

Axelera Server 250p

AI on your terms, powered by Europa

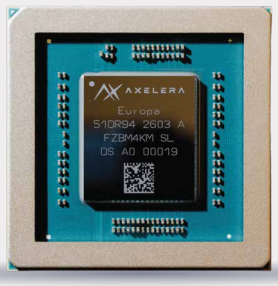
Purpose-built inference for enterprises and system builders running private generative AI on their own infrastructure, computer vision at scale or Physical AI. It all runs from a standard PCIe slot, without rebuilding your stack.

The Axelera Server 250p brings four Europa AIPUs, Axelera's second-generation AI processing units, to each PCIe card. It is purpose-built for high-performance inference in a small energy footprint: real-time answers, a system you own outright, and a lower total cost of ownership.

One card runs the full spectrum of your inference needs: reasoning, language, action models and advanced vision, all in one easy-to-deploy card. Pre-processing, post-processing and video decode stay on board, focusing the CPU utilization on business logic instead of image processing. And because it is a standard passively air-cooled PCIe card in a standard power envelope, inference deploys where your servers or systems already are: on-premises, air-gapped if required, with your models and data never leaving your infrastructure.

Key Features

- **Any modern AI workload.** LLMs, VLMs, agentic pipelines, CNNs, vision transformers and VLAs at INT8, INT16 or INT4 (mem. compression).
- **200+ optimized reference models** across 16 task families, 30 ready-to-run reference pipelines, and support for your own weights and models.
- **The whole pipeline on board.** Dedicated vector processors and a hardware HEVC decoder keep pre/post-processing and video decode off the host allowing for a cheaper host CPU.
- **No infrastructure rebuild.** Dual-slot, passive air cooling, standard PCIe. Same system, same cooling, same power budget.
- **AI you control.** Hardware Root of Trust and Secure Boot. Inference runs entirely on-card. For legal, financial services, healthcare, government and defense workloads where data cannot leave the building. Tokens are not spent in the cloud, but utilized directly within your control.
- **Zero lock-in.** RISC-V throughout, Voyager® SDK on GitHub, same toolchain, same developer workflow.



Key Technical Specifications

AIPU	4x Europa chips per card
AI Performance (INT8)	2516 TOPS
Precision	INT8, INT16 or INT4 (mem. compression)
AI Cores	32x 2nd-Gen, D-IMC with Vector Processing Unit, Improved transformer support
Memory	LPDDR5, 128 or 256 GB, 800 GB/s
Video Decode	HEVC/H.265, 64x FHD or 16x 4K@30FPS, 4x 4K@120FPS
Security	Secure Boot, Root of Trust
Form Factor	PCIe Dual Slot FHFL (Full-Height, Full-Length)
Host Interface	PCIe Gen4 x16 [x4/x4/x4/x4 bifurcation (4x8 GB/s)]
Thermal Solution	Passive heatsink (50 CFM min.)
Power Consumption	100-170 W in typical workloads
Operating Temperature	-10 to 45°C ambient

Easy to Integrate

VOYAGER. TOOLCHAIN

Axelera Wingman Builds it for You

- **Describe the pipeline, get a running code.** Axelera Wingman builds the pipeline, compiles it, runs it on the card and benchmarks the result, drawing on a curated set of workflows, tools and knowledge for Axelera hardware.
- **Need to port your existing project?** Axelera Wingman adapts your existing inference stack, including for non-Axelera hardware, and builds your Axelera pipeline, often from a single prompt.

Voyager SDK Gives You Direct Control

- **Fast end-to-end integration.** Build production-ready pipelines, optimized end-to-end, with very few lines of code.
- **Measure before you commit.** Model Zoo models let you evaluate performance, accuracy and power consumption on the card in minutes.
- **Direct hardware control.** Axelera Script (AxScript) allows you to port and optimize your own models.

Deploy Your Own Models and Weights

Deploy your own model or pair your own weights with a standard architecture. Quick start: every Model Zoo model comes with a short and clean Python recipe: quantization, compilation and end-to-end pipeline in one file, easy to repurpose.

A Simple, Feature-Rich SDK

200+ optimized reference models and ready-to-run reference pipelines in a continuously updated Model Zoo. This is a starting point because with Axelera Script you will be able to port and run your own models too. Representative model families below; see the current list at *https://docs.axelera.ai/sdk/reference/models/model-zoo*

- **Detection** Ultralytics YOLO family, YOLOX, YOLO-NAS, RetinaFace
- **Classification** ResNet, EfficientNet, MobileNet, DenseNet, RegNet, SqueezeNet, Inception, ResNeXt
- **Segmentation and Pose** Ultralytics YOLO models for instance segmentation, semantic segmentation and pose detection. U-Net.
- **Vision transformers** ViT and DINOv2 for instance segmentation, Depth-Anything
- **Vision language** Qwen3.5 0.7B, 2B, 4B, 9B, Qwen 3.6 MoE 35B-A3B, Qwen3.8 27B, LFM3-VL 2B
- **Language models** Qwen and LFM families, from sub-1B edge models to 35B-parameter mixture-of-experts
- **Specialist** Re-identification, depth estimation, license plate recognition, super-resolution, face recognition

Compiler and Tooling

PyTorch and ONNX models compile with automatic quantization and graph optimization. AxScript adds direct hardware control so you can port your own models. Full pre- and post-processing libraries included, with accuracy and performance evaluation built in.

Model authoring

- User Models PyTorch / ONNX
- AxGraph compiler
- AxScript

Runtime

- vLLM plugin — Scalable LLM Serving
- Pipeline Builder — for Computer Vision
- Ax Monitor
- Axelera Runtime
- Kubernetes plugin

To order the Axelera Server 250p,
Contact sales at axelera.ai/contact-us
Part numbers are listed in the product datasheet.

AXELERA ARTIFICIAL INTELLIGENCE