

10122026 Maia 300 October News Announcement

Systems approach to efficient AI inference

Rani Borkar – President, Azure Hardware Systems and Infrastructure

Monday, October 12, 2026

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.