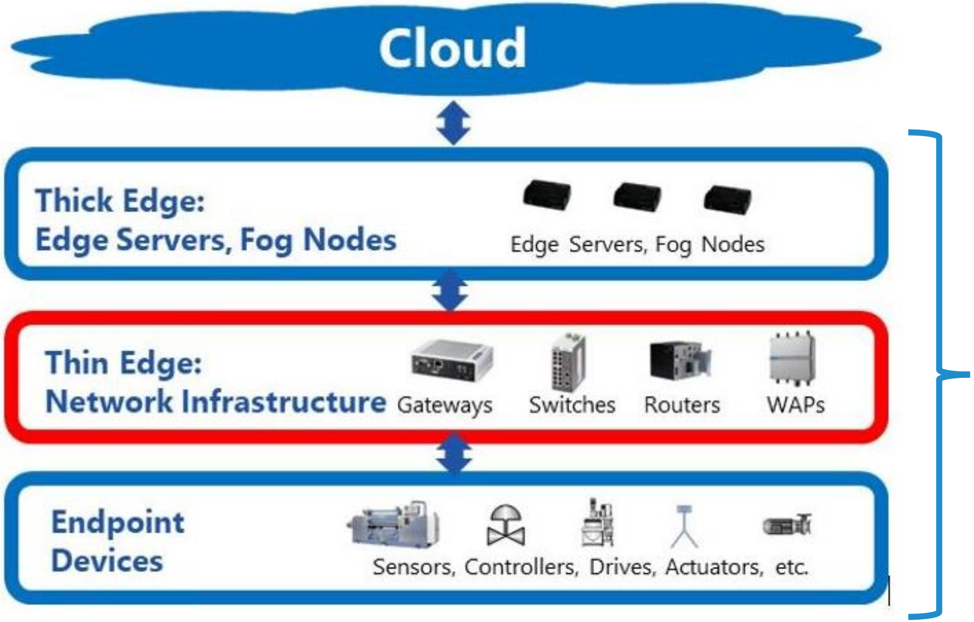
How Axelera AI Uses Digital Compute-in-Memory to Deliver Fast and Energy-Efficient Computer Vision

