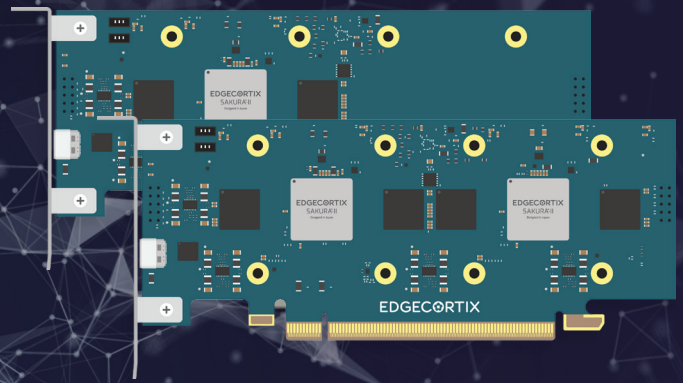




# EDGE CORTIX®

## SAKURA®-II PCIe Cards

*Energy-Efficient Edge AI:  
Vision to Generative AI*



## High Performance PCIe Form Factor for Edge AI Inferencing

SAKURA-II PCIe Cards are high-performance, up to 120 TOPS, edge AI accelerator solutions architected to run the latest vision and Generative AI models with market-leading energy efficiency and low latency.

EdgeCortex'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

**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 or 20W

**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:** 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

**Low Profile PCIe Cards:** Best choice high performance edge AI servers and appliances

## Technical Specifications

### Form Factor

Low profile, single slot PCIe (x16)

### Temp Range

-20C to 85C

### DRAM Bandwidth

68 GB/sec

### SINGLE SAKURA-II

#### Performance

60 TOPS (INT8)  
30 TFLOPS (BF16)

#### Power

10W (typical)

#### Interface

PCI Gen 3.0 x8

#### Onboard DRAM

16GB (2 banks of 8GB LPDDR4)

### DUAL SAKURA-II

#### Performance<sup>1</sup>

120 TOPS (INT8)  
60 TFLOPS (BF16)

#### Power

20W (typical)

#### Interface

PCI Gen 3.0 x8/x8  
(bifurcated)

#### Onboard DRAM

32GB (4 banks of 8GB LPDDR4)



# Fast and Easy Model Porting and System Integration

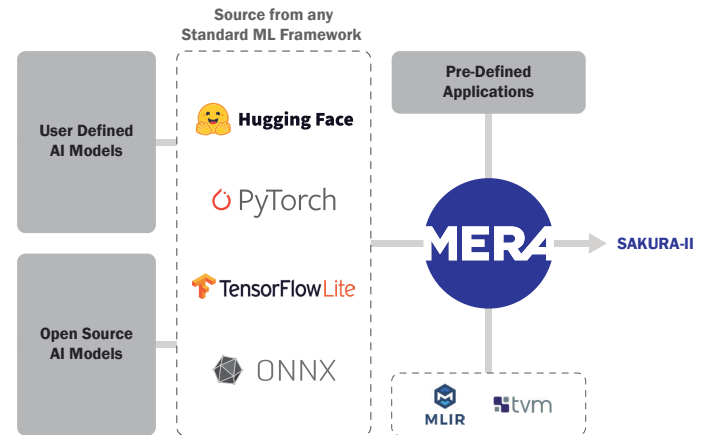
MERA provides the entire stack for edge AI inferencing from modeling to deployment with familiar neural network model workflows and supports easy integration with existing systems, reducing time-to-market.

## MERA Tools

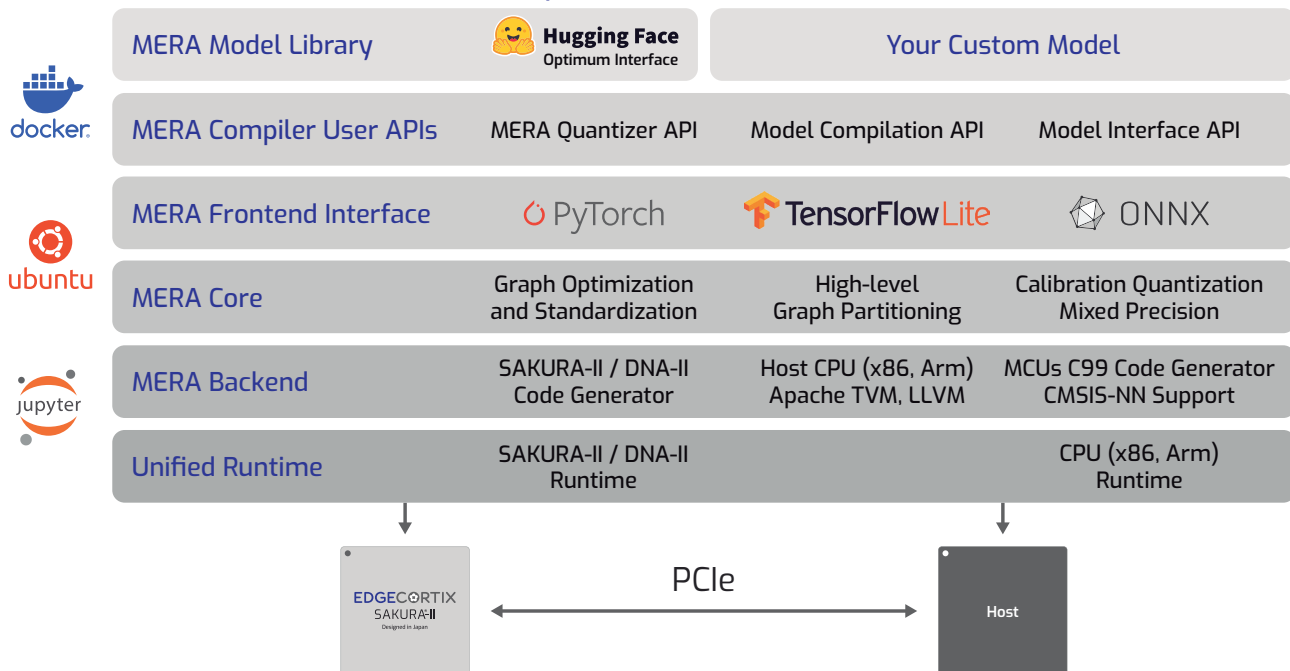
- Source models using Hugging Face, PyTorch, TensorFlow Lite, or ONNX
- Integrate and customize design using Python or C++
- MERA front end is open sourced with support for Apache TVM and MLIR

## 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

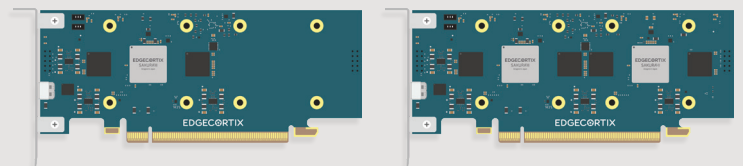


## MERA Compiler and Software Framework



Pre-Order a PCIe Card and Get Started!

[edgecortex.com/en/pre-order-sakura](https://edgecortex.com/en/pre-order-sakura)



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

