

10122026 Maia 300 October News Announcement

Maia 300 October News Announcement

Scott Guthrie – EVP, Cloud + AI Group

Rani Borkar – President, Azure Hardware Systems and Infrastructure

Saurabh Dighe – CVP, Azure Hardware Systems Planning and Architecture

Mustafa Suleyman – CEO, Microsoft AI

Jim Kleewein – Technical Fellow, Microsoft 365

Monday, October 12, 2026

INTRO VIDEO: The innovation around us is the work of an entire industry. Materials and machines. Chips and systems. Software and networks. Power, cooling, and data centers. We work at scales almost impossible to comprehend.

From structures measured in atoms to some of the largest and most complex infrastructure projects of our time. Again and again, we turn the seemingly impossible into something essential to everyday life.

Our achievements aren't just the technologies we create, but everything they make possible. The impact is all around us and the opportunity ahead is extraordinary.

As we enter the next era, what will be the defining measure of progress?

SCOTT GUTHRIE: Because Microsoft operates some of the world's largest AI services, we have a unique opportunity to learn directly from production workloads and apply those learnings across every layer of the technology stack.

Maia's already running production AI workloads across Microsoft. This includes Microsoft Copilot, GitHub Copilot, Microsoft Foundry, and Microsoft's own AI models. Maia 200 is the best performing AI infrastructure in our fleet, measured on real-world metrics for our AI services.

This is made possible by co-designing the complete stack from silicon to system to agent. Maia 300 delivers superior efficiency and economics. We're scaling Maia to gigawatt-scale deployments starting next year.

RANI BORKAR: From the very beginning, Maia has been built around one guiding principle: optimize the entire system for AI inference. That means starting with the outcome we want and designing the full stack around it.

Maia 300 advances every layer of that stack, from silicon and software to networking and data center infrastructure. That gives us more capacity to serve AI workloads, support more

demanding applications, and put more intelligence into the products and services people use every day. The economics of AI are increasingly determined not by peak compute performance, but how efficiently an entire system can generate tokens at scale.

So we designed Maia 300 end-to-end to be our most efficient, scalable, and reliable inference platform yet. The opportunity ahead is not simply to build more compute. It is to make advanced intelligence more accessible, affordable, and useful. Maia 300 is an important step toward that goal.

SAURABH DIGHE: Maia 300 really builds on the strong foundation of software-defined dataflow architecture that we introduced with Maia 200, but we did make major advances in five areas.

First, we moved to a multi-chiplet architecture onto the N2 process node from TSMC. Second, we went to HBM4 and we added substantially greater HBM capacity and bandwidth. We added new precision data types and MX precision data types for computation and KV storage. We did advance our next-generation of scale-up network.

And then finally, we built a larger, more power-efficient rack scale system.

And so together these changes increase the resources available for long-context decode-heavy inference from a single accelerator through the entire full stack rack and data center.

MUSTAFA SULEYMAN: At Microsoft AI, our mission is to build a superintelligence that truly works for humanity, and that means that it will make everyone smarter, healthier, and more capable. Built right here at Microsoft for the billions of people that use our products every day.

Over the last 15 years, the single biggest driver of progress in AI has been scale. We have to work across the full stack all the way down to the silicon. At that scale, we need a platform that we own and that we co-design with our models.

That's why running our Microsoft AI models on top of the Maia chip is a critical focus for our team. Chip-model co-design is critical for us to build frontier superintelligence.

All of the lessons that we have gained over the last few years of being able to put MAI models on top of the Maia 200 have gone straight into the design of the Maia 300, and we are only able to get that loop when the people building the models and the people behind the very systems and silicon that are driving them work as one.

JIM KLEWEIN: AI is a key part of the future of productivity and the infrastructure behind that AI matters more than ever. Maia 200 provides foundational AI infrastructure for experiences like Microsoft Copilot and Cowork, helping users summarize information, compose content, and complete multi-step tasks.

What's especially powerful is the connection between the teams building Microsoft 365 experiences and the teams building the infrastructure underneath them. We can understand how people use those experiences, where performance matters most, and what opportunities exist to improve efficiency, responsiveness, and reliability.

Those insights help shape the next generation of infrastructure. Maia 300 builds on the lessons we've learned operating AI services at scale. Ultimately, our goal is simple: help every person and every organization accomplish more.

RANI BORKAR: That is why we are building Maia as a multi-generational platform. We cannot meet that challenge simply by adding more resources. We must get more useful intelligence from every byte, every cycle, and every watt.

What matters is how broadly that input translates into productivity in the economy and value in people's lives. And that happens when AI moves into everyday work across industries and around the world.

To make that possible, AI must be accessible and at scale. Accessibility depends on efficiency. That is the real opportunity and that is exactly what Maia 300 is designed to deliver.