Bram Verhoef

Head of Machine Learning & Co-Founder

Axelera AI

Compute and Intelligence at Different Layers



The Edge

New AI Applications Are Emerging at the Edge

embedded VISION SUMMIT®

Retail

Inventory management

Cashier-less checkouts

Agriculture

Crop health monitoring

Automated pest control

Industrial

Quality control automation

Worker safety monitoring

Security

Traffic control systems

Intelligent surveillance

Health

Real-time diagnostics tools

Surgical tools & equipment

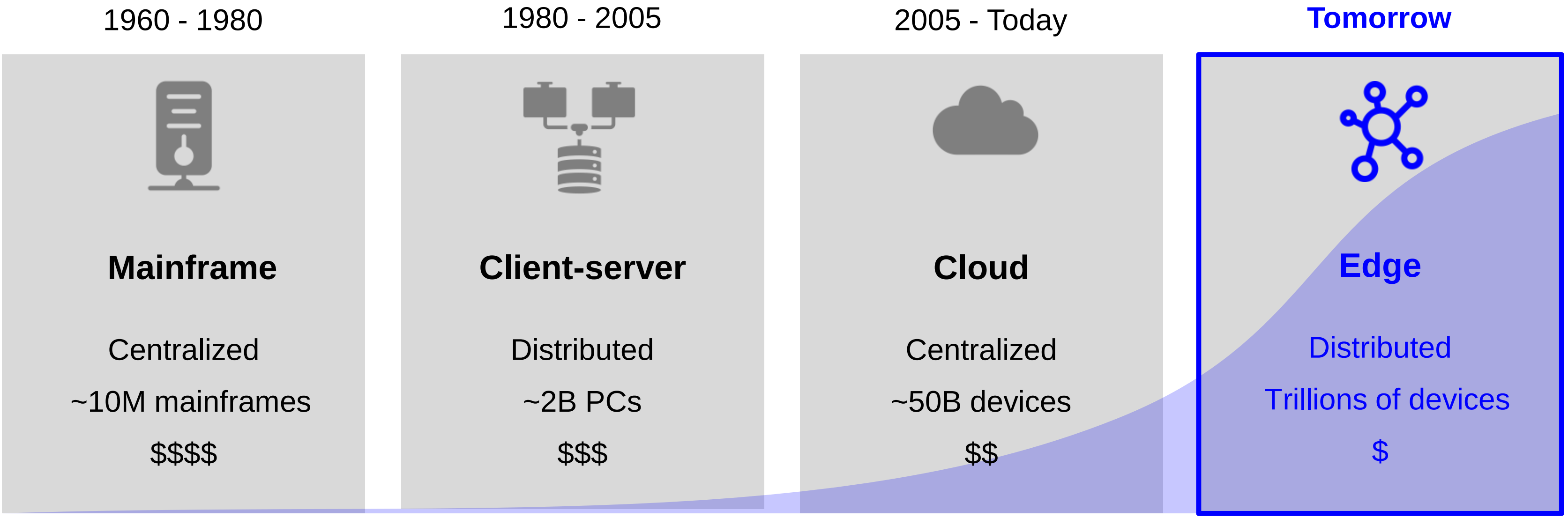
Auto

Driver assistance systems

Autonomous driving systems

AXELERA ARTIFICIAL INTELLIGENCE

AI Is Moving From the Cloud to the Edge



1960 - 1980	1980 - 2005	2005 - Today	Tomorrow
Mainframe	**Client-server**	**Cloud**	**Edge**
Centralized	Distributed	Centralized	Distributed
~10M mainframes	~2B PCs	~50B devices	Trillions of devices
$$$$	$$$	$$	$

Role tomorrow: Training and data storage

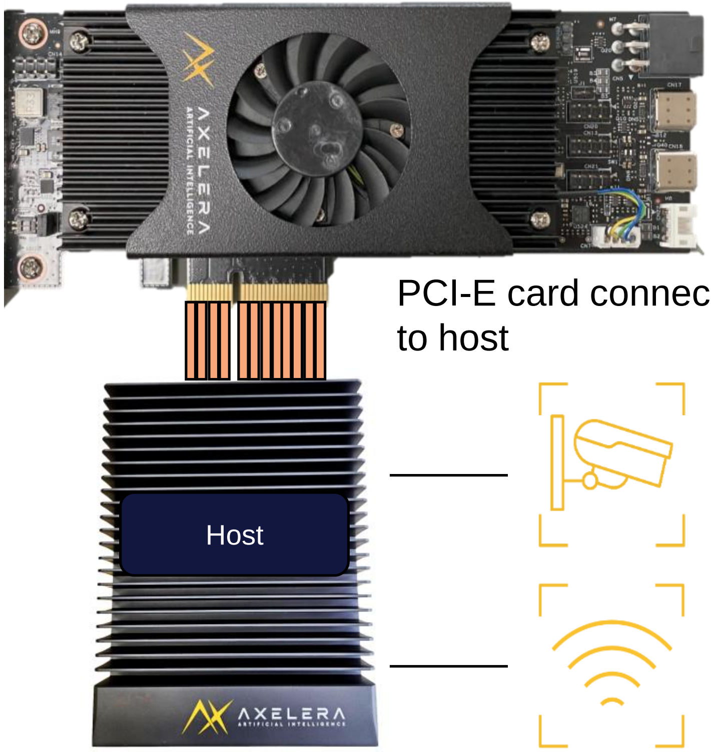
Role tomorrow: Sensing, inference & automation

Emerging AI edge applications require performance and accuracy, energy efficiency, and low price

Fast, Accurate, Energy-Efficient, and Cost-Effective AI Inference With Digital Compute-In-Memory (D-IMC)

Metis - AI Platform

- **AI edge inference accelerator**
  - M.2 module or PCIe card
- **Metis AIPU executes all tasks of an AI workload**
  - Offload complete network(s)
  - Not just individual layers
- **Easy-to-use software stack**
  - Voyager SDK combining compilation and quantization flow

