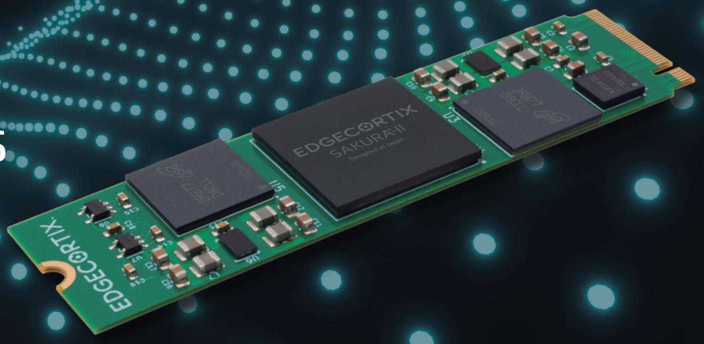
EDGECORTIX®

SAKURA®-II M.2 Modules

Energy-Efficient Edge AI: Vision to Generative AI



High Performance Small Form Factor Edge AI Inferencing

SAKURA-II M.2 modules are high-performance, 60 TOPS, edge AI accelerators 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.

Key Benefits

Small Form Factor: M.2 is the ideal size for space-constrained designs

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 10W

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

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

Large DRAM Capacity: Up to 16GB 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

Technical Specifications

Performance
60 TOPS (INT8)
30 TFLOPS (BF16)

Interface
PCI Gen 3.0 x4

DRAM Support
68 GB/sec

Power Consumption
10W (typical)

Module Height
D6 (3.2mm top, 1.5mm bottom)

Temp Range
-20C to 85C

Form Factor
M.2 2280 Key M

Onboard DRAM
8GB (2 banks of 4GB LPDDR4X) or 16GB (2 banks of 8GB LPDDR4X)

Fast and Easy Model Porting and System Integration

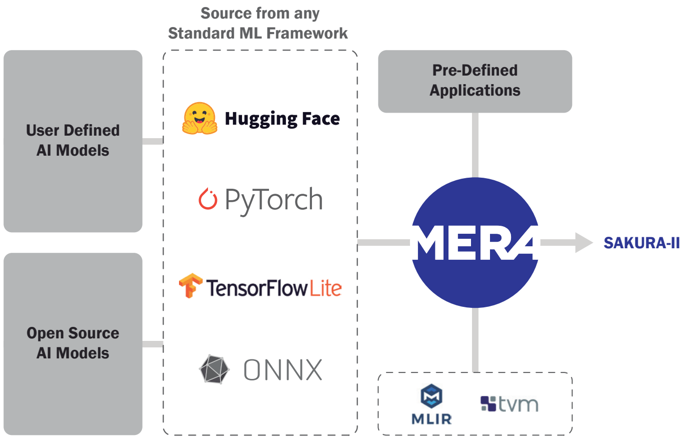
MERA provides the entire stack for edge AI inferencing from modeling to deployment with familiar neural network model workflows and supports easy integration with existing systems, reducing time-to-market.

