Microsoft

Primary Inference Role for Maia Family with Maia 300 Expected to Deliver Significant Performance Gains and Underpin Strong Ramp Plans

Overweight

MSFT, MSFT US
Price (12 Oct 26):$541.15

Software - Large Cap / Mid & Small Cap

Samik Chatterjee, CFA [AC]
(1-212) 622-0798
samik.x.chatterjee@jpmorgan.com

Jaiden R Patel
(1-212) 270-6953
jaiden.patel@jpmchase.com

Brian Hyska
(1-212) 622-2883
brian.hyska@jpmchase.com

Mashu Nishi
(1-212) 622-0078
mashu.nishi@jpmorgan.com
J.P. Morgan Securities LLC

Microsoft announced on October 12 that it plans to scale its Azure Maia accelerator family to multi-gigawatt deployments starting in 2027, which will include making Maia the primary inference system for Microsoft products and services. Alongside the deployment roadmap, the company introduced its third-generation Maia 300 system, with production deployments scheduled to begin in 2027. Microsoft expects Maia 300 to deliver 40% more tokens per megawatt versus leading GPU-based systems and up to 7x the rack-scale token throughput of Maia 200 in a full 96-accelerator configuration. Maia 200 is already serving production workloads, including Microsoft Copilot, with GitHub Copilot and Microsoft Foundry expected to follow by year-end. Key details of the announcement are outlined below.

- **Maia to become the primary inference system across Microsoft services leveraging the experience through prior generations; multi-gigawatt expansion starts in 2027.** Microsoft outlined a deployment plan encompassing the Maia family rather than Maia 300 alone, with the next-generation system entering production deployment in 2027. With Maia 200 already serving Microsoft Copilot and supporting the MAI team's development of next-generation models, while GitHub Copilot and Microsoft Foundry are expected to follow by the end of 2026, the company already has experience in relation to supporting OpenAI models, Microsoft's MAI models and leading open-source model families, and is looking to further scale the Maia family and make it the primary inference platform for the company in time.
- **Maia 300's higher performance and throughput are driven by a combination of innovations in software-defined dataflows and compute architecture as well as memory.** Maia 300's higher performance including 40% more tokens per MW, building further on the better performance of Maia 200 in relation to up to 1.7x more tokens per Watt on built-in AI capabilities relative to GPU infrastructure, is expected to be driven by changes in relation to both more efficient software defined dataflows, and with hardware innovation including chiplet architecture, leading to higher HBM bandwidth per Watt and scale-up networking at datacenter scale.
- **Reuters reports potential deployment of more than one million Maia 300 chips in 2027.** Beyond Microsoft's announced multi-gigawatt expansion, Reuters reports (see link to Reuters Article) that the company plans to deploy potentially more than one million Maia 300 chips next year to support internal AI services. The announcement today also follows reports earlier of Microsoft's plans to more than triple its global datacenter capacity by 2032, which in conjunction with large scale deployment of Maia family of chips will likely drive a more optimal cost outcome on its scaled infrastructure. We estimate a robust increase in datacenter capacity to 30+GW by 2030 from the company, the roadmap for a robust increase in capacity to be further solidified

See page 3 for analyst certification and important disclosures.

by the FY27E capex guide, which we expect to track to $210B vs. $175B for CY26.

- **Expect Maia's scaling roadmap to remain constrained to internal use cases.** We expect the deployments to remain focused on internal use cases rather than external compute revenue, which implies that the two avenues that investors will most specifically see the impact of the scaling of deployments will be in relation to capex intensity as well as cost of goods sold in Cloud software.
- **Positive implications to both Copilot performance as well as Microsoft Cloud margins.** The robust positioning of Maia as the primary inference system for the company will position the company to deliver better performance for Copilot with higher token throughput, while at the same time leveraging the throughput also towards lower cost per compute unit relative to the current landscape, implying opportunity for better gross margins for Microsoft Cloud which are tracking in the mid- to high 60% range.