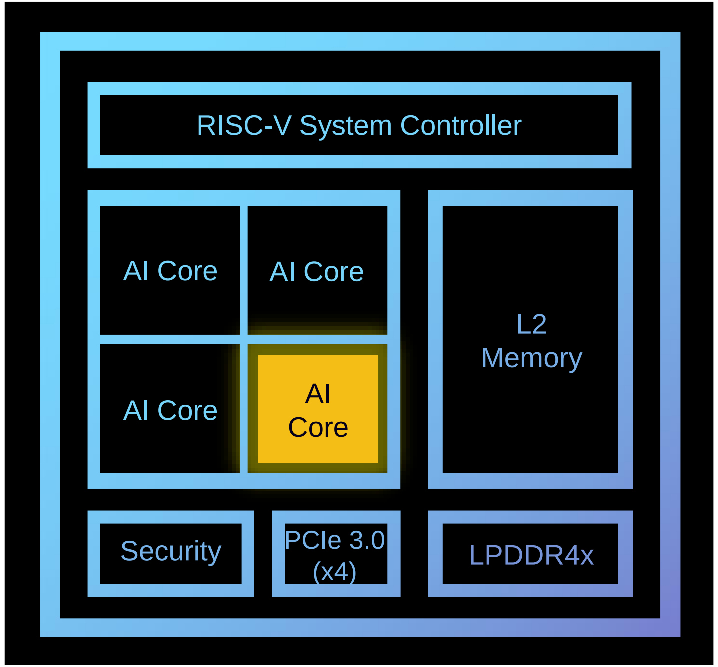

AXELERA ARTIFICIAL INTELLIGENCE
Host
AXELERA ARTIFICIAL INTELLIGENCE

PCI-E card connected to host

AI computer vision applications at the edge

Metis AI Processing Unit (AIPU)

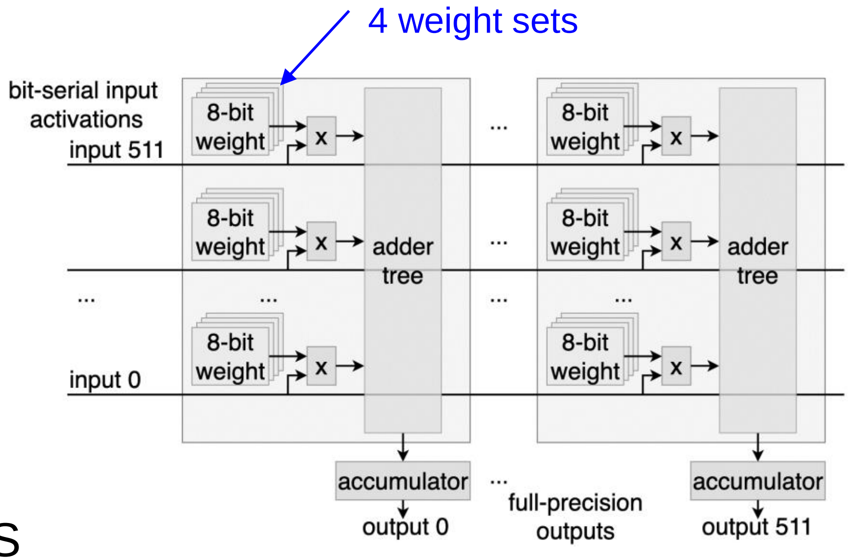
- **Quad-core System-on-Chip**
  - RISC-V controlled
  - Security
  - PCIe 3.0 4x link to host
  - LPDDR4x
  - Large on-chip SRAM capacity
- **AI-Core powered by D-IMC**
  - 52.4 TOPS @ INT8 (209.6 TOPS aggregate)
  - 15 TOPS/W energy efficiency



Digital In-Memory Computing (D-IMC)

