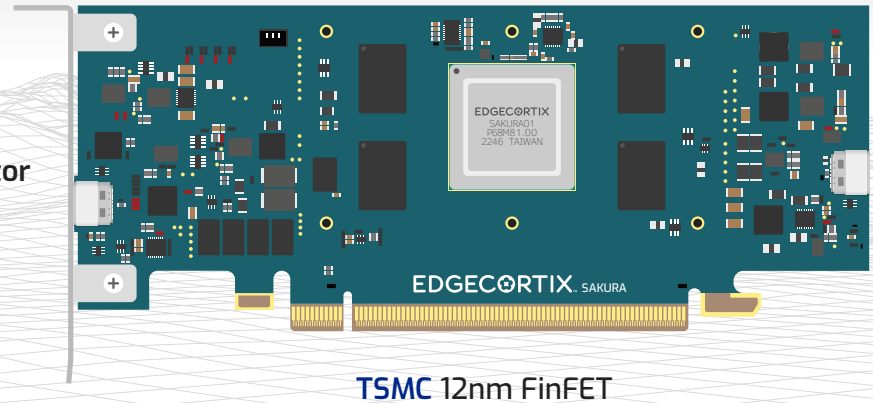
EDGECORTIX

Deep Learning Accelerator

SAKURA-I™
Dynamic Neural Accelerator
Framework



TSMC 12nm FinFET
Performance of 40 TOPs @ 800 MHz

EdgeCortix™ SAKURA-I: Energy-efficient Edge AI Co-processor

EdgeCortix SAKURA-I is a TSMC 12nm FinFET co-processor (accelerator) delivering class-leading compute efficiency and latency for edge artificial intelligence (AI) inference. It is powered by a 40 trillion operations per second (TOPS), single core **Dynamic Neural Accelerator® (DNA) Intellectual Property (IP)**, which is EdgeCortix's proprietary neural processing engine with built-in runtime reconfigurable data-path connecting all compute engines together. DNA enables the new SAKURA-I AI co-processor to run multiple deep neural network models together, with ultra-low latency, while preserving exceptional TOPS utilization. This unique attribute is key to enhancing the processing speed, energy-efficiency, and longevity of the system-on-chip, providing exceptional total cost of ownership benefits. The DNA IP is specifically optimized for inference with streaming and high-resolution data.

Key industrial segments where the SAKURA-I performance profile is ideally suited include:
transportation/autonomous vehicles • defense • security • 5G communications • augmented & virtual reality • smart manufacturing • smart cities • smart retail & robotics • all markets that require low power • low latency AI inference.

KEY FEATURES

40 TOPs
AI compute

INT8 Inference
(99% of FP32 accuracy)

Low Power
10-15W TDP

PCIe x16 dev. boards available

Hardware Architecture Overview

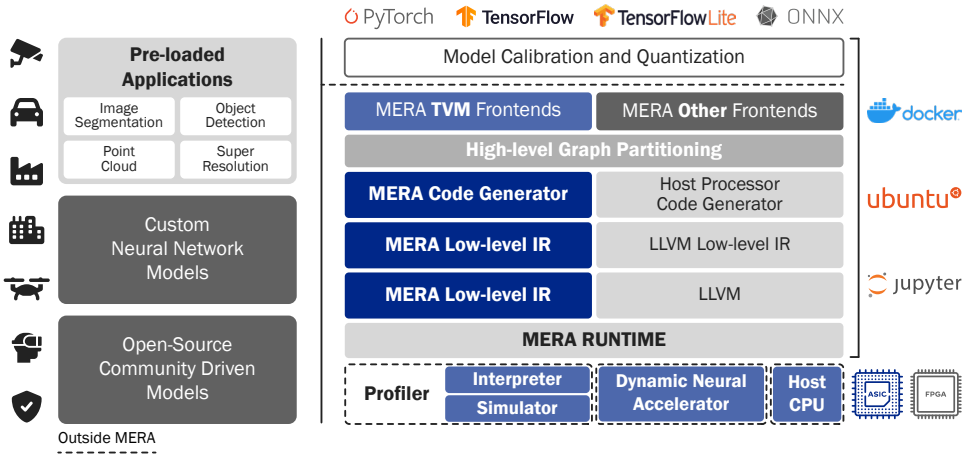
- Up to 40 TOPS (single chip) and 200 TOPS (multi-chip)
- Board Power @ 10W-15W
- Typical model Power consumption ~5W
- 2x64 LPDDR4x – 16 GB
- PCIe Gen 3 up to 16 GB/s bandwidth
- Two form factors: Low-profile PCIe & M.2
- Runtime-reconfigurable datapath

