Beyond the Edge: Revolutionizing AI Workloads with Axelera AI's Digital In-Memory Computing and RISC-V technology

About us

- Co-founded in July 2021, by Fabrizio Del Maffeo and Evangelos Eleftheriou, with 16 founding team members from IBM, ETH Zurich, IMEC, Bitfury AI, Google and Qualcomm.
- Our team has grown to 180+ people, including 60+ PhDs and are present in 16 countries.
- Our first product, Metis, is the most powerful AI processing unit for computer vision, with the best performance/price and efficiency on the market, and in 2025 we will expand our product line to generative AI applications.
- We have been delivering to customers since September 2023. We now have 25+ customers and are moving into mass production in Q4 2024.
- We have raised USD ~120M from leading deep-tech investors, institutions and European sovereign funds.

Our investors include

Innovation Industries · imec xpand · VERVE VENTURES · SAMSUNG CATALYST FUND · European Innovation Council · BITFURY · cdp · Fondo Nazionale Innovazione CDP Venture Capital Sgr · sfpi fpim

Our team come from

IBM · Lenovo · imec · Movidius · ASUS · intel · Google · Ambarella · ETH zürich · McKinsey & Company · arm · DELL · blaize · GRAPHCORE · BCG

Opportunity

Retail

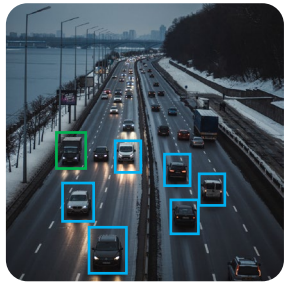
Customer flow analysis
Inventory management
Cashier-less checkouts

Agriculture

Crop health monitoring
Automated pest control
Agricultural robotics

Industrial

Quality control automation
Worker safety monitoring
Automated material handling



Security

Traffic control systems
Intelligent surveillance
Access control systems

Healthcare

Remote patient monitoring
Real-time diagnostics tools
Surgical tools and equipment

Automotive

Driver assistance systems
Autonomous driving systems
Pedestrian safety systems

Computer vision at the edge is generating real value across a range of industries **today**

Opportunity

Enterprise server

General purpose systems used to run business applications and services

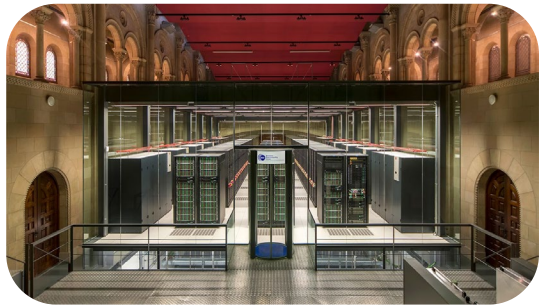
E.g. Real-time analysis of medical imaging data, like X-rays, MRIs and CT scans

Datacenter

Scalable and flexible resources designed to host and manage vast amounts of data and apps

E.g. inference on image and text generation models (e.g. ChatGPT)

HPC



Designed for complex, computation-intensive tasks, optimized for parallel processing

E.g. simulations, scientific research, weather modeling, large-scale data analysis

But our technology can scale up to address the trillion-dollar industries of **tomorrow** and work across different **computing environments**

Solution

METIS

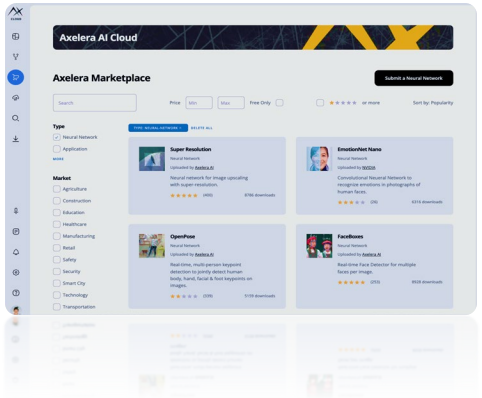
First Gen
(2024)

Modules

Cards

Boards

Systems



AI Processing Unit (AIPU)

The most powerful and efficient AI accelerator with our **Digital in-Memory Computing** technology and RISC-V ISA