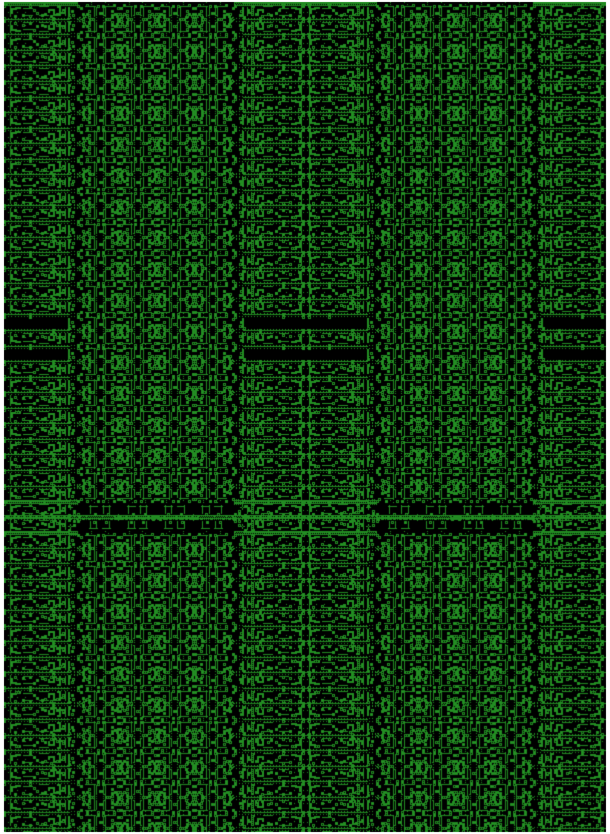
SRAM-based D-IMC

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



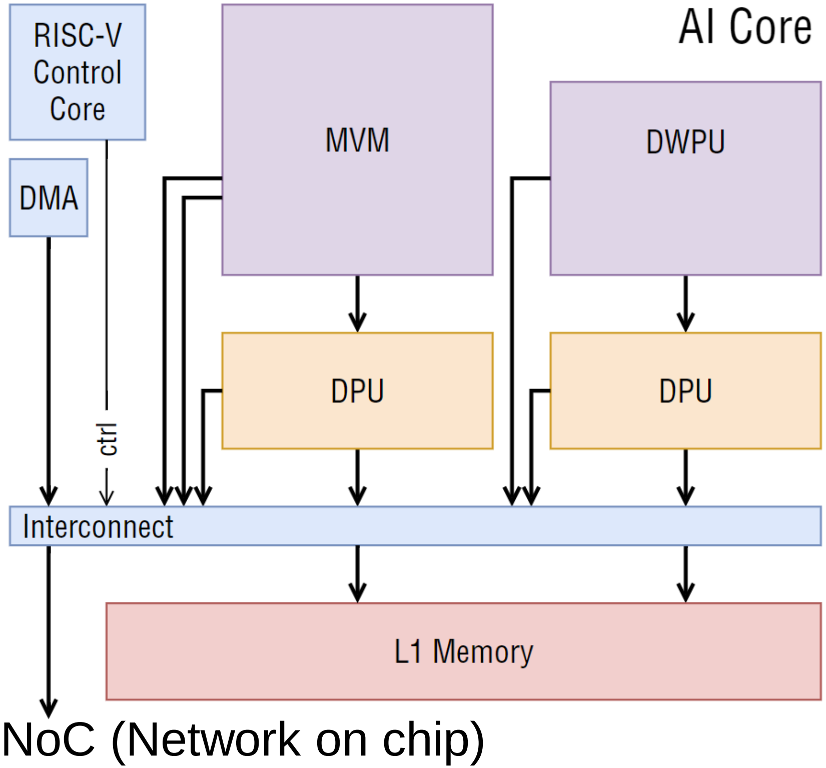
D-IMC Differentiating Improvements

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



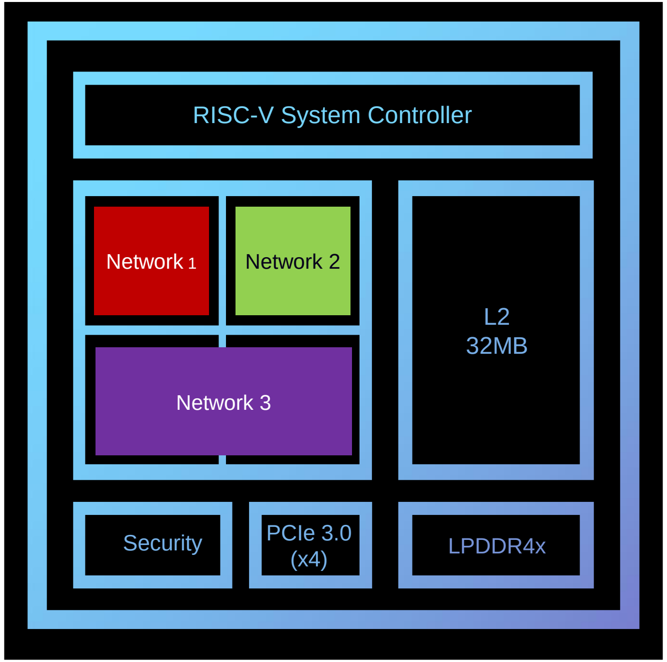
AI Core – Key Components

- **Matrix-Vector Multiplier (MVM)**
  - D-IMC based
  - 512 inputs x 512 outputs (4 weight sets)
  - INT8 inputs and weights
- **Data Processing Unit (DPU)**
  - Element-wise vector operations
  - Apply activation functions
- **Depth-Wise Processing Unit (DWPU)**
  - Depth-wise convolution
  - Pooling and Up-sampling
- **4 MiByte L1 SRAM**
- **RISC-V control core**



AI Core – Deployment Scenarios

- **A single AI core**
  - Can execute all layers of a neural network
  - Eliminates need for external interactions
  - MVM
- **Flexibile deployment of multiple AI cores**
  - Manage different neural networks independently
    - In multi-network applications
  - Jointly tackle a workload to enhance throughput
  - Work on same neural network to reduce latency



