

Microsoft unveils Maia 100 AI accelerator and Cobalt 100 Arm CPU for Azure

Source copy kept by wheresthe.ai for the story "Microsoft unveils Maia 100 AI accelerator and Cobalt 100 Arm CPU for Azure" (<https://wheresthe.ai/d/microsoft-maia-cobalt-chips-2023-11/>).

Source: <https://news.microsoft.com/source/features/ai/in-house-chips-silicon-to-service-to-meet-ai-demand/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 70288cf943a81a24a6d6b2ea4327b001d912da4ae708a2c41db6c2d9e1d5db97

Tucked away on Microsoft's Redmond campus is a lab full of machines probing the basic building block of the digital age: Silicon. This multi-step process meticulously tests the silicon, in a method Microsoft engineers have been refining in secret for years.

Today at Microsoft Ignite the company unveiled two custom-designed chips and integrated systems that resulted from that journey: the Microsoft Azure Maia AI Accelerator, optimized for artificial intelligence (AI) tasks and generative AI, and the Microsoft Azure Cobalt CPU, an Arm-based processor tailored to run general purpose compute workloads on the Microsoft Cloud.

The chips represent a last puzzle piece for Microsoft to deliver infrastructure systems – which include everything from silicon choices, software and servers to racks and cooling systems – that have been designed from top to bottom and can be optimized with internal and customer workloads in mind.

The chips will start to roll out early next year to Microsoft's datacenters, initially powering the company's services such as Microsoft Copilot or Azure OpenAI Service. They will join an expanding range of products from industry partners to help meet the exploding demand for efficient, scalable and sustainable compute power and the needs of customers eager to take advantage of the latest cloud and AI breakthroughs.

1 of 5 The Microsoft Azure Maia 100 AI Accelerator is the first designed by Microsoft for large language model training and inferencing in the Microsoft Cloud. Image courtesy of Microsoft. 2 of 5 The Microsoft Azure Cobalt 100 CPU is the first developed by Microsoft for the Microsoft Cloud. Image courtesy of Microsoft. 3 of 5 This "cold plate" is attached to the surface of a Maia 100 AI Accelerator. The plate has channels through which liquid is circulated to absorb and transport heat. Photo by John Brecher for Microsoft. 4 of 5 A system-level tester at a Microsoft lab in Redmond, Washington, mimics conditions that a chip will experience inside a Microsoft datacenter. The machine rigorously assesses each chip under real world conditions to ensure it meets performance and reliability standards. Photo by John Brecher for Microsoft. 5 of 5 Chips are being assessed in a system-level tester at a Microsoft lab in Redmond, Washington, that simulates their use in real production conditions before being installed into a server. Photo courtesy of Microsoft.

"Microsoft is building the infrastructure to support AI innovation, and we are reimagining every aspect of our datacenters to meet the needs of our customers," said Scott Guthrie, executive vice president of Microsoft's Cloud + AI Group. "At the scale we operate, it's important for us to optimize and integrate every layer of the infrastructure stack to maximize performance, diversify our supply chain and give customers infrastructure choice."

Optimizing every layer of the stack

Chips are the workhorses of the cloud. They command billions of transistors that process the vast streams of ones and zeros flowing through datacenters. That work ultimately allows you to do just about everything on your screen, from sending an email to generating an image in Bing with a simple sentence.

Much like building a house lets you control every design choice and detail, Microsoft sees the addition of homegrown chips as a way to ensure every element is tailored for Microsoft cloud and AI workloads. The chips will nestle onto custom server boards, placed within tailor-made racks that fit easily inside existing Microsoft datacenters. The hardware will work hand in hand with software – co-designed together to unlock new capabilities and opportunities.

The end goal is an Azure hardware system that offers maximum flexibility and can also be optimized for power, performance, sustainability or cost, said Rani Borkar, corporate vice president for Azure Hardware Systems and Infrastructure (AHSI).

“Software is our core strength, but frankly, we are a systems company. At Microsoft we are co-designing and optimizing hardware and software together so that one plus one is greater than two,” Borkar said. “We have visibility into the entire stack, and silicon is just one of the ingredients.”

Rani Borkar, corporate vice president for Azure Hardware Systems and Infrastructure (AHSI) at Microsoft. Photo courtesy of Microsoft.

At Microsoft Ignite, the company also announced the general availability of one of those key ingredients: Azure Boost, a system that makes storage and networking faster by taking those processes off the host servers onto purpose-built hardware and software.

To complement its custom silicon efforts, Microsoft also announced it is expanding industry partnerships to provide more infrastructure options for customers. Microsoft launched a preview of the new NC H100 v5 Virtual Machine Series built for NVIDIA H100 Tensor Core GPUs, offering greater performance, reliability and efficiency for mid-range AI training and generative AI inferencing. Microsoft will also add the latest NVIDIA H200 Tensor Core GPU to its fleet next year to support larger model inferencing with no increase in latency.

The company also announced it will be adding AMD MI300X accelerated VMs to Azure. The ND MI300 virtual machines are designed to accelerate the processing of AI workloads for high range AI model training and generative inferencing, and will feature AMD’s latest GPU, the AMD Instinct MI300X.

By adding first party silicon to a growing ecosystem of chips and hardware from industry partners, Microsoft will be able to offer more choice in price and performance for its customers, Borkar said.

“Customer obsession means we provide whatever is best for our customers, and that means taking what is available in the ecosystem as well as what we have developed,” she said. “We will continue to work with all of our partners to deliver to the customer what they want.”

Co-evolving hardware and software

The company’s new Maia 100 AI Accelerator will power some of the largest internal AI workloads running on Microsoft Azure. Additionally, OpenAI has provided feedback on Azure Maia and Microsoft’s deep insights into how OpenAI’s workloads run on infrastructure tailored for its large language models is helping inform future Microsoft designs.