Dynamic Neural Accelerator® IP

- Optimized for INT8 and batch size 1
- Relatively large on-chip memory – 20 MB
- Maximises compute utilization exploiting multiple degrees of parallelism defined by software
- Extreme low-latency (< 4 ms) on demanding workloads, like Yolov3, Yolov5, Point-Cloud processing based AI etc.

Product Description

MERA Compiler and Software Framework



EdgeCortix SAKURA-I AI Co-processor enabled devices are supported by the heterogeneous compiler and software framework - EdgeCortix MERA that can be installed from a public pip repository, enabling seamless compilation and execution of standard or custom convolutional neural networks (CNN) developed in industry-standard frameworks. MERA has built-in integration with Apache TVM, and provides simple API to seamlessly enable deep neural network graph compilation and inference using the DNA AI engine in SAKURA-I. It provides profiling tools, code-generator and runtime needed to deploy any pre-trained deep neural network after a simple calibration and quantization step. MERA supports models to be quantized directly in the deep learning framework, e.g., Pytorch or TensorflowLite.

DETAILED FEATURE LIST

Drop-in Replacement for GPUs

- Python and C++ interfaces
- PyTorch, ONNX and TensorFlowLite natively supported
- No need for retraining
- Supports high-resolution inputs

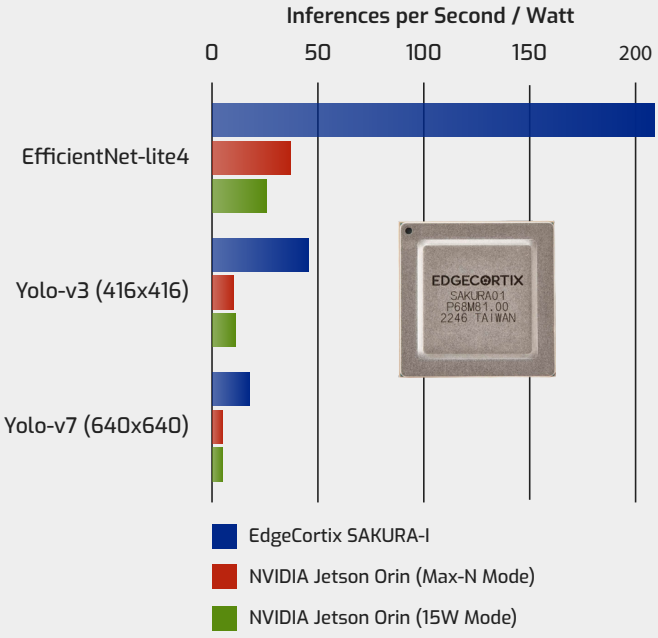
INT8 / BF16 bit Quantization

- Post-training calibration and quantization
- Support for deep learning framework built-in quantizers
- Preserve high accuracy

Built-in Simulator

- Deploy without the SAKURA-I device, simulating inference within x86 environment
- Estimate inference latency & throughput under different conditions

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



* EdgeCortix SAKURA-I is benchmarked vs. NVIDIA Jetson Orin under different power modes. NVIDIA Jetson Orin is expected to be a TSMC 7nm device, while SAKURA-I is TSMC 12nm device. End to end latency is measured under batch size 1, in all cases. All models deployed at INT8 without any sparsity tricks. All measurements on NVIDIA Jetson Orin were compiled by the vendor's latest tools. All SAKURA-I measurements were compiled and deployed on hardware using EdgeCortix MERA v1.3 software.