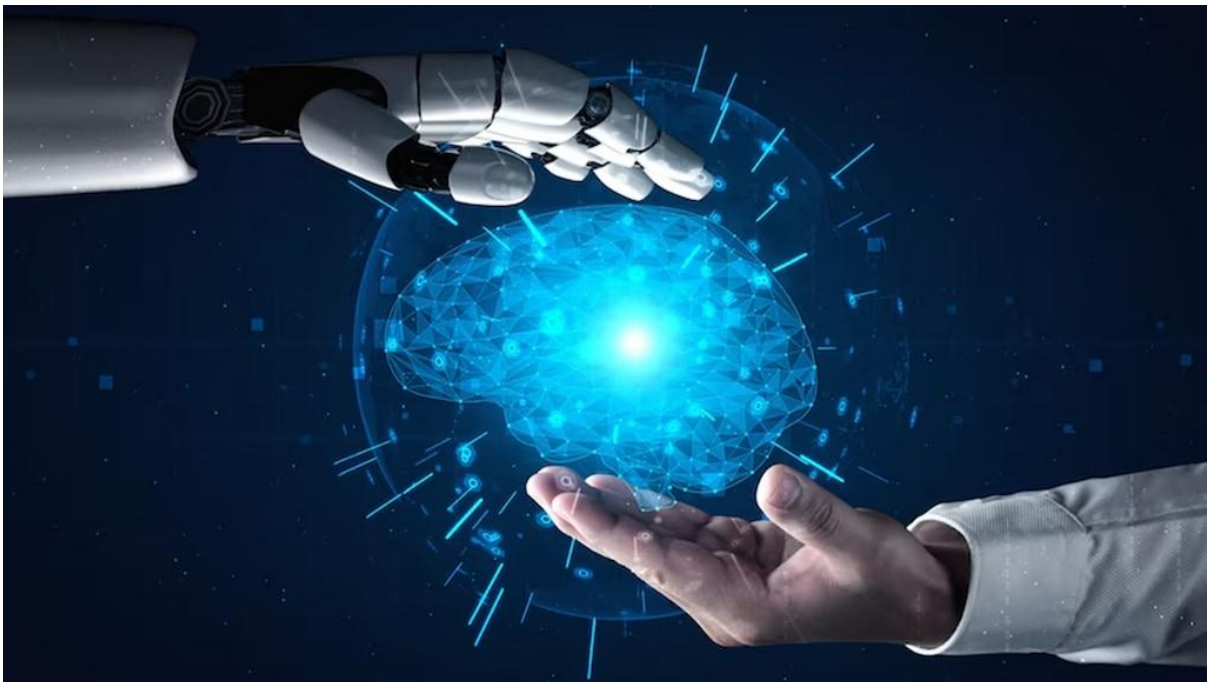
Quote by Arun Moral, Managing Director, Primus Partners

Published in Financialexpress
Jan 2, 2025 | 05:30 IST

AMD, Intel target AI inferencing to rival Nvidia

Gartner projects that by 2028, over 80% of workload accelerators in data centres will be dedicated to AI inferencing, compared to just 40% in 2023.

Authored by Jatin Grover



Unlike AI training, which is concentrated in data centres, inferencing is expected to take place closer to users on edge devices. (Image/Freepik)

Read on:https://www.financialexpress.com/life/technology-amd-intel-target-ai-inferencing-to-rival-nvidia-3705297/

Article Content:

The race for dominance in AI computing is intensifying as AMD and Intel set their sights on AI inferencing, a critical and fast-growing segment poised to disrupt Nvidia's monopoly.

With enterprises increasingly prioritising cost-effective and energy-efficient solutions, the landscape of AI hardware is undergoing a shift. Nvidia has long held a commanding lead in AI computing, fueled by its cutting-edge graphic processing units (GPUs) that dominate AI training workloads. However, industry analysts say the emergence of AI inferencing – the process of deploying trained AI models to make real-time predictions or decisions – could redefine the competitive dynamics in this market.

AI inferencing represents the stage where businesses begin to see tangible returns on their AI investments, and AMD and Intel are positioning themselves to capitalise on this opportunity.

While Nvidia GPUs remain the gold standard for AI training, their high cost – up to five times more than AMD and Intel alternatives – poses a challenge for enterprises looking to scale inferencing operations. “AI inferencing will become a larger market than training over time, and both AMD and Intel are positioning their GPUs and CPUs to capitalise on this transition,” said Sunil Gupta, co-founder and CEO of Yotta.

The demand for power-efficient, cost-effective solutions is driving enterprises to consider AMD and Intel’s offerings. Both companies are expected to roll out dedicated chipsets optimised for inferencing tasks, increasing the pressure on Nvidia to lower its prices. Lisa T Su, president and CEO of AMD, highlighted the company’s confidence in its inferencing capabilities during a recent earnings call. “The $5 billion we’re projecting for 2024 data centre GPU revenue reflects early traction primarily in inferencing, thanks to the MI300’s memory capacity and bandwidth optimisation,” Su said.

Nvidia continues to dominate AI compute, with its data centre revenue reaching $30.8 billion in the July-September quarter, dwarfing AMD’s $3.5 billion. Yet, AMD and Intel see inferencing as their chance to challenge Nvidia’s supremacy. According to Karan Kirpalani, chief product officer at Neysa, the growing prominence of inferencing could significantly alter the competitive landscape. “As AI becomes as pervasive as electricity or the internet, inference workloads across consumer and enterprise ecosystems will vastly outpace training,” Kirpalani said. “This shift will amplify the revenue and innovation opportunities associated with inference hardware.”

Intel, though still a minor player in the GPU market, is leveraging its CPU expertise to carve out a niche in inferencing. Jeongku Choi, research analyst at Counterpoint Research, explained that Intel and AMD can leverage their power-efficient, cost-effective hardware to challenge Nvidia in this space.

Unlike AI training, which is concentrated in data centres, inferencing is expected to take place closer to users on edge devices. These include smartphones, autonomous vehicles, and IoT systems. Applications like real-time traffic analysis in cars or personalised recommendations on smartphones are just some of the areas where inferencing is set to flourish. “Data is the fuel of generative AI, but its training phase has limits,” said Arun Moral, managing director at Primus Partners. “Once we cross 5 trillion data points, only synthetic data will remain, reducing the scope for training. This makes inferencing the future growth driver for AI hardware.”

Gartner projects that by 2028, over 80% of workload accelerators in data centres will be dedicated to AI inferencing, compared to just 40% in 2023.

Nvidia is not sitting idle. The company has expanded its portfolio to include ARM-based CPUs and optimised GPU platforms for inferencing, signaling its commitment to maintaining dominance in this segment. “Inferencing is incredibly complex,” said Jensen Huang, Nvidia’s president and CEO. “It requires high accuracy, low latency, and high throughput simultaneously, which makes it very challenging to build efficient systems. But we are seeing significant growth in this area.”