Microsoft introduces its own chips for AI

November 16, 2023

Online Desk: Microsoft on Wednesday announced a duo of custom-designed computing chips, joining other big tech firms that – faced with the high cost of delivering artificial intelligence services – are bringing key technologies in-house.

Microsoft said it does not plan to sell the chips but instead will use them to power its subscription software offerings and as part of its Azure cloud computing service.

At its Ignite developer conference in Seattle, Microsoft introduced a new chip, called Maia, to speed up AI computing tasks and provide a foundation for its $30-a-month “Copilot” service for business software users, as well as for developers who want to make custom AI services.

The Maia chip was designed to run large language models, a type of AI software that underpins Microsoft’s Azure OpenAI service and is a product of Microsoft’s collaboration with ChatGPT creator OpenAI.

Microsoft and other tech giants such as Alphabet are grappling with the high cost of delivering AI services, which can be 10 times greater than for traditional services such as search engines.

Microsoft executives have said they plan to tackle those costs by routing nearly all of the company’s sprawling efforts to put AI in its products through a common set of foundational AI models. The Maia chip, they said, is optimized for that work.

“We think this gives us a way that we can provide better solutions to our customers that are faster and lower cost and higher quality,” said Scott Guthrie, the executive vice president of Microsoft’s cloud and AI group.

Microsoft also said that next year it will offer its Azure customers cloud services that run on the newest flagship chips from Nvidia and Advanced Micro Devices Microsoft said it is testing GPT 4 – OpenAI’s most advanced model – on AMD’s chips.

“This is not something that’s displacing Nvidia,” said Ben Bajarin, chief executive of analyst firm Creative Strategies.

He said the Maia chip would allow Microsoft to sell AI services in the cloud until personal computers and phones are powerful enough to handle them.

“Microsoft has a very different kind of core opportunity here because they’re making a lot of money per user for the services,” Bajarin said.

Microsoft’s second chip announced Tuesday is designed to be both an internal cost saver and an answer to Microsoft’s chief cloud rival, Amazon Web Services.

Named Cobalt, the new chip is a central processing unit made with technology from Arm Holdings Microsoft disclosed on Wednesday that it has already been testing Cobalt to power Teams, its business messaging tool.

But Microsoft’s Guthrie said his company also wants to sell direct access to Cobalt to compete with the “Graviton” series of in-house chips offered by Amazon Web Services.

“We are designing our Cobalt solution to ensure that we are very competitive both in terms of performance as well as price-to-performance (compared with Amazon’s chips),” Guthrie said.

AWS will hold its developer conference later this month, and a spokesman said that its Graviton chip now has 50,000 customers.

“AWS will continue to innovate to deliver future generations of AWS-designed chips to deliver even better price-performance for whatever customer workloads require,” the spokesman said after Microsoft announced its chip.

Microsoft gave few technical details that would allow gauging the chips’ competitiveness versus those of traditional chipmakers. Rani Borkar, corporate vice president for Azure hardware systems and infrastructure, said both are made with 5-nanometer manufacturing technology from Taiwan Semiconductor Manufacturing Co.

She added that the Maia chip would be strung together with standard Ethernet network cabling, rather than a more expensive custom Nvidia networking technology that Microsoft used in the supercomputers it built for OpenAI.

“You will see us going a lot more the standardization route,” Borkar told Reuters.