SAKURA™-I
Edge AI Accelerator

Delivering High Performance at Low Power & Low Latency

EdgeCortix SAKURA-I is a TSMC 12nm FinFET co-processor (accelerator) delivering efficient compute and low latency for edge artificial intelligence (AI) inference. It is powered by a 40 TOPS (dense), single core Dynamic Neural Accelerator® (DNA), which is EdgeCortix's proprietary neural processing engine with built-in runtime reconfigurable data-path effectively connecting all compute engines together.

SAKURA-I runs multiple deep neural network models together, providing exceptional TOPS utilization at ultra-low latency. This capability is key for consolidated workloads, enhanced processing speed and lower energy at reduced total cost of ownership.

Efficiency

Compute
Up to **4X TOPS utilization** vs. GPUs and TPUs

Energy
Up to **7X better** (IPS/W) vs. existing solutions

Latency
Market leading **real-time Batch 1** processing

SAKURA-I Key Metrics

Peak Processing:	40 TOPS
Data Format:	INT8
Compute Efficiency:	Execution flow is runtime configurable; and achieves up to 90% of peak processing on real-world workloads
Latency:	Batch Size 1
On-chip Memory:	20MB
External Memory Support:	2x ports of x64b LPDDR4x
Host Interface:	PCIe Gen 3.0 x16

SAKURA-I Low Profile Board Key Metrics

Form Factor:	Low Profile PCIe (68.90 × 167.65 × 20.32mm)
External Memory:	16GB (2x banks of 8GB LPDDR4)
Host Interface:	PCIe Gen 3.0 x16
Board Power:	10W - 12W

Key industrial segments where the SAKURA-I performance profile is ideally suited include:

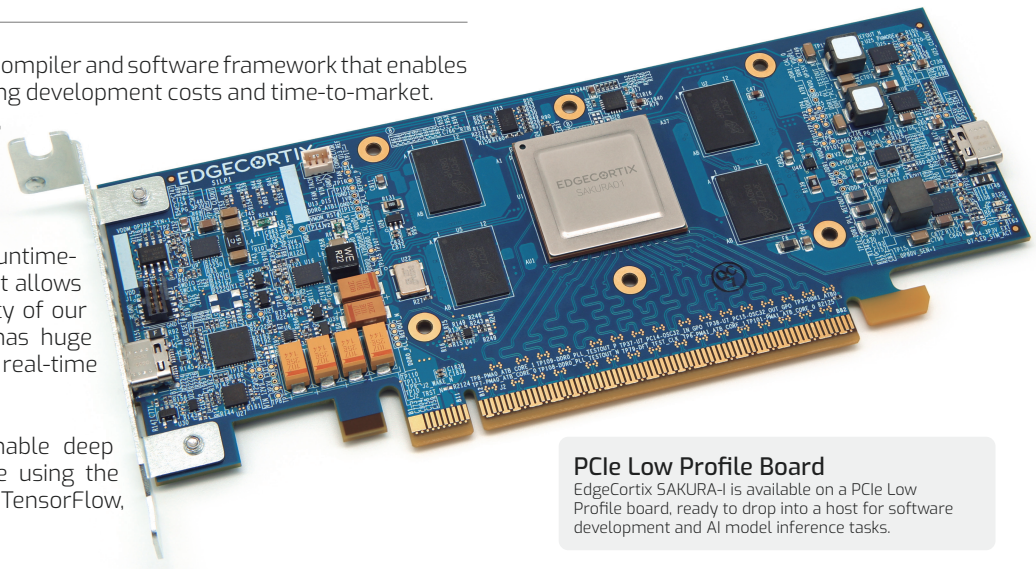
- Transportation/Autonomous Vehicles
- Defense/Aerospace/Security
- 5G Communications
- Augmented & Virtual Reality
- Smart Manufacturing/Robotics
- Smart Cities
- Smart Retail
- Drones & Robotics

Product Description

SAKURA-I is supported by MERA, a heterogeneous compiler and software framework that enables inference offloads from hosts, significantly reducing development costs and time-to-market. This combination enables seamless compilation, execution, and hardware acceleration of standard or custom convolutional neural networks (CNNs) developed in industry-standard frameworks.

Dynamic Neural Accelerator (DNA) is a novel runtime-reconfigurable neural processing architecture that allows us to significantly increase the compute efficiency of our AI devices as compared to typical GPUs. This has huge benefits for lower power yet high performance, real-time processing.

MERA provides a simple API to seamlessly enable deep neural network graph compilation and inference using the DNA AI engine in SAKURA-I, using PyTorch, ONNX, TensorFlow, or TensorFlow Lite.



PCIe Low Profile Board
EdgeCortix SAKURA-I is available on a PCIe Low Profile board, ready to drop into a host for software development and AI model inference tasks.

