BittWare
a molex company

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 trillion operations per second (TOPS), 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

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

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 • < 4 ms on intensive inference workloads
On-chip Memory:	20MB
External Memory:	2x 64b LPDDR4
Host Interface:	PCIe Gen 3.0 x16

S1LP1 Board Key Metrics

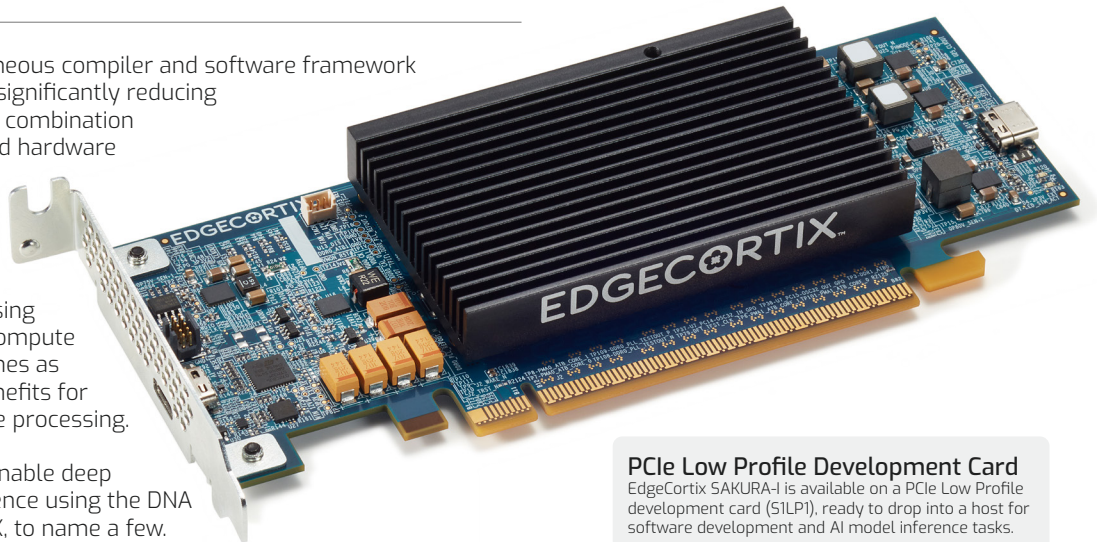
Form Factor:	Low Profile PCIe (68.90 × 167.65 × 20.32mm)
External Memory:	16GB (2x banks of 8GB LPDDR4)
Host Interface:	PCIe Gen 3.0 x16
Board Power:	10W - 12W

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 (CNN) developed in industry-standard frameworks.

Dynamic Neural Accelerator (DNA) is a novel runtime-reconfigurable neural processing architecture that allows us to increase the compute efficiency of our AI chips, by more than 5 times 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. Tensorflow, and ONNX, to name a few.



PCIe Low Profile Development Card
EdgeCortix SAKURA-I is available on a PCIe Low Profile development card (S1LP1), ready to drop into a host for software development and AI model inference tasks.

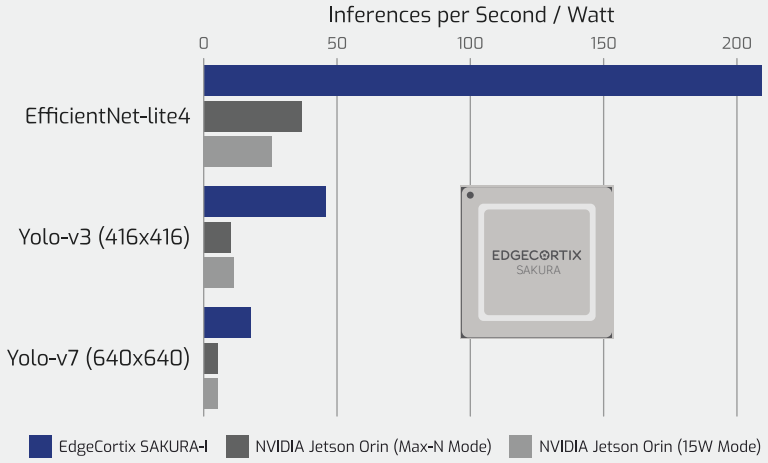
SAKURA-I Key Benefits/Features

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, TensorFlow and ONNX natively supported

