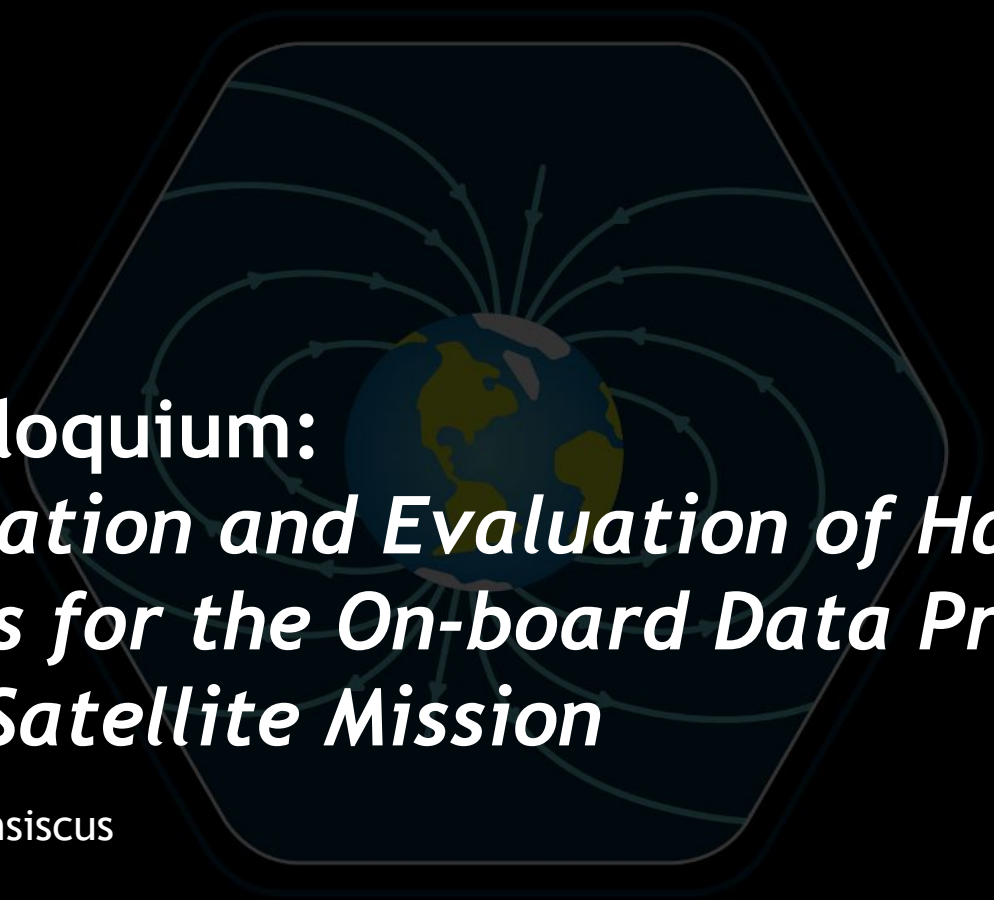
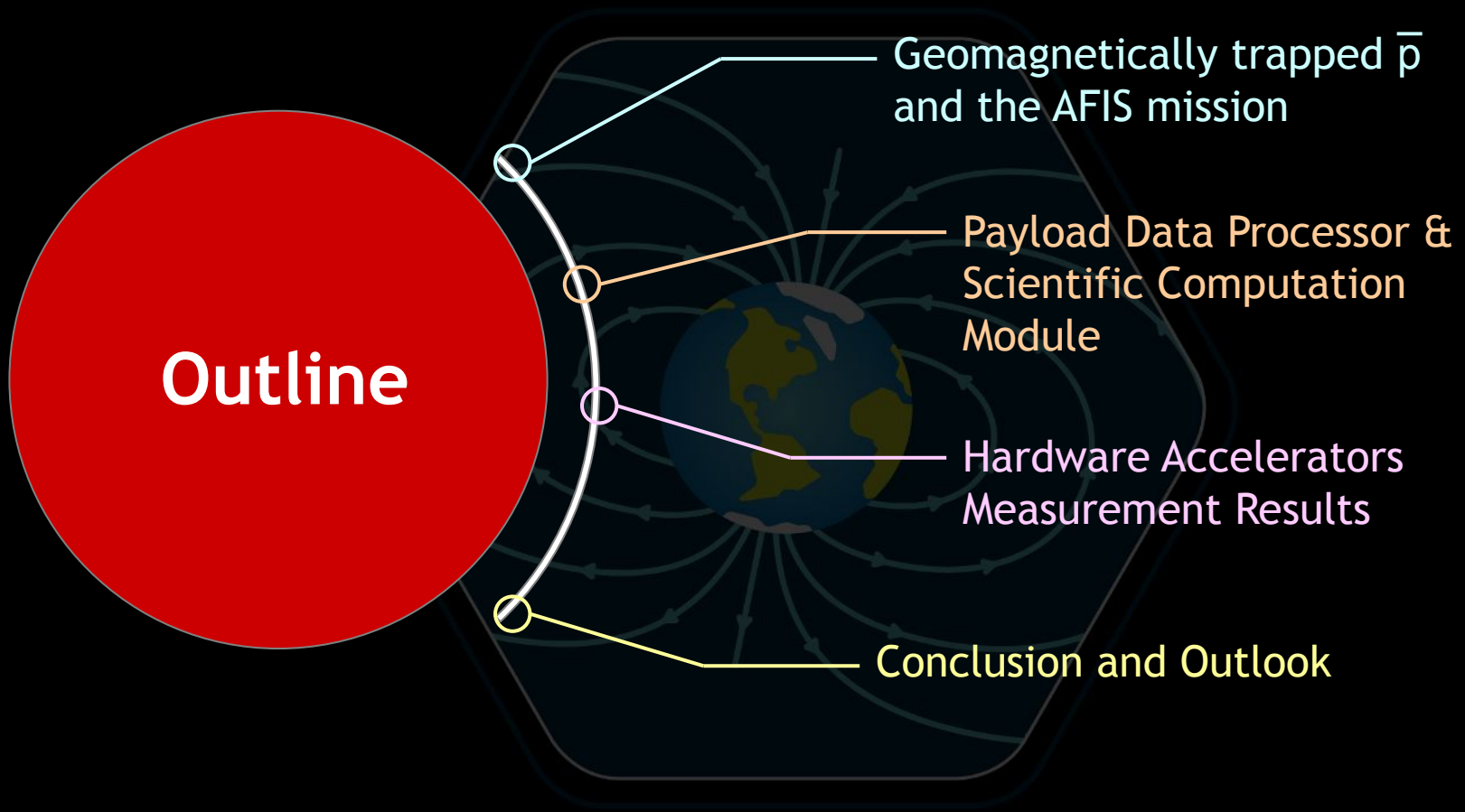
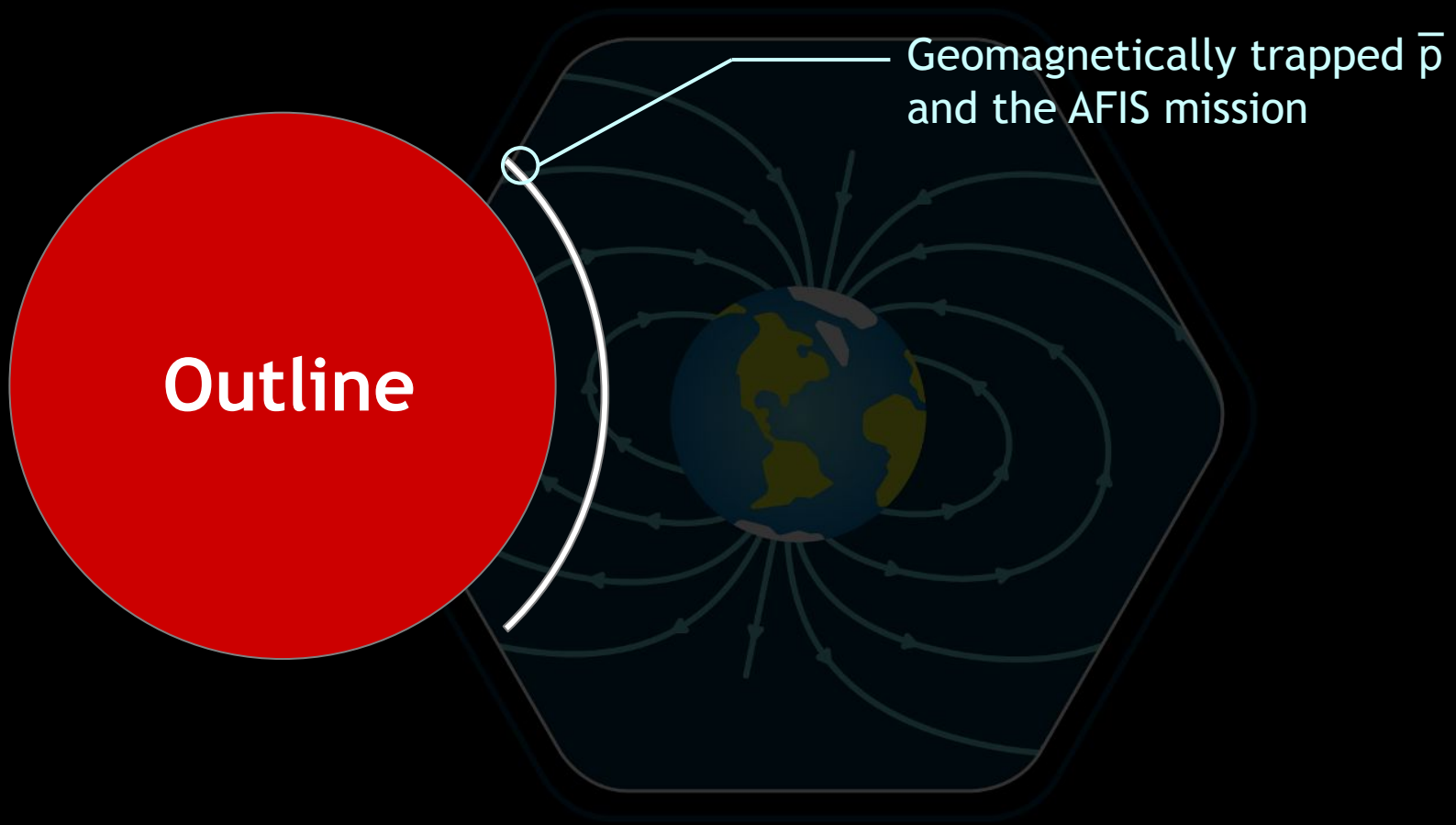


Master's Colloquium:
Characterization and Evaluation of Hardware Accelerators for the On-board Data Processing of the AFIS Satellite Mission

Limodya, Vernando Fransiscus
Munich, April 24th 2024

Outline



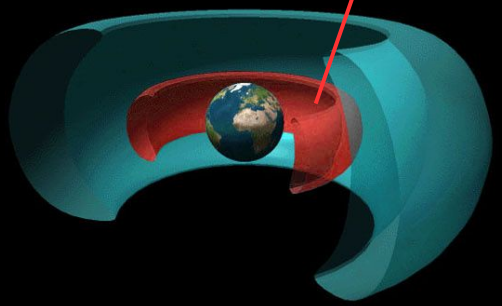

Geomagnetically trapped $\bar{p}$
and the AFIS mission
Outline

Geomagnetically Trapped Antiprotons

PAMELA a Payload for Antimatter Matter Exploration and Light–nuclei Astrophysics

The PaMeLa Mission [1]

inner van Allen radiation belt



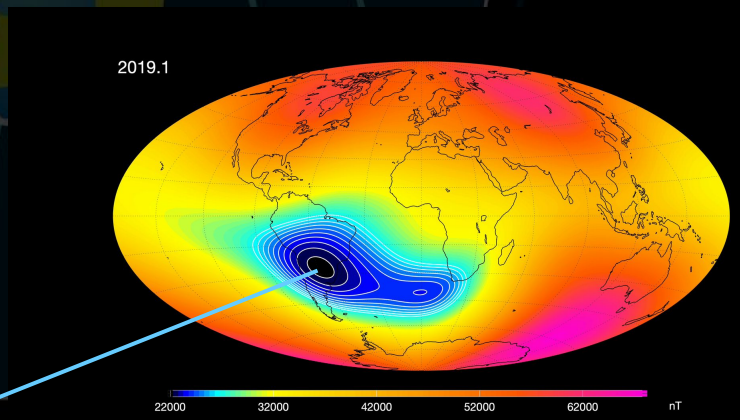
An illustration of the van Allen radiation belts [2]

weaker magnetic field
→ the inner belt dips down to a lower altitude

The discovery of geomagnetically trapped cosmic ray antiprotons

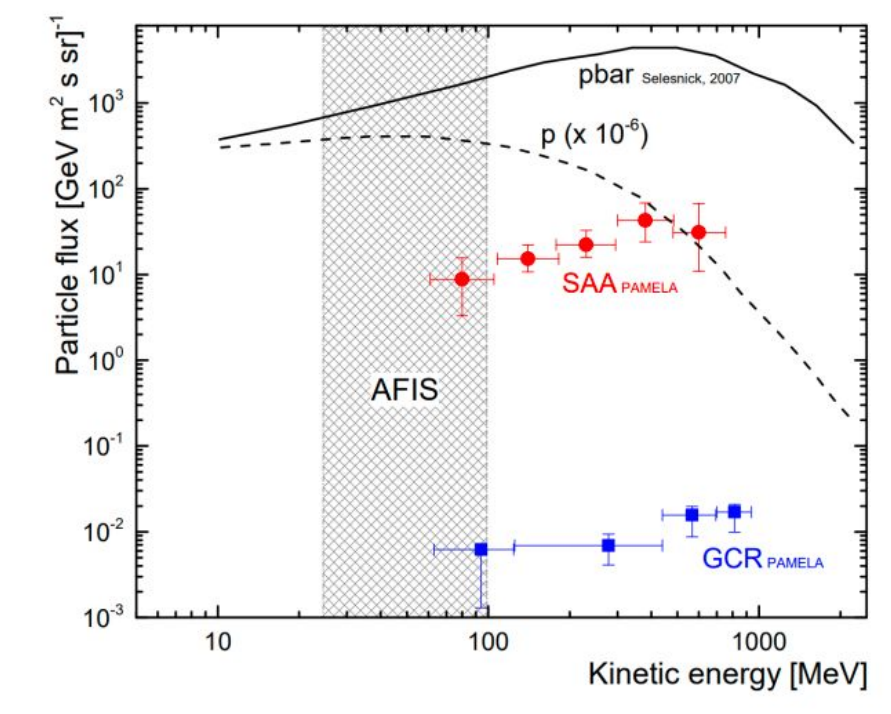
O. Adriani[1,2], G. C. Barbarino[3,4], G. A. Bazilevskaya[5], R. Bellotti[6,7], M. Boezio[8], E. A. Bogomolov[9], M. Bongi[2], V. Bonvicini[8], S. Borisov[10,11,12], S. Bottai[2], A. Bruno[6,7,18], F. Cafagna[6], D. Campana[4], R. Carbone[4,11], P. Carlson[13], M. Casolino[10], G. Castellini[14], L. Consiglio[4], M. P. De Pascale[10,11], C. De Santis[10,11], N. De Simone[10,11], V. Di Felice[10], A. M. Galper[12], W. Gillard[13], L. Grishantseva[12], G. Jerse[8,15], A. V. Karelin[12], M. D. Kheymits[12], S. V. Koldashov[12], S. Y. Krutkov[9], A. N. Kvashnin[5], A. Leonov[12], V. Malakhov[12], L. Marcelli[10], A. G. Mayorov[12], W. Menn[16], V. V. Mikhailov[12], E. Mocchiutti[8], A. Monaco[6,7], N. Mori[1,2], N. Nikonov[9,10,11], G. Osteria[4], F. Palma[10,11],

Significant flux of antiprotons is reportedly detected in the South Atlantic Anomaly (SAA)



The location of the SAA [3]

PAMELA Results



The results from the PAMELA mission [4]

- Measured $\bar{p}$ flux in the SAA is 3 orders higher than the flux measured outside of the SAA (GCR)
- Selesnick, et. al. – two main mechanisms
  - Direct proton-antiproton production

$$p + A \rightarrow p + \bar{p} + p + X$$

  - CRANbarD → **dominant mechanism**

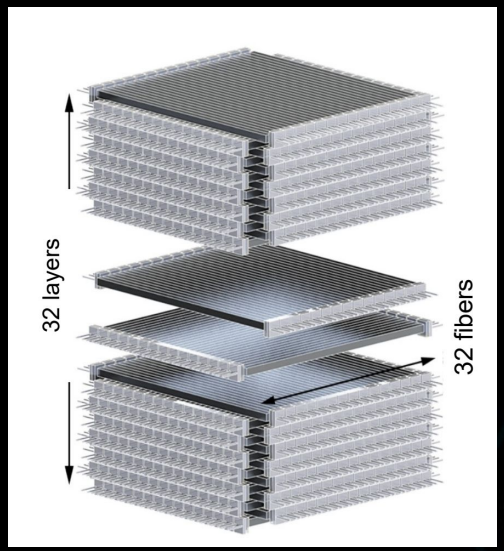
$$p + A \rightarrow p + n + \bar{n} + X$$

$$\bar{n} \rightarrow \bar{p} + e^+ + \nu_e$$

- Measurements off by 2 orders of magnitude to theoretical predictions

In order to adjust the theory and experiment, data from lower energies are required.