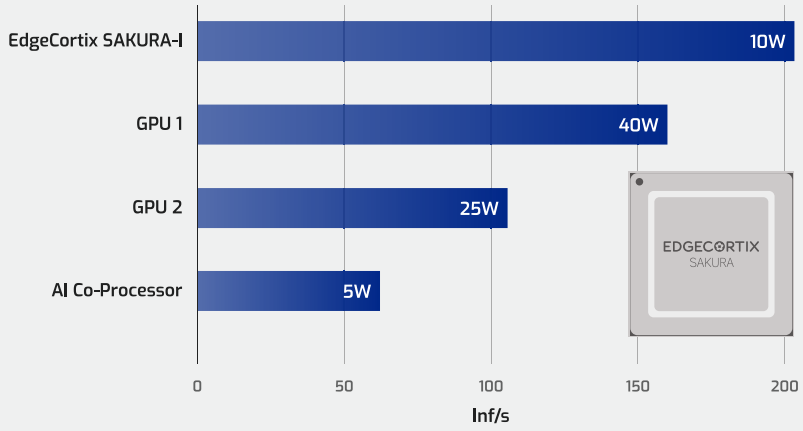
SAKURA-I Key Benefits

Efficient Edge Inferencing Alternative to GPUs
- Lower power
- Lower latency
- Higher compute efficiency, up to 90% of peak TOPS
  - Comparable to GPUs/TPUs running at 120-160 TOPS
- No need for retraining
- Python and C++ interfaces
- PyTorch, ONNX, TensorFlow, and TensorFlow Lite natively supported

Real-time Processing
- Optimized for streaming data
- Batch 1 workloads with higher efficiency
- Runtime configurable execution flow

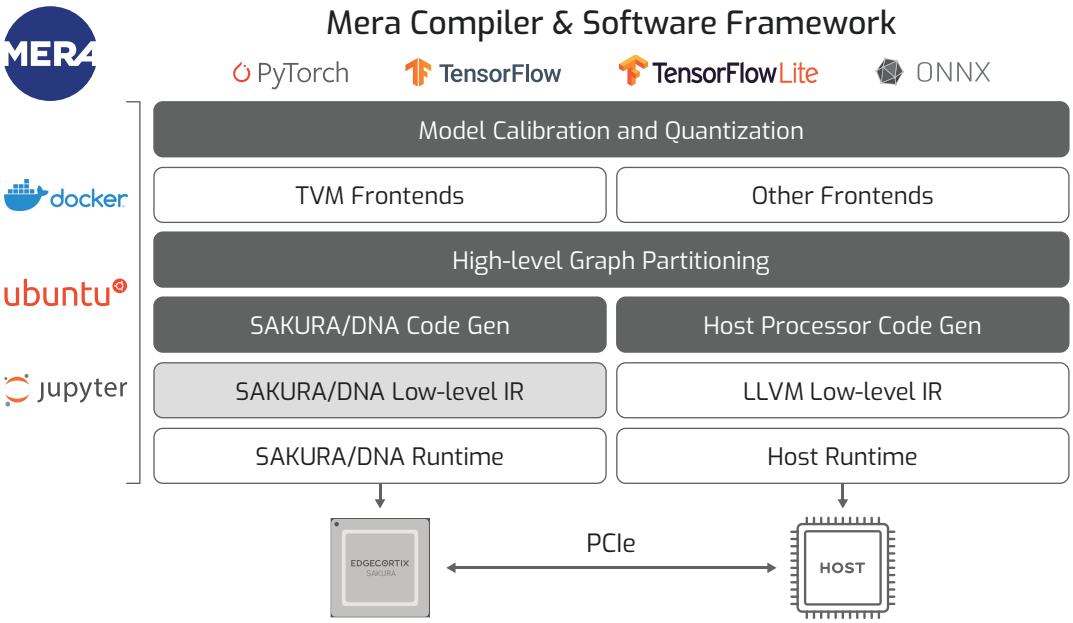
Dedicated AI Accelerator/Co-processor
- Easy to integrate with existing systems
- Standard PCIe interconnect with I/O and Host

Industry-Leading Performance and Efficiency

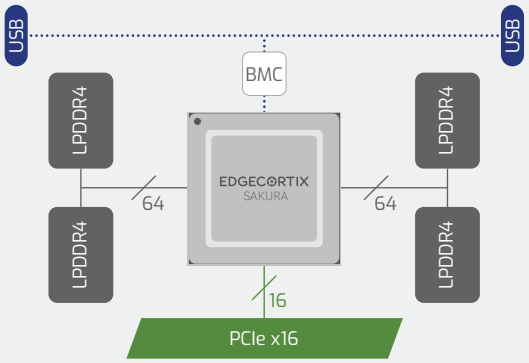


Inferences/Sec - YOLOv5m, 640x640, Batch=1

Mera Compiler & Software Framework



SAKURA-I Low Profile Board Diagram



Download MERA and test SAKURA today