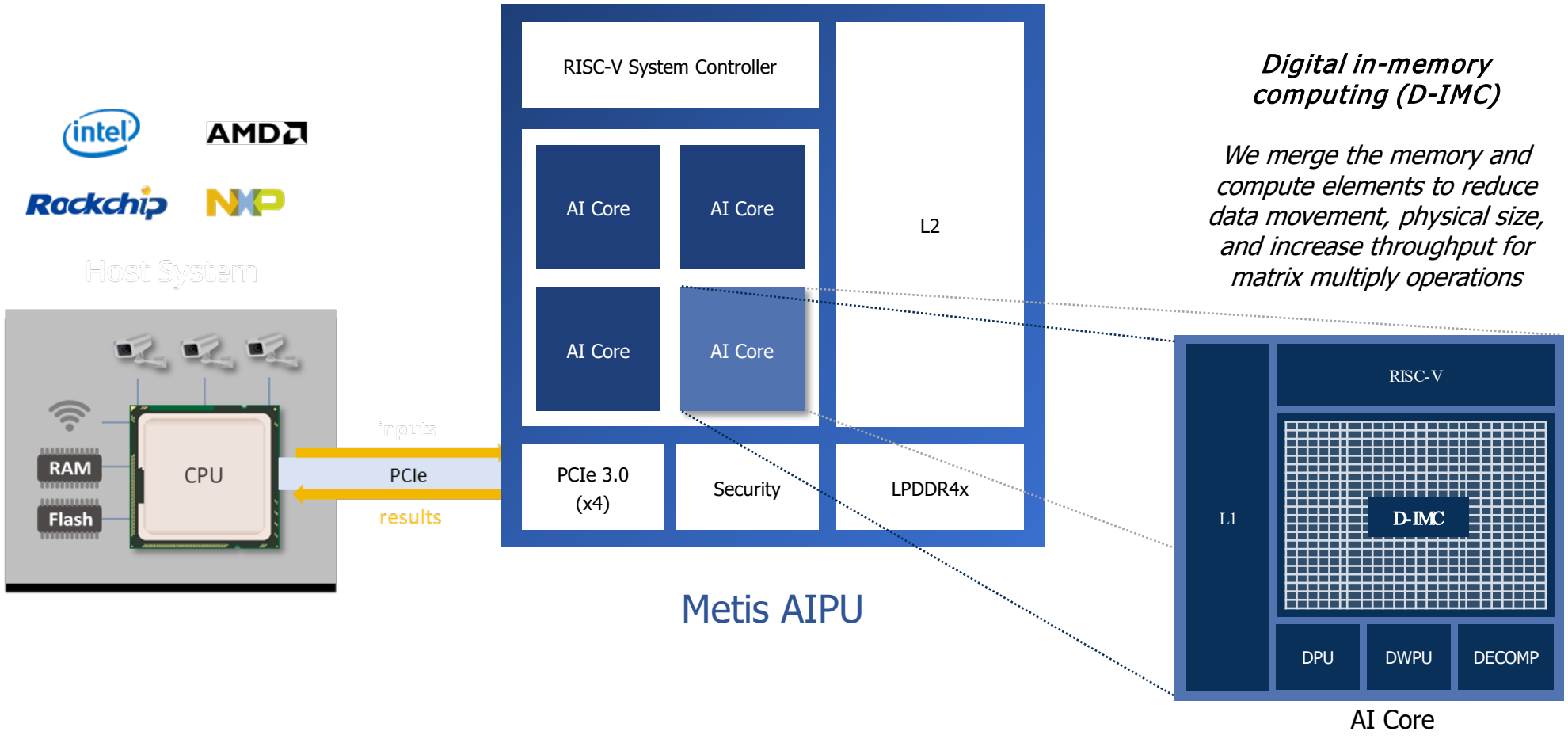
AI Accelerators Card & Systems

Edge-native hardware powered by an Axelera AIPU to enable instant field installation and faster time-to-market

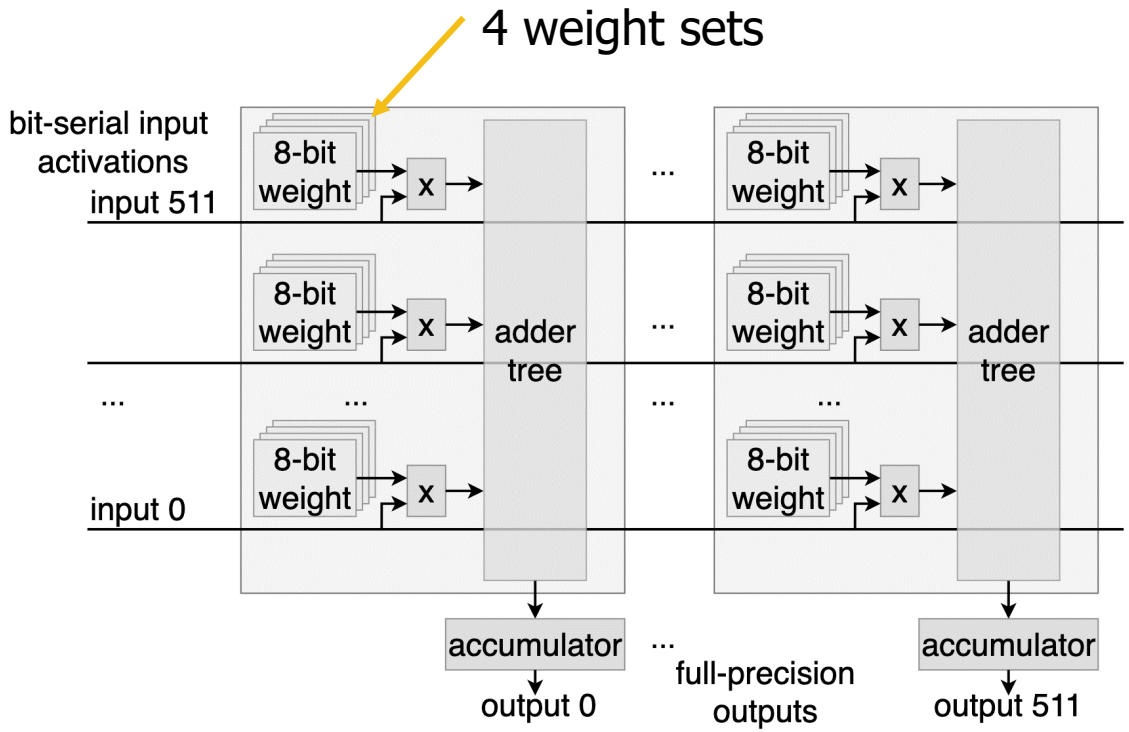
Voyager SDK

Integrated **AI software stack** designed to simplify application development, optimization and deployment

Metis AIPU


intel
AMD
Rockchip
NXP
Host System
RAM
Flash
CPU
inputs
PCIe
results
RISC-V System Controller
AI Core
AI Core
AI Core
AI Core
L2
PCIe 3.0
(x4)
Security
LPDDR4x
Metis AIPU
Digital in-memory computing (D-IMC)
We merge the memory and compute elements to reduce data movement, physical size, and increase throughput for matrix multiply operations
L1
RISC-V
D-IMC
DPU
DWPU
DECOMP
AI Core

