EDGECORTIX®
SAKURA®-II AI Accelerator

Energy-Efficient Edge AI:
Vision to Generative AI

High Performance Low Power Edge AI Inferencing

SAKURA-II is a high-performance, 60 TOPS, edge AI accelerator architected to run the latest vision and Generative AI models with market-leading energy efficiency and low latency.

EdgeCortix's MERA compiler and software framework provides a robust platform for deploying the latest AI inference models quickly and easily, in an application agnostic manner.

SAKURA-II is available in multiple form-factors enabling flexible system integration, easy evaluation, and fast time-to-market.

Key Benefits

Optimized for Generative AI: Supports multi-billion parameter Generative AI models like Llama 2, Stable Diffusion, DETR, and ViT within a typical power envelope of 8W

Efficient AI Compute: Achieves more than 2x the AI compute utilization of other solutions, resulting in exceptional energy efficiency

Enhanced Memory Bandwidth: Up to 4x more DRAM bandwidth than competing AI accelerators, ensuring superior performance for LLMs and LVMs

Large DRAM Capacity: Support for up to 32GB of DRAM, enabling efficient processing of complex vision and Generative AI workloads

Real-Time Data Streaming: Optimized for low-latency operations with Batch=1

Arbitrary Activation Function Support: Hardware-accelerated approximation provides enhanced adaptability

Advanced Precision: Software-enabled mixed-precision provides near FP32 accuracy

Efficient Data Handling: Integrated tensor reshaper engine minimizes host CPU load

Sparse Computation: Reduces memory footprint and optimizes DRAM bandwidth

Power Management: Advanced power management enables ultra-high efficiency modes

SAKURA-II Offering

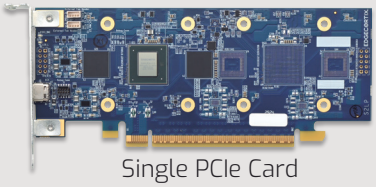
Silicon Device

19 x 19 BGA

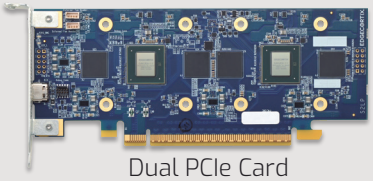
M.2 Module and PCIe Cards

Solutions for quick integration and fast time-to-market

M.2 2280 Key M Module



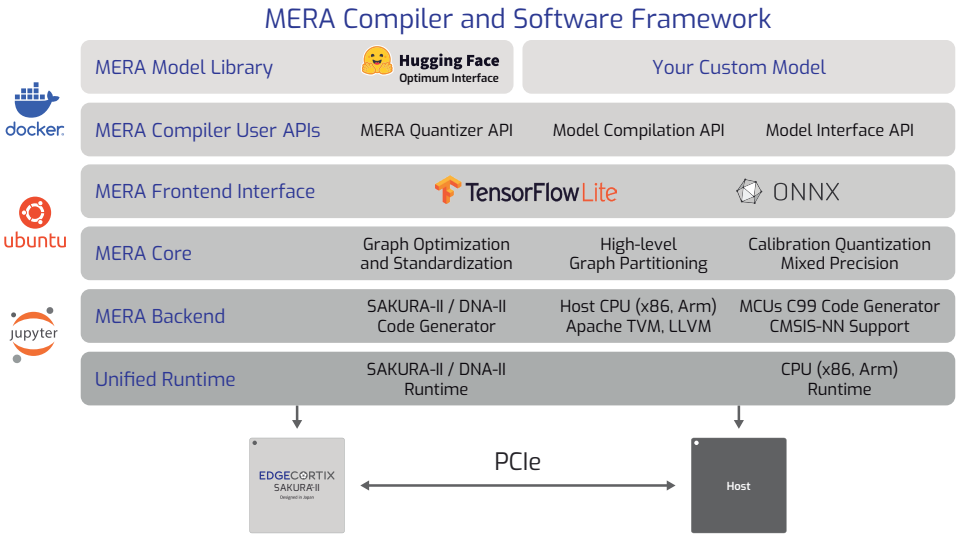
Single PCIe Card



Dual PCIe Card

MERA Fast and Easy Model Porting and System Integration

MERA provides the entire stack for edge AI inferencing from modeling to deployment with familiar neural network model workflows and supports easy integration with existing systems, reducing time-to-market.



MERA Tools

- Source models using Hugging Face, TensorFlow Lite, or ONNX
- Integrate and customize design using Python or C++
- MERA front end is open sourced with support for Apache TVM and MLIR

Model Resources

- **Model Zoo:** Pre-trained, optimized AI inference models
- Support for popular Generative AI models, including Llama-2, Stable Diffusion, Whisper, DETR, DistillBert, DINO and ViT
- Post training model calibration and quantization

Technical Specifications

Performance
60 TOPS (INT8)
30 TFLOPS (BF16)

DRAM Support
Dual 64-bit LPDDR4X
(8/16/32GB total)

DRAM Bandwidth
68 GB/sec

On-chip SRAM
20MB

Compute Efficiency
Up to 90% utilization

Temp Range
-40C to 105C

Power Consumption
8W (typical)

Package
19mm x 19mm BGA

Learn more about SAKURA-II

edgecortix.com/sakura

Key SAKURA-II Market Segments

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