“Since first partnering with Microsoft, we’ve collaborated to co-design Azure’s AI infrastructure at every layer for our models and unprecedented training needs,” said Sam Altman, CEO of OpenAI. “We were excited when Microsoft first shared their designs for the Maia chip, and we’ve worked together to refine and test it with our models. Azure’s end-to-end AI architecture, now optimized down to the silicon with Maia, paves the way for training more capable models and making those models cheaper for our customers.”

The Maia 100 AI Accelerator was also designed specifically for the Azure hardware stack, said Brian Harry, a Microsoft technical fellow leading the Azure Maia team. That vertical integration – the alignment of chip design with the larger AI infrastructure designed with Microsoft’s workloads in mind – can yield huge gains in performance and efficiency, he said.

“Azure Maia was specifically designed for AI and for achieving the absolute maximum utilization of the hardware,” he said.

Meanwhile, the Cobalt 100 CPU is built on Arm architecture, a type of energy-efficient chip design, and optimized to deliver greater efficiency and performance in cloud native offerings, said Wes McCullough, corporate vice president of hardware product development. Choosing Arm technology was a key element in Microsoft’s sustainability goal. It aims to optimize “performance per watt” throughout its datacenters, which essentially means getting more computing power for each unit of energy consumed.

These servers inside a datacenter in Quincy, Washington, are the first to be powered by the Microsoft Azure Cobalt 100 CPU. Photo by John Brecher for Microsoft.

“The architecture and implementation is designed with power efficiency in mind,” he said. “We’re making the most efficient use of the transistors on the silicon. Multiply those efficiency gains in servers across all our datacenters, it adds up to a pretty big number.”

Custom hardware, from chip to datacenter

Before 2016, most layers of the Microsoft cloud were bought off the shelf, said Pat Stemen, partner program manager on the AHSI team. Then Microsoft began to custom build its own servers and racks, driving down costs and giving customers a more consistent experience. Over time, silicon became the primary missing piece.

The ability to build its own custom silicon allows Microsoft to target certain qualities and ensure that the chips perform optimally on its most important workloads. Its testing process includes determining how every single chip will perform under different frequency, temperature and power conditions for peak performance and, importantly, testing each chip in the same conditions and configurations that it would experience in a real-world Microsoft datacenter.

The silicon architecture unveiled today also lets Microsoft not only enhance cooling efficiency but optimize the use of its current datacenter assets and maximize server capacity within its existing footprint, the company said.

For example, no racks existed to house the unique requirements of the Maia 100 server boards. So Microsoft built them from scratch. These racks are wider than what typically sits in the company's datacenters. That expanded design provides ample space for both power and networking cables, essential for the unique demands of AI workloads.

A custom-built rack for the Maia 100 AI Accelerator and its "sidekick" inside a thermal chamber at a Microsoft lab in Redmond, Washington. The sidekick acts like a car radiator, cycling liquid to and from the rack to cool the chips as they handle the computational demands of AI workloads. Photo by John Brecher for Microsoft.

Such AI tasks come with intensive computational demands that consume more power. Traditional air-cooling methods fall short for these high-performance chips. As a result, liquid cooling – which uses circulating fluids to dissipate heat – has emerged as the preferred solution to these thermal challenges, ensuring they run efficiently without overheating.

But Microsoft's current datacenters weren't designed for large liquid chillers. So it developed a "sidekick" that sits next to the Maia 100 rack. These sidekicks work a bit like a radiator in a car. Cold liquid flows from the sidekick to cold plates that are attached to the surface of Maia 100 chips. Each plate has channels through which liquid is circulated to absorb and transport heat. That flows to the sidekick, which removes heat from the liquid and sends it back to the rack to absorb more heat, and so on.

The tandem design of rack and sidekick underscores the value of a systems approach to infrastructure, McCullough said. By controlling every facet — from the low-power ethos of the Cobalt 100 chip to the intricacies of datacenter cooling — Microsoft can orchestrate a harmonious interplay between each component, ensuring that the whole is indeed greater than the sum of its parts in reducing environmental impact.

Microsoft has shared its design learnings from its custom rack with industry partners and can use those no matter what piece of silicon sits inside, said Stemen. "All the things we build, whether infrastructure or software or firmware, we can leverage whether we deploy our chips or those from our industry partners," he said. "This is a choice the customer gets to make, and we're trying to provide the best set of options for them, whether it's for performance or cost or any other dimension they care about."

Microsoft plans to expand that set of options in the future; it is already designing second-generation versions of the Azure Maia AI Accelerator series and the Azure Cobalt CPU series. The company's mission remains clear, Stemen said: optimize every layer of its technological stack, from the core silicon to the end service.

"Microsoft innovation is going further down in the stack with this silicon work to ensure the future of our customers' workloads on Azure, prioritizing performance, power efficiency and cost," he said. "We chose this innovation intentionally so that our customers are going to get the best experience they can have with Azure today and in the future."

Related resources:

[Read more: Microsoft delivers purpose-built cloud infrastructure in the era of AI](#)

[Read more: Azure announces new AI optimized VM series featuring AMD's flagship MI300X GPU](#)

[Read more: Introducing Azure NC H100 v5 VMs for mid-range AI and HPC workloads](#)

[Read more: AI chips are getting hotter. A microfluidics breakthrough goes straight to the silicon to cool up to three times better.](#)

Learn more: [Microsoft Ignite](#)

Top image: A technician installs the first server racks containing Microsoft Azure Cobalt 100 CPUs at a datacenter in Quincy, Washington. It's the first CPU designed by Microsoft for the Microsoft Cloud. Photo by John Brecher for Microsoft.

Tags:

AI