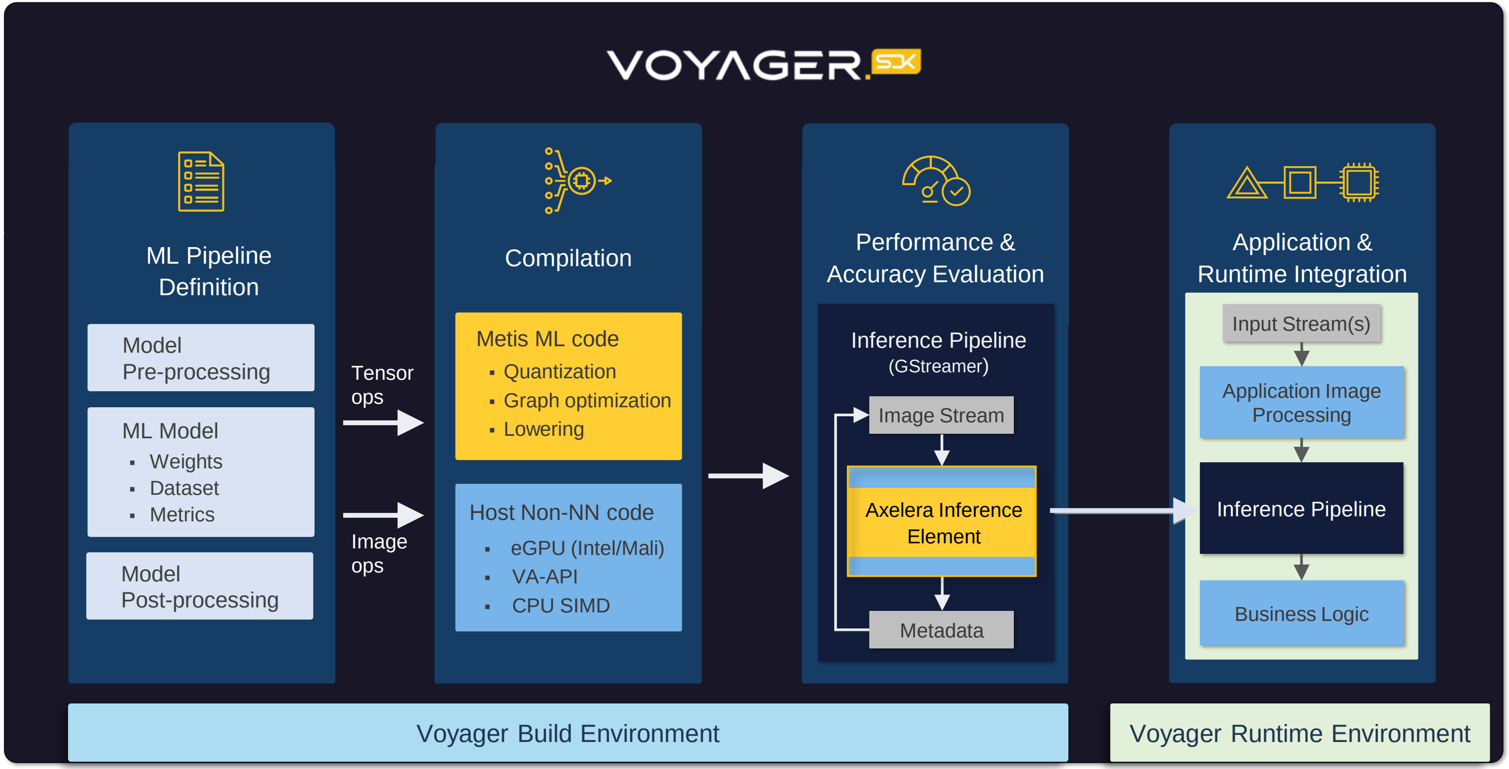
© 2024 Axelera AI

Software Development Flow

Trained Model

- PyTorch
- TensorFlow
- ONNX

Model Zoo
Sample
Pipelines



VOYAGER SDK

ML Pipeline Definition

Model
Pre-processing

ML Model
- Weights
- Dataset
- Metrics

Model
Post-processing

Tensor
ops

Image
ops

Compilation

Metis ML code
- Quantization
- Graph optimization
- Lowering

Host Non-NN code
- eGPU (Intel/Mali)
- VA-API
- CPU SIMD

Performance & Accuracy Evaluation

Inference Pipeline
