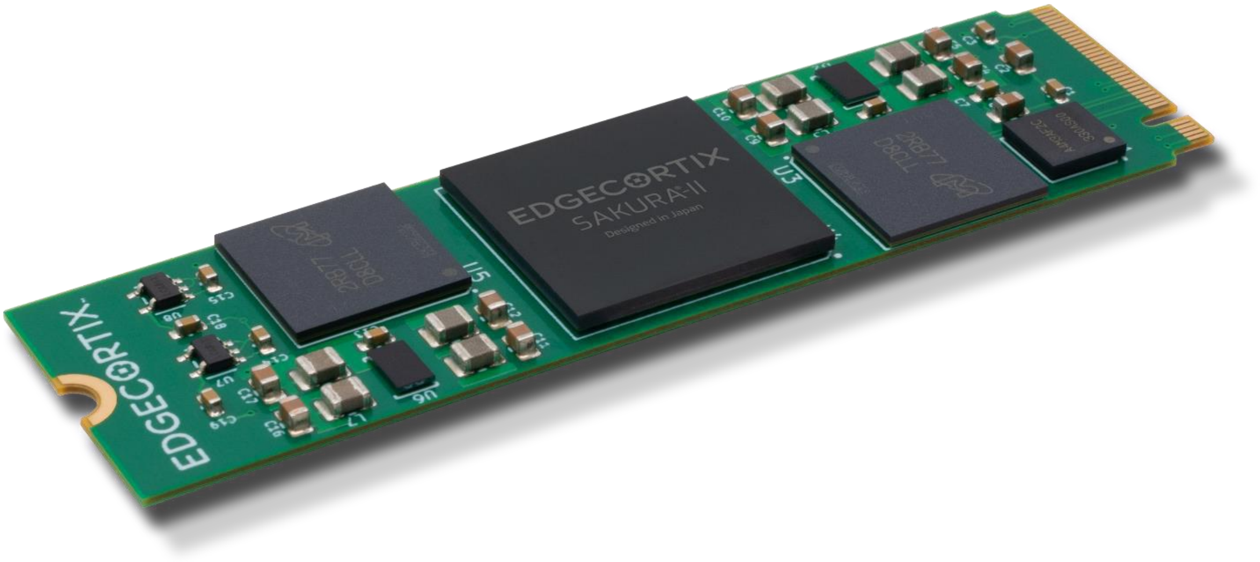
MEDIA COVERAGE REPORT: Nikkei XTech

Nikkei XTech:
EdgeCortix develops semiconductor for edge inference, delivering 60 TOPS at 8W
7/03/2024
By Ryunosuke Kubota

EdgeCortix (Chuo-ku, Tokyo), a semiconductor startup, has developed a new AI (artificial intelligence) accelerator capable of 60 TOPS (60 trillion operations per second) with as little as 8W of power. Compared to NVIDIA's GPU (image processing semiconductor), it is specialized for edge inference and is characterized by its low cost and high power efficiency. They will sell their products starting at US$249 (about 40,000 yen at US$1 = 157 yen) and ship by the end of 2024.



EdgeCortix announced it on May 22, 2024. Compared to its predecessor, it has about 20 TOPS more computing power at about the same power.

One reason for the high-power efficiency is the system's ability to reconfigure the datapath for each constant calculation. It is said that it can efficiently compute and process large language models (LLMs) with high sparsity. Similar methods are used by competitors for learning and inference, and compared to these competitors, **"we are unique in that we specialize in inference,"** says Tim Vehling, EVP of Global Sales for EdgeCortix.

The previous generation product was optimized for 8-bit integer arithmetic (INT8), but the new product supports 16-bit floating-point number arithmetic (BF16). The system can process INT8 and BF16 while switching between them, which should improve accuracy while shortening the arithmetic processing time. For example, we could switch to the highly accurate BF16 only for facial recognition on surveillance cameras.

In addition to LLM, the company is also targeting image recognition for smart cities and factory automation (FA). They support AI models that employ network architectures such as "Transformer" and "Convolutional Neural Network (CNN)". **"In image recognition applications, CNN has been superior or equal to Transformer in the past, but the situation is reversing around 2023,"** Vehling said.

Like its predecessor, the new product uses Taiwan Semiconductor Manufacturing Company's (TSMC) 12nm generation manufacturing process. **"After the TSMC subsidiary, JASM (Japan Advanced Semiconductor Manufacturing) Kumamoto plant starts mass production, we plan to use their semiconductors,"** said Vehling.

They are equipped with either 8 or 16 Gbytes of DRAM. The new products are available in two types: a module type and a PCI Express (PCIe) type.

EdgeCortix is a startup founded in Japan in 2019 by Sakya Dasgupta, originally an engineer at IBM and Microsoft (USA). Renesas Electronics and other companies are investing in the company. As for the reason for establishing the company in Japan, the company states that **"Japan lacks startups in semiconductors and other areas but has government support (for the edge AI field) and geopolitical advantages."**

- Translation prepared by EdgeCortix.
- Nikkei Crosstec by Ryunosuke Kubota - 2024.07.03 - full original Japanese article: **https://xtech.nikkei.com/atcl/nxt/news/24/01002/**
-
-
-
-