News Magnifier / PFN Accelerates AI Semiconductor Development, Specializing in Deep Learning

The Daily Industrial News: 9/23/2024

Semiconductors are essential for developing and running AI models, with NVIDIA's GPUs dominating. Now, more companies are creating AI-specific devices to compete and differentiate themselves to capitalize on the expanding AI market. GPUs offer strong computing power, but they aren't optimized for AI-specific tasks. As LLM demands grow, the higher computing power has led to higher power consumption when using GPUs, posing a significant challenge.

Preferred Networks (PFN) in Tokyo is leveraging its software expertise to develop semiconductors optimized for deep learning. By shifting functions like network control to software and enhancing computational capabilities, PFN's devices are more specialized for deep learning, reducing power consumption. PFN will provide servers as a cloud service within 2024, and will run its own AI software along with a new SDK. PFN is also jointly developing 2-nanometer semiconductors (using Samsung Electronics fab which will open in 2025) with Internet Initiative Japan (IIJ) and the Japan Advanced Institute of Science and Technology, and plans to commercialize them after 2026.

EdgeCortix (Chuo-ku, Tokyo), which provides edge AI semiconductors, also focuses on power efficiency, and has developed devices specialized for AI processing, using a software-centered design. In 2024, EdgeCortix will begin mass production of SAKURA-II, which supports multi-billion parameter AI models, and CEO Sakyasingha Dasgupta says, "There is strong demand."

According to JEITA, the global demand for AI will reach 211 billion dollars (~30 trillion yen) in 2030, about 20X of the 2023 demand. As the market expands, there will likely be a movement to find a balance between computing power and power consumption.

Summary prepared by EdgeCortix.

- Full Original Japanese Article: **https://www.nikkan.co.jp/articles/view/00725269**