電波新聞 DEMPA DIGITAL

Toward Cloudless Computing with ‘Edge AI’, Developing Next-Generation Chips — EdgeCortix

Written and compiled by Semiconductor Navi

While generative AI improves operational efficiency, it also poses challenges such as processing delays due to cloud connectivity and increased power consumption at data centers. Amid these challenges, emerging fabless semiconductor company EdgeCortix (Tokyo, Chuo-ku) is gaining attention with its edge AI technology that performs AI processing directly on devices.

The company is developing AI accelerators specialized for edge computing. It has created a proprietary architecture called "DNA" for its NPU (Neural Processing Unit), a processor designed specifically for AI workloads, and is rolling out AI-enabled devices such as “SAKURA”.

Demonstration of image segmentation with Generative AI on a Raspberry Pi

The latest solution, "SAKURA-II", delivers up to 60 TOPS (trillion operations per second) of processing power while maintaining low power consumption under 8 watts. With enhanced memory bandwidth, it efficiently handles complex computations. Going forward, the company plans to offer modules compatible with storage interface standards such as M.2 and PCIe, with mass production scheduled for the latter half of 2025.

An accelerator PCIe card implementing 'SAKURA-II'

In November last year, the company was selected for the Next-Generation Communication Project by the New Energy and Industrial Technology Development Organization (NEDO) and received a subsidy of 4 billion yen to begin developing energy-efficient AI accelerators. It is currently advancing the development of "SAKURA-X," a communication solution compatible with Beyond 5G/6G.

The SAKURA-X utilizes "chiplet" technology, which integrates multiple chips into a single package, allowing for a four-chip configuration with processing power of up to 1200 TOPS. The first phase, running through 2026, will focus on developing the initial chips, followed by a second phase from 2026 onward to build a platform that integrates generative AI with RAN (Radio Access Network).

Mr. Vehling, Executive VP of Global Sales, talking about 'SAKURA-X'

According to Tim Vehling, Executive VP of Global Sales, *"By capitalizing on our expertise in developing devices that combine high performance with low power consumption, we aim to deliver the best solutions for emerging areas like AI-RAN."*

Translation prepared by EdgeCortix

- 電波新聞デジタル / DEMPA Digital
- Publication date: May 19, 2025
- Full original Japanese article: https://dempa-digital.com/article/660669
-