METIS® M.2 CARD

The Most Performant M.2 Edge AI Accelerator

Security

Industry 4.0

Retail

Mobility

Logistics
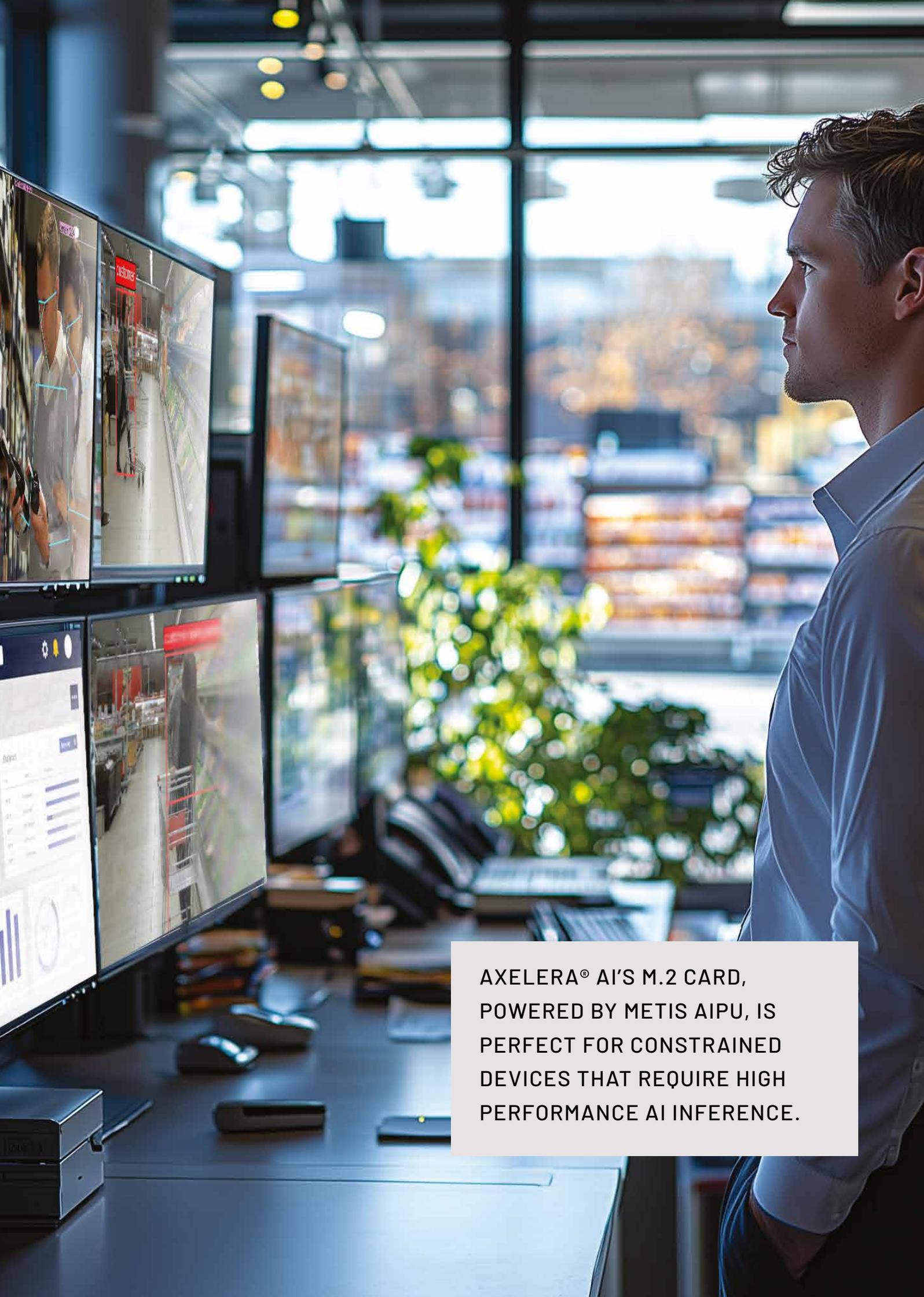
Robotics

Medical

Hospitality

Utilities

Agritech

AXELERA®
ARTIFICIAL INTELLIGENCE

AXELERA® AI'S M.2 CARD, POWERED BY METIS AIPU, IS PERFECT FOR CONSTRAINED DEVICES THAT REQUIRE HIGH PERFORMANCE AI INFERENCE.

