

Microsoft and AMD are reportedly teaming up to combat Nvidia's AI dominance



Enlarge / AMD's Instinct AI accelerators aren't as widely used as competing products from Nvidia.

reader comments

59 with

The hardware that has enabled and driven the recent explosion of generative AI products has primarily been manufactured by Nvidia. That's thanks in part to its dominant position in the consumer and workstation graphics markets, plus the well-established CUDA libraries and its AI-accelerating Tensor cores.

Although Microsoft uses "tens of thousands" of Nvidia GPUs in its Azure data centers, Bloomberg reports that the company is also working with AMD to improve the AI capabilities of its

GPUs. Although extremely light on details, the report says Microsoft is contributing “support” and “engineering resources” to AMD to improve its products’ performance in AI workloads.

Bloomberg also says that AMD is working on a Microsoft-developed AI accelerator named “Athena,” which The Information reported details on last month. A Microsoft spokesperson, however, flatly denied that AMD was involved in Athena’s development.

The upside for Microsoft is that a more competitive market for AI-accelerating hardware can save on server costs as the company looks to build AI features into more and more of its products. For example, Microsoft is reportedly working on a private version of ChatGPT to offer to businesses that could cost “10 times” as much to run as the regular version of ChatGPT; if the server hardware required to run these generative AI models is cheaper, Microsoft could lower its prices, its costs, or both to make these products more appealing.

Advertisement

The two companies have kind-of-sort-of collaborated before; “Surface Edition” Ryzen processors have been used in some older versions of Microsoft’s Surface PCs, though not in any of the most recent Surface devices. Like the SQ-series chips from Qualcomm that Microsoft uses in the Arm versions of the Surface Pro, the Surface Edition Ryzen processors weren’t much different from regular Ryzen processors—they’re the same chips AMD sold to other manufacturers but with Microsoft-assisted optimizations to the “firmware, drivers, and software stack.”

Although it’s not representative of how each company’s server-grade GPUs perform, recent tests from Tom’s Hardware illustrate how AMD’s GPU architectures struggle with AI workloads. In testing with the Stable Diffusion image generator, AMD’s current Radeon RX 7900 XTX flagship performed

slower than Nvidia's RTX 4090, 4080, and 4070 Ti, even though the RX 7900 XTX is faster than the RTX 4080 in most games. Even the entry-level RTX 3050 managed to beat all of AMD's previous-generation RX 6000-series cards (software optimizations could help these cards perform better, but the fact that the software was Nvidia-optimized by default says a lot about the current state of things).

<https://arstechnica.com/?p=1937083>