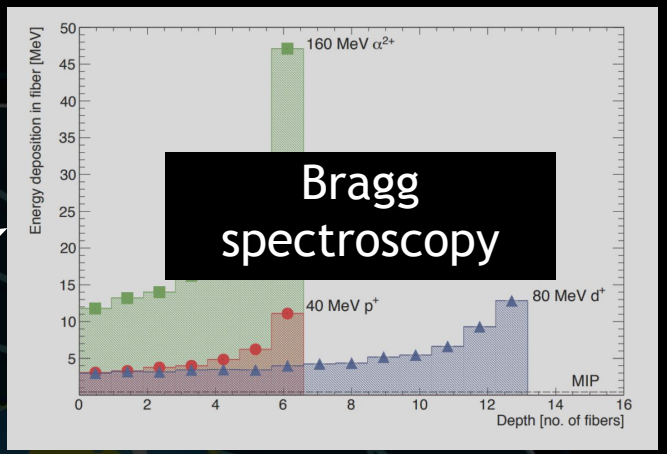
AFIS (Antiproton Flux in Space)

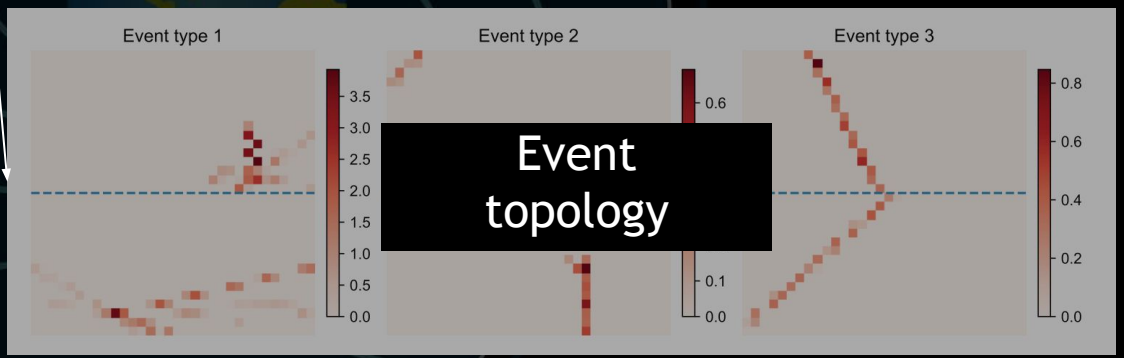


The Active Detection Unit (ADU) for the AFIS Mission [5]

32 layers, each contain 32 plastic scintillating fibers, oriented 90 degrees with respect to its adjacent neighbour

particle identification using





Bragg Spectroscopy

- Energy deposition per unit length given by the Bethe-Bloch formula

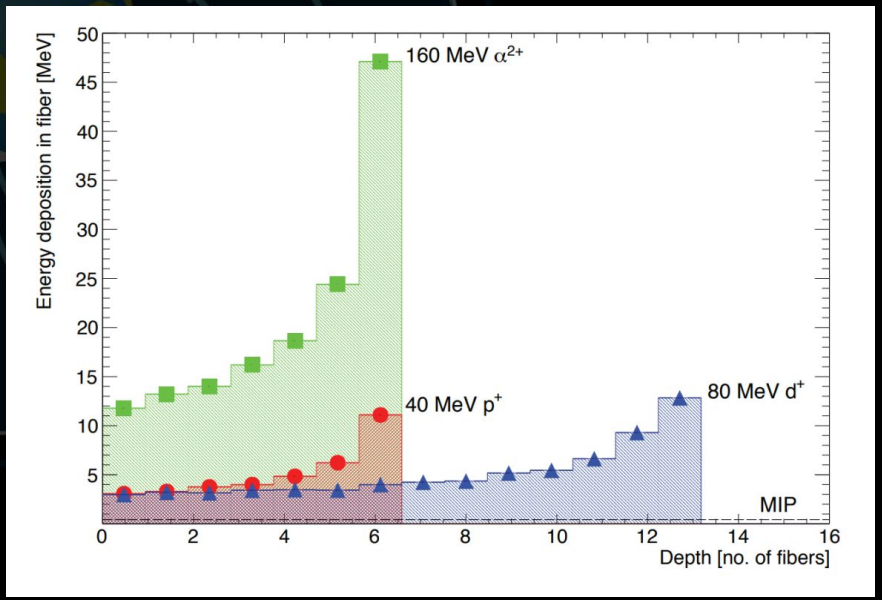
$$-\left\langle \frac{dE}{dx} \right\rangle = \frac{4\pi}{m_e c^2} \frac{nZ^2}{\beta^2} \left(\frac{e^2}{4\pi\epsilon_0} \right)^2 \left[\ln\left(\frac{2m_e c^2 \beta^2}{I(1-\beta^2)} \right) - \beta^2 \right]$$

- Particle passes through the detector material

→ particle velocity decreases

→ energy deposition per unit length increases

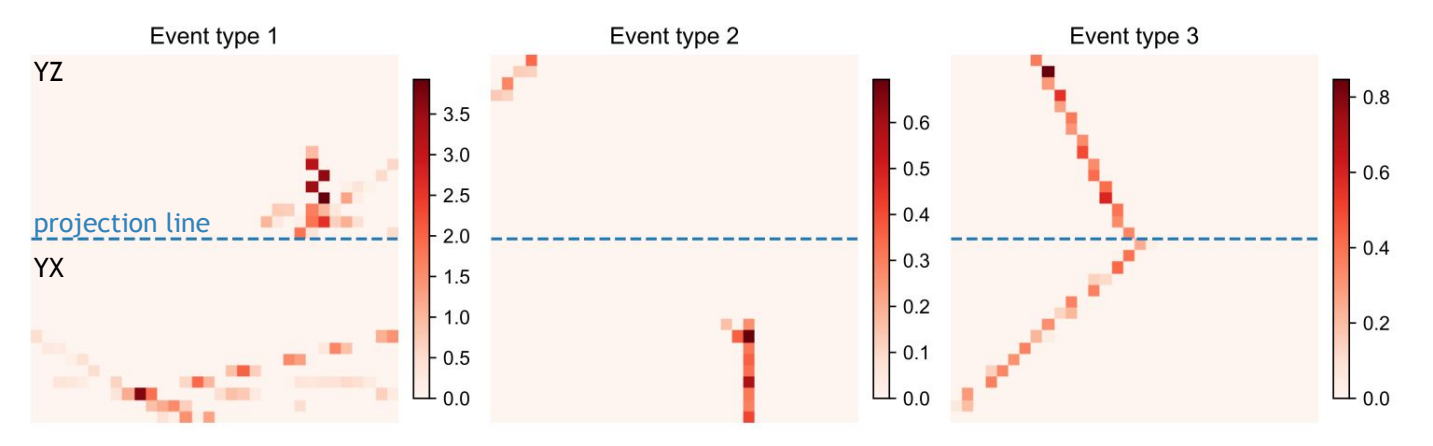
→ results in Bragg curves



Bragg Curves [5]

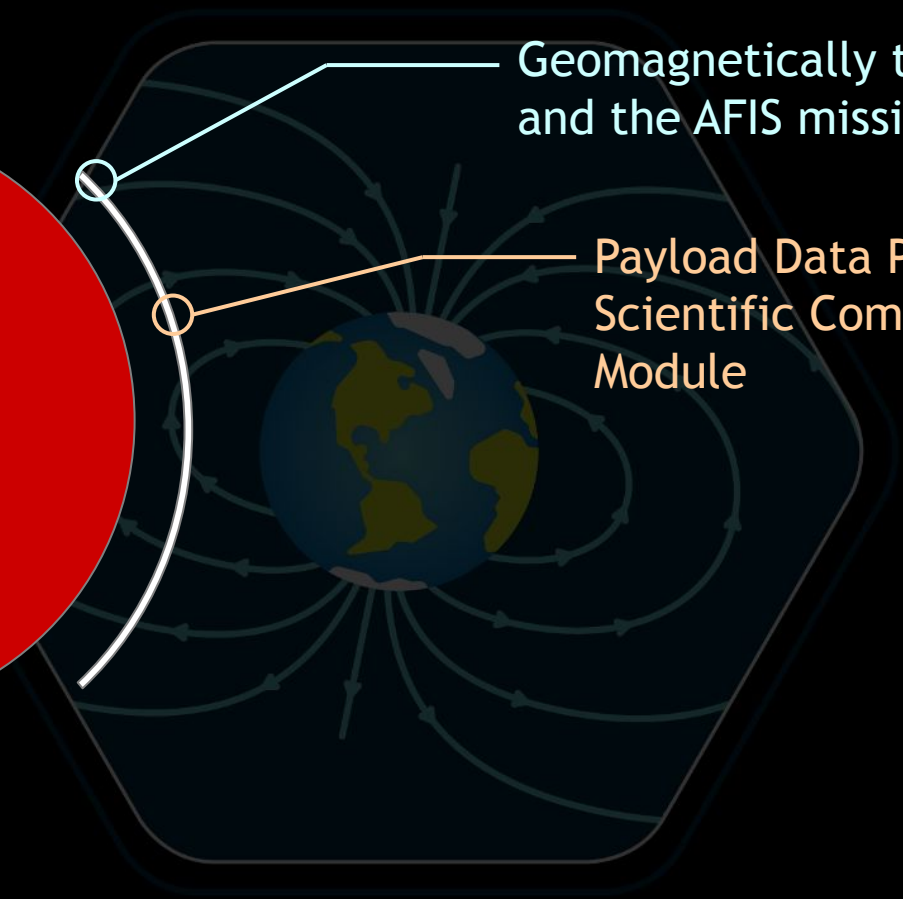
Event Topology

One event has the size of 1024 x 12 bits
Pixel value = energy in MeV



low-energetic $\bar{p}$	low-energetic p	high-energetic $\bar{p}/p$
$\bar{p}$ passes through detector → annihilation CS ↑	p completely stopped in the detector material, but no annihilation	particle not stopped inside the detector → can be both from $\bar{p}/p$

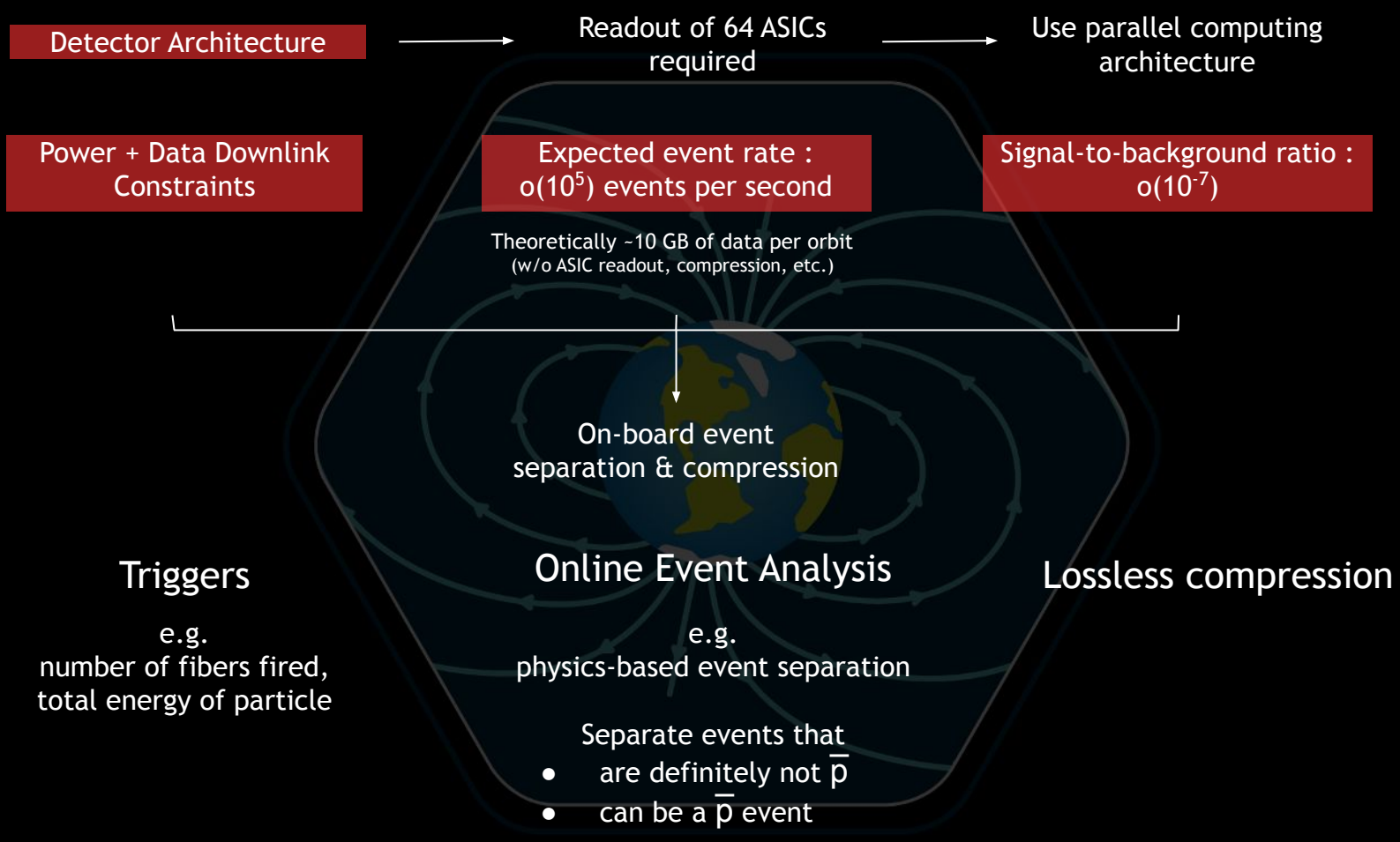
Outline



Geomagnetically trapped $\bar{p}$ and the AFIS mission

Payload Data Processor & Scientific Computation Module

PDP & Processing Chain



Scientific Computation Module

We can integrate specialized hardware (**hardware accelerator**) for neural networks

