EE Times Japan

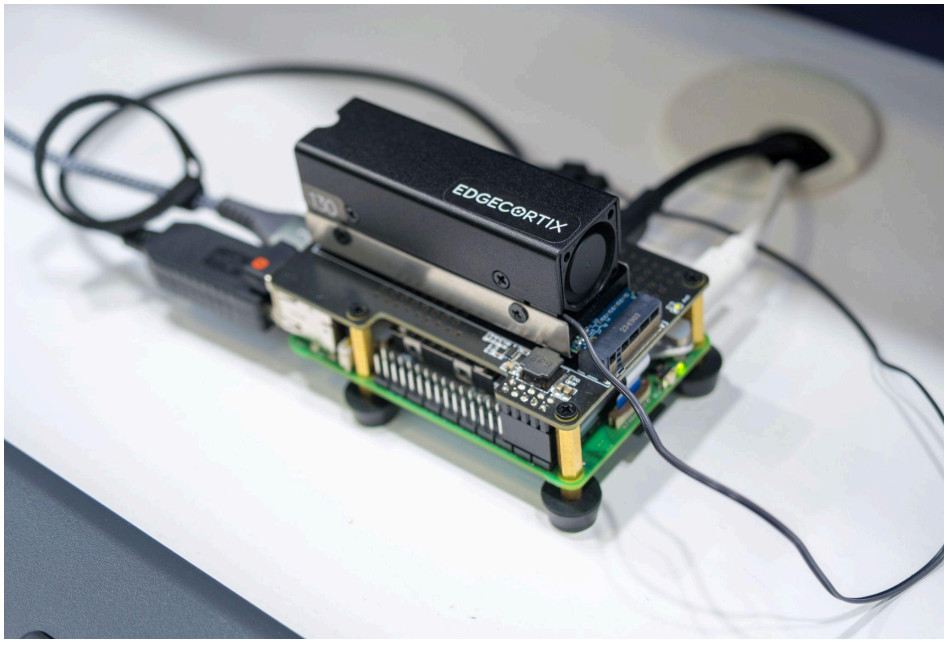
EdgeCortix Showcases a Generative AI Model Running on a Raspberry Pi

Published on April 18, 2025, at 11:00 AM, by Mayuko Murao

At the 9th AI Expo TOKYO Spring, EdgeCortix showcased a demo combining a Raspberry Pi with its AI accelerator "SAKURA-II" to run Transformer models. The company also presented a demonstration running three Generative AI models on a single card.

At the 9th AI Expo Tokyo Spring (April 15–17, 2025, at Tokyo Big Sight), EdgeCortix showcased a demo combining its AI accelerator with a Raspberry Pi to run Transformer models.

Specifically, the company combined EdgeCortix's AI accelerator "SAKURA-II" M.2 module with a Raspberry Pi 5 to run a segmentation model called "SegFormer" based on Transformer architecture. They say the accuracy is BF16 and the AI performance is 60 TOPS/10W. Example business applications include drones, robotics, and embedded systems, where low power consumption is essential.



The actual demo unit combines the Raspberry Pi 5 and SAKURA-II. The Raspberry Pi and SAKURA-II are connected via an M.2 conversion module. A heatsink and fan are mounted on top.

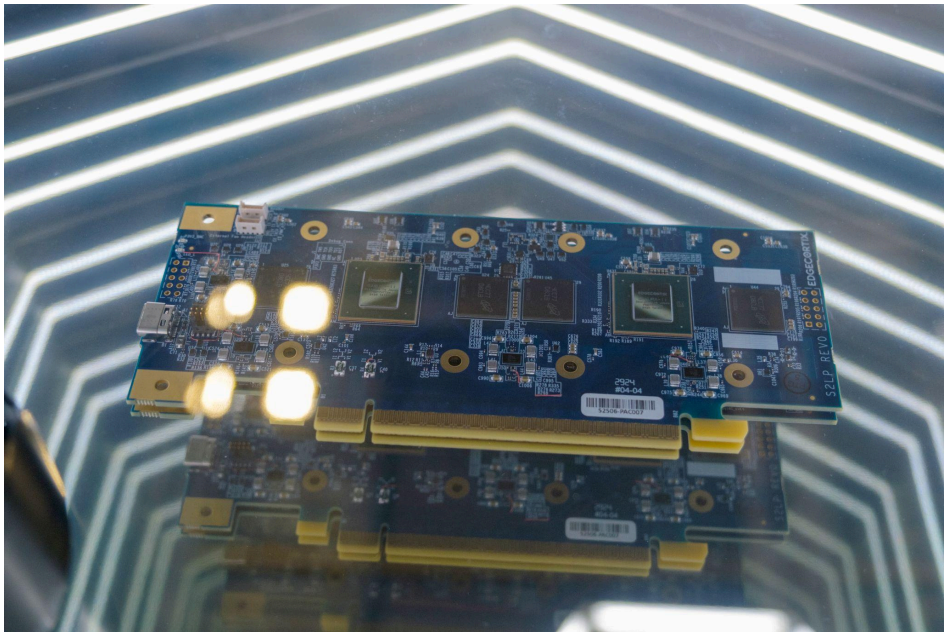
A demo running the segmentation model "SegFormer" using the Raspberry Pi and SAKURA-II.

Since Raspberry Pi's are widely used for PoC (Proof of Concept), the benefit of combining it with the SAKURA-II is that it enables low-cost PoC for edge AI applications, according to EdgeCortix. This demo was previously showcased at the embedded systems exhibition "embedded world 2025" held in Nuremberg, Germany, in March 2025. According to Tim Vehling, Executive Vice President of Global Sales at EdgeCortix, **"The Raspberry Pi is widely used in Europe, so the demo attracted a lot of attention."** This is the first time the demo has been presented at an exhibition in Japan.

Raspberry Pi is increasingly being used not only for proof-of-concept (PoC) projects but also in mass-produced industrial equipment. Mr. Vehling explained that the solution combining a Raspberry Pi and SAKURA-II can, of course, be used in mass production as well. Currently, SAKURA-II is being shipped as samples, with mass production scheduled to begin in the second half of 2025.

Running Three AI Models on a Single Card

A multimodal demo was also showcased, running three Generative AI models — a vision model, a language model, and a segmentation model — using the SAKURA-II. Two SAKURA-II chips were used to demonstrate segmentation with SegFormer, as well as image-to-text generation using the vision model ViT and the language model GPT-2. This demo is intended for high-end applications such as urban management systems and surveillance cameras.



The board used in the demo featuring two SAKURA-IIs.

A view of the demo in action. The PC used in the demo is ASRock's "DeskMeet X600," equipped with an AMD processor, the "Ryzen 7 7700."

Translation prepared by EdgeCortix

- EE Times Japan (2025/04/18). Full original Japanese article:
- https://eetimes.itmedia.co.jp/ee/articles/2504/18/news122.html
-