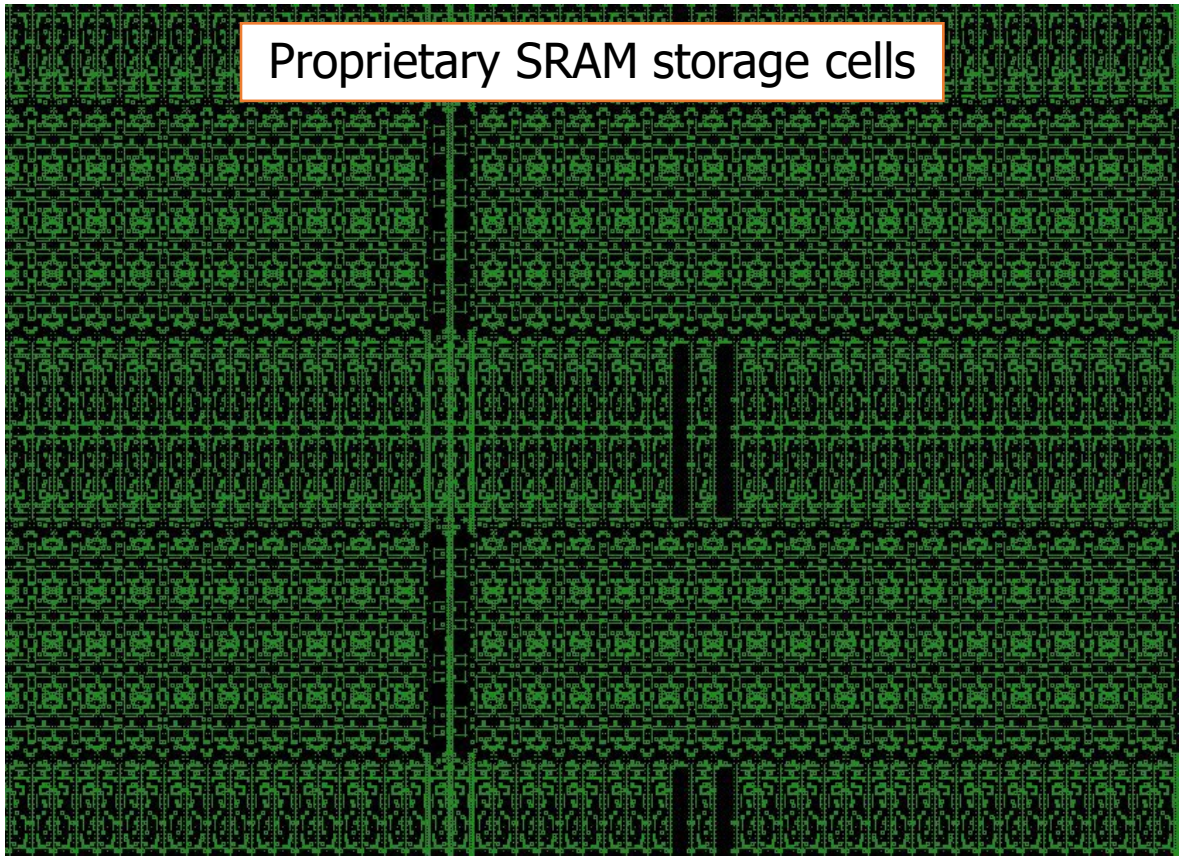
SRAM-based D-IMC



Commentary

- Interleaved weight-storage and compute units in an extremely dense fashion
- INT8 activations / weights, with INT32 accumulation to maintain full precision
- Immune to noise and memory non-idealities affecting analog IMC precision
- Technology commensurate with CMOS scaling to low lithography nodes

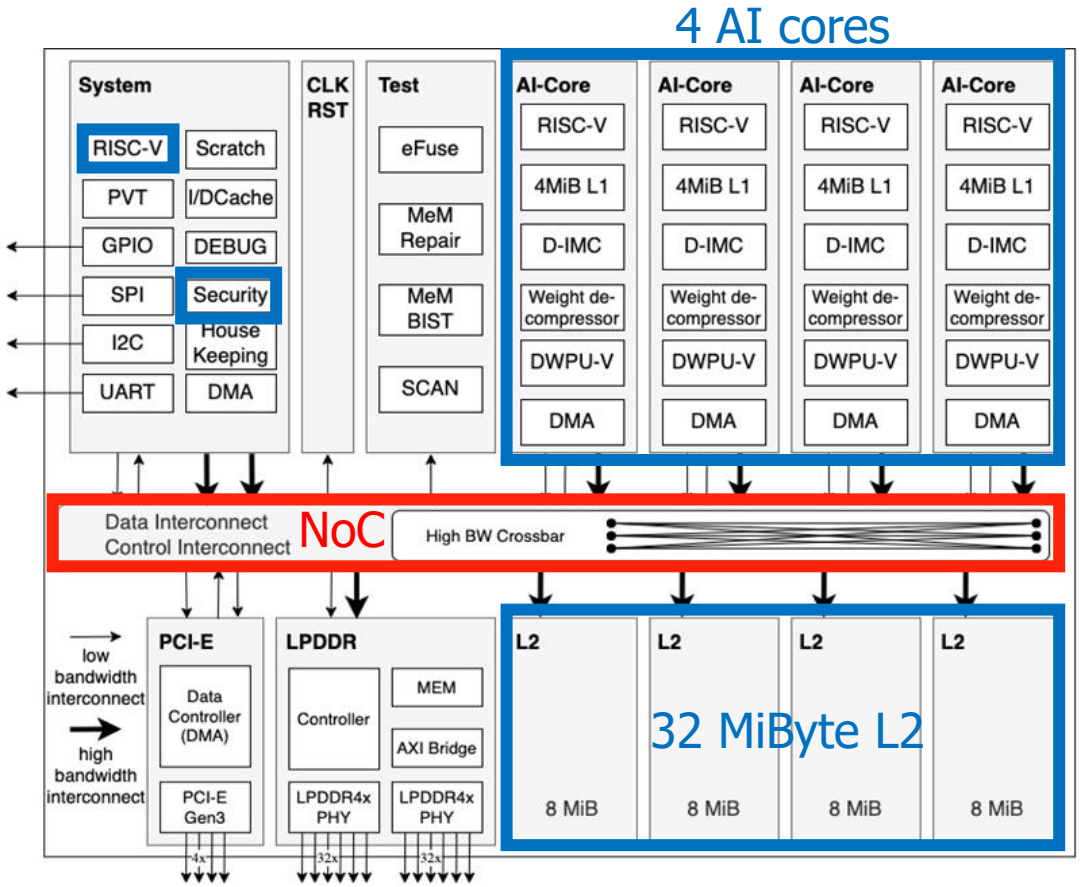
SRAM-based D-IMC



Commentary

- Stores multiple weight sets in computational memory
  - Enhances IMC storage density
  - Allows accumulation up to 16k inputs
  - Enables simultaneous processing and weight reloading
- Activity gating and clock gating. Maintains high energy efficiency at low utilization
- Ensures full-precision accumulation
  - Negligible accuracy loss compared to FP32
  - Use of post-training quantization; no need for retraining