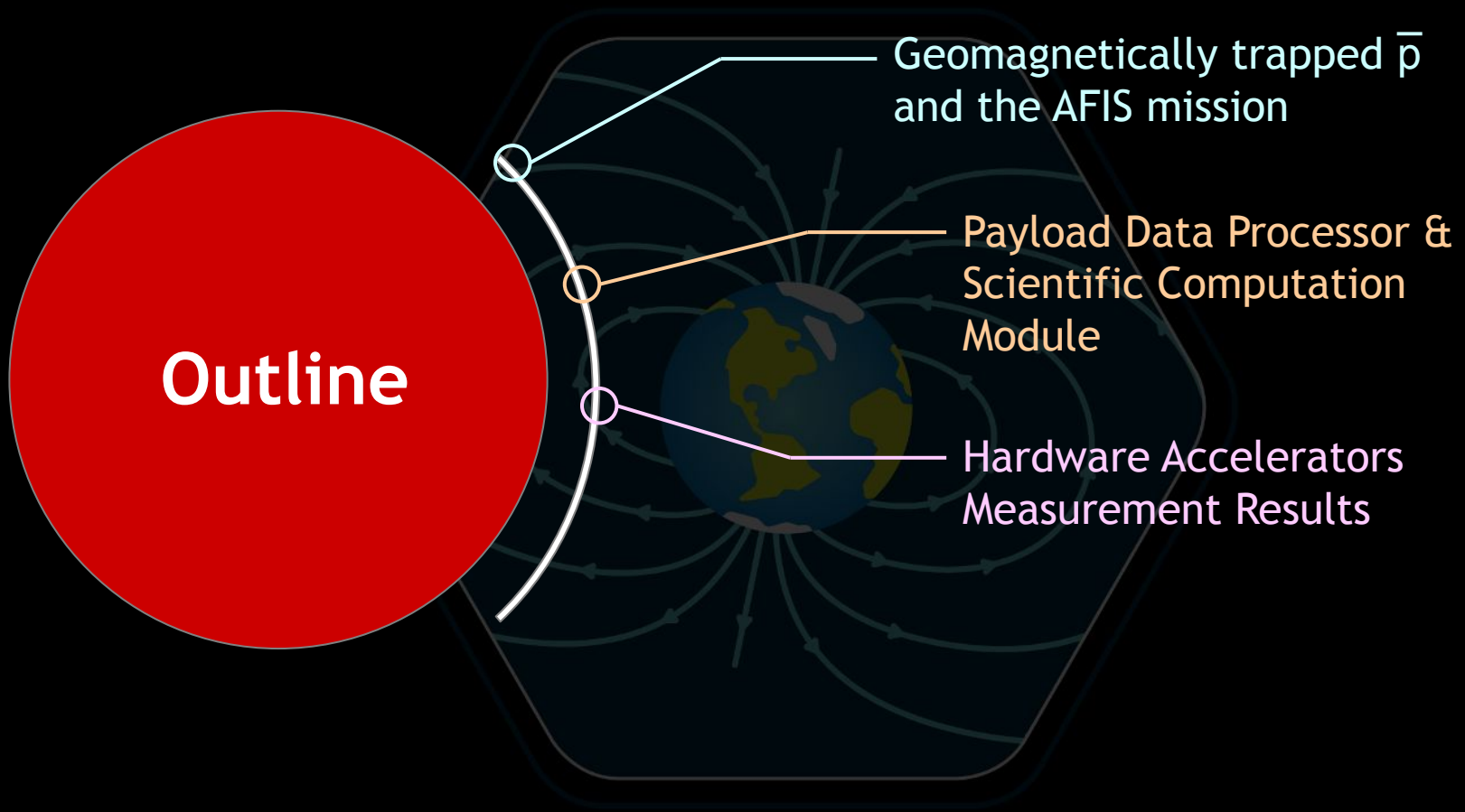
Evaluation of hardware accelerators based on :

- Throughput
- Power Consumption
- Integrability to the FPGA
- Model requirements

→ Optimize models for each hardware accelerator

Analysis of different NN architectures

Outline

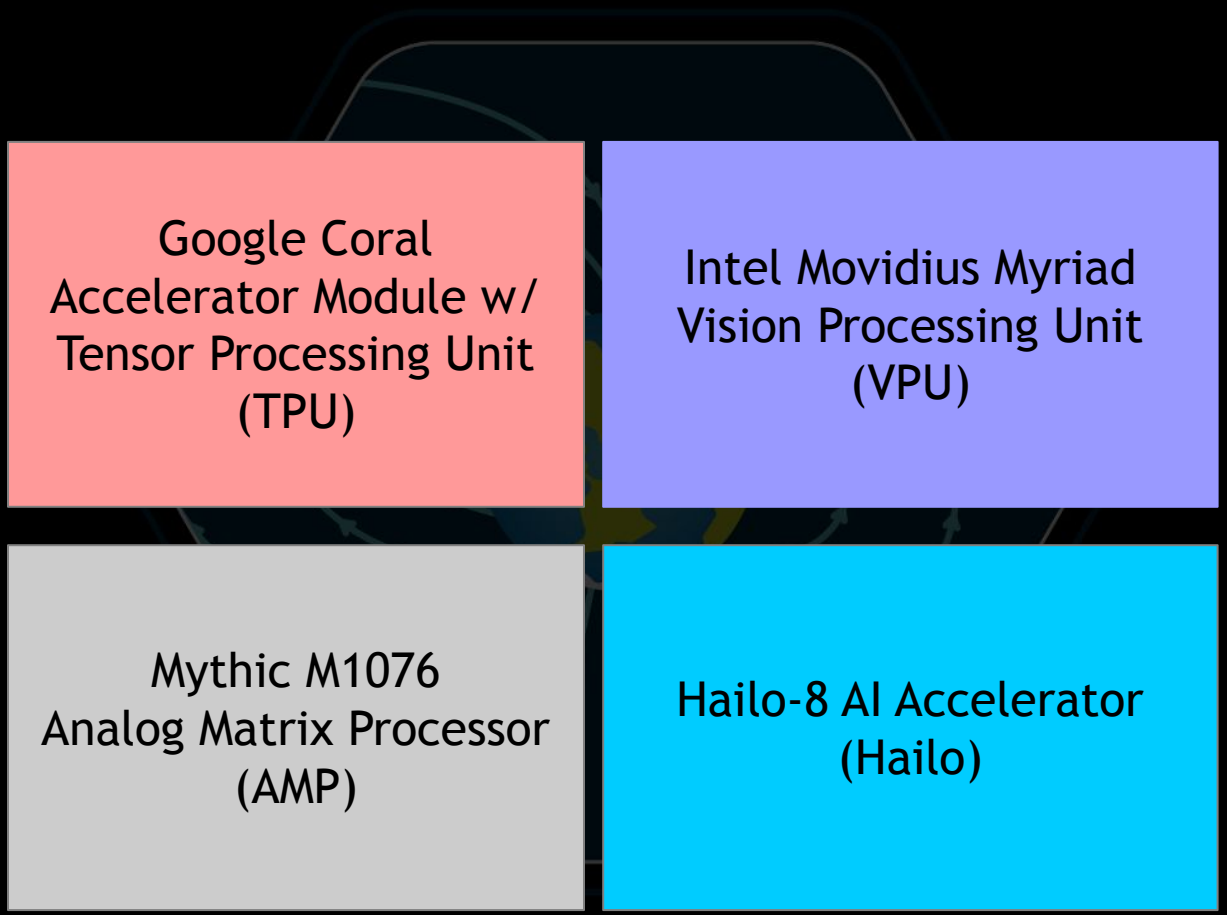


Geomagnetically trapped $\bar{p}$ and the AFIS mission

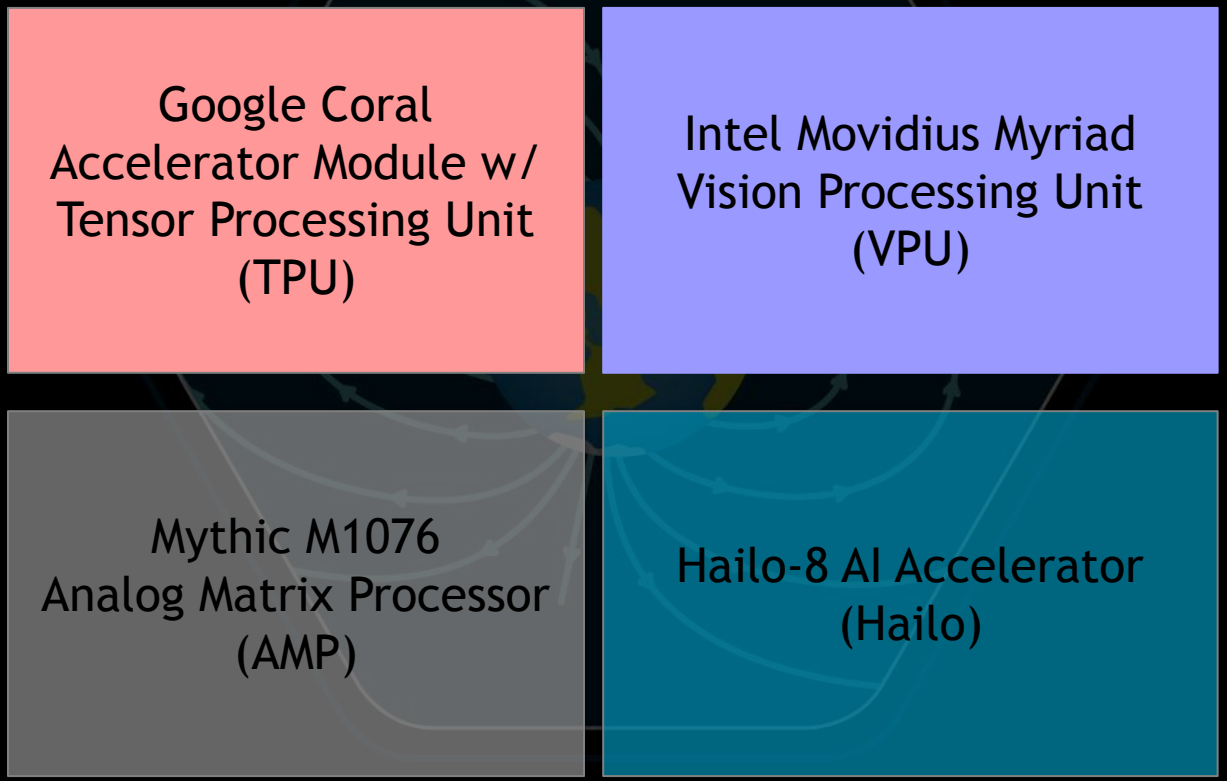
Payload Data Processor & Scientific Computation Module

Hardware Accelerators Measurement Results

Hardware Accelerators

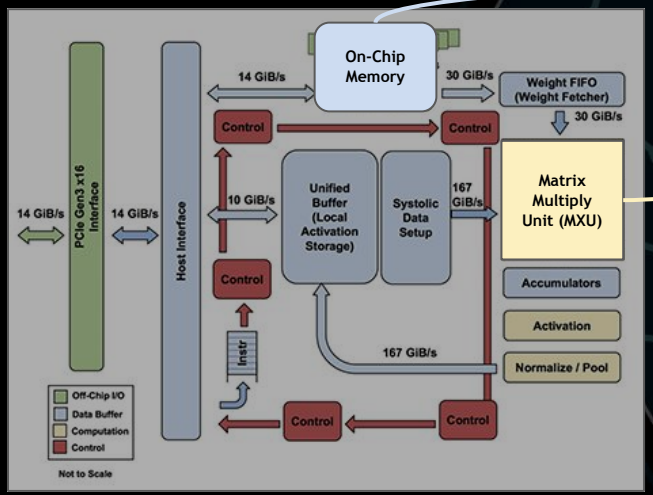


Hardware Accelerators



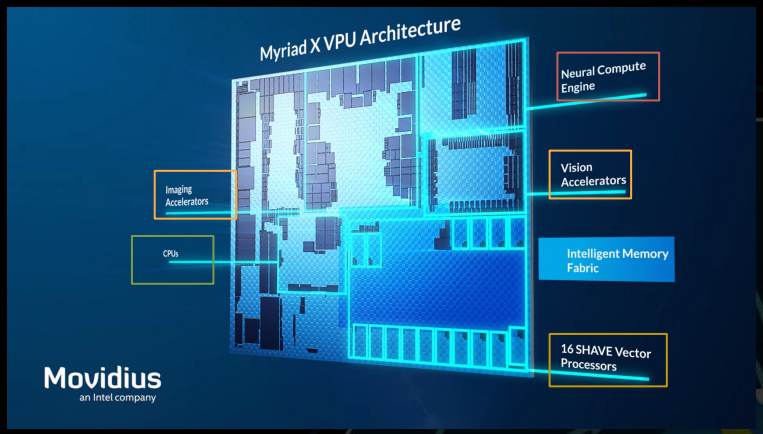
Coral Accelerator Module

Format	Multi-chip module, comprising of a Tensor Processing Unit (TPU) and a Power Management IC (PMIC)
Max. clock frequency	500 MHz
On-Chip Memory	8 MB SRAM
Theoretical Performance	**4096 (= 64 x 64) Multiply-Accumulates (MACs)** → max. 4 TOPS
Quantization	INT8
Interfaces	PCIe Gen2 x1, USB 3.1
Radiation Tolerance	TID tests – 5-year lifetime in LEO

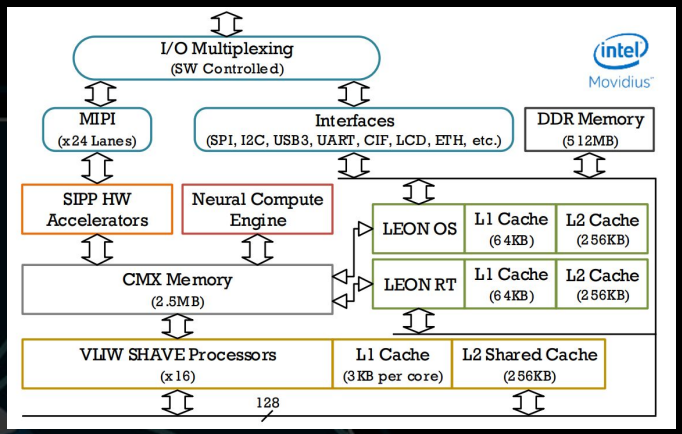


Computing architecture of TPU [7]

Vision Processing Unit



Myriad X VPU high level architecture [8]

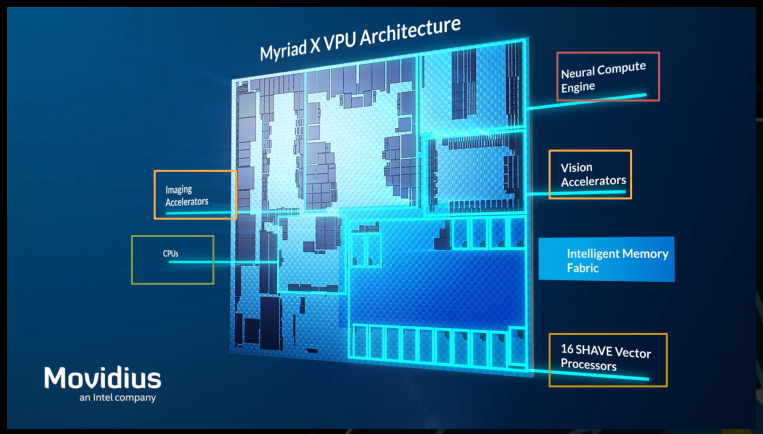


Myriad X VPU computing architecture [9]

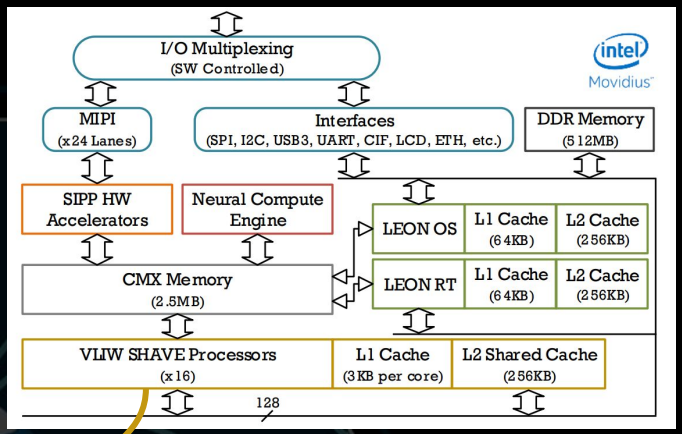
Format	System-on-Chip
Base clock frequency	700 MHz
On-Chip Memory	512 MB DDR Memory
Interfaces	USB 3.1, PCIe Gen 3, Quad SPI, I2C, 16 MIPI lanes

Theoretical Performance	Max. 1 TOPS
Quantization	Floating Point 16
Radiation Tolerance	Average SEFIs per day, 0.083 due to heavy ions & 0.0035 due to protons

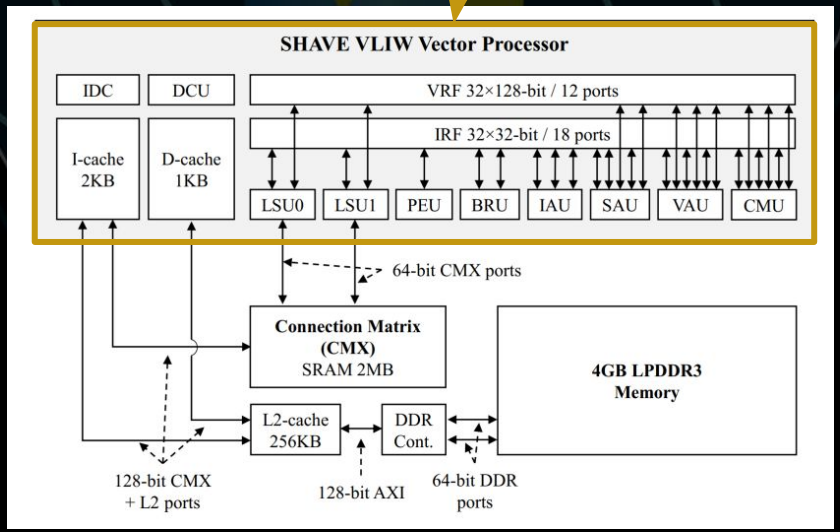
Vision Processing Unit



Myriad X VPU high level architecture [8]

