The End of NVIDIA's Dominance

By Shun Okuhama and Toshiyuki Omori

Currently, GPUs (graphics processing units) are widely used as AI chips for developing generative AI (artificial intelligence). It is unlikely that GPUs will disappear from AI development anytime soon. However, 2025 is shaping up to be the year when the era of GPU dominance comes to an end. This is due to the emergence of new ASICs (application-specific integrated circuits) designed specifically for AI processing. It also signals the end of NVIDIA's overwhelming dominance in the market.

GPUs were originally developed for graphics processing, but today they are used across a wide range of applications, including AI, simulations, and finance. In contrast, AI-specific ASICs are expected to replace many of the roles currently performed by GPUs in the future. This is hinted at by the historical evolution of chips used for cryptocurrency mining.

Mining is the process required to issue new cryptocurrency. 'Bitcoin', the most well-known cryptocurrency, requires massive computational power for mining. In the past, CPUs (central processing units) handled these calculations. However, due to their inefficiency, GPUs, which can execute mining computations at a much higher speed, eventually took over.

The issue with GPUs is their massive power consumption, or rather, their poor energy efficiency. To address this problem, ASICs were introduced, allowing mining to be carried out efficiently with lower power consumption. Today, almost all Bitcoin mining is done using ASICs.

AI chips are expected to follow a similar trajectory. In the past, AI computations were performed using CPUs. However, as GPUs proved to be much faster for AI processing, they replaced CPUs. Yet, power consumption remains a major issue for GPUs in AI applications, making it likely that ASICs will eventually become the mainstream solution.

The Main Applications are Training and Inference

Let's take a look at specific AI chips. First, GPUs are dominated by NVIDIA, which holds an overwhelming market share. Although it doesn't match NVIDIA, AMD's GPUs are steadily gaining market share as well.

On the other hand, various companies are developing ASICs for AI. Major cloud service providers, in particular, are developing AI chips to enable AI processing on their own cloud platforms. A pioneering example is Google's "TPU" (Tensor Processing Unit). Mark Lohmeyer, Vice President and General Manager overseeing AI infrastructure in Google's cloud division,

stated, "We have been developing TPUs for about 10 years, and we are now on the sixth generation."

Meanwhile, the AI chips of Amazon Web Services (AWS) in the U.S. are divided into "Trainium" for training and "Inferentia" for inference. Training refers to the learning process for developing foundation models in generative AI, while inference refers to the use of pre-trained foundation models. Training includes inference as part of the process and generally requires more computation than inference. However, AWS has also introduced the use of Trainium for high-speed inference.

Japanese companies also show their presence. In July 2024, SoftBank Group announced that it had acquired the UK-based AI chip developer Graphcore.

Preferred Networks (PFN) has developed "MN-Core 2" for both training and inference, and began offering it to other companies in 2024. Additionally, the development of "MN-Core L1000," which is specialized for high-speed inference, has also commenced. Takahiro Ogura, Director of the AI Computing Division at the company, stated, "This product significantly increases memory bandwidth, which has been a major bottleneck in improving inference performance. We aim to provide it by 2026."

Specializing in inference in embedded and other devices and edge computing is EdgeCortix. The company will begin full-scale mass production of its AI chip, "SAKURA-II," in early 2025. **Founder and CEO Sakyasingha Dasgupta confidently stated, *"It is a high-performance, low-power chip specifically designed for edge applications."***

U.S.-based Cerebras Systems develops massive AI chips. Unlike conventional semiconductor chips, which are manufactured by dividing a large silicon wafer into multiple sections, Cerebras offers wafer-scale chips. While initially designed for training purposes, the company has recently positioned its technology as suitable for high-speed inference as well.

Translation prepared by EdgeCortix

- Nikkei Computer (2025/01/9). Full original Japanese article: https://bizboard.nikkeibp.co.jp/houjin/cgi-bin/nsearch/md_pdf.pl/0000510152.pdf?NEWS_ID=0000510152&CONTENTS=1&bt=NC&SYSTEM_ID=HO&BZB_DATE_TOKEN=f072df9d9140c7aa595b98736245b337ae6c9ab79ae590cfec7dde042720aa9ee62458dd7deefe0cd5bd8bc3bf953669
- 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.