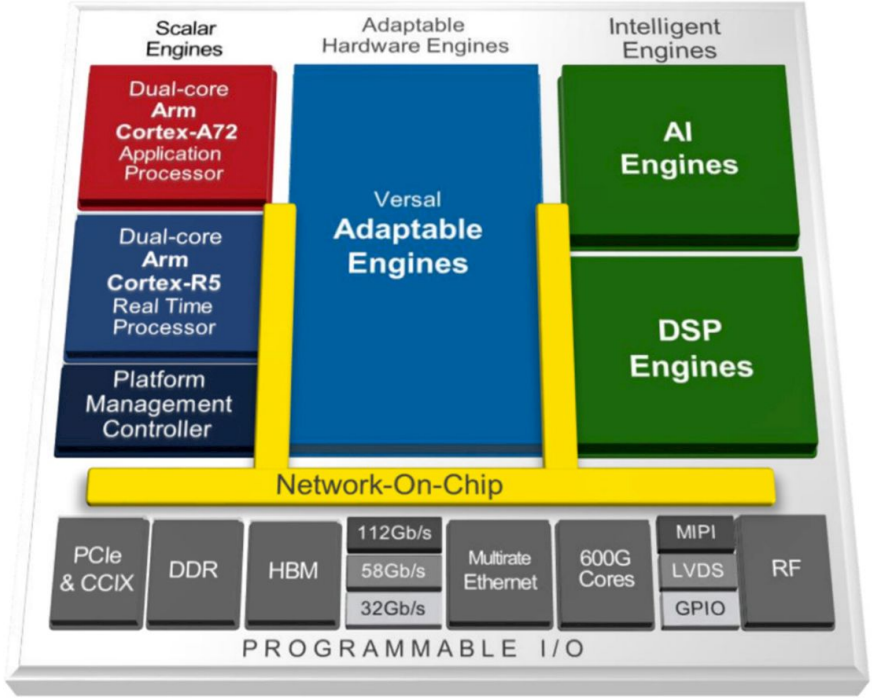
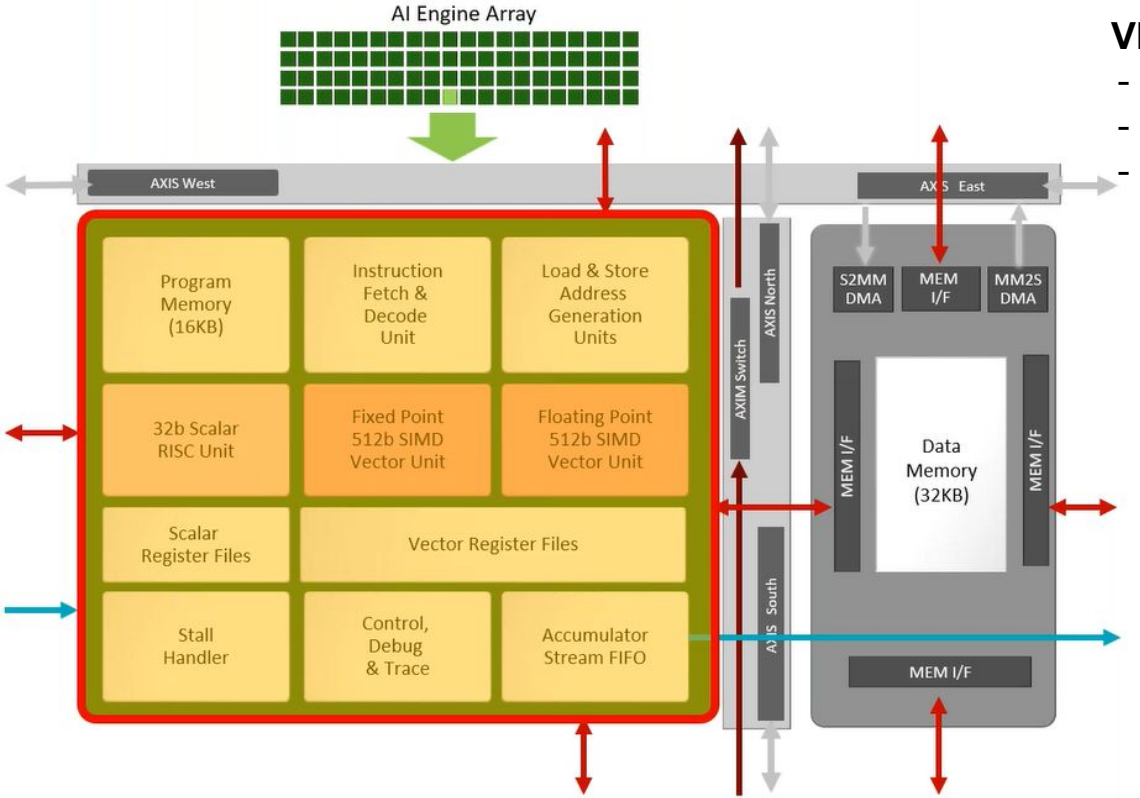
Xilinx Versal AI Core series