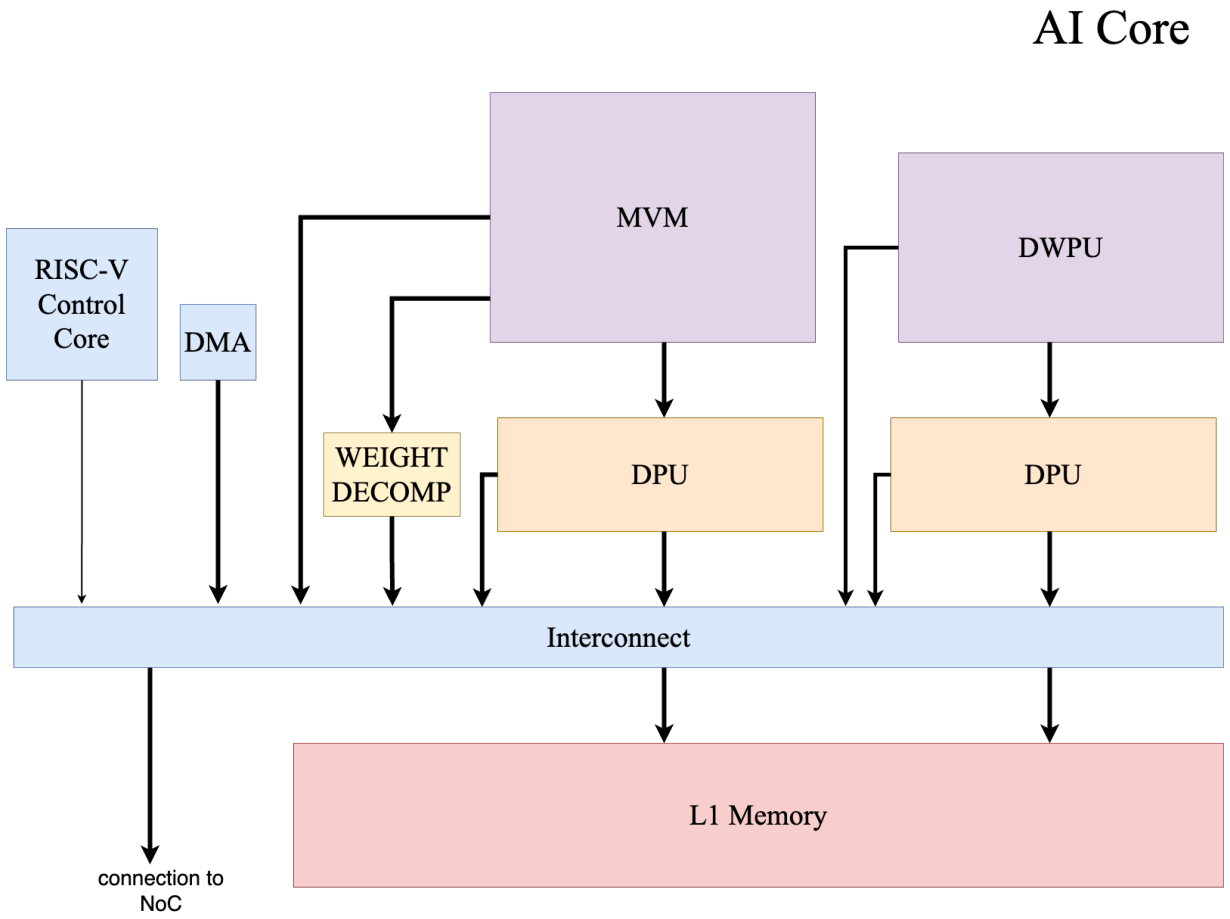
AIPU SoC Architecture



Commentary

- AI-Core
  - Self-sufficient compute engine for concurrent network execution
- RISC-V system controller
  - Boots chip, interfaces with peripherals, manages AI cores with a real-time OS
- Security module
  - Secure boot and weight/data encryption
- 32 MiByte L2 SRAM
  - 52 MiByte on-chip memory in total
- Interconnected through Network-on-Chip (NoC)
  - 1 Tbit/s bandwidth to shared memories
  - Ensures AI cores will not stall