(GStreamer)

Image Stream

Axelera Inference
Element

Metadata

Application & Runtime Integration

Input Stream(s)

Application Image
Processing

Inference Pipeline

Business Logic

Voyager Build Environment

Voyager Runtime Environment

🟨 Runs on Metis

🟦 Runs on host CPU/GPU (x86 / ARM)

AXELERA
ARTIFICIAL INTELLIGENCE

Metis AIPU SoC Performance

Deviation from FP32 accuracy

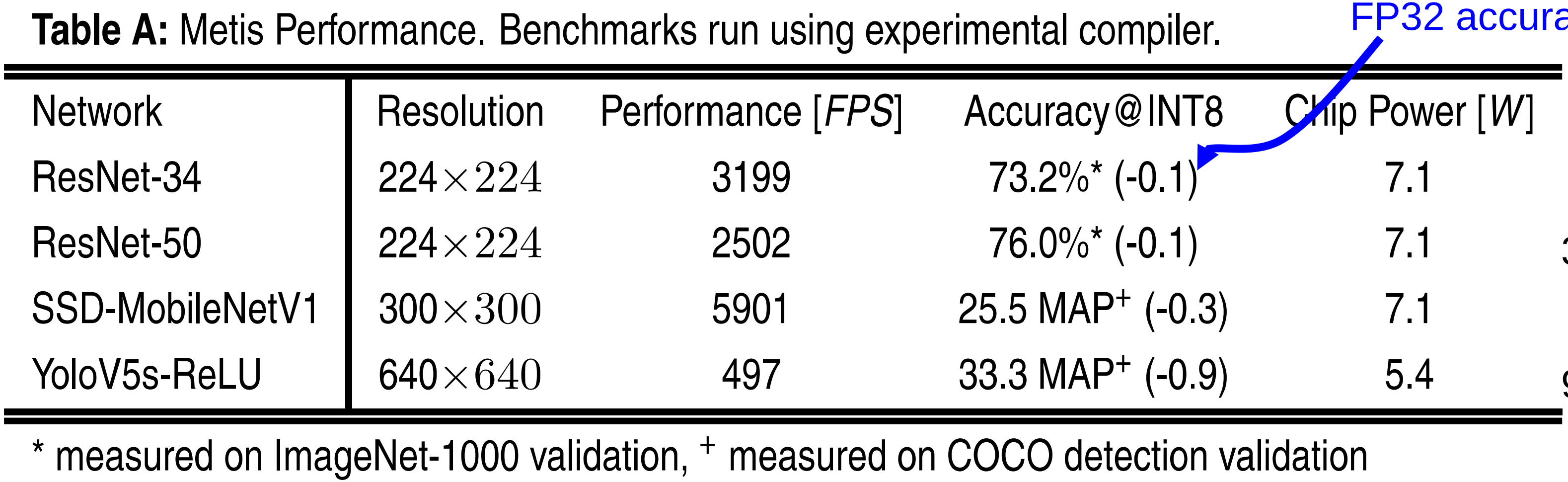


Table A: Metis Performance. Benchmarks run using experimental compiler.