AI Engine Tile

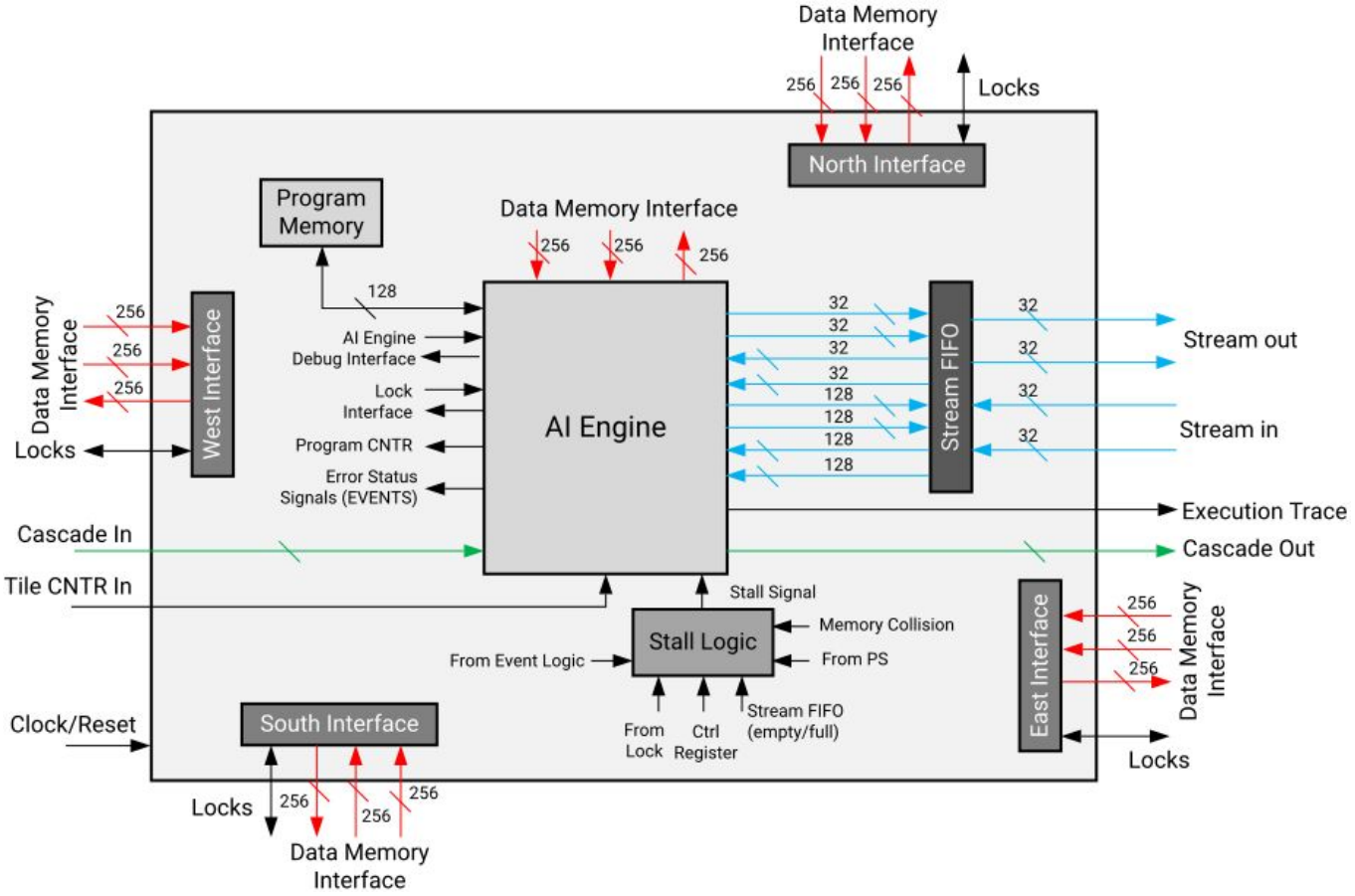


VLIW processing units

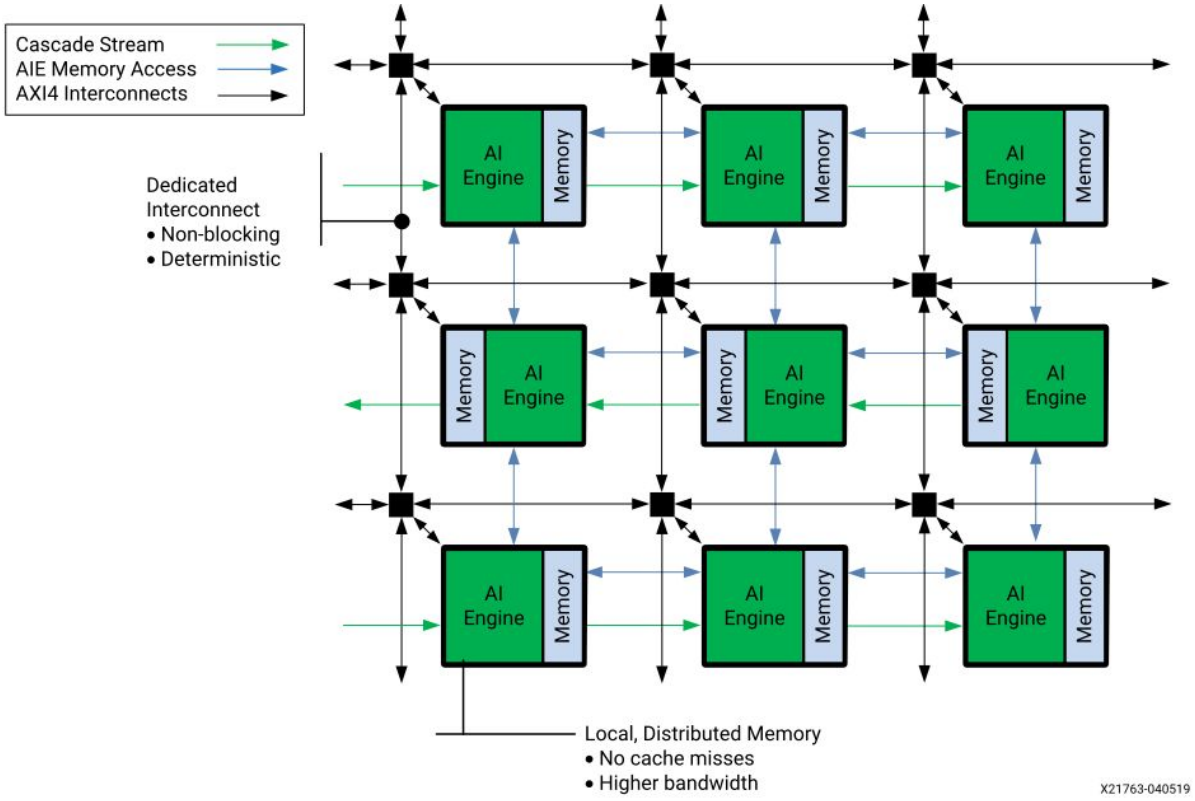
- 400 engines arranged in 2d-array
- running at >1GHz
- 512b vector unit:
  - floating point: 8 multiply-accumulates per cycle
  - fixed point:

X Operand	Z Operand	Output	Number of GMACs @ 1 GHz
8 real	8 real	48 real	128
16 real	8 real	48 real	64
16 real	16 real	48 real	32
16 real	16 complex	48 complex	16
16 complex	16 real	48 complex	16
16 complex	16 complex	48 complex	8
16 real	32 real	48/80 real	16
16 real	32 complex	48/80 complex	8
16 complex	32 real	48/80 complex	8
16 complex	32 complex	48/80 complex	4
32 real	16 real	48/80 real	16
32 real	16 complex	48/80 complex	8
32 complex	16 real	48/80 complex	8
32 complex	16 complex	48/80 complex	4
32 real	32 real	80 real	8
32 real	32 complex	80 complex	4
32 complex	32 real	80 complex	4
32 complex	32 complex	80 complex	2
32 SPFP	32 SPFP	32 SPFP	8

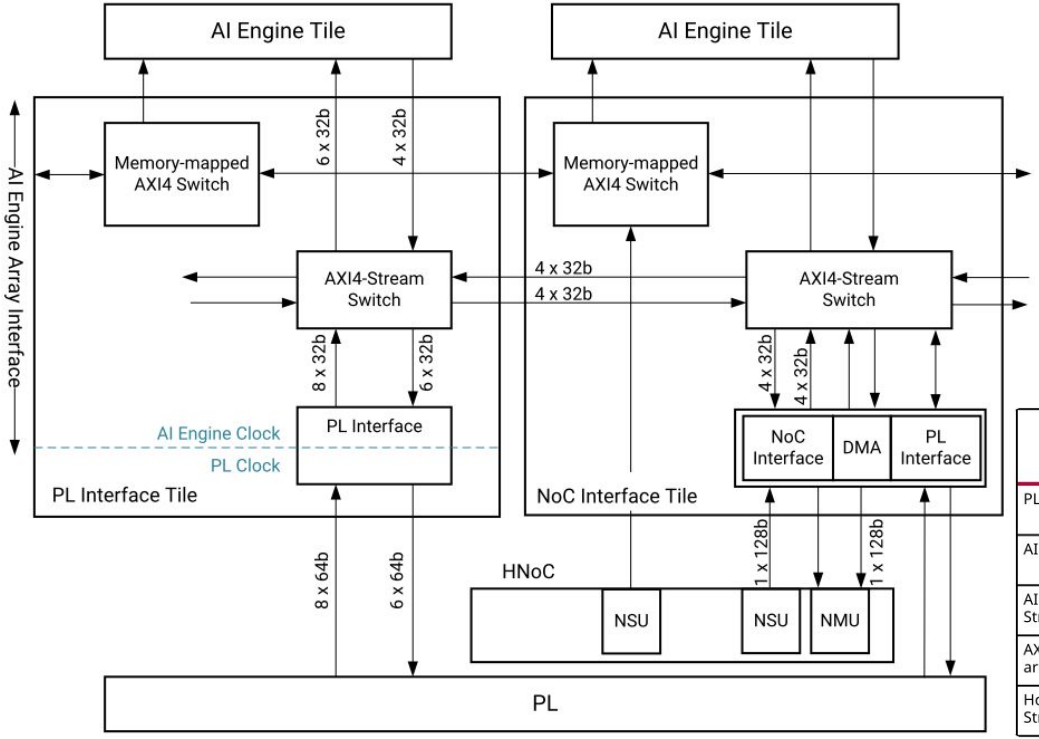
AI Engine Tile: Interfaces



AI Engine Array



AI Engine Array - PL interface



Aggregate bandwidth

In VC1902 39 PL Interfaces

- PL-to-AIE: 1.3 TB/s
- AIE-to-PL: 1.0 TB/s

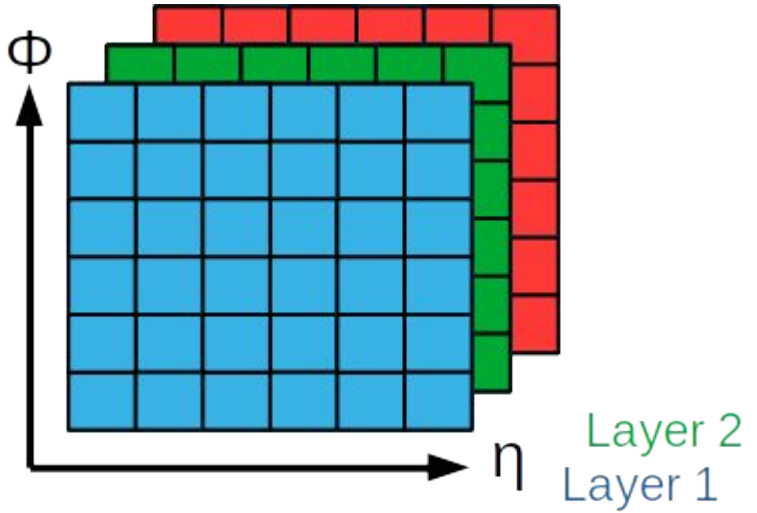
Connection Type	Number of Connections	Data Width (bits)	Clock Domain	Bandwidth per Connection (GB/s)	Aggregate Bandwidth (GB/s)
PL to AI Engine array interface	8	64	PL (500 MHz)	4	32
AI Engine array interface to PL	6	64	PL (500 MHz)	4	24
AI Engine array interface to AXI4-Stream switch	8	32	AI Engine (1 GHz)	4	32
AXI4-Stream switch to AI Engine array interface	6	32	AI Engine (1 GHz)	4	24
Horizontal interface between AXI4-Stream switches[1]	4	32	AI Engine (1 GHz)	4	16

Practical Application

- Neural Network for ATLAS Trigger application on FPGA (fFEX)
- Utilize AI Engines in the Design
- Basic Idea:
  - Frame particle identification as object detection
  - Treat calorimeter as an image

ATLAS Calorimeter Structure

- Calorimeter has layered cell structure
- Energy deposits are associated to a value in η and Φ
- Rough correspondence between calorimeter cells and pixels of an image
- All physics analysis is based on this information combined with tracking

