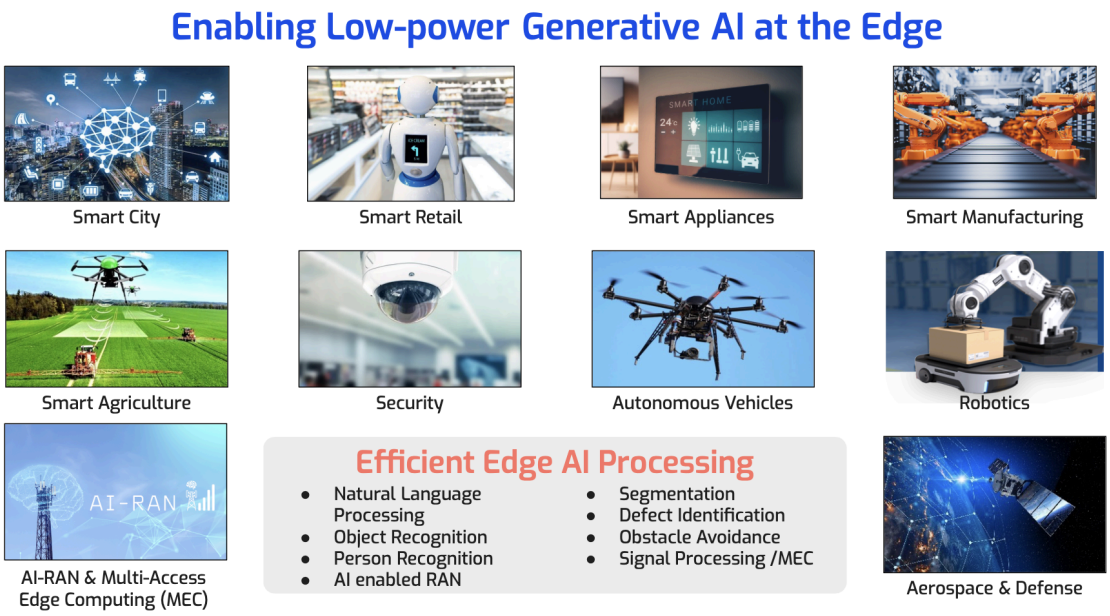
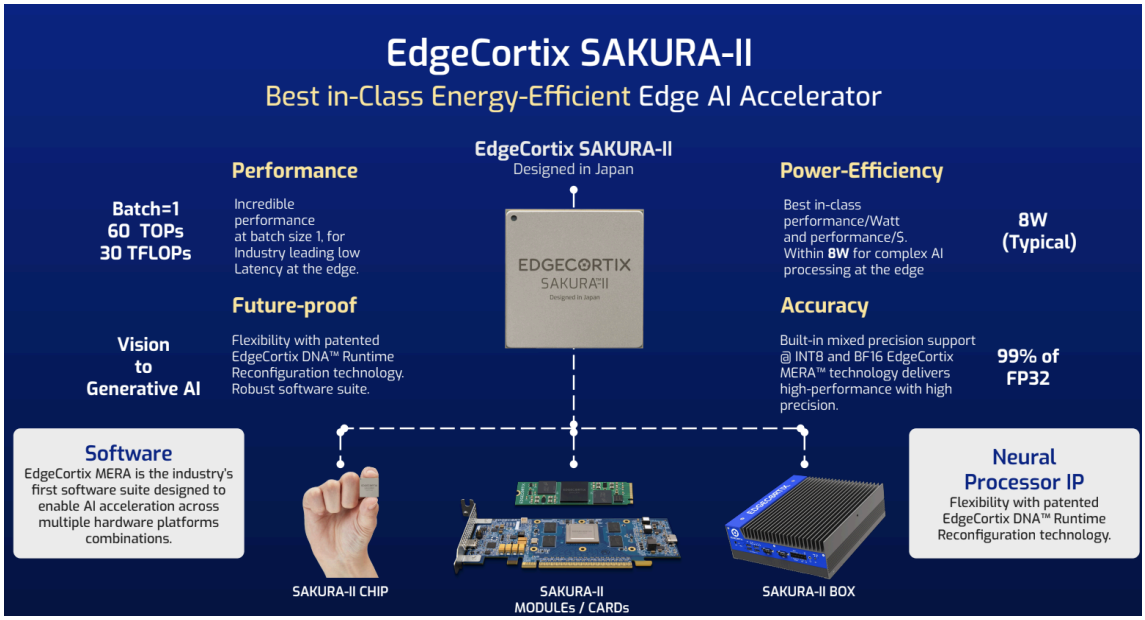
MEDIA COVERAGE REPORT: MONOist AI EXPO