MERA Tools

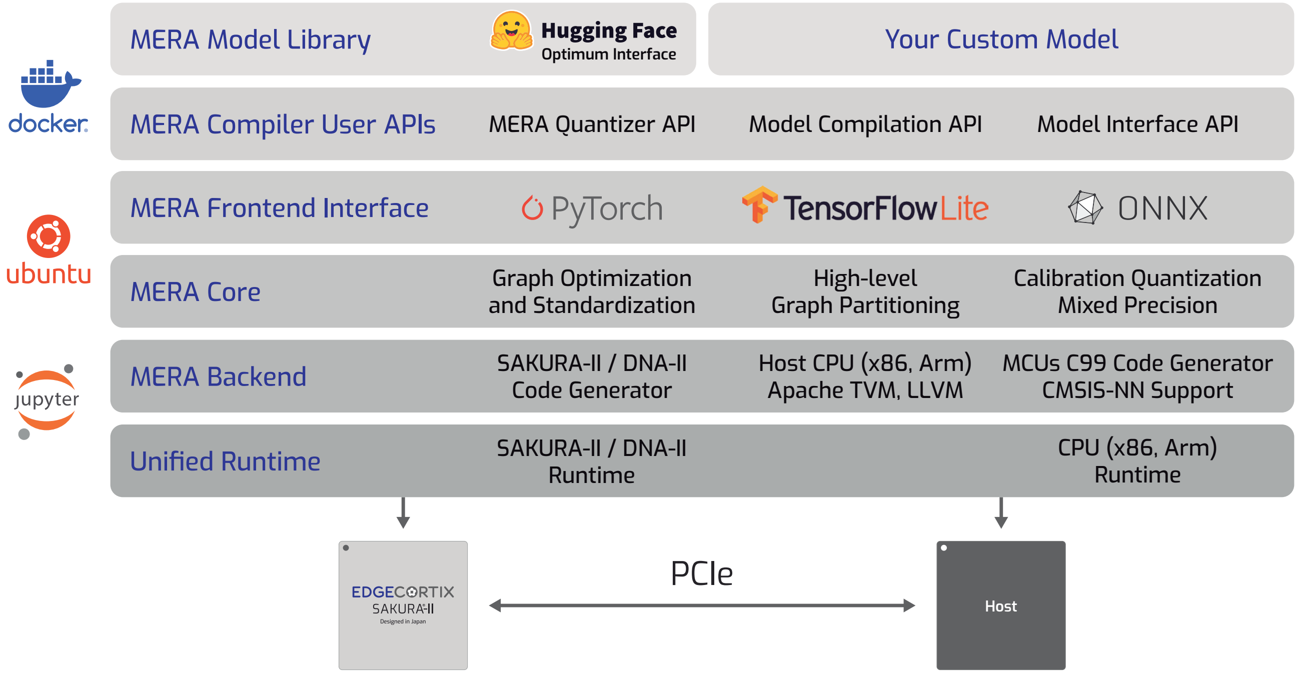
- Source models using Hugging Face, PyTorch, TensorFlow Lite, or ONNX
- Integrate and customize design using Python or C++
- MERA front end is open sourced with support for Apache TVM and MLIR

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

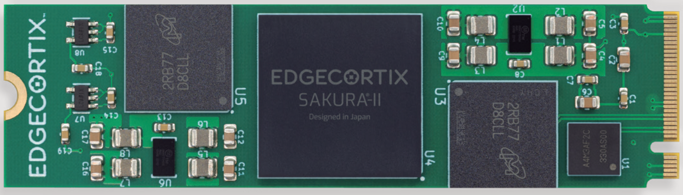


MERA Compiler and Software Framework


docker
ubuntu
jupyter
MERA Model Library
Hugging Face Optimum Interface
Your Custom Model
MERA Compiler User APIs
MERA Quantizer API
Model Compilation API
Model Interface API
MERA Frontend Interface
PyTorch
TensorFlowLite
ONNX
MERA Core
Graph Optimization and Standardization
High-level Graph Partitioning
Calibration Quantization Mixed Precision
MERA Backend
SAKURA-II / DNA-II Code Generator
Host CPU (x86, Arm) Apache TVM, LLVM
MCUs C99 Code Generator CMSIS-NN Support
Unified Runtime
SAKURA-II / DNA-II Runtime
CPU (x86, Arm) Runtime
EDGECORTIX SAKURA-II Designed in Japan
PCIe
Host

Pre-Order an M.2 Module and Get Started!

edgecortix.com/en/pre-order-sakura



© EdgeCortix 2024 All Rights Reserved | EdgeCortix, Dynamic Neural Accelerator, and SAKURA are registered trademarks of EdgeCortix, Inc. All other products are the trademarks or registered trademarks of their respective holders. | **July 2024: Letter**