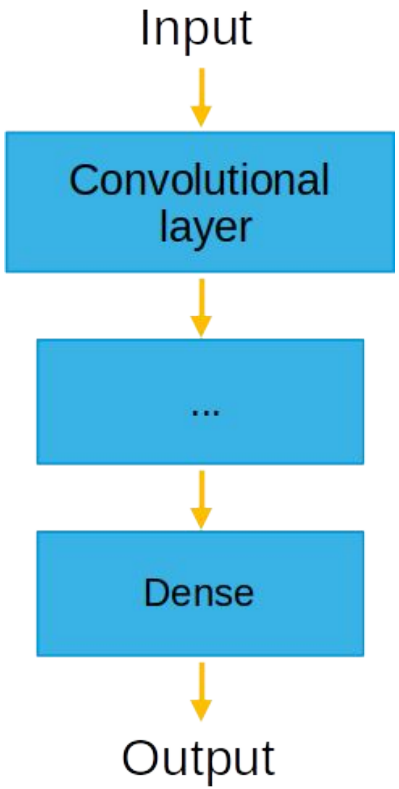


CALONet

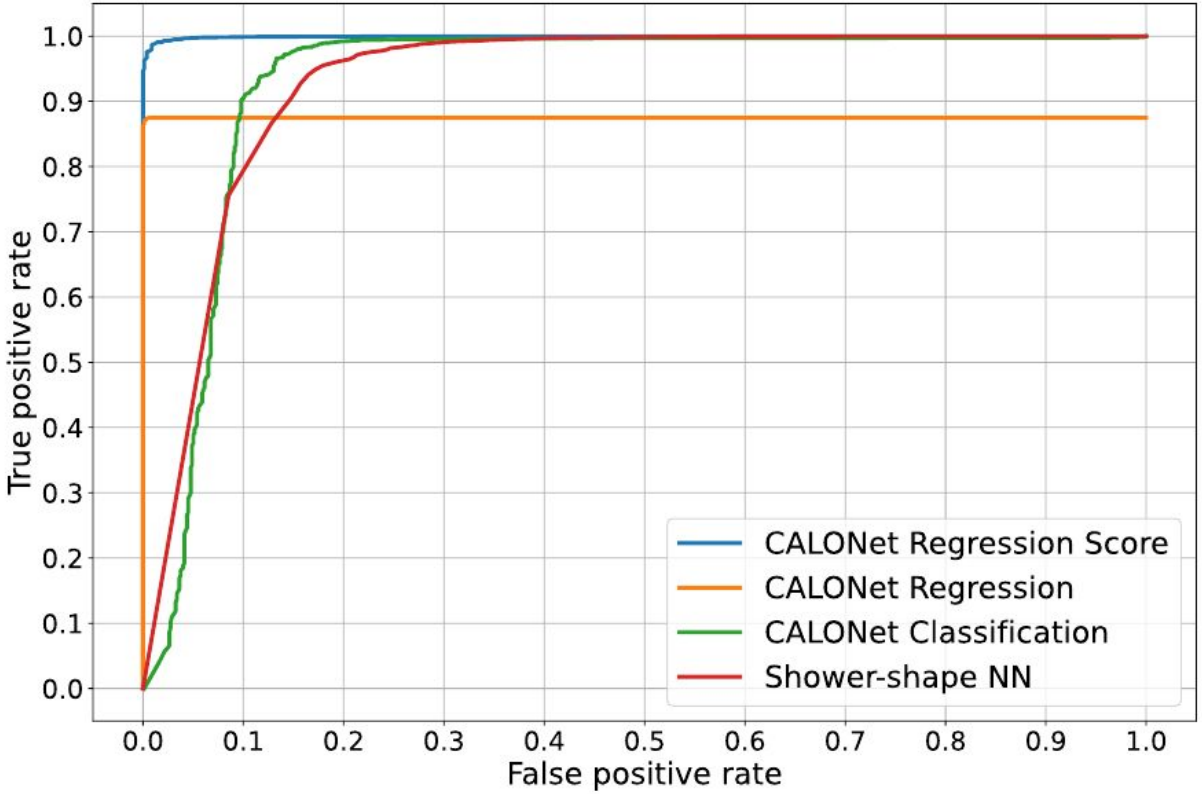
- Based on YOLO-architecture:
  - Divide image into grid and locate objects inside grid cells
  - Very fast algorithm
- Small region in η and Φ
- Proof of principle:
  - Predict electrons and their location in the calorimeter
  - Simple architecture

Architecture

- Regression
  - Conv2D → **Dense** → Dense
  - 60.000 parameters
- Classification
  - Conv2D → **MaxPool** → Dense
  - 400.000 parameters



Offline Results



DPU Implementation

- Xilinx IP core: Deep-learning Processing Unit
- Xilinx default method for neural network implementation
- Optimizes accuracy and latency in multiple steps
- Final outcome: 33 µs (for 60.000 parameters)
- Also: mini model 30 µs (3 parameters)
- DPU has large bottleneck
- Not intended for smaller networks at ultralow latency
- Optimized for general purpose implementation of larger networks

Summary & Conclusion

- AI Engines are highly capable computation units for neural networks
- Can utilize them using:
  - Mapping trained NN to AIEs via Vitis AI and the DPU IP core
  - or hardcoding in Vitis via C++

- The DPU shows comparably bad latencies for small networks
  - Can be used for larger networks with less stringent timing constraints
- Hardcoding can provide much lower latency