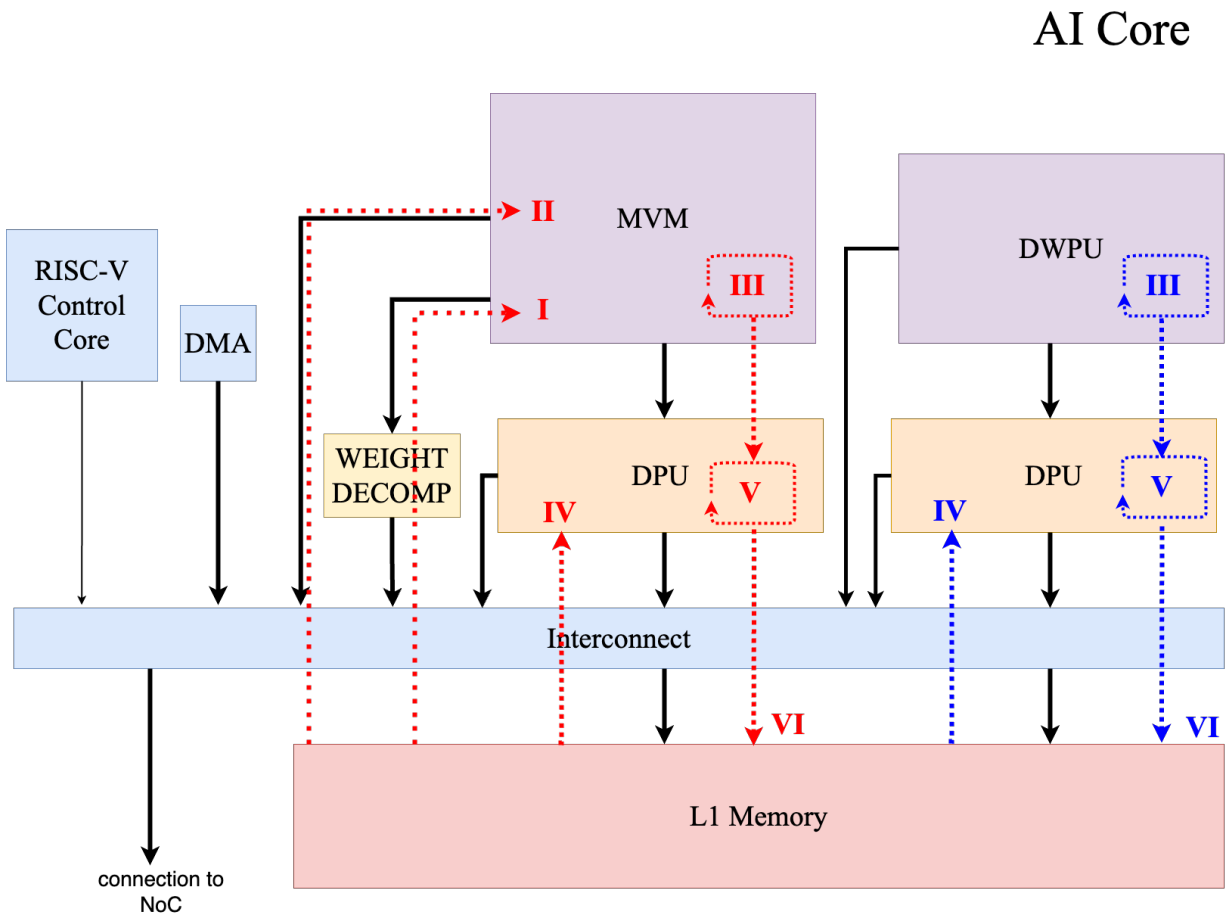
AI Core | Key Components



Commentary

- Matrix-Vector Multiplier: D-IMC based
- Data Processing Unit
  - Element-wise vector operations
  - Apply activation functions
- Depth-Wise Processing Unit
  - Depth-wise convolution
  - Pooling and Up-sampling
- Weight Decompression Unit
- 4 MiByte L1 SRAM
- RISC-V control core

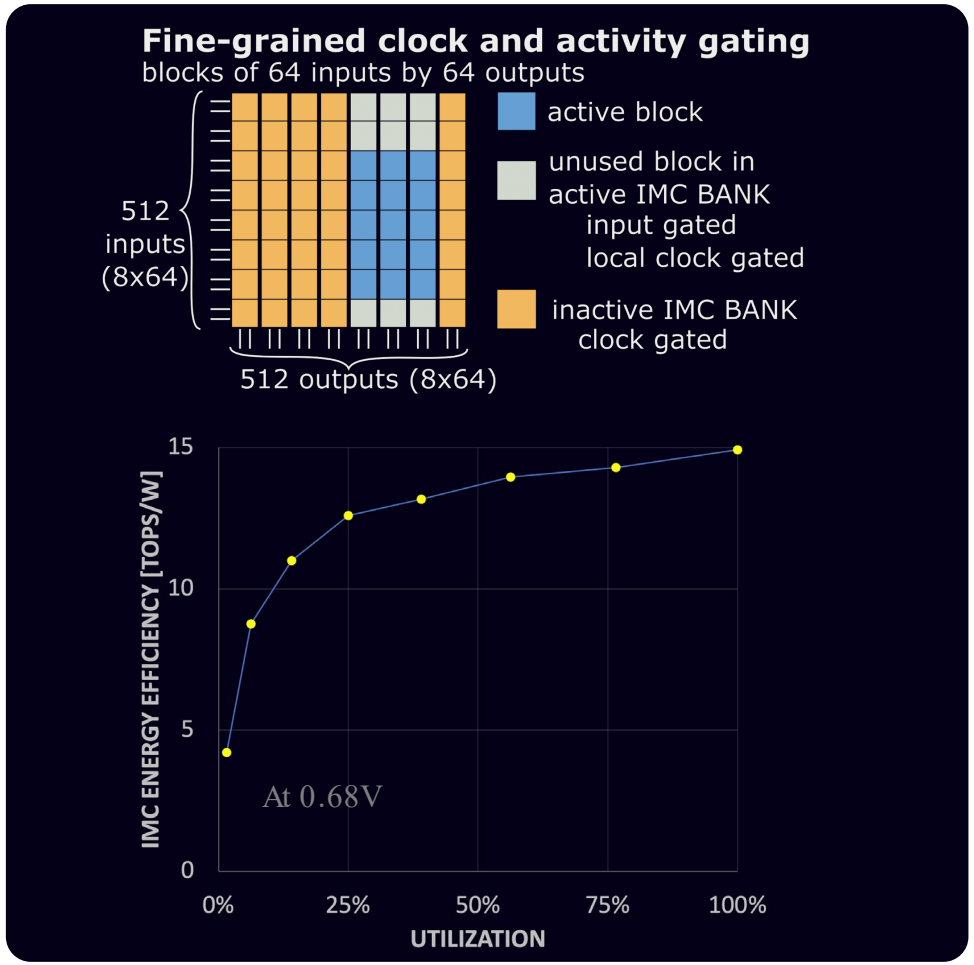
AI Core | Operational Model



Commentary

- Dataflow engine: RISC-V controlled
- Dual high-throughput streaming data paths
  - One for MVM
  - One for DWPU
  - Can operate fully in parallel
- Background weight loading
  - Write weights for next operation
  - In parallel with operation
  - Enabled by multiple weight sets
  - On-the-fly weight decompression

