Arm’s first CPU ever will plug into Meta’s AI datacenters later this year

After decades of only licensing its chip designs for others to use, UK-based Arm revealed the first chip it’s producing on its own, and the first customer. Dubbed the Arm AGI CPU, it’s another chip designed for inference, or running the cloud processing for AI tools like AI agents that can continue to spawn more and more tasks to run at once. The first company in line to use it is Meta, which has reportedly struggled to launch its own AI chips.

Meta says it’s both the lead partner and co-developer, and plans to work on “multiple generations” of the datacenter CPUs, for use along with hardware from other vendors like Nvidia and AMD. Arm customers like Amazon AWS, Microsoft, Google, Marvell, Nvidia, Samsung, and others included congratulatory notes with the announcement. However, Qualcomm, which said it had achieved “complete victory” over Arm with a court ruling last fall in their case over the terms of licensing agreements, was not one of them.

Financial terms of the deal weren’t revealed, nor was the number of chips Meta plans to use from Arm, which is currently owned by Softbank.

According to Arm, its new chip runs on the Neoverse platform used by AWS Graviton, Nvidia Vera, Microsoft, and other AI chips, with up to 136 cores per CPU and 64 CPUs per air-cooled server rack. It says the AGI CPU can get twice as much performance per watt as traditional x86 CPUs while reducing memory bottlenecks, taking advantage of the design’s long-

running efficiency advantages.

Other customers lined up for Arm’s chip include Cerebras, Cloudflare, F5, OpenAI, Positron, Rebellions, SAP, and SK Telecom. Arm cloud AI head Mohamed Awad told CNBC that its aim is to be an option for companies that can’t afford to make their own in-house processor.

https://www.theverge.com/ai-artificial-intelligence/899823/arm-agi-cpu-meta