METIS® M.2 - KEY FEATURES

- High performance M.2 card feature Metis AIPU to enable state-of-the-art AI inference in small footprint devices.

- A single board can run inference on multiple cameras as well as support many independent parallel neural networks.

- A wide range of end-to-end AI pipelines and models are available out of the box.

- Hassle free evaluation and software integration thanks to Voyager® SDK.

- Uncompromised prediction accuracy thanks to advanced quantization tools.

KEY TECHNICAL SPECIFICATIONS

Form Factor	M.2 2280 M-key
Host Interface	PCIe Gen3 x4 – 4 GB/s bidirectional
AIPU (AI Processing Unit)	1x Metis AIPU
AIPU Memory	1 GB DRAM
Peak INT8 TOPS	214
Operating temperature [1]	-20 to +70°C
Thermal solution	Optional standalone active cooling
Typical Application Power	3.5-9 W
Security Features	Secure Boot, Root of Trust

[1] Extended temperature variant [-40 to +85°C] available upon request.

PROVEN IN KEY MARKETS

Companies in multiple market segments have already adopted Metis M.2-based AI acceleration for different applications such as:

Surveillance: provide multi-camera AI processing capabilities in small form factors for easy deployment in shops, building sites, factories, venues.

Robotics: enable advanced cooperation between human and robots thanks to high resolution advanced gesture recognition, pose estimation and human intention decoding.

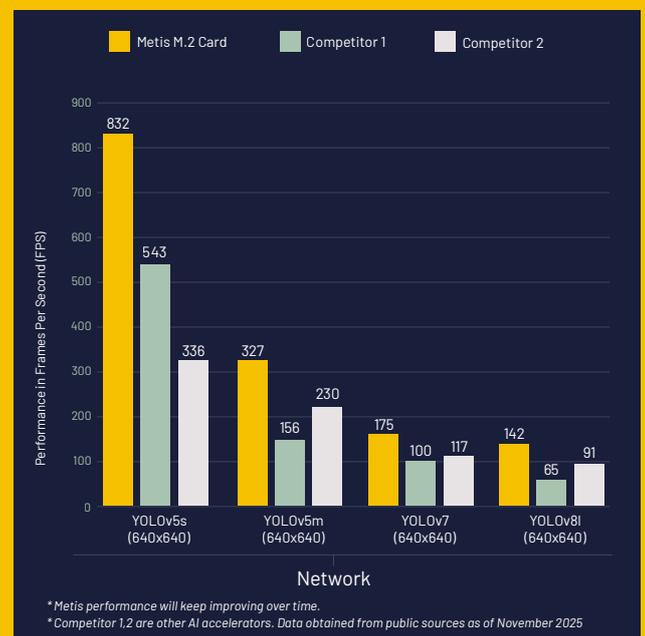**Medical:** improve accuracy of real time diagnostics, monitoring, imaging and analysis.

Agritech: enhance precision in agricultural practices with crop analysis, smart irrigation/pest control, automated harvesting.

Drones: deliver real-time video processing for navigation, object detection and surveillance in lightweight, battery-powered UAV systems.

PERFORMANCE BENCHMARK (FPS). HIGHER IS BETTER

Legend: Metis M.2 Card, Competitor 1, Competitor 2

Performance in Frames Per Second (FPS) vs Network

- YOLOv5s (640x640): 832, 543, 336
- YOLOv5m (640x640): 327, 156, 230
- YOLOv7 (640x640): 175, 100, 117
- YOLOv8l (640x640): 142, 65, 91

*Metis performance will keep improving over time.
*Competitor 1,2 are other AI accelerators. Data obtained from public sources as of November 2025

EASY TO INTEGRATE

Axelera® AI's Metis® technology integrates seamlessly with host CPUs based on both x86 and ARM architectures. Our team actively tests different systems from vendors making it easy for embedded developers to prototype AI applications.