MONOist Embedded Development Interview:
Japanese startup takes on "generative AI at the Edge", achieving 60 TOPS AI processing performance at 8W
6/19/2024

As generative AI gains attention, there's a growing need to run it on edge devices. EdgeCortix, a Japanese startup founded in 2019 by Sakyasingha Dasgupta, addresses this need with its SAKURA-II AI accelerator, offering 60 TOPS of AI processing performance and 8W power consumption. Generative AI models are larger and more complex than traditional machine learning models, making it challenging to ensure sufficient processing performance on edge devices. EdgeCortix, headquartered in Tokyo with an R&D center in Kawasaki, is a major player in the edge AI accelerator market.



EdgeCortix specializes in developing efficient AI accelerators and software for edge devices. In April 2023, they launched their first AI accelerator, SAKURA-I, with 40 TOPS of AI processing performance. However, the rise of generative AI like ChatGPT highlighted the need for more powerful AI accelerators. In response, EdgeCortix released SAKURA-II on May 22, 2024, with 60 TOPS performance and 8W power consumption, optimized for generative AI models such as Llama2 and Stable Diffusion. Tim Vehling, EVP of Global Sales, noted the importance of supporting mixed-precision models and high memory bandwidth to meet the demands of edge generative AI.



EdgeCortix's core technology includes the DNA (Dynamic Neural Accelerator) architecture, which reconfigures compute engine connections at runtime, and MERA, a compiler supporting various AI frameworks. SAKURA-I and SAKURA-II AI accelerators align with this platform. MERA supports DNA and major processor architectures like Arm, Intel, AMD, and RISC-V, enabling heterogeneous processing. Renesas Electronics has invested in EdgeCortix, utilizing MERA in its AI products.

Critical Edge AI Accelerator Requirements

Size: Space-constrained design option (M.2 Module)

Power: Low power, high efficiency (8W typical)

Cost: Effectively priced, high performance/cost ratio

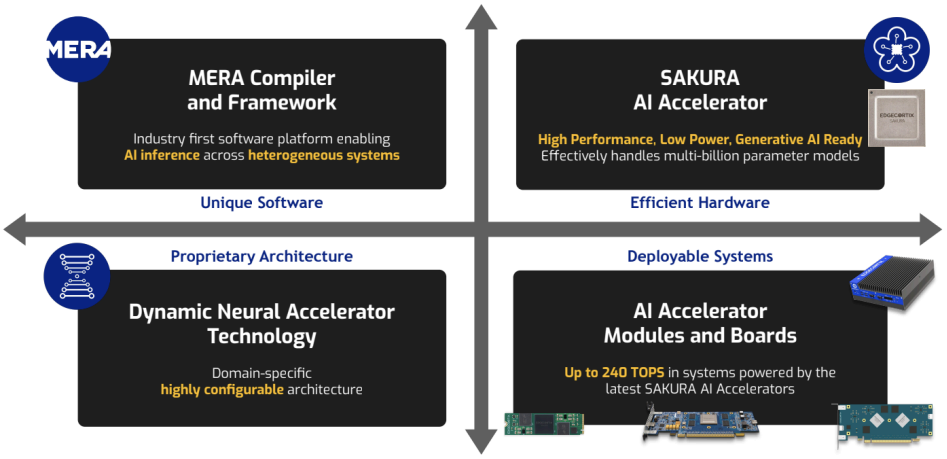
Memory Capacity: Low-power DRAM at high bandwidth critical for Generative AI designs.