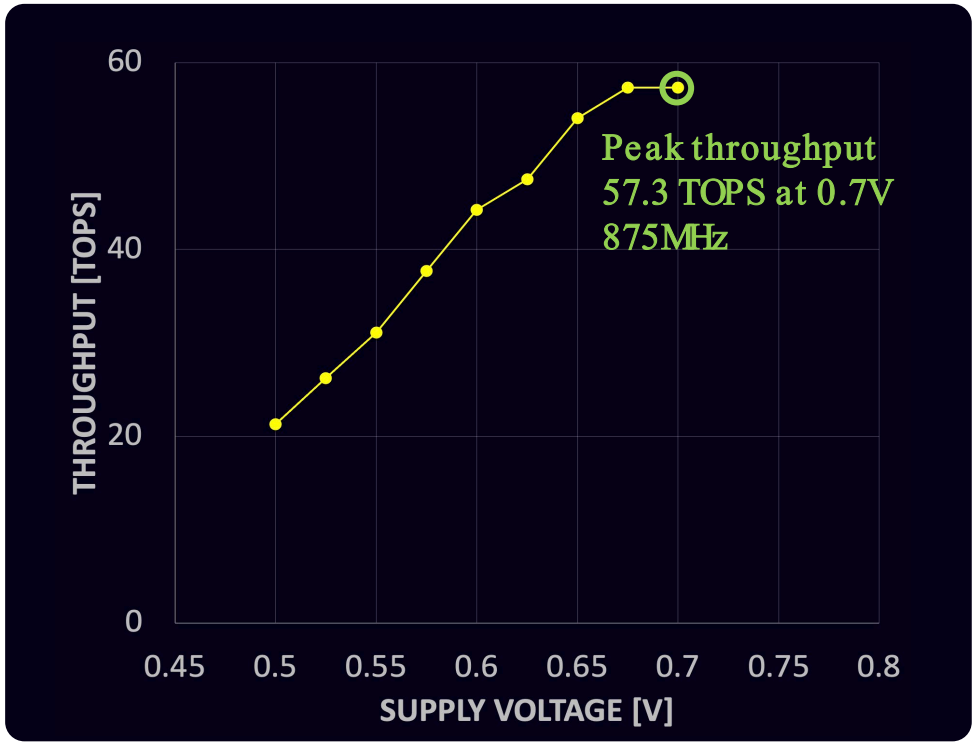
Clock & Activity Gating

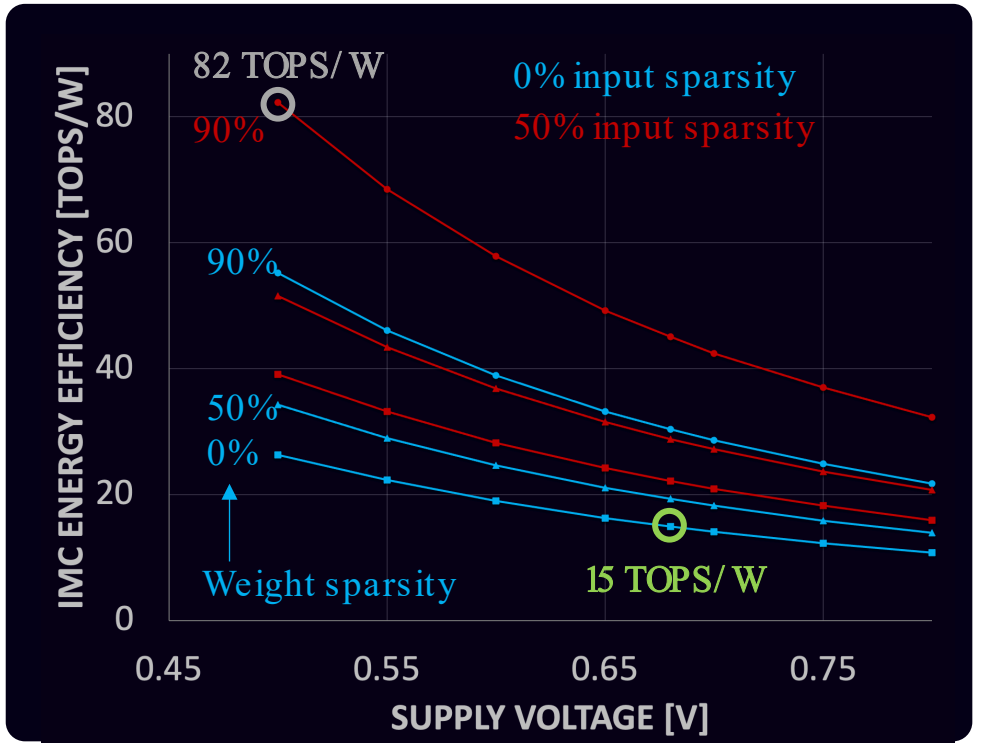


Commentary

- Bank gating
  - If all 64 outputs of an IMC bank pair are unused. Entire bank clock gated
- Block gating
  - If block is not used
    - ✓Block is clock-gated
    - ✓Inputs are silenced
- Energy efficiency high
  - Even at low utilization

Throughput & Energy Efficiency





82 TOPS/W under high sparsity conditions at reduced throughput

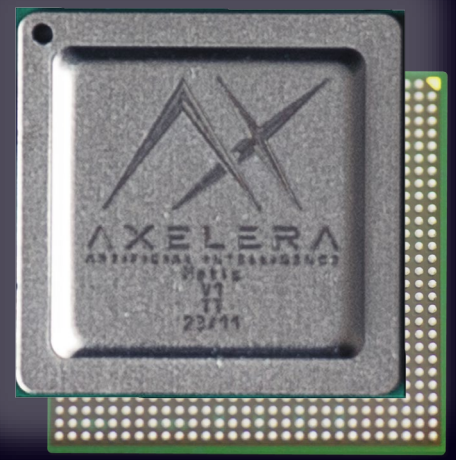
15 TOPS/W for random uniform activations and weights (no sparsity)

Metis AIPU Spec

Peak performance	210 TOPs @ INT8 (0.8 GHz)
# of AI Cores	4 x AIPU ▪ (Int4), Int8 ▪ 16MB L1 SRAM
Internal memory	32MB L2 SRAM 200GB/s aggregate BW
IMC efficiency	15 TOPs/W @ INT8
External memory	LPDDR4x, 34GB/s
Communication bus	PCIE 4x Gen3
Security module	Silex Security IP
Video decoder	-
Pre/post-processing	-
System controller	RISC-V
Node geometry	TSMC N12