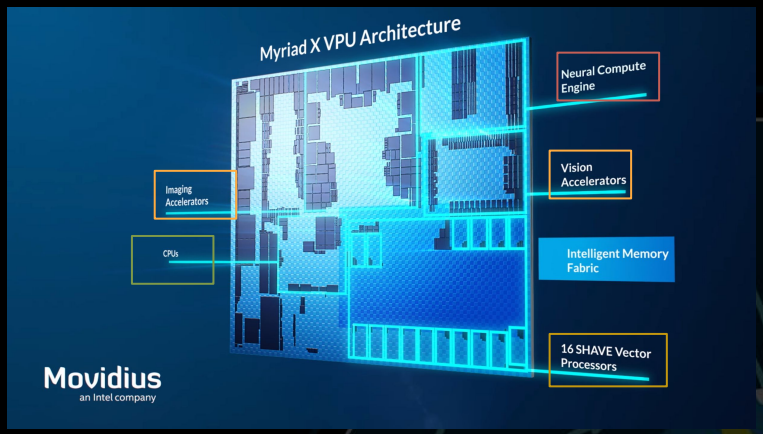


Myriad X VPU computing architecture [9]

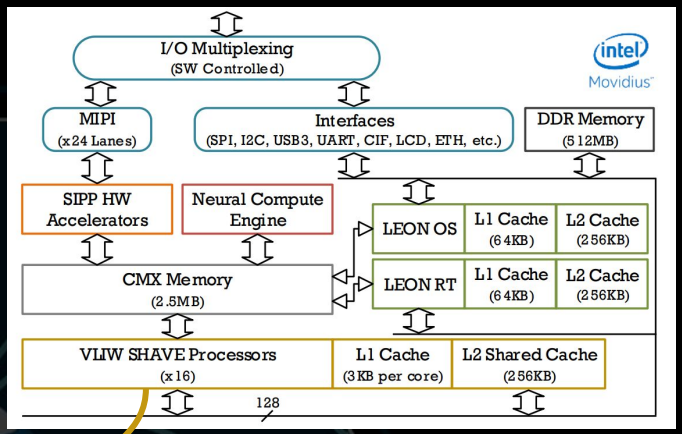


The computing architecture of the Myriad VPU 2, the predecessor [10]

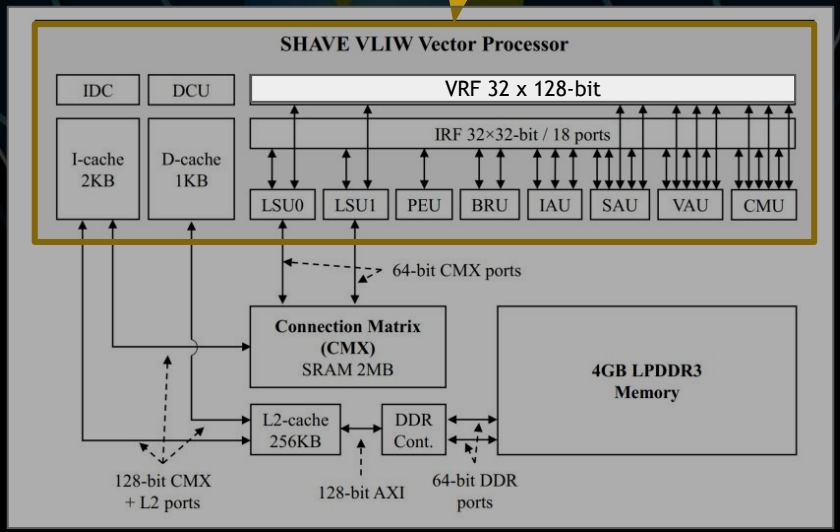
Vision Processing Unit



Myriad X VPU high level architecture [8]

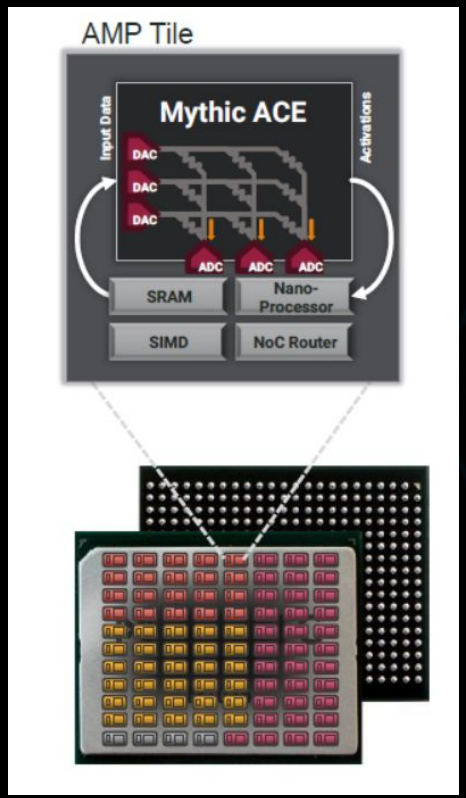


Myriad X VPU computing architecture [9]



The computing architecture of the Myriad VPU 2, the predecessor [10]

Analog Matrix Processor



M1076 Analog Matrix Processor [11]

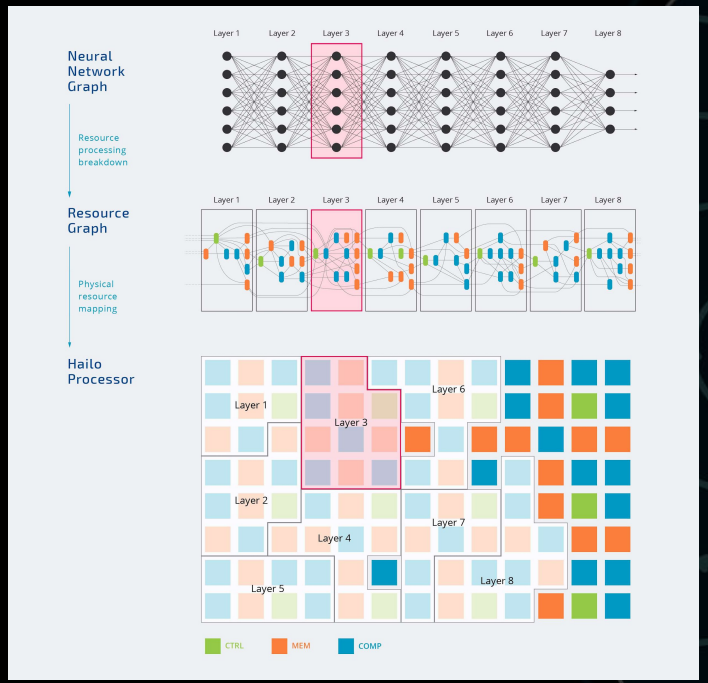
- ACE (Analog Compute Engine), where potentiometers are arranged in crosspoint arrays
- Multiplication based on Ohm's Law

$$I = G \cdot U$$

product (current) — weights (conductance) — input values (voltage)

- Addition based on 1st Kirchoff's Law → row currents add up

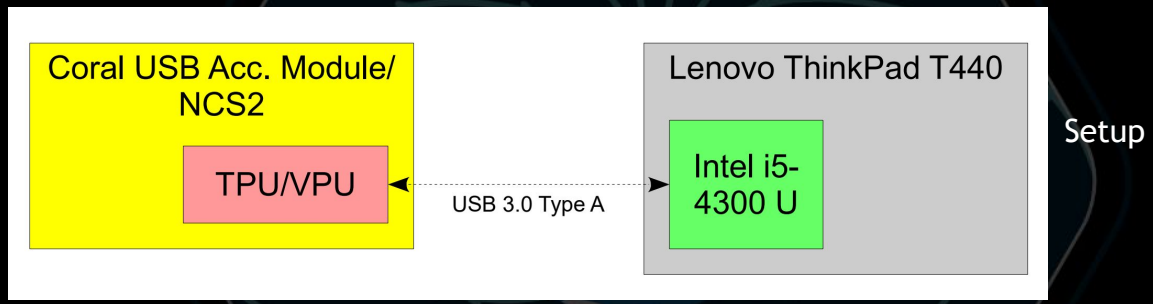
Hailo-8



Hailo-8 AI Accelerator [12]

- Consist of tiles that are responsible for control, memory and computations
- Resource assignment is made to optimize the dataflow

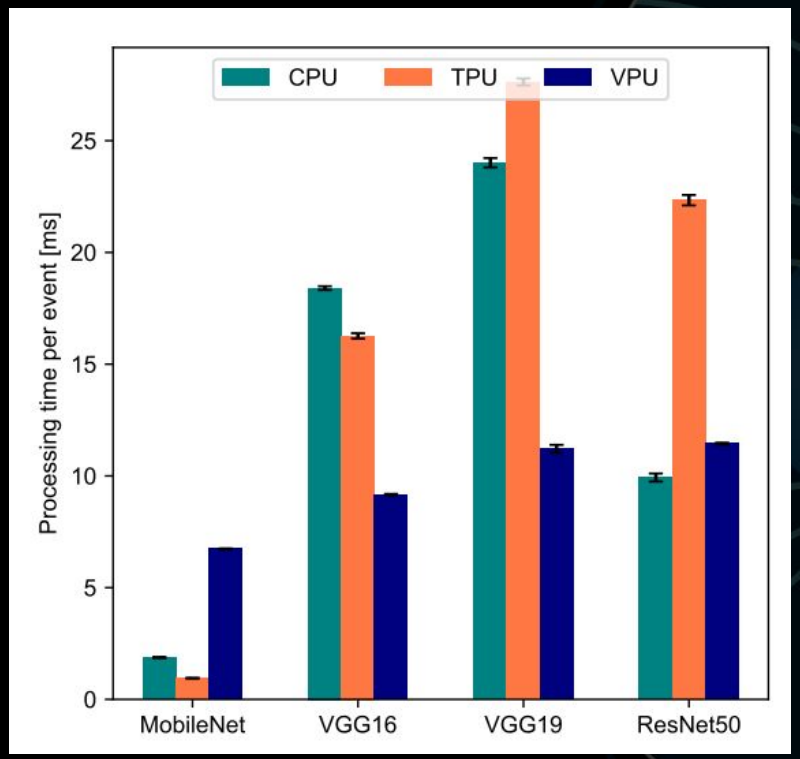
Measurement Method



Setup

As input : 10^4 simulated proton events with 10 repetitions

Common NN Architectures



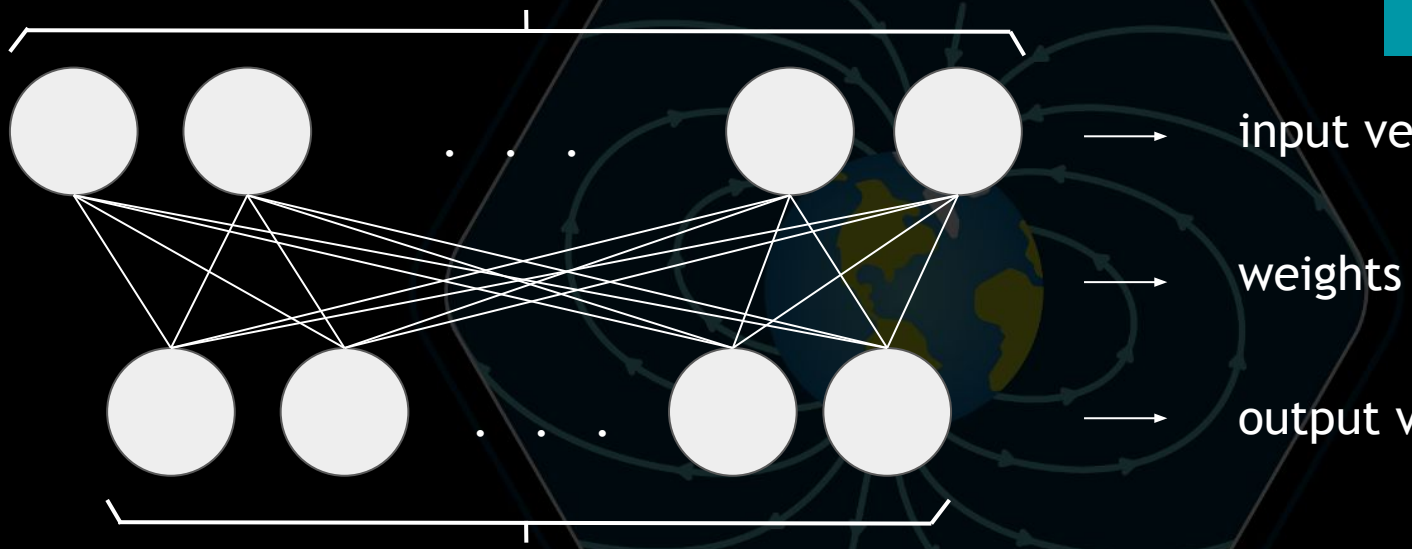
- No direct correlation between model size to performance

→ An analysis of the computing architecture is needed

Matrix Multiplications

NN architecture

Feed forward NN with 2 layers



$$y = Wx + b$$

***b* is bias vector**

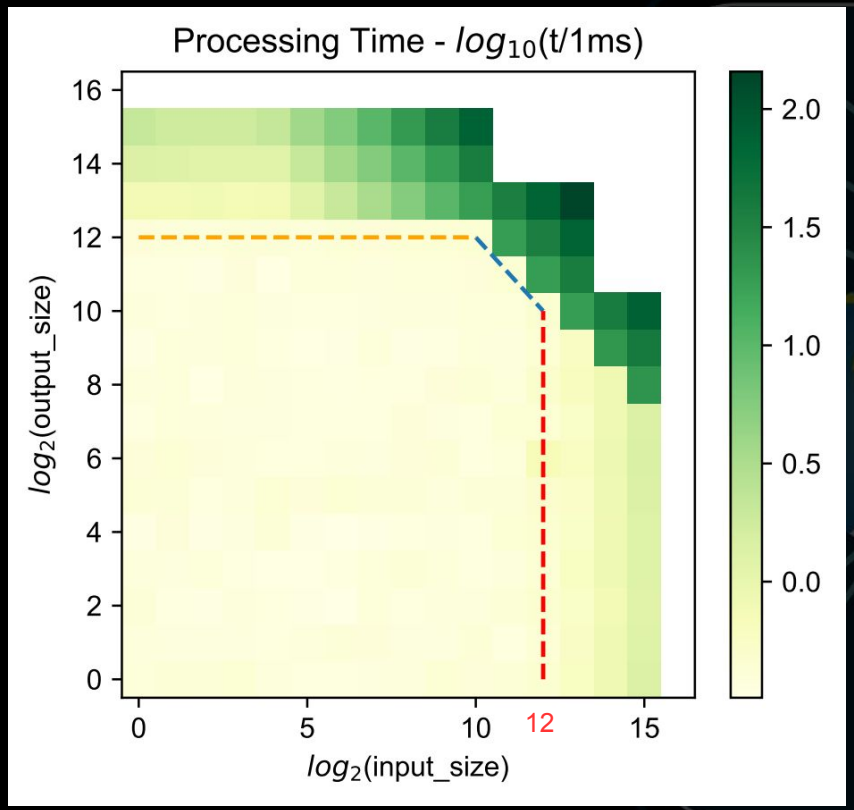
input vector ***x***

weights matrix **W**

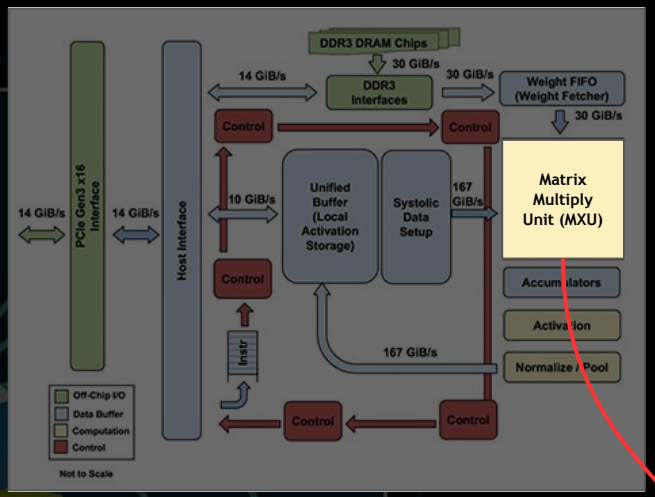
output vector ***y***

The number of MACs per output neuron = input_size
Total number of MACs = input_size * output_size

