AXELERA SERVER 150p

Unmatched Performance for AI Inference at Scale

WHAT IS

- Security
- Industry 4.0
- Retail
- Mobility
- Logistics
- Robotics
- Medical
- Hospitality
- Utilities
- Agritech

AXELERA® ARTIFICIAL INTELLIGENCE

PROVEN IN KEY MARKETS

Companies in multiple market segments have already adopted Axelera Server-based AI acceleration. What are they using it for?

AXELERA® SERVER 150p CARD, POWERED BY METIS AIPU, OFFERS THE HIGHEST PERFORMANCE INFERENCE ACCELERATION ON THE MARKET, COMBINING EASE OF USE, POWER EFFICIENCY, AND SCALABILITY.

Security: reduce the time to detect and resolve incidents (abandoned baggage, intrusion, fall) thanks to high resolution, high throughput processing of tens of camera feeds.

Industry 4.0: improve accuracy and speed in defect detection and quality control. Increase worker safety with automated PPE control.

Retail: improve operational efficiency and customer experience with in-store customer behavior, stock monitoring and automated checkout systems.

Mobility: real-time vehicle detection, identification and tracking from multiple traffic cameras to enhances tolling, enforcement and parking management.

Logistics: monitor the movement of goods and personnel to improve operational efficiency and safety of logistic centers by improving resource allocation and detecting safety hazards.

AXELERA SERVER 150p - KEY FEATURES

- The highest-performance AI-accelerator Server 150p card in the market for edge AI applications. Powered by four Metis AIPU.
- A single board can run inference on dozens of cameras as well as support multiple 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	PCIe single slot, full height, 3/4 length
Host Interface	PCIe Gen3.0 x16 – 16 GB/s bidirectional
AIPU (AI Processing Unit)	4 Metis AIPUs
AIPU Memory	8 to 64 GB DRAM
Peak INT8 TOPS	856
Operating temperature	0 to 60°C
Thermal solution	Passive and active air cooling
Security Features	Secure Boot, Root of Trust
Typical Power	30 W - 58 W

WORLD-CLASS PERFORMANCE

Axelera outperforms other competing AI accelerators all while maintaining a superior power efficiency. At the same time, Axelera meets much larger and expensive GPU performance while providing a massive efficiency boost.

PERFORMANCE BENCHMARKS

See how Axelera Server 150p Card compares against other AI accelerators across real-world neural networks. Scan to view the latest results.

https://axelera.ai/axelera-products/server#150p

EASY TO INTEGRATE

Axelera 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:

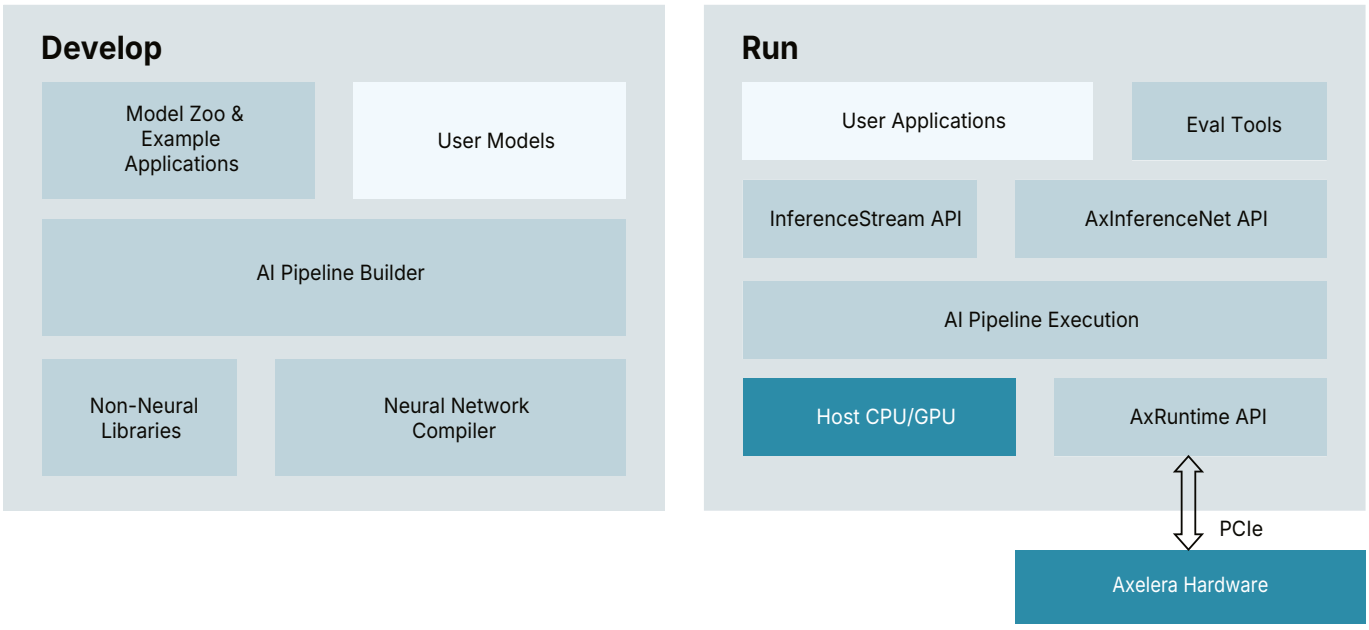
- **Great out-of-the-box experience:** The SDK's built-in tools and models allows evaluating Axelera performance, accuracy and power consumption in a few minutes.
- **Fast end-to-end integration path:** The SDK provides a high-level pipeline description framework that allows building optimized end-to-end 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.



SCAN ME

Ordering information

Part numbers are listed in the product datasheet.
To order Axelera Server 150p card, please visit:
https://store.axelera.ai/products/

AXELERA ARTIFICIAL INTELLIGENCE