Network	Resolution	Performance [*FPS*]	Accuracy@INT8	Chip Power [*W*]
ResNet-34	224×224	3199	73.2%* (-0.1)	7.1
ResNet-50	224×224	2502	76.0%* (-0.1)	7.1
SSD-MobileNetV1	300×300	5901	25.5 MAP$^+$ (-0.3)	7.1
YoloV5s-ReLU	640×640	497	33.3 MAP$^+$ (-0.9)	5.4

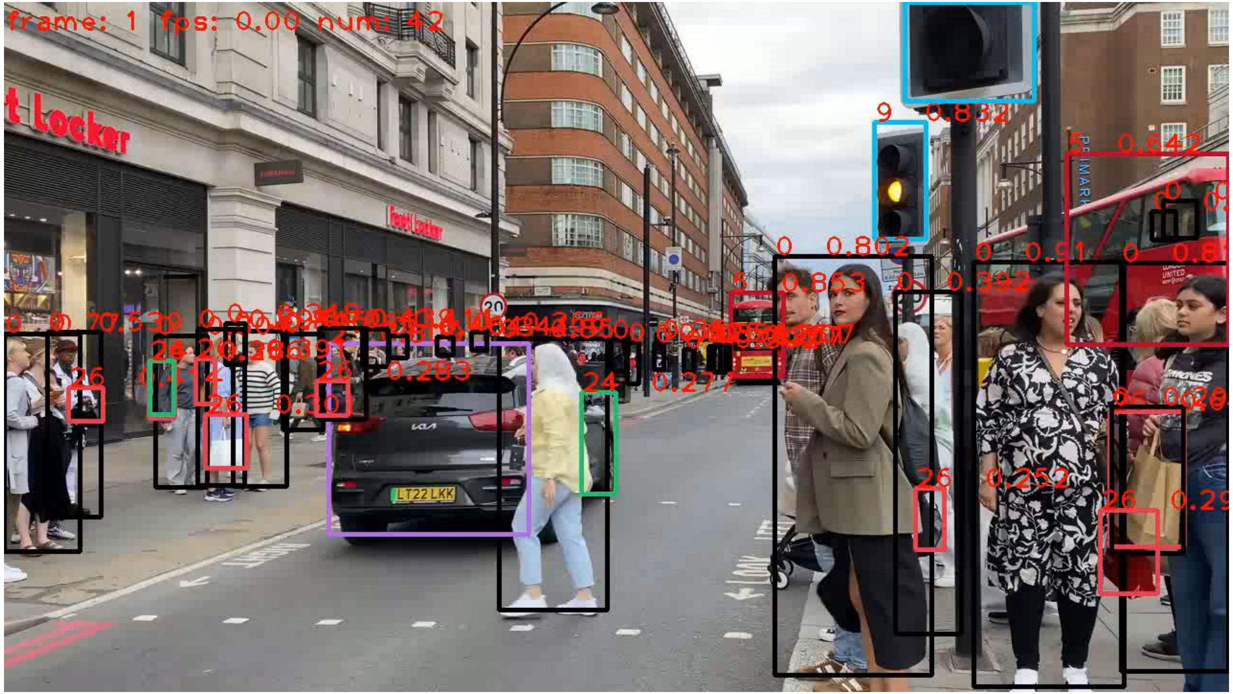
* measured on ImageNet-1000 validation, $^+$ measured on COCO detection validation

