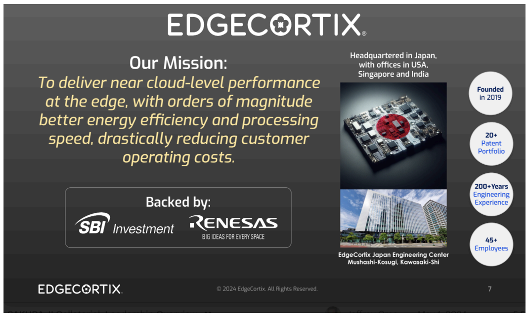
# MEDIA COVERAGE REPORT: EE Times Japan AI EXPO

EE Times Japan:
**EdgeCortix Announces SAKURA-II: A Japanese-developed accelerator that can be used for generative AI on embedded devices**
6/14/2024

EdgeCortix, a Japanese startup focused on edge AI accelerators, introduced the SAKURA-II platform, highlighted by its capability to implement both convolutional and transformer models effectively. Showcased at the 8th AI Expo Spring in Tokyo Big Sight from May 22-24, 2024, the SAKURA-II generative AI accelerator is tailored for embedded devices, supporting various AI models from vision to large language models (LLMs) with high power efficiency. EdgeCortix aims to advance the adoption of generative AI in embedded systems.



EDGECORTIX

Our Mission:
*To deliver near cloud-level performance at the edge, with orders of magnitude better energy efficiency and processing speed, drastically reducing customer operating costs.*

Backed by:
SBI Investment
RENESAS
BIG IDEAS FOR EVERY SPACE

Headquartered in Japan, with offices in USA, Singapore and India

EdgeCortix Japan Engineering Center
Mushashi-Kosugi, Kawasaki-Shi

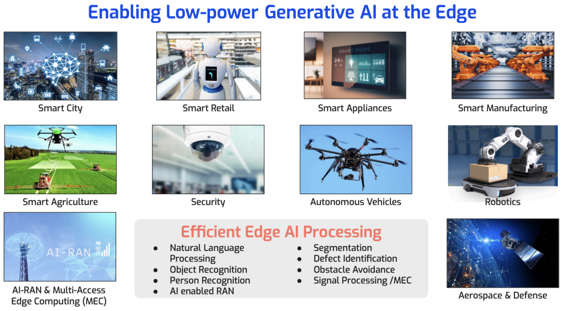
Founded in 2019

20+ Patent Portfolio

200+Years Engineering Experience

45+ Employees

EDGECORTIX © 2024 EdgeCortix. All Rights Reserved. 7



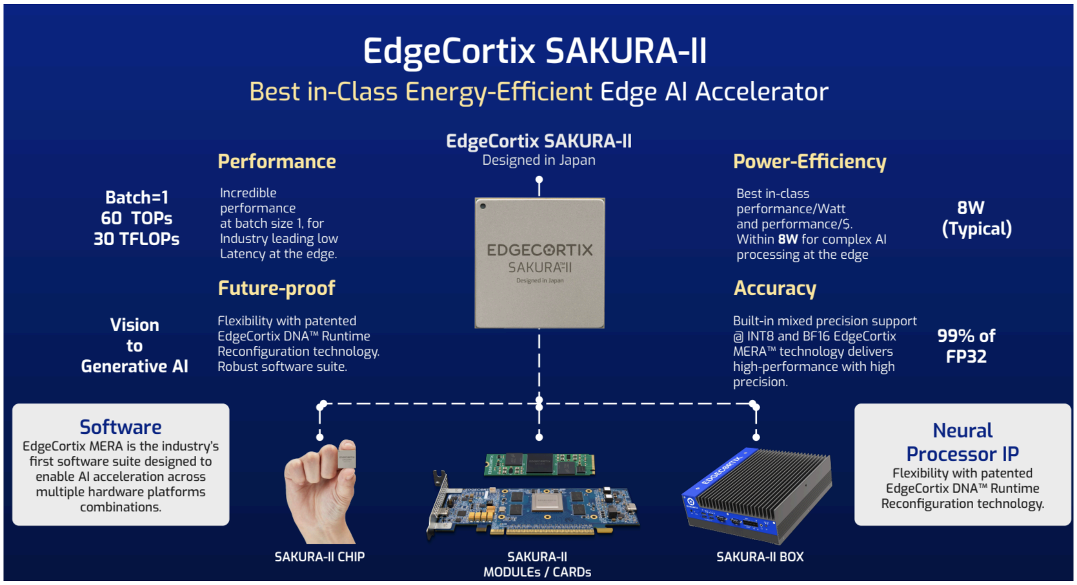
Enabling Low-power Generative AI at the Edge

Smart City
Smart Retail
Smart Appliances
Smart Manufacturing
Smart Agriculture
Security
Autonomous Vehicles
Robotics
AI-RAN & Multi-Access Edge Computing (MEC)
Aerospace & Defense

