BIZCOMMUNITY

OpenAI is getting into the hardware business in the AI arms race

By Lindsey Schutters

29 Jan 2024

"There's no way to get there without a breakthrough," said OpenAI CEO Sam Altman about the journey to the future of artificial intelligence. "It motivates us to go invest more in fusion." The consumer AI pioneer was speaking at a Bloomberg event on the sidelines of the World Economic Forum. Concerns about the 85TWh energy cost of running AI engines were raised in a peer reviewed study published in Joule at the end of 2023.

AI hardware could be a bigger power consumption problem than crypto mining. Source: Microsoft Designer.

Those numbers were drawn from an analysis of hardware order trends, with Nvidia – the early AI hardware winner – projecting that it will ship 1,5 million server chips per year over the next 3 years.

This hardware volume was at least partially corroborated by Meta executive chairman and CEO Mark Zuckerberg's bold proclamations in a recent interview with The Verge.

"We've come to this view that, in order to build the products that we want to build, we need to build for general intelligence," Zuckerberg said in the interview. "I think that's important to convey because a lot of the best researchers want to work on the more ambitious problems."

Google releases Gemini, says it's the next generation AI model

Lindsey Schutters 6 Dec 2023

Meta and Microsoft emerged as the joint biggest customers for Nvidia's H100 GPUs in 2023 with analysts estimating that the two tech giants purchased 300,000 units – which is equivalent to the shopping carts of the rest Nvidia's top 10 customers combined. That list includes Google, Amazon, Oracle and Tencent.

> *Just in case you don't want to click over to the other sites, Big Zuck update- Open sourcing will continue*
>
> *- Currently training LLama 3*

- AI + Metaverse
- Will have 350,000 H100s and ~600 H100 equivalents of compute ��
-Ideal AI formfactor is ��□ pic.twitter.com/xJSi7yVzXe
— Alex Volkov (Thursd/AI) (@altryne) January 18, 2024

Hive mind

One of the coveted features of the H100 Tensor Core GPU is its dedicated Transformer Engine, designed to handle trillion-parameter language models. These colossal machine learning models are a cornerstone of natural language processing (NLP), a field of AI that focuses on computer-human language interaction that has become the cornerstone of the current AI revolution.

These systems are also equipped with Nvidia's NVLink Switch System, a high-speed interconnect technology that enables efficient communication between multiple GPUs. This allows for the connection of up to 256 H100s which enables boost to exascale workloads, which demand at least a billion billion (10^18) calculations per second.

This translates to a 30x speed boost for large language models (LLMs) compared to the previous generation GPUs. This acceleration leads to faster processing of language-based tasks, quicker responses, and more efficient operation.

Nvidia is not the only game in town, however. AMD unveiled the Instinct MI300X at the close of 2023, with bold claims that it surpassed H100 GPU in several tests.

According to AMD, the MI300X was up to 60% quicker in a direct 8 to 8 server comparison. Nvidia responded, saying that AMD had benchmarked the H100 incorrectly and claimed that its GPU is 2x faster.

AMD found a niche selling into the high-performance computing (HPC) space with Asus, Dell, Gigabyte, HP, Lenovo, and Supermicro among its clients.Meta's AI accelerator stockpile is an enticing lure for engineering talent in the space and that creates an existential threat for its LLM competitors like OpenAI.

"The biggest companies that started off with the biggest leads are also, in a lot of cases, the ones calling the most for saying you need to put in place all these guardrails on how everyone else builds AI," Zuckerberg explained.

"I'm sure some of them are legitimately concerned about safety, but it's a hell of a thing how much it lines up with the strategy."

Scare tactics

The OpenAI CEO is on record calling for stricter regulations on AI development and visited South Korea on the weekend to explore potential collaborations with leaders of the country's semiconductor industry. His visit was a tour of Samsung's chip

fabrication plants in Pyeongtaek and meetings with top executives from Samsung and SK Hynix.

The visit comes in the wake of a surge in interest in AI applications since the release of OpenAI's ChatGPT, which has led to a significant demand for computing power and processors. Altman has previously stated that there are not enough chips to meet his company's needs.

> *OpenAI CEO Sam Altman is visiting leaders of South Korea's semiconductor industry, as the AI pioneer weighs an ambitious move into chip production https://t.co/UnrfpTOz4X— Bloomberg Technology (@technology) January 26, 2024*

While the specific objectives of Altman's trip are not clear, he has been working to raise funds to establish a network of factories for semiconductor manufacturing. This visit could potentially signal a move towards a more significant role for OpenAI in chip production and less reliance on Microsoft's Nvidia-powered Azure servers.

Microsoft uses OpenAI's models for its Copilot on Nvidia's GPUs and is tuning the models for its upcoming Maia 100 AI accelerator. The Maia 100, introduced in November, is an AI accelerator specifically designed for inferencing tasks, which involve providing answers to customer queries.

Maia is Microsoft's first AI accelerator, purpose-built to handle intensive AI workloads like training large AI models and running generative AI services.

On the other hand, Google has built its entire AI infrastructure around its Tensor Processing Units (TPUs). TPUs are custom-built to speed up machine learning workloads and are used to accelerate neural network machine learning, using Google's own TensorFlow software. They are designed as matrix processors specialised for neural network workloads.

This Stencil app is disabled for this browser.

Developers:

- ES5 builds are disabled **during development** to take advantage of 2x faster build times.

While Altman's actions can be viewed in an altruistic light that will benefit the development of the full AI value chain, a cynical read of the situation highlights a play to gain a bigger piece of the investment pie.

His appeals for regulation, for instance can be understood as an effort to raise the walls around the OpenAI garden. While his declarations around the energy breakthroughs that are needed also come after he invested in nuclear fusion company Helion, which has a power purchase deal in place with Microsoft – OpenAI's biggest investor.

ABOUT LINDSEY SCHUTTERS

Lindsey is the editor for ICT, Construction&Engineering and Energy&Mining at Bizcommunity
- Microsoft unveils Copilot+ PCs, sends Qualcomm stock through the roof - 20 May 2024
- #ATW24: Mukuru's customer-first approach behind Africa longevity - 20 May 2024
- #ATW24: Minister Gungubele calls for connectivity and inclusion - 17 May 2024
- Google uses AI to move Search posts for web publishers - 15 May 2024
- OpenAI debuts GPT-4o as a multimodal, more personal AI - 14 May 2024

View my profile and articles...