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

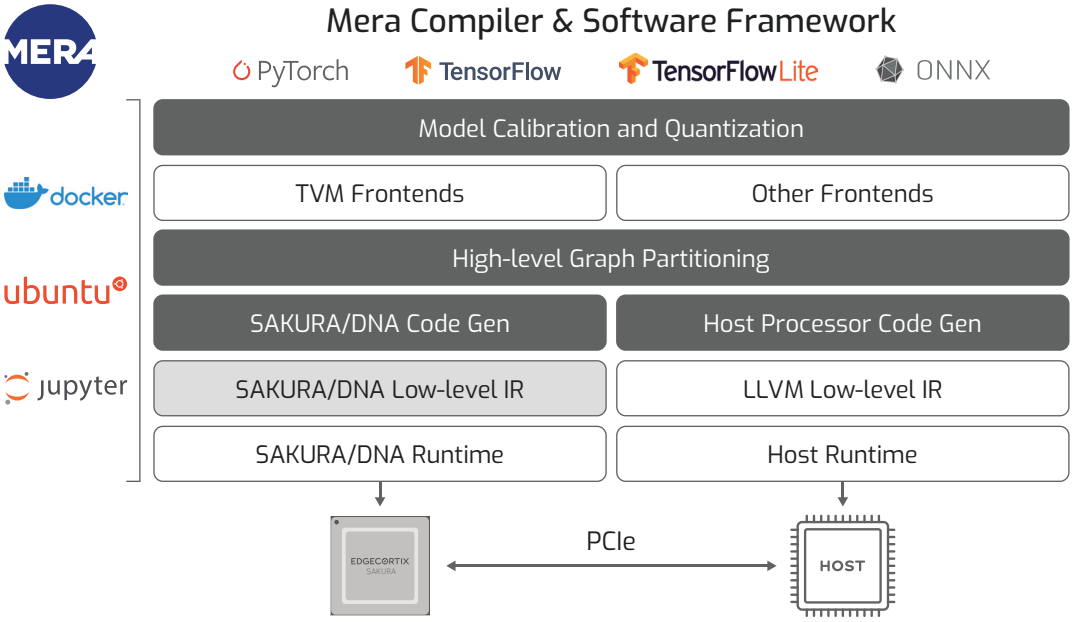
Dedicated AI Accelerator/Co-processor
- Easy to integrate with existing systems
- Standard PCIe interconnect with I/O and Host

SAKURA-I Performs on average 3X-8X better than NVIDIA GPUs*

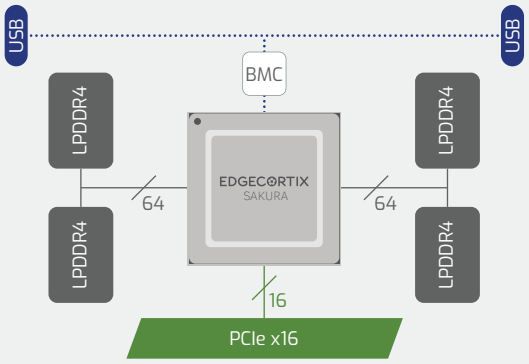


*EdgeCortix SAKURA-I is benchmarked vs. NVIDIA Jetson Orin under different power modes. All metrics are batch size 1. Models deployed at INT8 without any sparsity. NVIDIA Jetson Orin metrics compiled using latest tools. SAKURA-I metrics compiled and deployed on hardware using MERA v1.3 software.

Mera Compiler & Software Framework



S1LP1 Board Diagram



Download MERA and test SAKURA today

BittWare a molex company

SAKURA™-I Edge AI Accelerator
To learn more visit **bittware.com/edgecortix**