Matrix Multiplications



TPU



if over the red line, the number of MACs > 64 x 64

↓

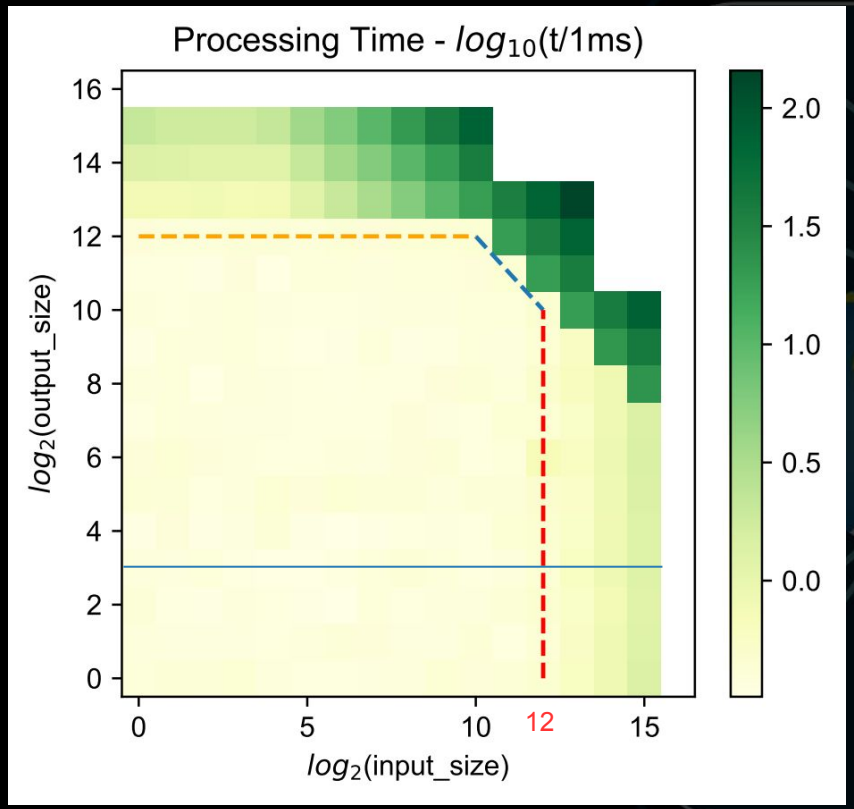
several data cycles for one output neuron, since the MXU has only 64 x 64 MAC units

↓

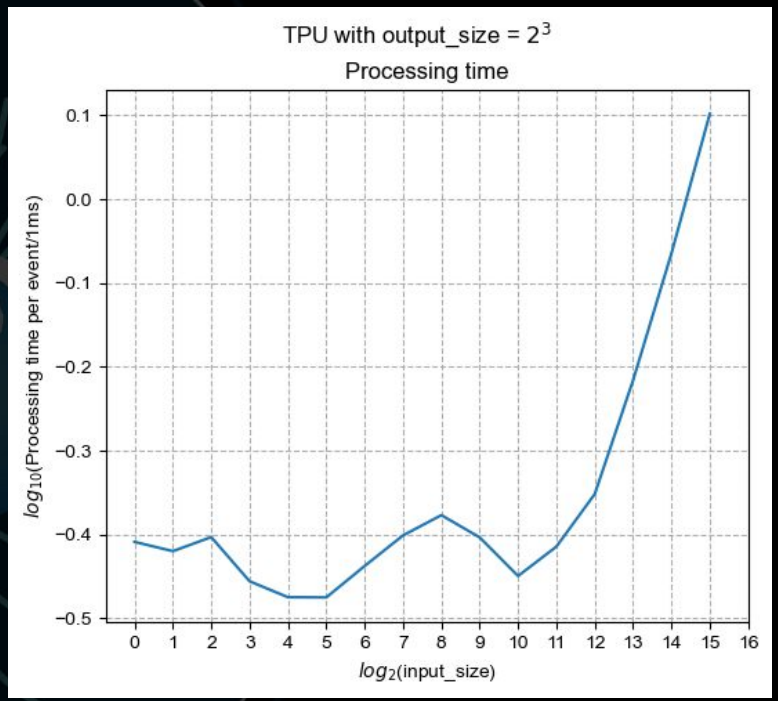
more processing time needed

– – – one output neuron requires 64 x 64 MAC operations

Matrix Multiplications

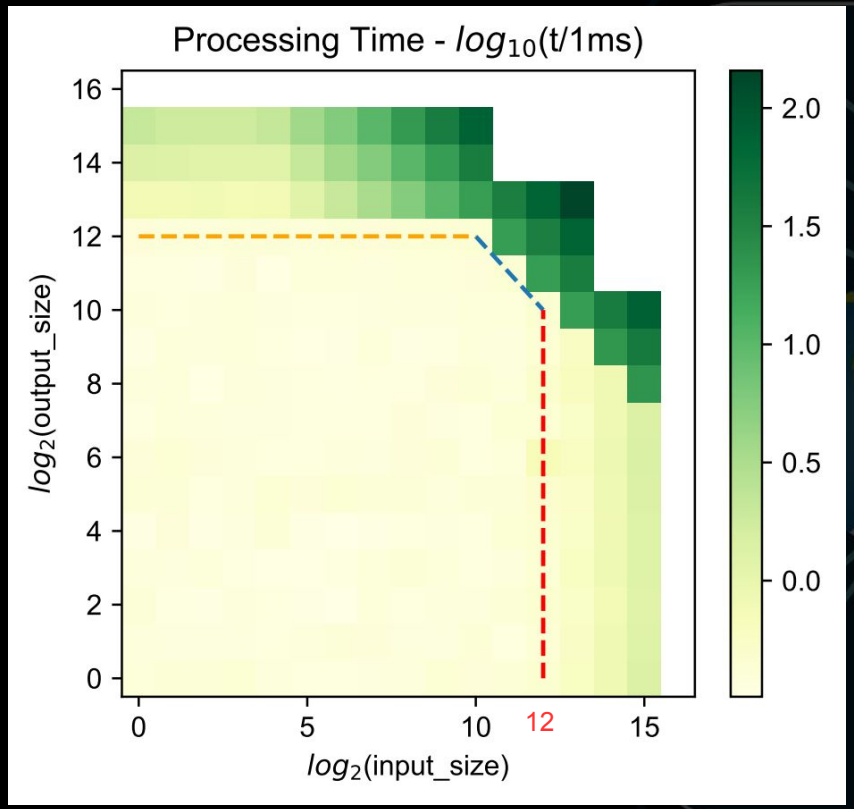


TPU

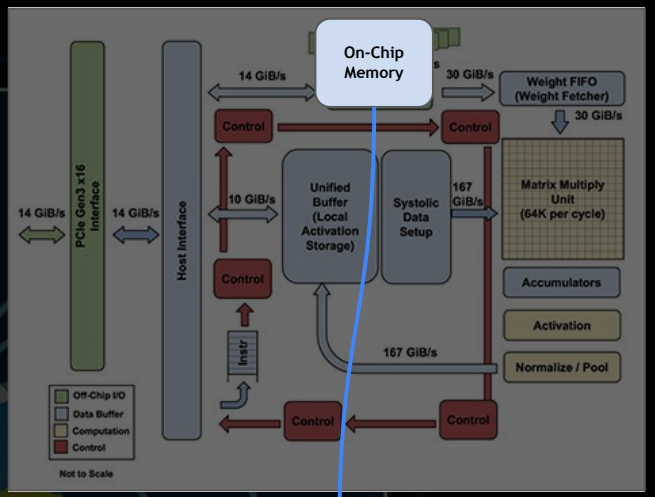


one output neuron requires 64 x 64 MAC operations

Matrix Multiplications



TPU



over the blue line, model weights have the size > 8 MB

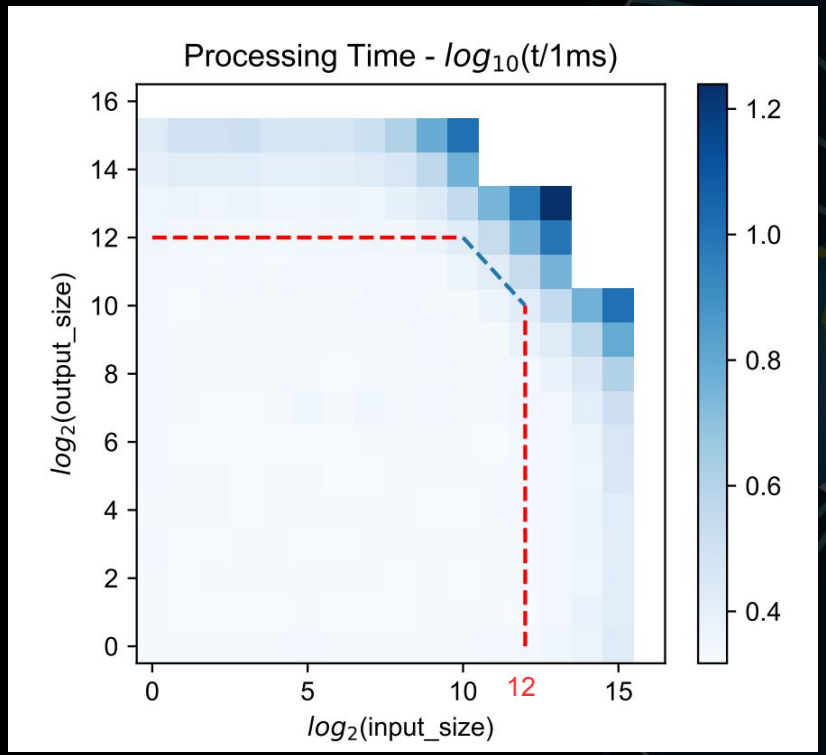
weights are stored externally in the host's RAM

transmission bottleneck from host's RAM

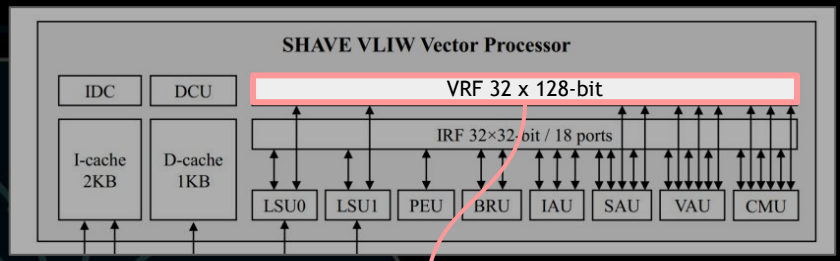
(significantly) more processing time needed

– – – Model size is < ~8 MB

Matrix Multiplications



VPU



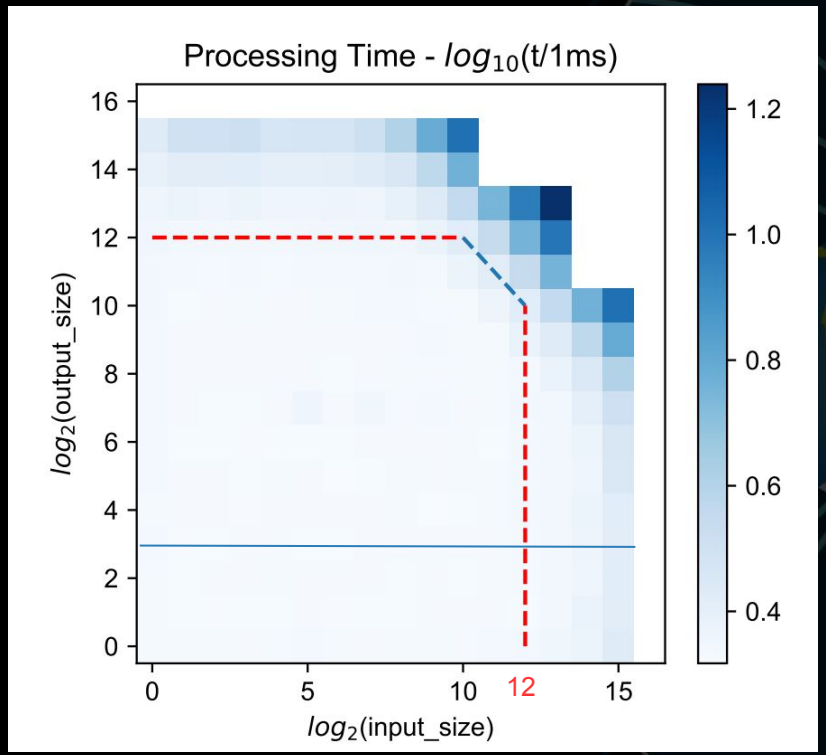
Total VRF capacity = 16 (SHAVEs) x 32 x 128 bit
= 4096 * 16 bit = 4096 Floating Point 16 values

Like the TPU, only 4096 Multiplications or Accumulations can be performed in one instruction cycle