VOYAGER SDK

Tools Suite



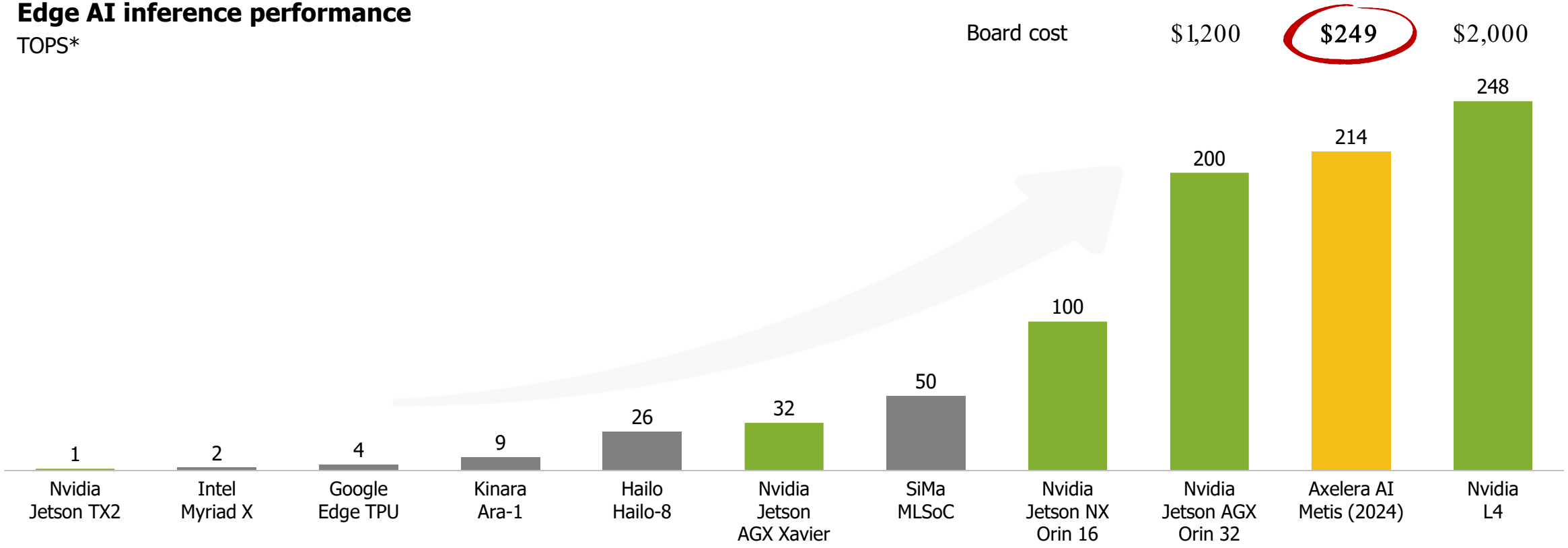
METIS

First Gen (2024)

Performance

Edge AI inference performance

TOPS*



* Tera Operations Per Second (TOPS) is a measure of computational performance, and it quantifies the number of trillion operations (such as additions or multiplications) that a processing unit can perform in one second. TOPS presented here are what was reported in official datasheets. Nvidia performance reported as 'Sparse TOPS' (2x 'Dense TOPS')

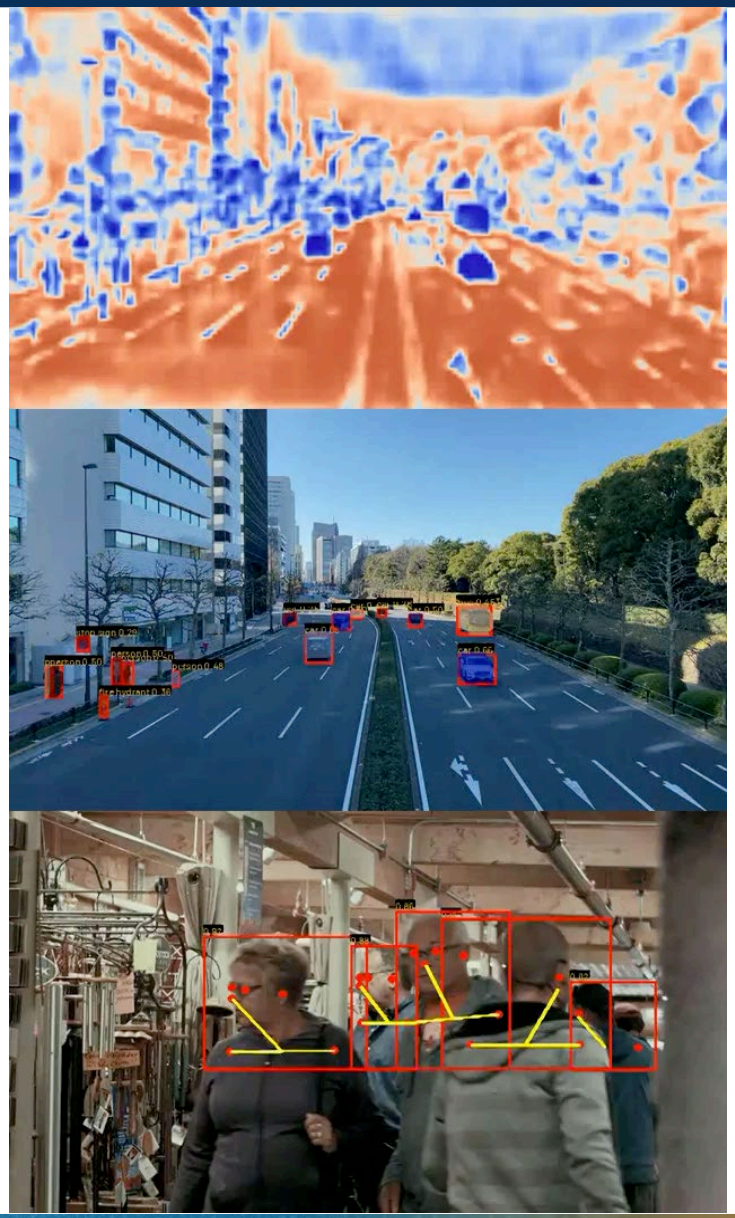
Performance

Deviation from FP32 accuracy

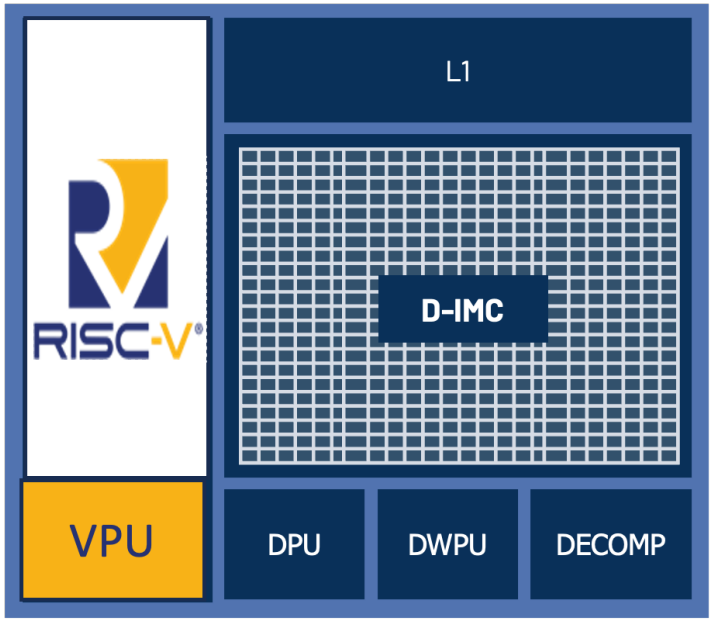
Metis AIPU performance: Benchmarks

Network	Resolution	Metis AIPU [FPS]	Energy Efficiency [FPS/W]	Accuracy @INT8
ResNet-50	(224x224)	3155 fps	394 fps/W	80.69%* (-0.16)
MobileNet-SSD1	(300x300)	5395 fps	771 fps/W	mAP 25.52+ (-0.21)
YoLoV5m	(640x640)	369 fps	46 fps/W	mAP 44.04+ (-1.09)

