



Hailo-8™ Mini PCIe AI Acceleration Module

An accelerator module for AI applications delivering data center class performance to edge devices

Hailo-8™ Mini PCIe AI Acceleration Module

The Hailo-8™ Mini PCIe is an accelerator module for AI applications based on Hailo-8™ AI processor.

Compatible with the PCI Express Mini (mPCIe) form factor, the module provides 13 tera-operations per second (TOPS) and industry-leading power efficiency.

The module can be plugged into an existing edge device with an mPCIe Full-Mini socket to execute deep learning inferencing in real time and with low power consumption, for a broad range of market segments.

Target Markets



Smart City



Smart Home



Smart Retail



Industry 4.0

KEY Features and Benefits

- Plug-in AI acceleration module for bringing data center class performance to edge devices
- Provides 13 TOPS with best-in-class power efficiency
- Fast time-to-market using standard form factor module
- Enabling real-time, low latency and high-efficiency AI inferencing on edge devices
- Leveraging Hailo's comprehensive Dataflow Compiler and its support for standard AI frameworks, customers can easily and quickly port their Neural Network model
- Comes with a unique high-performance applications infrastructure that shortens development time with visualization and out-of-the-box results



Ordering Information

Purchase the Hailo-8™ mPCIe Module [here](#)

Technical Specifications

- Form Factor: PCI Express Mini Card Full-Mini F1
- Dimensions: 30 x 50.95 mm
- Interface: PCIe Gen-3.0, 1-lane
- Supported AI Frameworks: TensorFlow and ONNX
- Supported OS: Linux and Windows (coming soon)
- Supported host hardware: x86 based architecture or ARM AArch64 based architecture
- Certification: CE, FCC

Featuring Hailo-8™ AI Processor

The Hailo-8™ AI processor offers up to 13 tera-operations per second (TOPS) in the mPCIe form factor, significantly outperforming all other edge AI processors. The AI processor's area and power efficiency are far superior to other leading solutions by an order of magnitude.

With an architecture that takes advantage of the core properties of neural networks, edge devices can now run deep learning applications at full scale more efficiently, effectively, and sustainably than traditional solutions, while significantly lowering costs.

Hailo Dataflow Compiler