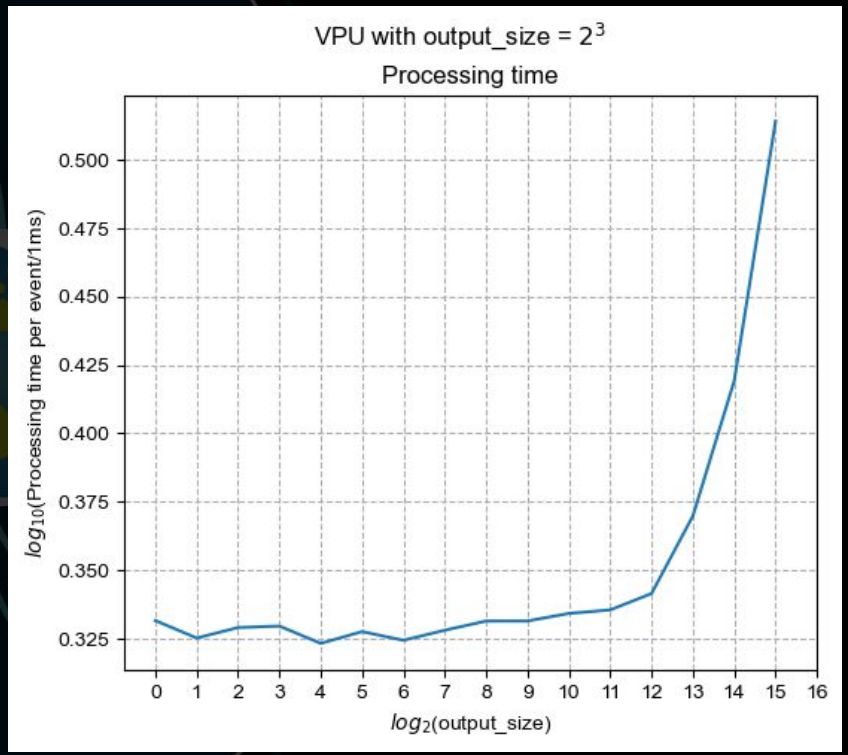
More processing time

– – – one output neuron requires 4096 MACs

Matrix Multiplications



VPU

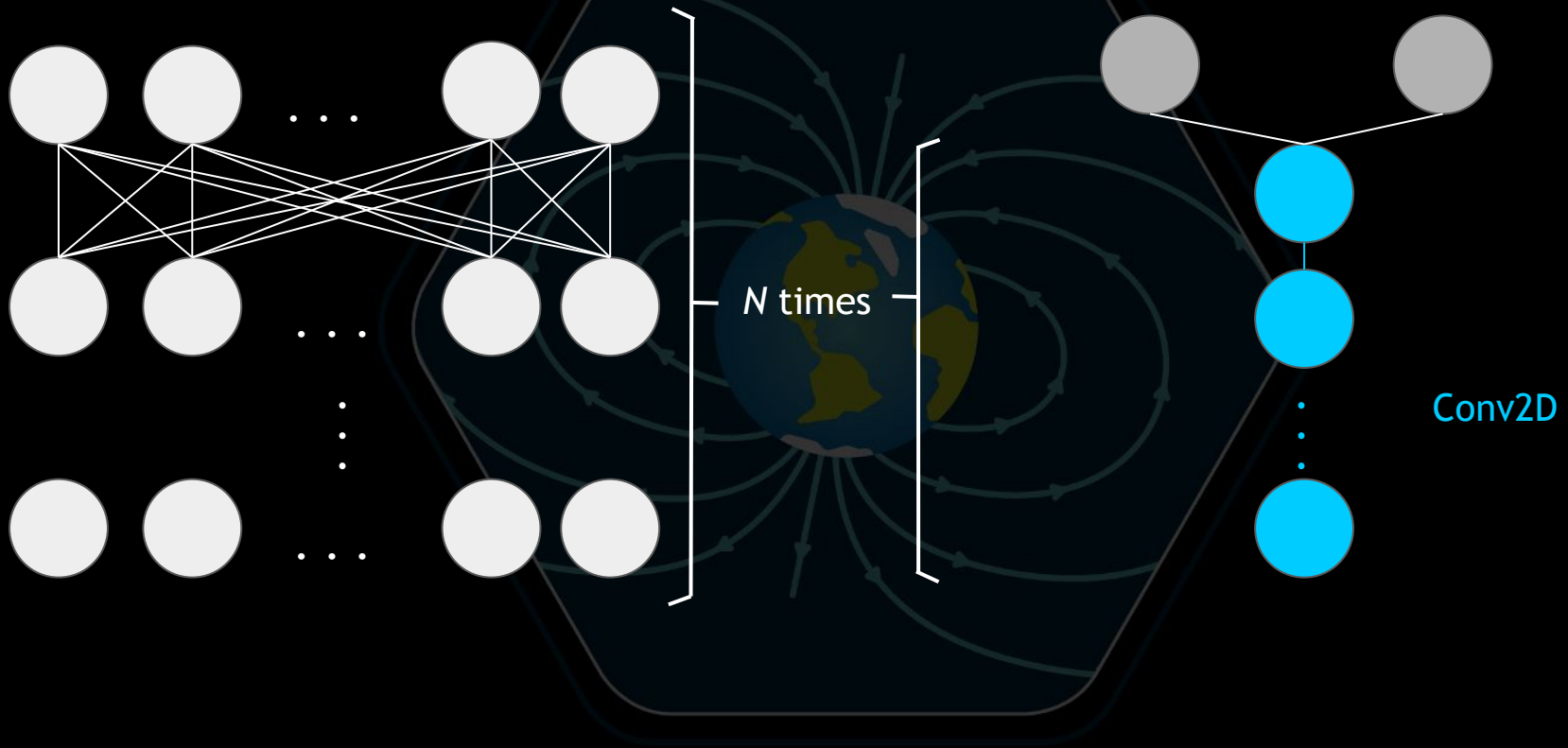


– – – one output neuron requires 4096 MACs

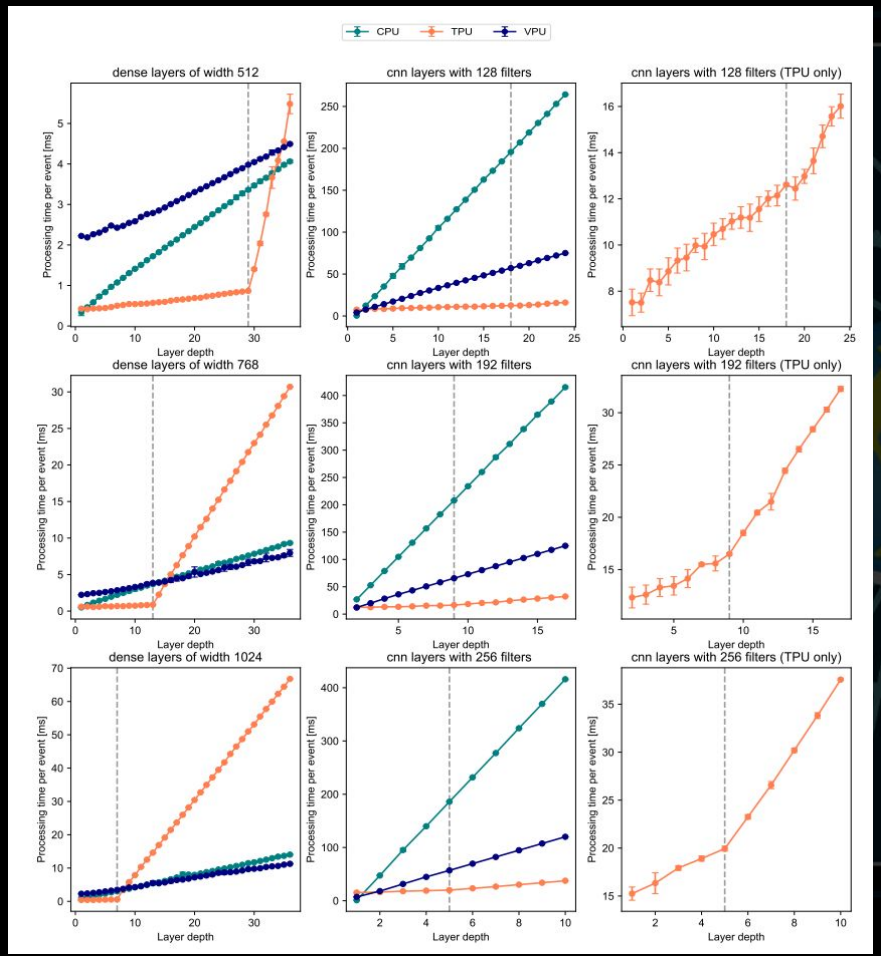
The effect of NN depth

Fully connected layers with same width connected in series

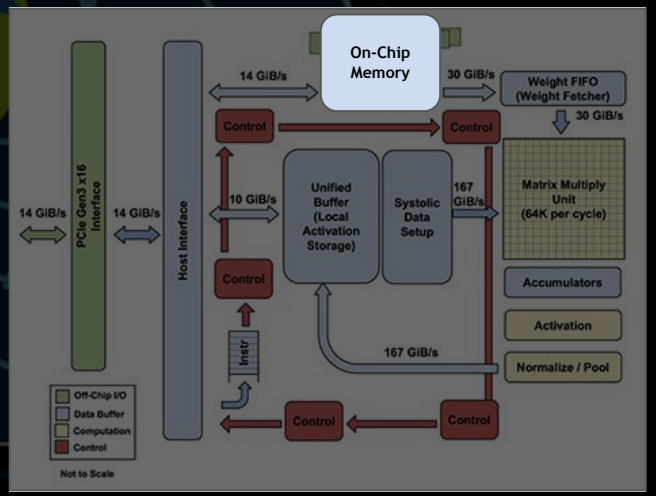
Convolutional layers with same kernel size and number of filters connected in series



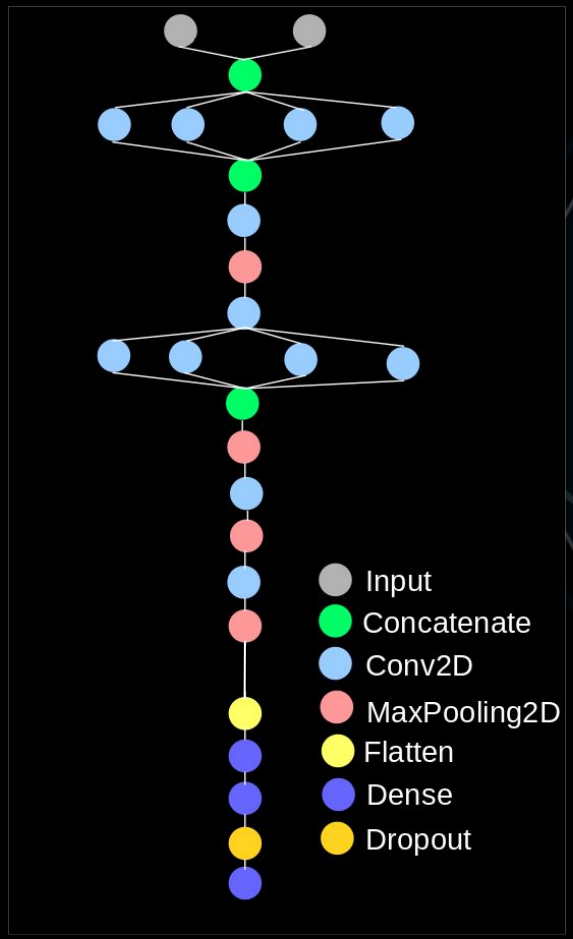
The effect of NN depth



- Processing time per event increases linearly with depth
- Grey line is where the model size exceeds 8 MB
- “Knee” for the case of TPU observed at grey line
  - Change in slope not as apparent for convolutions



1x1-Convolution as Dimensionality Reduction

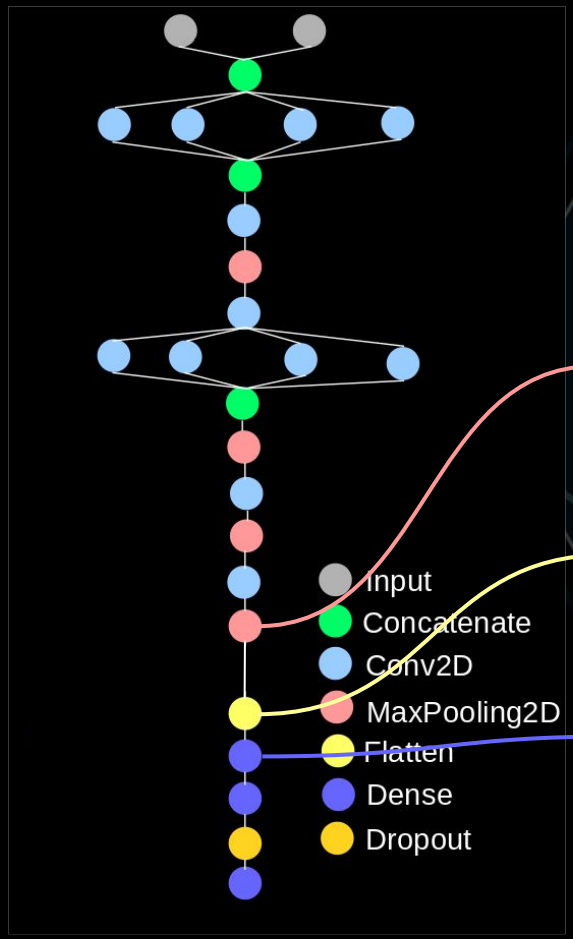


PID model → identification of ion type

Input : 2 images with the size of 16 x 32 pixels (YX & YZ)
Output : (1,26) vector → Z-th index = probability that the event is from an ion with atomic number Z

Might have similar architecture to the (later) model used for $\bar{p}/p$ event separation

1x1-Convolution as Dimensionality Reduction

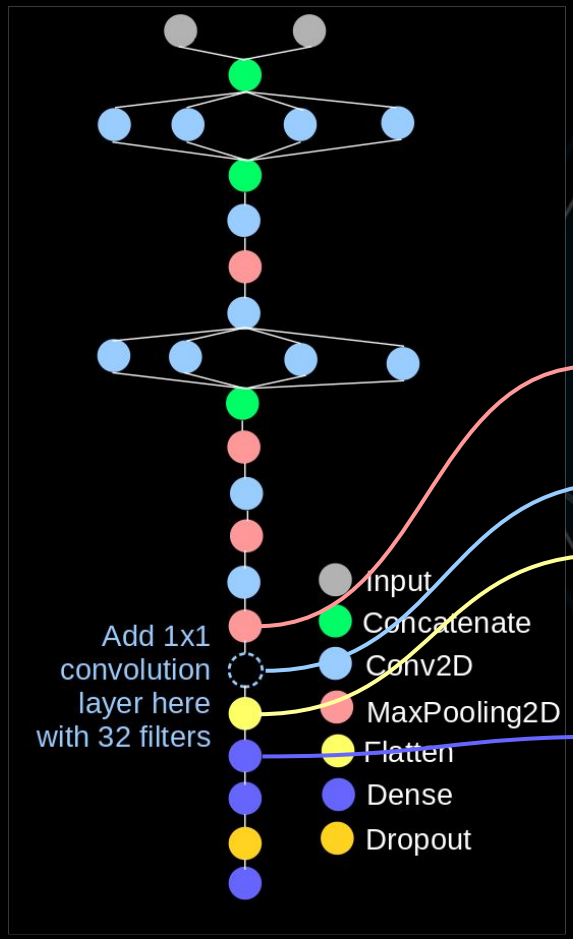


Output dimensions : (4,20,64)

Output dimensions : (4*20*64,) = (5120,)

Input size for the next dense (fully connected layer) = $5120 > 2^{12}$ → suboptimal

1x1-Convolution as Dimensionality Reduction



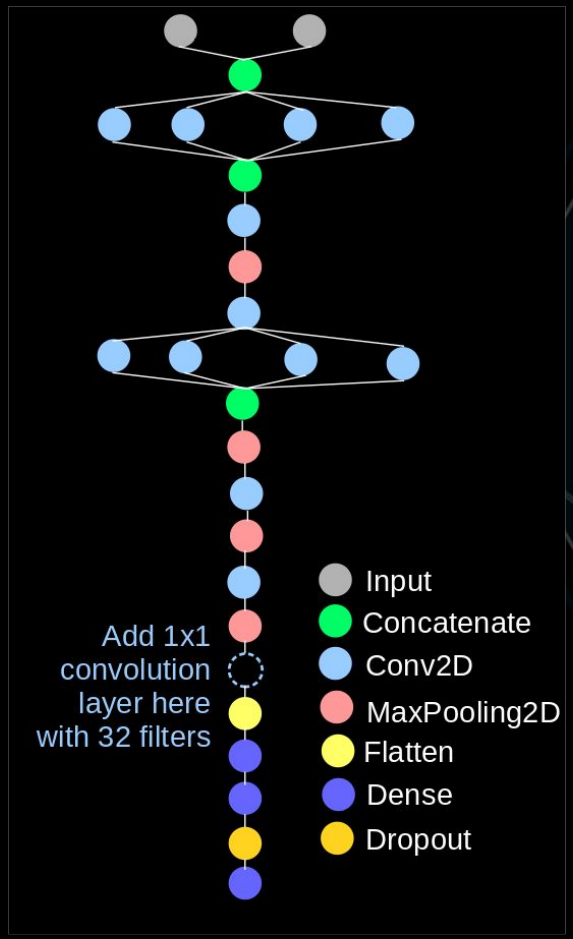
Output dimensions : (4,20,64)

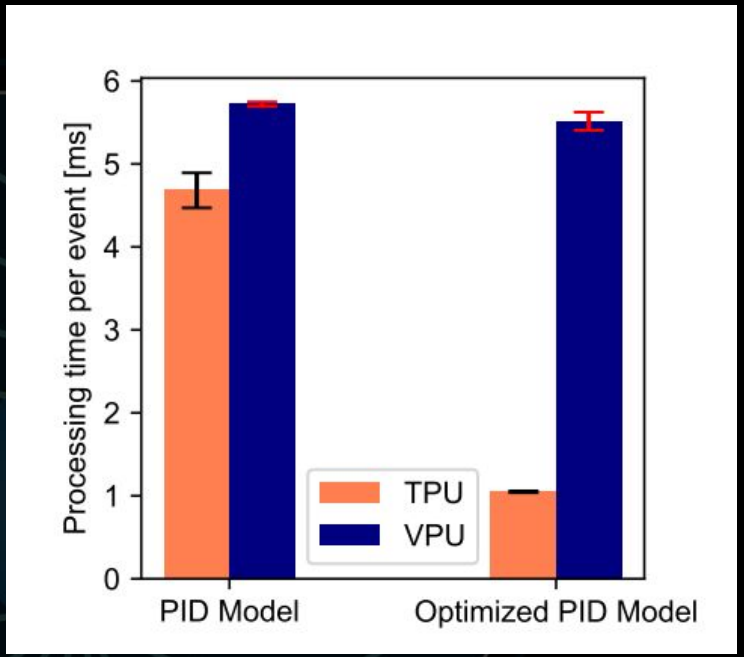
Output dimensions : (4,20,32)

Output dimensions : (4*20*32,) = (2560,)

Input size for the next dense (fully connected layer) = $2560 < 2^{12} = 4096$ → OK!