Performance:

- **Large Models:** Ability to efficiently run LLMs (Large Language Models) and large visual perception models.
- **Hardware Designed for Generative AI:** Ability to run multimodal models with built-in mixed precision support to balance accuracy and speed
- **Effective Software:** Powerful Compiler and Software Framework, enabling models to be quickly deployed across heterogeneous systems

The second-generation DNA and MERA have been optimized for SAKURA-II, manufactured using TSMC's 12nm process. Memory bandwidth is increased to 68GB/s, making SAKURA-II capable of handling generative AI models. MERA now integrates with HuggingFace for transformer models.

Software Driven Unified Platform Delivering Highest Efficiency



Combining the AI Accelerator with Flexible Software to Deploy Power Efficient Solutions

Fast and Easy Model Porting and System Integration

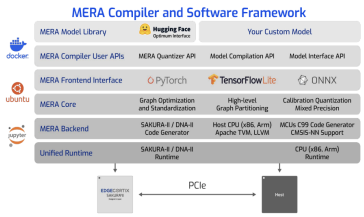
MERA provides the entire stack for edge AI inferencing from modeling to deployment with familiar neural network model workflows and supports easy integration with existing systems, reducing time-to-market.

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

Example Models Include:

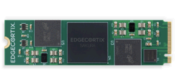
Transformer Models		Convolutional Models	
DETR	TinyLama (HF) - 1.1B	ResNet 18	MonoDepth- MiDaS
DINO	Phi-2 (HF) - 3B	ResNet 50/101	U-Net
Whisper Encoder / Decoder	Open-Llama2 (HF) - 7B	Big YoloV3	MoveNet
DistillBERT	CodeLlama (HF) - 7B	TinyYolo V3	DeepLab
DistillBert-SST2	Mistral-v0.2 (HF) - 7B	Yolo V5/V6/V8	MobileNet V1-V2
Nano-GPT	Llama3 -8B	YoloX	MobileNetV2-SSD
GPT-2 - 150M	ViT (HF) / CLIP / Mobile-ViT	EfficientNet-Lite	GladNet
Distil-GPT-2 (HF)	ConvNextV1/V2 (HF)	EfficientNet-V2	ABPN
GPT-2 (HF) - 117M	SegFormer	SFA3D	SCI
GPT-2 (HF) - medium / large	Roberta-Emotion		
GPT-2 - XL (HF) - 1.5B	StableDiffusion V1.5		

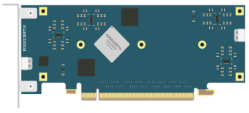
Bring 100's of models with built in HuggingFace Integration — Hugging Face

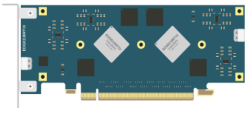
SAKURA-II will be available in various configurations in the second half of 2024, starting at $249 for the 8GB M.2 module. Higher demand configurations include a potential four-chip PCIe card with 240 TOPS performance and 272 GB/s memory bandwidth.

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

	M.2 Modules	PCIe Cards	
	Ideal for space-constrained designs	Standard PCIe form factor Single	Dual
SAKURA-II AI Accelerator	Single SAKURA-II 60 TOPS, 30 TFLOPS	Single SAKURA-II 60 TOPS, 30 TFLOPS	Two SAKURA-II 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 D6 Height (3.2mm top, 1.5mm bottom)	Low profile, single slot PCIe cards Provided with half- and full-height brackets and selectable active or passive heat sink	







Summary prepared by EdgeCortix.

• Full Original Japanese Article: https://monoist.itmedia.co.jp/mn/articles/2406/19/news075.html