VOYAGER.SDK

Thanks to Voyager® Software Development Kit (SDK), users have a simple software integration path for AI inference at the edge:
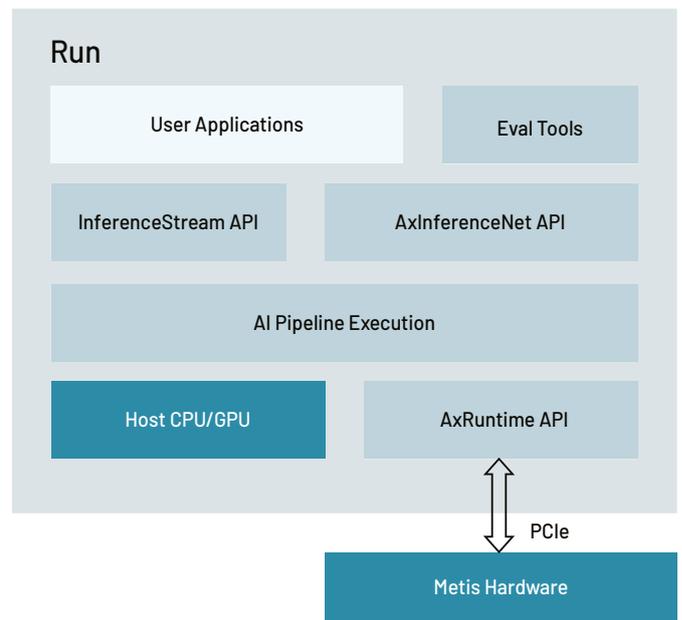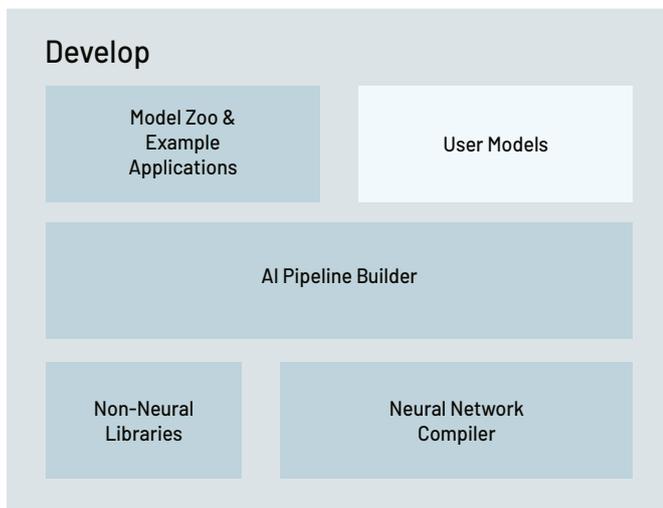
- **Great out-of-the-box experience:** The SDK's built-in tools and models allows evaluating Metis performance, accuracy and power consumption in a few minutes.

- **Fast end-to-end integration path:** The SDK provides a high-level pipeline description framework that allows building optimized end-to-end AI applications with custom inputs, datasets, models and business logic with very few lines of code.

- **Low-level knobs and APIs:** For users that have their own pipelines and software infrastructure, the SDK includes low-level APIs to directly control the inference hardware.

Voyager is a simple yet feature rich SDK:

- Large **Model Zoo** supporting, among others:
  - Small Language Models (Phi3-mini, Llama-3.1 8B etc.)
  - Image Classification (EfficientNet, RestNet etc.)
  - Object Detection (YOLO models, RetinaFace etc.)
  - Semantic Segmentation (U-Net FCN)
  - Instance Segmentation (YOLO models)
  - Keypoint Detection (YOLO models)

- Compiler support for models from Pytorch and ONNX. The compiler automatically manages quantization and graph optimization without user intervention and achieves optimal performance and accuracy.

PyTorch   ONNX

- Libraries including all pre- and post-processing required to run end-to-end pipelines: scaling; cropping; normalization; format conversion; nonmaximal suppression (NMS) and more.

- A YAML description file is used to automatically generate the AI pipelines. The pipeline can then be run as a plugin to GStreamer or within an inference server.

- Built-in tools to test accuracy and performance of models running on Metis AIPU.

Develop

Model Zoo & Example Applications	User Models
AI Pipeline Builder	
Non-Neural Libraries	Neural Network Compiler

Run

User Applications	Eval Tools
InferenceStream API	AxInferenceNet API
AI Pipeline Execution	
Host CPU/GPU	AxRuntime API

PCIe

Metis Hardware