Efficient Edge AI Processing

- Natural Language Processing
- Object Recognition
- Person Recognition
- AI enabled RAN
- Segmentation
- Defect Identification
- Obstacle Avoidance
- Signal Processing /MEC

EdgeCortix, founded in 2019 and based in Japan, provides a comprehensive offering of IP, hardware, and software solutions. In April 2023, EdgeCortix launched its first AI accelerator, SAKURA-I, followed by the newly developed successor, SAKURA-II. The new SAKURA-II platform includes: DNA (Dynamic Neural Accelerator) IP, featuring a dynamically reconfigurable processor architecture, the SAKURA-II AI accelerator integrated with DNA IP, the MERA compiler and software framework, and M.2 modules and PCIe cards equipped with these AI accelerators.



EdgeCortix SAKURA-II

Best in-Class Energy-Efficient Edge AI Accelerator

EdgeCortix SAKURA-II
Designed in Japan

Batch=1
60 TOPs
30 TFLOPs

Performance
Incredible performance at batch size 1, for Industry leading low Latency at the edge.

Power-Efficiency
Best in-class performance/Watt and performance/$. Within **8W** for complex AI processing at the edge

8W (Typical)

Vision to Generative AI

Future-proof
Flexibility with patented EdgeCortix DNA™ Runtime Reconfiguration technology. Robust software suite.

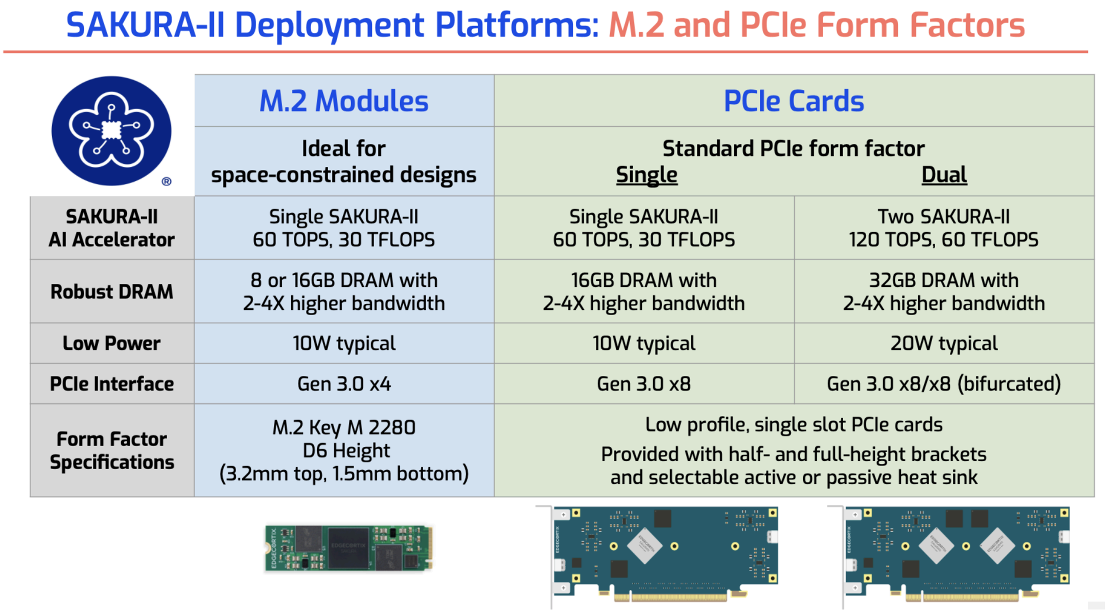
Accuracy
Built-in mixed precision support @ INT8 and BF16 EdgeCortix MERA™ technology delivers high-performance with high precision.

99% of FP32

Software
EdgeCortix MERA is the industry's first software suite designed to enable AI acceleration across multiple hardware platforms combinations.

Neural Processor IP
Flexibility with patented EdgeCortix DNA™ Runtime Reconfiguration technology.

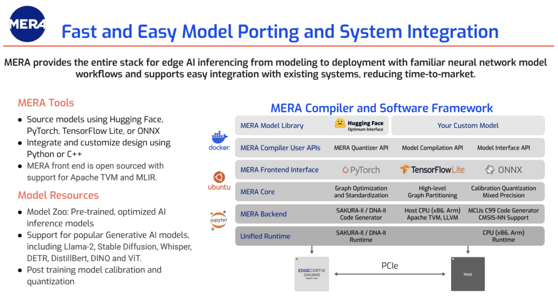
Tim Vehling, Executive Vice President of Global Sales at EdgeCortix, notes a shift in edge AI towards transformer-based inference models, moving beyond CNNs like ResNet and YOLO to meet growing demands for generative AI. EdgeCortix developed SAKURA-II in response, available in M.2 and PCIe form factors, achieving up to 60 TOPS with INT8 and 30 TFLOPS with BF16 while maintaining low power consumption (8W for the AI accelerator chip, 10W for the board configuration). This supports higher accuracy and efficiency, crucial for edge AI applications. SAKURA-II interfaces with DRAM at a peak bandwidth of 68 GB/s, significantly enhancing performance for edge AI tasks, especially with increasingly complex AI models. EdgeCortix's MERA software framework supports a wide range of models, including large language models (LLMs) and new transformers, tailored for optimal edge AI efficiency.