1x1-Convolution as Dimensionality Reduction



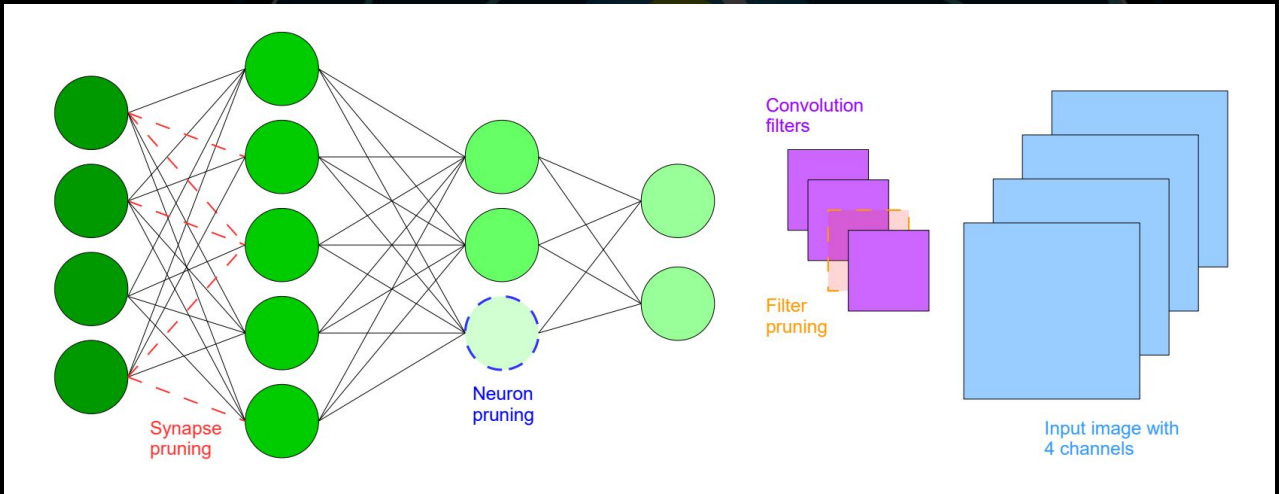


- A significant improvement is observed for the TPU
- Slight improvement for the VPU is observed → the effect of adding another layer for the VPU is larger than the effect of reducing the model width

Model Compression Techniques

Lower Quantization — Lower quantization than the one required from hardware?

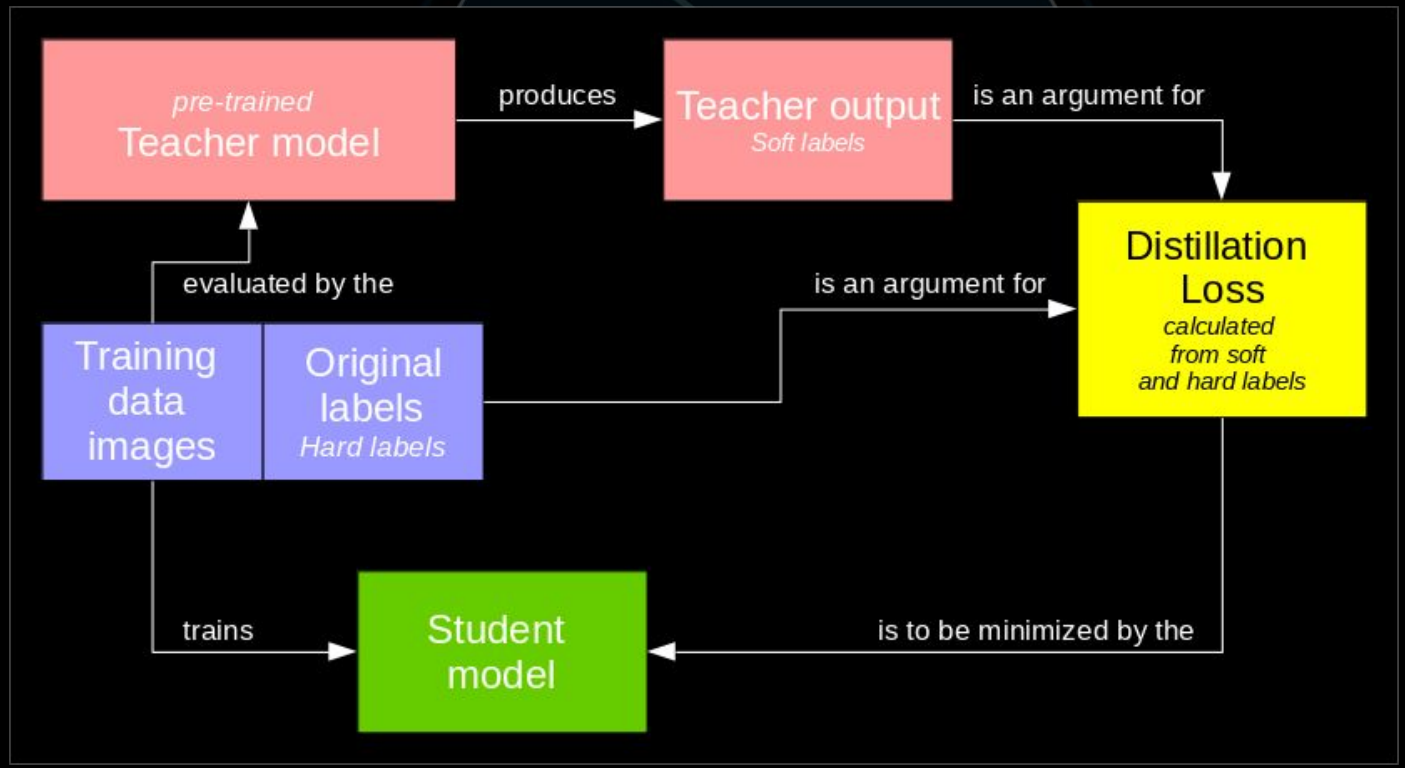
Structural Pruning — No improvement observed for all hardware accelerators



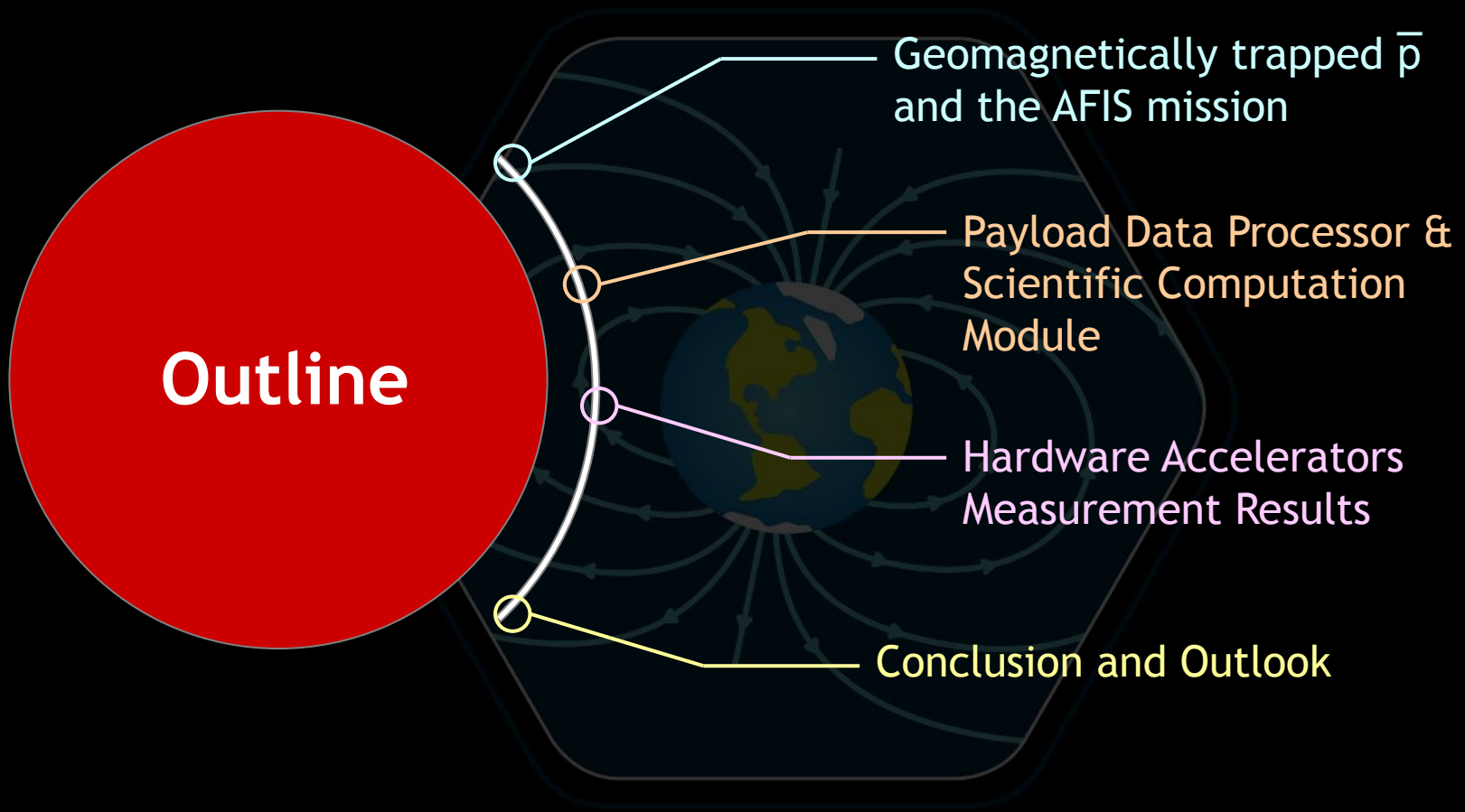
Model Compression Techniques

Knowledge Distillation

Promising to compress any model to a student model, tailored for each hardware



Outline



Conclusion & Outlook

- Available hardware accelerators
  - Power consumption 1-2 W
  - Might be useful if FPGA resources do not suffice for NNs
  - Interfacing still unknown
  - Need performance data for FPGA as reference

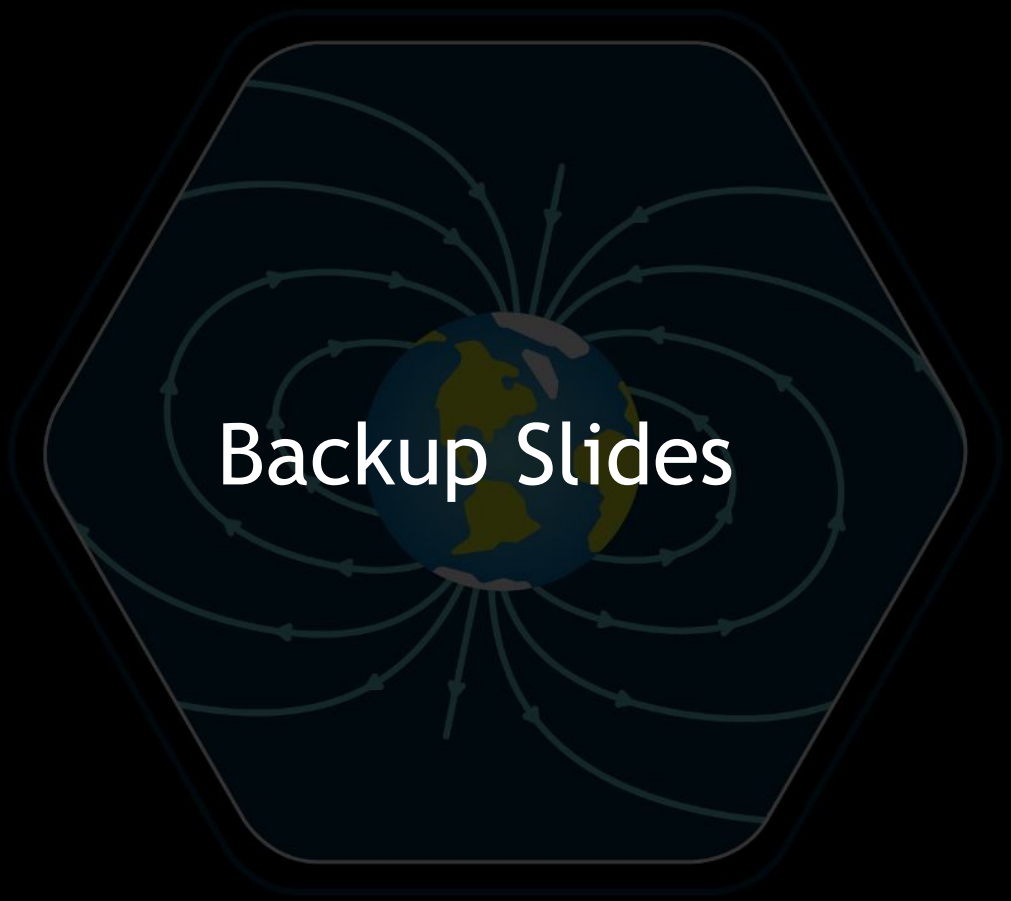
- NN architecture
  - TPU + VPU → layer width should be <= 4096 neurons
  - TPU → smaller and “narrow” models, VPU → larger models, but preferably less layers

- Compression methods
  - Structural pruning does not affect throughput for TPU and VPU
  - Knowledge distillation is promising for both TPU and VPU → to be investigated in the future


Thank you!

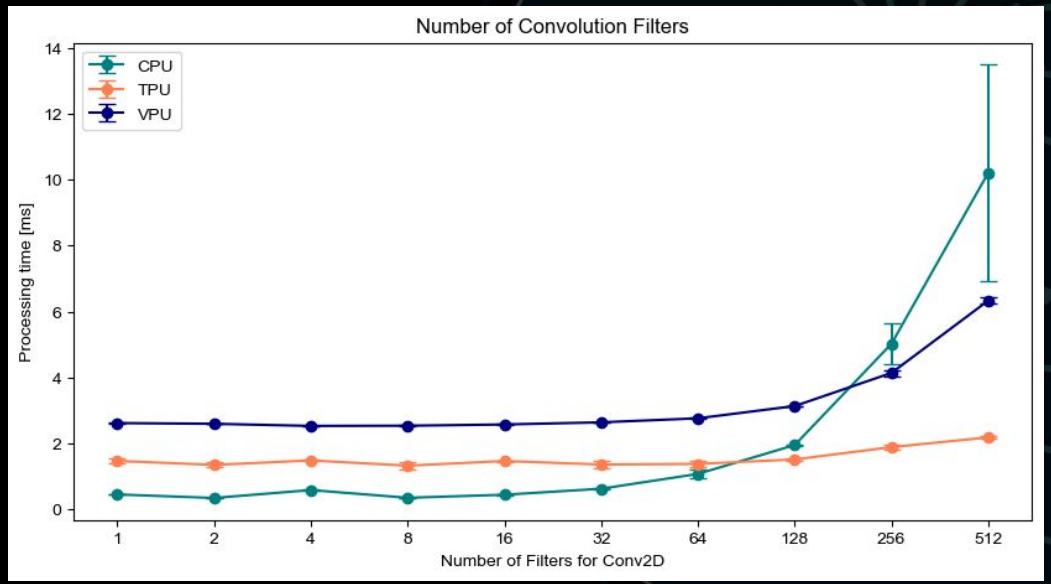
Citations

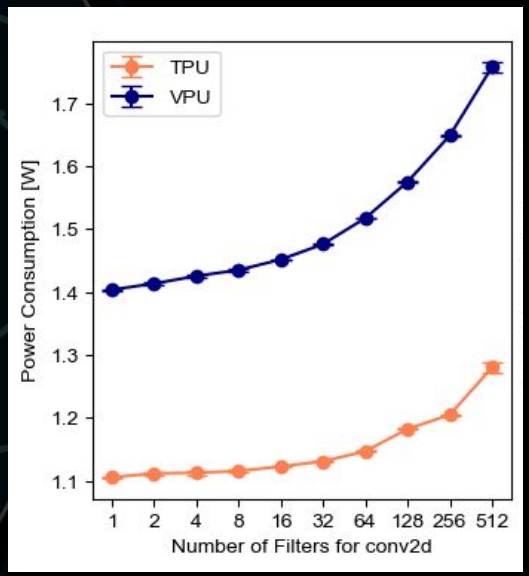
[1] https://pamela-web.web.roma2.infn.it/
[2] http://www.nasa.gov/content/goddard/van-allen-probes-reveal-zebra-stripes-in-space
[3] https://www.esa.int/ESA_Multimedia/Videos/2020/05/Development_of_the_South_Atlantic_Anomaly
[4] T. Pöschl. »Modeling and Prototyping of a Novel Active-Target Particle Detector for Balloon and Space Applications – Master's Thesis«. Technical University of Munich, Mar. 2015.
[5] Martin Losekamm et al. »The AFIS Detector: Measuring Antimatter Fluxes on Nanosatellites«. In: Proceedings of the International Astronautical Congress, IAC. Vol. 5. Sept. 2014. DOI: 10.13140/RG.2.1.4996.0405.
[6] M. Tanabashi et al. (Particle Data Group), (2018) Phys. Rev. D 98, 030001
[7] Jouppi et al. In-Datacenter Performance Analysis of a Tensor Processing Unit. 2017. arXiv: 1704.04760 [cs.AR].
[8] https://www.anandtech.com/show/11771/intel-announces-movidius-myriad-x-vpu
[9] Petrongonas, Evangelos & Leon, Vasileios & Lentaris, George & Soudris, Dimitrios. (2021). ParalOS: A Scheduling & Memory Management Framework for Heterogeneous VPUs. 221-228. 10.1109/DSD53832.2021.00043.
[10] Sergio Rivas-Gomez et al. »Exploring the Vision Processing Unit as Co-Processor for Inference«. In: 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). 2018, pp. 589–598. DOI: 10.1109/IPDPSW.2018.00098.
[11] https://codasip.com/wp-content/uploads/2021/03/Codasip_case-study_Mythic.pdf
[12] https://www.cnx-software.com/2020/10/07/learn-more-about-hailo-8-ai-accelerator-and-understanding-ai-benchmarks/