354 FPS/W

92 FPS/W

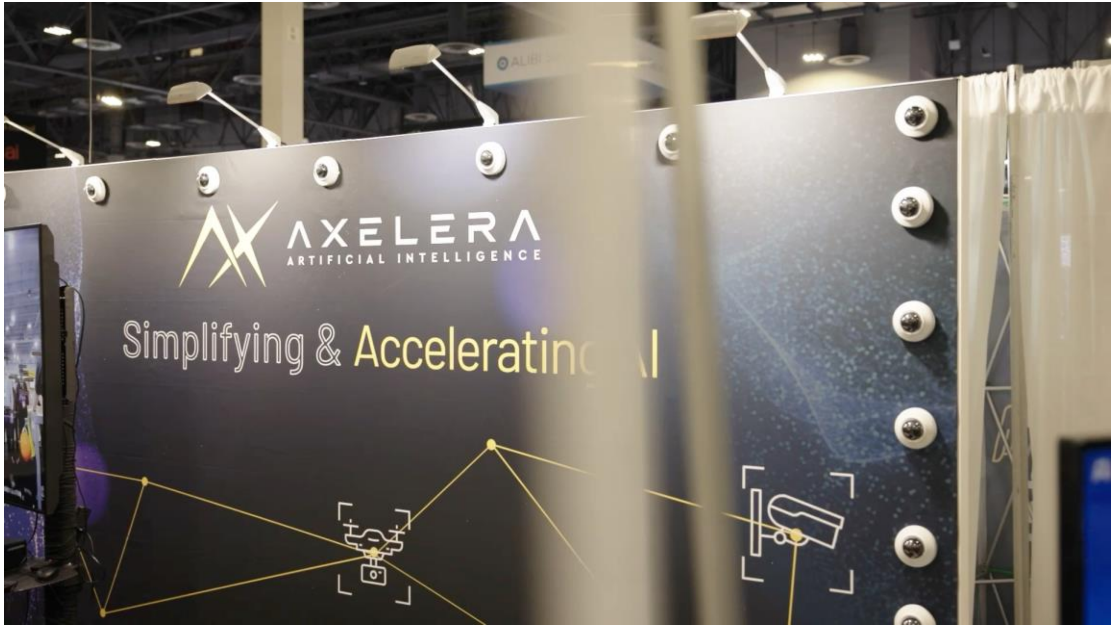
YOLOv5s on Metis – Demo Preview

496 FPS
YoloV5s
inference
@640x640



Running YoloV5s on 24 Streams on a Single Metis Chip

24 RTSP streams
15FPS/stream
1 Metis Chip



Product Line-Up

Modules

Metis M.2

159 USD

AI acceleration to systems with an M.2 2280M slot where space is at a premium

Cards

Metis PCIe

212 USD

PCIe cards with 1x or 4x Metis AIPUs for Edge Servers where AI performance and flexibility is a priority

Boards

Single Board Computer

Price upon request

ARM (Rockchip RK3588)
For stand-alone and compact form factor embedded systems

Systems

Partner products

Price upon request

x86 Edge Servers, Industrial PC’s
Ready to use devices for edge or near edge processing where out-of-the-box systems are needed

Evaluation Kits to get stated

Edge Server PC

Dell Precision 3460XE

Edge Server PC

Lenovo ThinkStation P360

Industrial PC

Advantech ARC-3534

Industrial PC

Advantech MIC-770

Embedded ARM

Firefly ITX-3588J

Metis Evaluation Kits	
Edge Host Systems	Dell Precision 3460XE SFF Core i7 LENOVO ThinkStation P360 ULTRA Core i5 Advantech ARC-3534B Core i5, Industrial PC Advantech MIC-770v3W Core i5, Industrial PC Firefly ITX-3588J, 8-core ARM, embedded
AI Acceleration	Axelera Metis PCIe, 214 TOPS (int8)
PCIe	PCIe 3.0 (x4), HHHL size, 64 x 168 x 40 (mm)
ML frameworks	PyTorch / ONNX / TensorFlow (via ONNX) Axelera Voyager SDK
Neural Networks	Detection: YOLOv5s / m / l / YOLOv7 / SSD-MobileNetV2 Classification: Resnet-50 / MobileNetV2 / and more Pre-compiled optimized models and compiler supported
OS	Ubuntu Desktop v22.04, v20.04 (w/ Docker)

Summing Up: Powerful, Efficient and Cost-Effective AI

- Metis AIPU SoC is an innovative and advanced digital compute-in-memory inference solution for optimized AI computer vision applications
- Metis delivers fast, energy-efficient, cost-effective and accurate AI inference
- Voyager SDK supports deep learning out-of-the-box

Metis evaluation kits available now to get started

Resources

- https://www.axelera.ai
- Products: https://www.axelera.ai/ai-acceleration-hardware-products
- Metis: https://www.axelera.ai/metis-aipu
- Voyager SDK: https://www.axelera.ai/ai-software
- Evaluation Kits: https://www.axelera.ai/metis-evaluation-kit

Thank You!

Visit us at the Axelera booth (#510)!!!