*Measured on ImageNet-1000 validation
+Measured on COCO detection validation



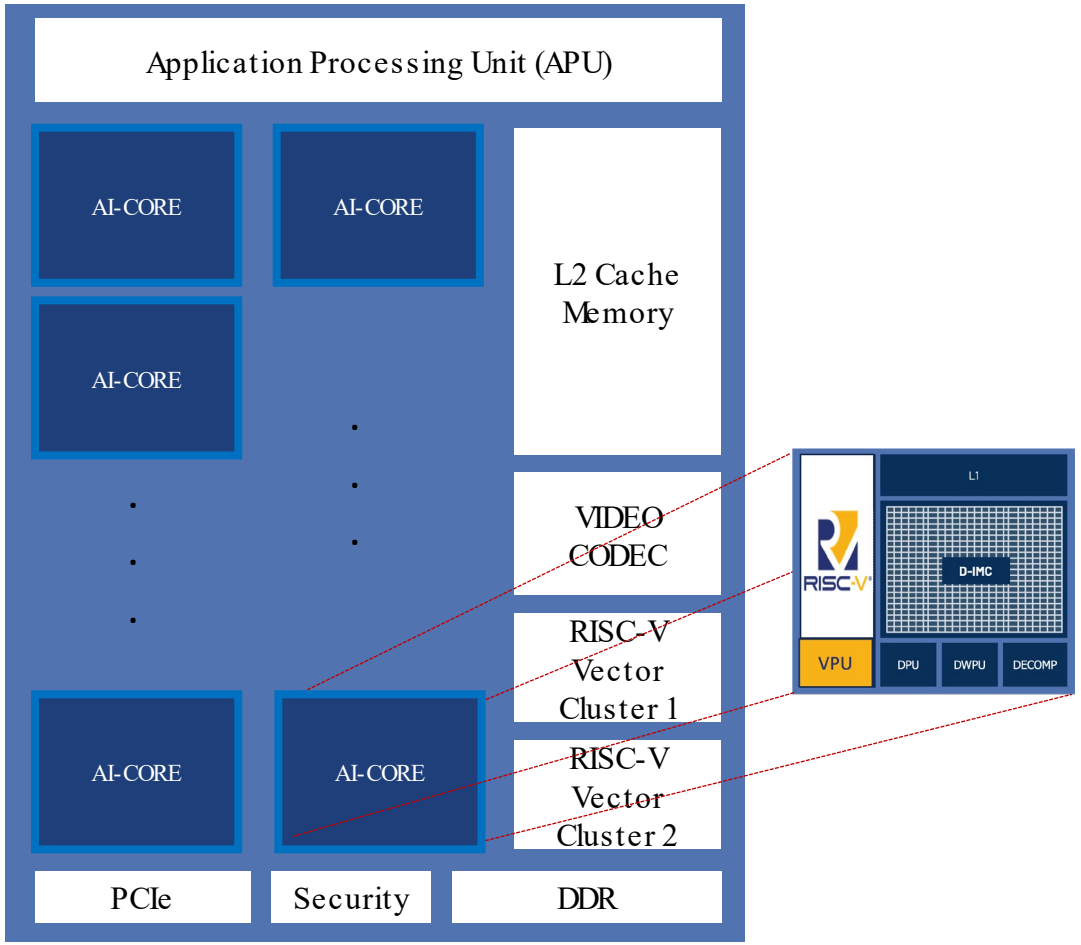
Scaling Up | Integrating RISC-V into the datapath



Commentary

- CVA6: Open-source 64-bit RISC-V core with support for an application-class profile
  - Single-issue, 6-stage, in-order CPU
- Vector Processing Unit (VPU):
  Our proprietary vector extension
  - Builds on the existing RISC-V ISA
  - Vector ISA is agnostic to vector size
  - Enables general compute kernels
  - Flexibility and future-proof

Scaling Up | Design considerations



Commentary

- **Generation Two**: Supports multiple high-speed video streams and medium-sized LLMs, e.g., LLAMA-2/3 7/8B or LLAMA-2 13B, to be deployed on edge servers
  - Massively parallel compute
  - Energy-efficient in-memory computing
  - Large on-chip SRAM
  - Massive on-chip bandwidth
  - High-capacity external memory
  - High-Speed External Bandwidth
  - Support for multi-device pipeline parallelism

Scaling Up | Performance

AIPU evolution: From Metis to Generation Two

Network	Speedup vs Axelera METIS
MobileNetv3	2.9x
ResNet-50	3.9x
SSD-MobileNetV1	2.7x
SSD-ResNet34	3.7x
YoloV5s	3.8x
YoloV5m	3.8x
YoloV8s	5.8x
PHI3	6.3x > 500 output tokens/s, small batch
LLAMA3-8B	6.2x > 3000 output tokens/s, large batch

Scaling initiatives

Our second-generation chip can achieve up to **4x improvement** in performance over Metis based on:

1. Moving to a smaller node geometry
2. Improving the design of the AI core
3. Improving the memory hierarchy
4. Doubling the number of AI cores

Interested in learning more?

The products we sell

M.2 AI Acceleration Cards

- **Form factor:** M.2 2280 M-key
- **AIPU:** 1x Metis AIPU
- **Peak performance:** 100 TOPS
- **RAM:** 2GB of LPDDR4x
- **Connection:** PCIe 3.0 x4
- **Power spec:** Max 15W (typical 7W)

PCIe AI Acceleration Card

- **Form factor:** PCIe CEM (half or full)
- **AIPU:** 1-4x Metis AIPUs*
- **Peak performance:** 214 - 856 TOPS
- **RAM:** 4GB of LPDDR4x
- **Connection:** PCIe 3.0 x4-16*
- **Power spec:** Max 50-200W*

All-in-one AI Systems

- **Peak performance:** 214 TOPS
- **Host device types:** x86 and ARM
- **Systems:** Various

How to reach us

1. Meet us here at AIHW Summit
2. Reach out by email directly evangelos.eleftheriou@axelera.ai
3. Visit our website at axelera.ai and contact our sales team for more information on our products

AXELERA
ARTIFICIAL INTELLIGENCE