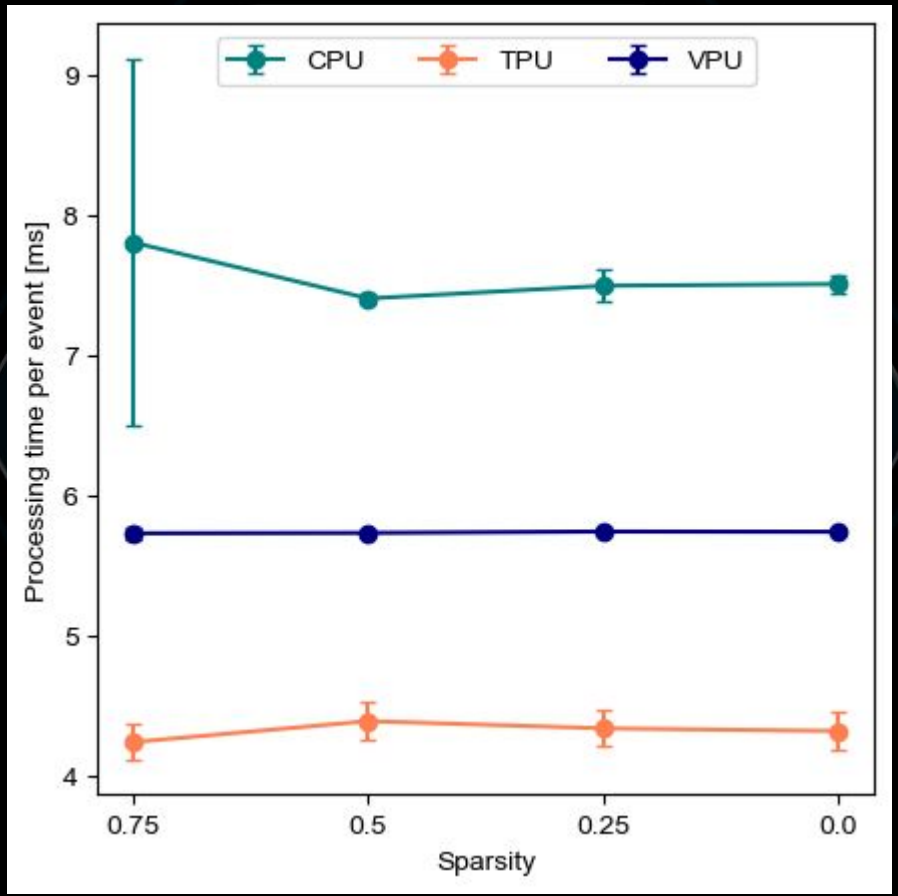
Backup Slides

Different number of Channels

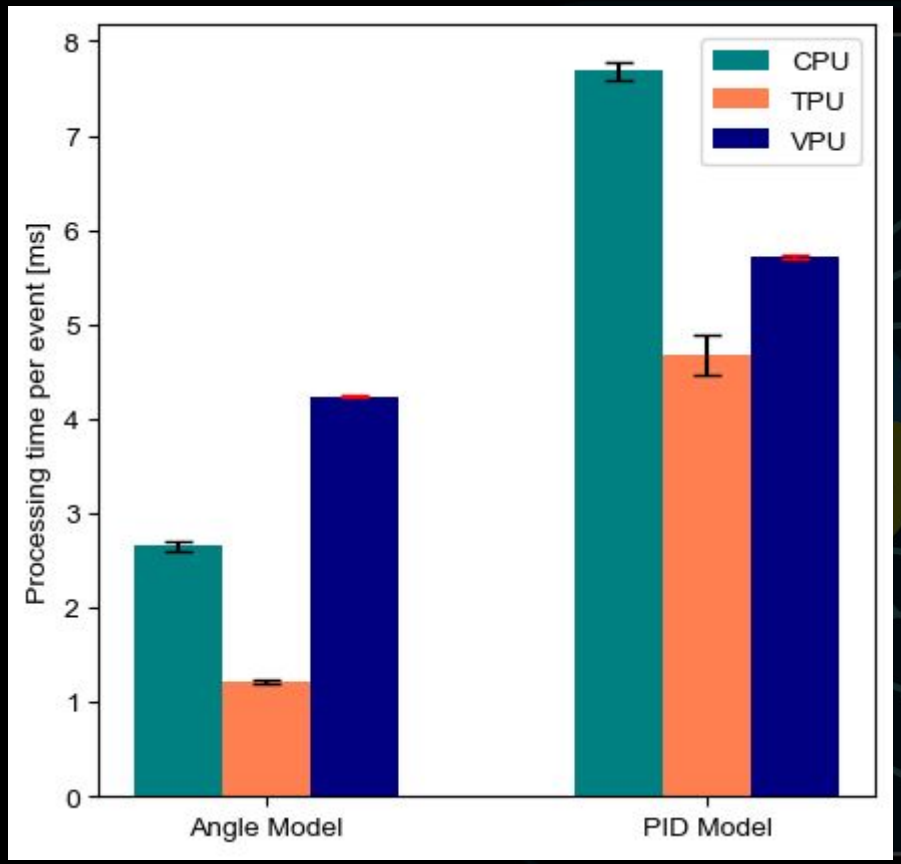


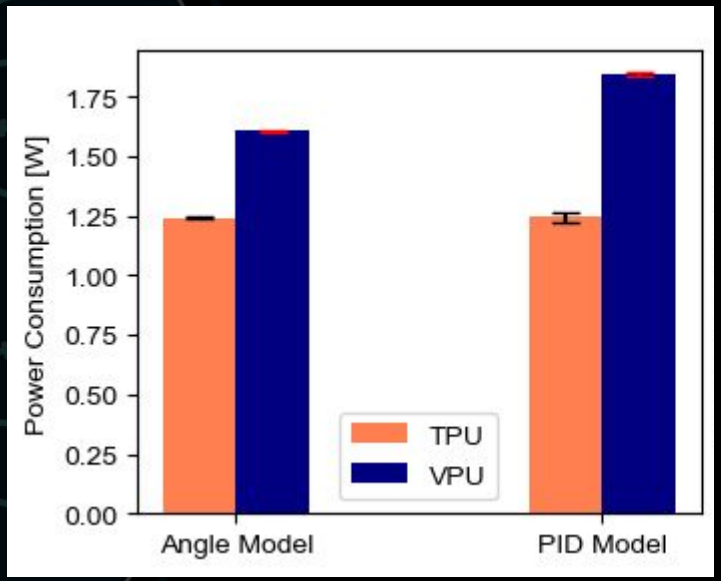


Sparsity

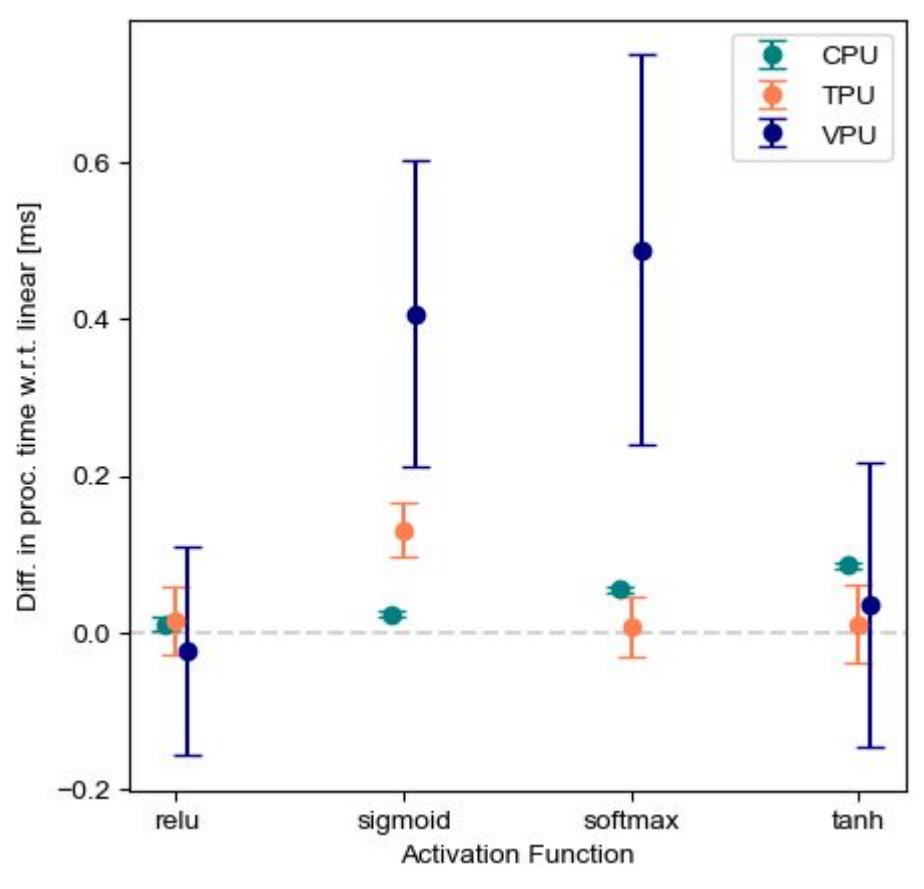

CPU
TPU
VPU
Processing time per event [ms]
9
8
7
6
5
4
0.75
0.5
0.25
0.0
Sparsity

Angle & PID Model

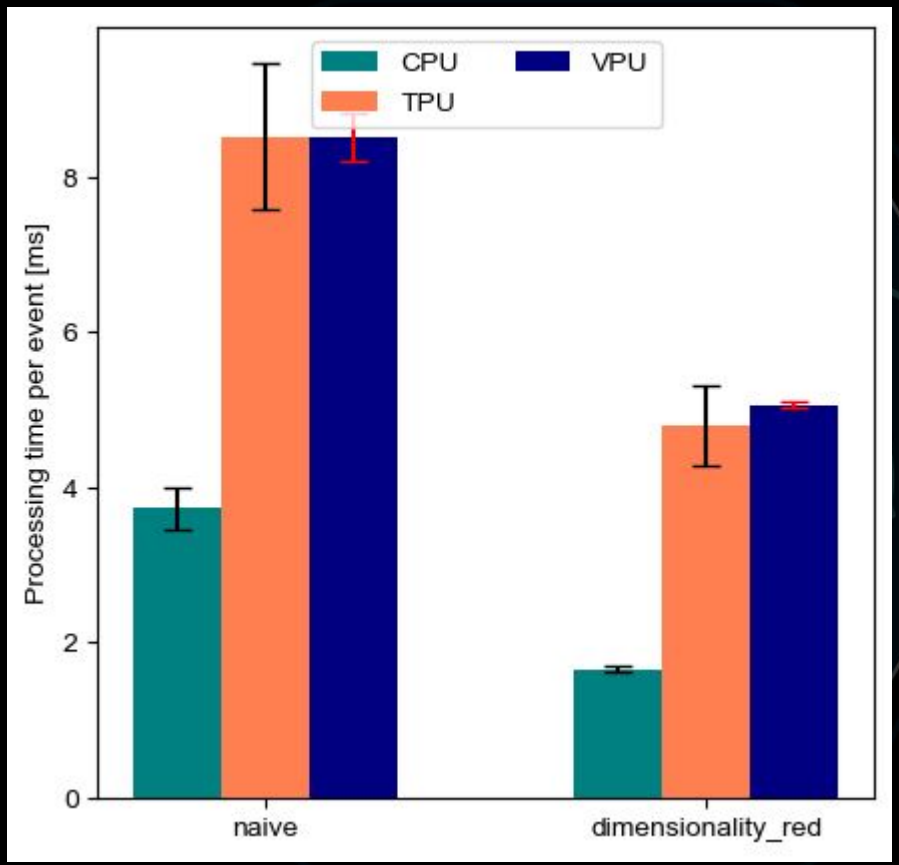




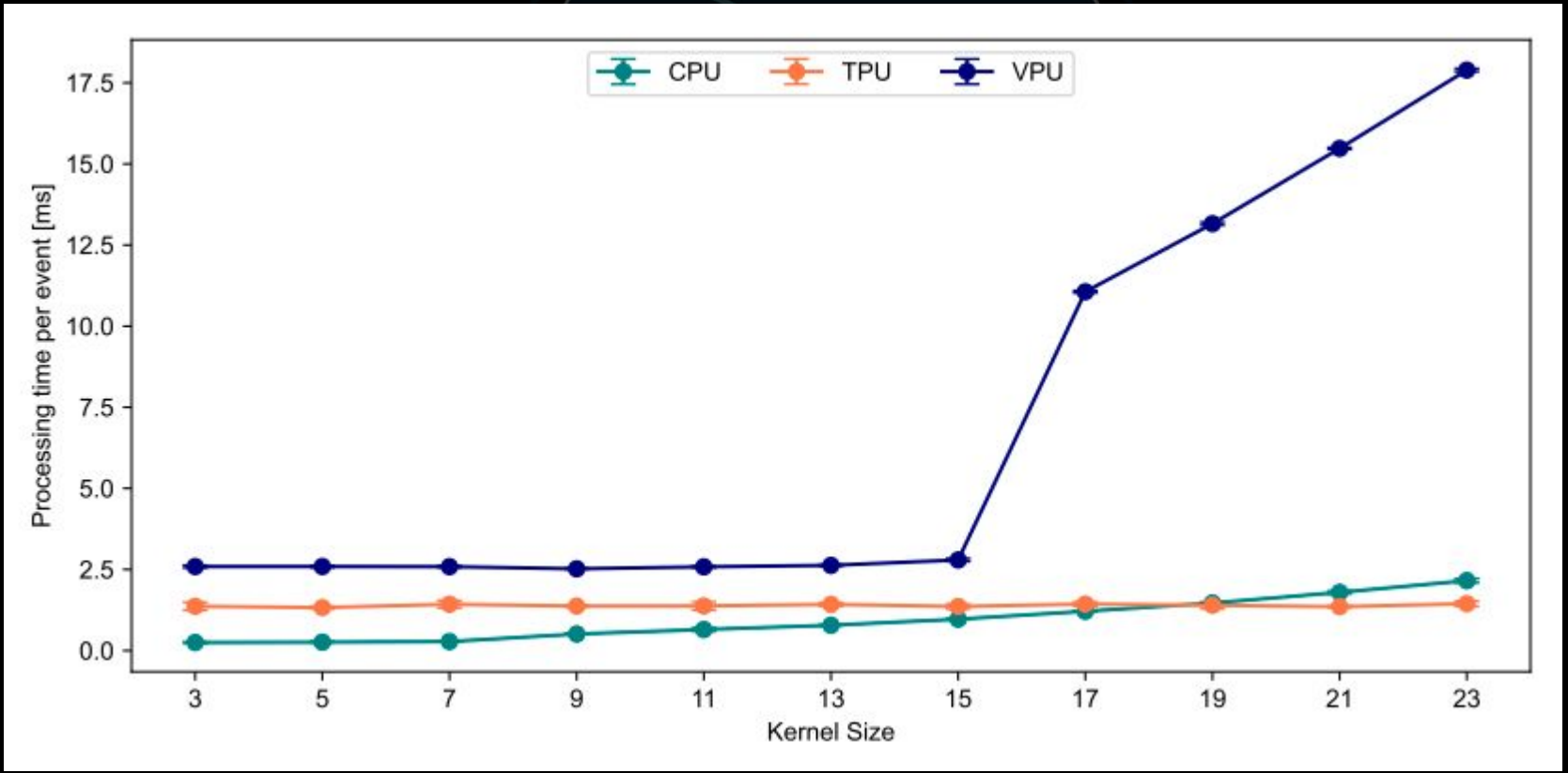
Activations



Inception Module



Kernel Sizes > 16 are suboptimal for VPU



Kernel Sizes > 16 are suboptimal for VPU