## SAKURA-II Deployment Platforms: M.2 and PCIe Form Factors

| | M.2 Modules | PCIe Cards | |
|---|---|---|---|
| | Ideal for space-constrained designs | Standard PCIe form factor<br>**Single** | **Dual** |
| **SAKURA-II AI Accelerator** | Single SAKURA-II<br>60 TOPS, 30 TFLOPS | Single SAKURA-II<br>60 TOPS, 30 TFLOPS | Two SAKURA-II<br>120 TOPS, 60 TFLOPS |
| **Robust DRAM** | 8 or 16GB DRAM with 2-4X higher bandwidth | 16GB DRAM with 2-4X higher bandwidth | 32GB DRAM with 2-4X higher bandwidth |
| **Low Power** | 10W typical | 10W typical | 20W typical |
| **PCIe Interface** | Gen 3.0 x4 | Gen 3.0 x8 | Gen 3.0 x8/x8 (bifurcated) |
| **Form Factor Specifications** | M.2 Key M 2280<br>D6 Height<br>(3.2mm top, 1.5mm bottom) | Low profile, single slot PCIe cards<br>Provided with half- and full-height brackets and selectable active or passive heat sink | |

At the exhibition, EdgeCortix demonstrated the capabilities of SAKURA-I with several impactful demos. Firstly, using the YOLO v5s model, a single SAKURA-I processed object detection from video feeds of 16 cameras at 500 frames per second (fps), consuming 8.2W. Secondly, SAKURA-I employed the ABPN model to enhance video resolution from 360x640 to 1080x1920 in real-time, achieving a threefold increase in clarity. Lastly, SAKURA-I ran multiple AI models simultaneously, including YOLO v5m for object detection and mono Depth for depth estimation from RGB and thermal cameras, showcasing its multi-model processing capability efficiently managed by 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, 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.

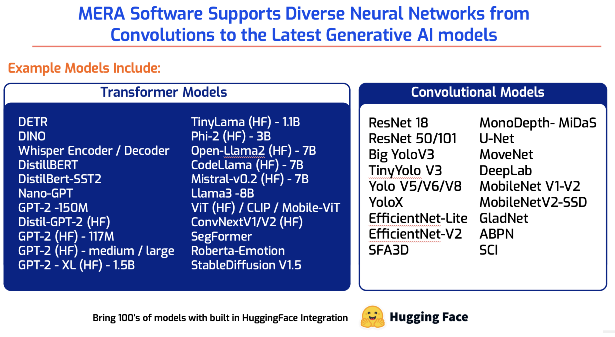
Model Resources

- 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

| | | | |
|---|---|---|---|
| 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 |

docker
ubuntu
jupyter
EDGECORTIX SAKURA-II
PCIe
Host



## MERA Software Supports Diverse Neural Networks from Convolutions to the Latest Generative AI models

Example Models Include:

Transformer Models

DETR
DINO
Whisper Encoder / Decoder
DistillBERT
DistilBert-SST2
Nano-GPT
GPT-2 -150M
Distil-GPT-2 (HF)
GPT-2 (HF) - 117M
GPT-2 (HF) - medium / large
GPT-2 - XL (HF) - 1.5B
TinyLama (HF) - 1.1B
Phi-2 (HF) - 3B
Open-Llama2 (HF) - 7B
CodeLlama (HF) - 7B
Mistral-v0.2 (HF) - 7B
Llama3 -8B
ViT (HF) / CLIP / Mobile-ViT
ConvNextV1/V2 (HF)
SegFormer
Roberta-Emotion
StableDiffusion V1.5

Convolutional Models

ResNet 18
ResNet 50/101
Big YoloV3
TinyYolo V3
Yolo V5/V6/V8
YoloX
EfficientNet-Lite
EfficientNet-V2
SFA3D
MonoDepth- MiDaS
U-Net
MoveNet
DeepLab
MobileNet V1-V2
MobileNetV2-SSD
GladNet
ABPN
SCI

Bring 100's of models with built in HuggingFace Integration

Summary prepared by EdgeCortix.

- Full Original Japanese Article: https://eetimes.itmedia.co.jp/ee/articles/2406/14/news021.html