AI Semiconductor Startup Challenging NVIDIA Sets Sights on 'Next-Generation Communications and Space' with Edge Inference

Ryunosuke Kubota＝Nikkei xTech

EdgeCortix CEO Sakyasingha Dasgupta, PhD. says that the company will accelerate the development of AI semiconductors for telecommunications.

Chuo, TOKYO — EdgeCortix Inc., an AI semiconductor startup is embarking on the development of next-generation telecommunications and space sectors. They are offering semiconductors capable of processing generative AI on the edge (device side) with lower power consumption compared to GPUs (graphics processing semiconductors), and have also received support from the Ministry of Economy, Trade, and Industry. They plan to develop new products for next-generation telecommunications by 2026 and are negotiating for adoption in the space sector, including with lunar transport services like ispace.

EdgeCortix Business Strategy
Exploring the market for edge AI processing in next-generation wireless communications and space exploration vehicles with power-efficient semiconductors.
Developing chiplet-integrated semiconductors for AI-RAN by 2026.
In the telecommunications sector, the Ministry of Economy, Trade, and Industry is providing a 4 billion yen subsidy, while in the space sector, collaboration with ispace is under consideration.

Leveraging the advantage of being able to perform generative AI processing with low power consumption without the need to interact with data centers.

Targeting AI-RAN and Lunar Explorers

In the next-generation communication sector, the company aims to target the AI-RAN (Radio Access Network) market, which uses AI to optimize communication. By 2026, they are planning to develop AI semiconductors for telecom operators that can perform up to 1600 TOPS (1.6 trillion operations per second) of computational processing with a power consumption of less than 50W.

In the space sector, they aim to apply technologies such as image recognition to control the landing positions of spacecraft on the moon and other planets. EdgeCortix CEO Sakyasingha Dasgupta, PhD. says, "We received positive results from the space product evaluation tests conducted with NASA (the National Aeronautics and Space Administration). The results of this test have led to inquiries from multiple customers regarding the company's semiconductors. Discussions for adoption will be ramping up moving forward.

AI inference processing at the edge has primarily been considered for areas such as factory automation (FA), robotics, autonomous driving, and smart cities. In fact, according to Mr. Dasgupta, there has been strong interest in the company's AI semiconductors for robotics in Japan and for smart cities in Europe and the U.S. With significant potential for growth in next-generation communications and space sectors, the company aims to make early moves to establish a presence in these fields.

EdgeCortix History	
2019	Founded in Japan by former employees of Microsoft.
2022	The first-generation semiconductor for inference, "SAKURA-I," has been made available.
2023	Approximately 3 billion yen has been raised from SBI Investment and others.
2024	The second-generation semiconductor for inference, "SAKURA-II," has been made available.
2026 (**Planned**)	With support from the Ministry of Economy, Trade, and Industry (METI), a chiplet-type semiconductor for AI-RAN has been developed.
After 2026	A chiplet-type semiconductor has been demonstrated for next-generation 5G base stations.

Although founded in 2019 as a startup, they have already commercialized two generations of AI semiconductors.

Achieving Power Efficiency through Dynamic Circuit Reconfiguration

The AI semiconductor from EdgeCortix adopts a data flow type architecture, which is commonly referred to as a non-Von Neumann architecture. By reconstructing the data path for each calculation and dynamically rewriting the circuit configuration, it efficiently handles the complex computation flows required by generative AI. This makes it possible to eliminate the need for frequent communication with memory, unlike the von Neumann-type semiconductors such as the GPUs from NVIDIA, thus significantly reducing power consumption associated with data transfer. As a result, it becomes easier to handle inference processing for large language models (LLMs) on the edge, which has traditionally been managed by GPUs in data centers.

AI Semiconductor Operating at 30W for Open RAN

For next-generation communications, the aim is to improve power efficiency by more than five times compared to conventional AI-RAN systems, which are composed of GPUs and other components. AI-RAN is a technology that uses AI to enable coordinated operation among multiple base stations, optimally controlling communication areas. This allows for low-latency, high-capacity communications.

EdgeCortix's AI-RAN technology development was selected for a project by the New Energy and Industrial Technology Development Organization (NEDO) in November 2024. The development period is from 2024 to 2029, with a grant of 4 billion yen from the Ministry of Economy, Trade and Industry.

This project will develop a next-generation AI semiconductor with a chiplet integration approach, called "SAKURA-X". It will integrate CPU (Central Processing Unit), NPU (Neural Processing Unit) and NOC (Network on Chip) optimized for AI processing, and other components into a single package in the form of chiplets (individual semiconductor chips).

It is intended for use in Open RAN, which constructs wireless networks by combining base stations from various vendors based on open specifications. Among the components that make up the system, such as the Radio Unit (RU), Distributed Unit (DU), and Centralized Unit (CU), they will develop semiconductors specifically for the DU.

It is said that the power consumption of GPUs used in AI-RAN has been over 700W. The plan for SAKURA-X is to reduce this to around 30 to 50W.

Aiming for ispace's Lunar Transport Service

In the space sector, the goal is to acquire customers for the flagship AI semiconductor product, called "SAKURA-II." Discussions are ongoing with ispace, which handles lunar transport services, regarding adoption, and several other companies are considering its introduction. Space is an environment where securing power is challenging, making it easier to leverage the power-efficient features of EdgeCortix's AI semiconductors compared to GPUs.

There is another advantage in the space sector: the high radiation resistance of the company's semiconductors. In space, semiconductor circuits exposed to radiation can malfunction due to soft errors, causing data being processed to be lost and preventing correct calculation results from being obtained.

EdgeCortix's AI semiconductors avoid this issue with a mechanism that dynamically reconfigures the circuits. Since information such as the "weights" used for inference is updated with each inference process, even if an error occurs once, it is less likely to affect subsequent calculations.

NASA released the results of radiation resistance tests using EdgeCortix's first-generation AI semiconductor, "SAKURA-I," in March 2024. According to EdgeCortix, a total of three radiation resistance tests were conducted over seven months, and it was demonstrated that the semiconductor exhibited relatively high resistance. The Vice President of Defense & Space Technology, Mr. Stanley Crow said, "It was fortunate that the needs of the space sector aligned with the characteristics of our products."

EdgeCortix was founded in Japan in 2019 by Mr. Dasgupta, who has a background working at Microsoft in the U.S., IBM Japan, and the RIKEN Institute. Mr. Dasgupta emphasized the importance of "government subsidies for the AI semiconductor field," which influenced his decision to start the company in Japan.

SBI Investment, Renesas Electronics, and others are investing in the company, and its employees include engineers from Samsung Electronics in South Korea and NASA. As of December 2024, the company has around 120 customers, nearly half of them are based in Japan.

Translation prepared by EdgeCortix

- Nikkei X-Tech (2025/01/22 05:00). Full original Japanese article: https://bizboard.nikkeibp.co.jp/el_xtech/atcl/nxt/column/18/00001/10131/?ST=p_bizboard&bzb_pt=0&BZB_DATE_TOKEN=916975167593ddf9595b98736245b337ae6c9ab79ae590cfec7dde042720aa9ee62458dd7deefe0cd5bd8bc3bf953669
- Copyrights and other intellectual property rights to articles, photographs, charts, headlines, and other information (hereinafter referred to as “Information”) provided through the Service belongs to the providers of such Information.
- Unauthorized reproduction of Information provided by this service is prohibited.
- The service may not be used by any other third party other than the subscriber, regardless of the method, with or without compensation.
- Copyright © Nikkei Inc. All Rights Reserved.