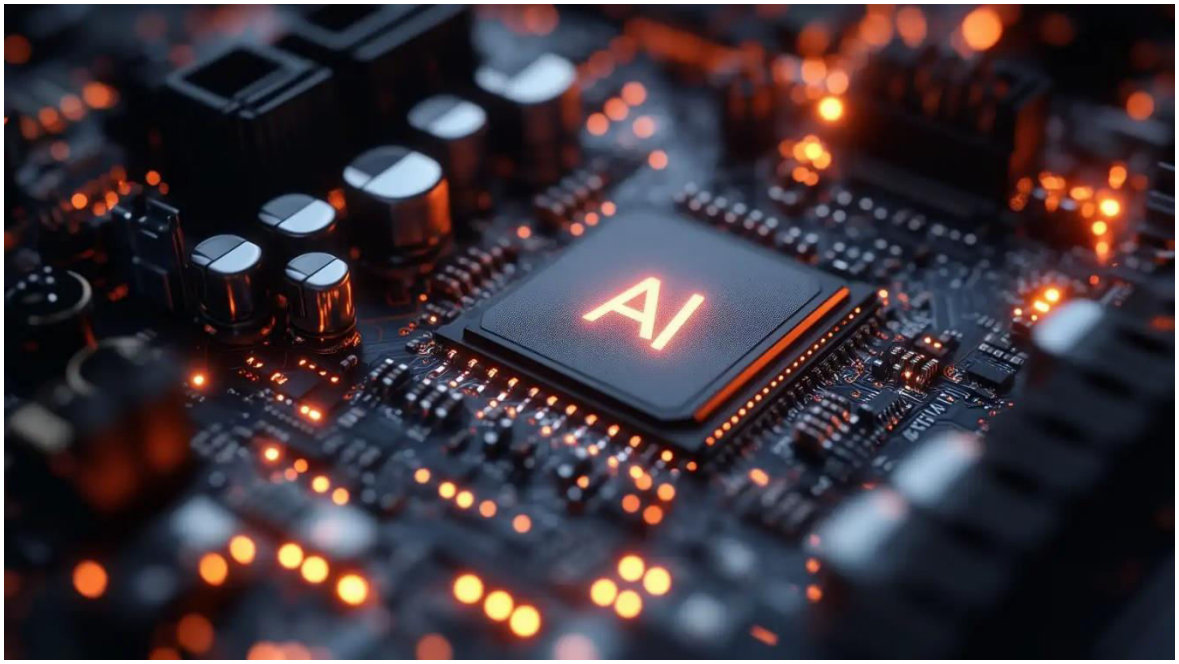
Quote by Devroop Dhar, Co-Founder and Managing Director, Primus Partners

Published in Network World
October 7th, 2026

AMD to ramp up CPU, GPU supply in 2027 as AI demand surges

The capacity ramp could improve AI compute availability, but supply chain constraints are likely to keep capacity tight and prices elevated through much of 2027.

Authored by Nidhi SIngal



Read on: AMD to ramp up CPU, GPU supply in 2027 as AI demand surges | AI infrastructure, GPU supply, data centres, chip packaging, cloud computing

Article Content:

As the demand for AI infrastructure continues to outstrip available capacity, AMD is planning a substantial increase in CPU and GPU supply in 2027. During her recent visit to Taiwan, AMD CEO Lisa Su confirmed that the company is working to expand supply across its ecosystem and is now planning capacity three to five years ahead.

Su told reporters that AMD has increased its supply and capacity in 2026 and is going to substantially increase supply in 2027, as per Reuters.

Su was in Taiwan on October 5-6 with the aim of ramping up AMD's supply for CPU and GPU production to meet growing AI demand. Su had reportedly held meetings with Foxconn and was also scheduled to meet TSMC as part of discussions with local partners. Following her meetings in Taiwan, Su is now visiting South Korea, where she

held meetings with Samsung Electronics' Device Solutions (DS) to broaden the partnership, among other companies.

For CIOs, this ramp-up could eventually improve access to AI compute and add another option to Nvidia's dominant infrastructure, but it is unlikely to end supply constraints overnight.

AMD's supply ramp depends on its partners:

AMD may have the demand and product roadmap, but scaling supply will depend on how much capacity these manufacturing and memory partners can deliver.

"AMD designs the chips, but relies on TSMC for advanced chip manufacturing and packaging, and SK Hynix and Samsung for HBM memory. Its 2027 supply increase therefore depends heavily on these partners adding capacity on time. Any delays in new wafer, packaging or HBM capacity could constrain AMD's GPU ramp," said Pareekh Jain, CEO at EIIRTrend & Pareekh Consulting.

Even though AMD has committed to a substantial supply increase in 2027, the execution is dependent on multiple factors.

First being the yield and efficiency of HBM4. "As AMD transitions its next-generation architecture to HBM4, it becomes dependent on highly complex architectural changes. HBM4 shifts to a logic base die, requiring deep collaboration with memory giants like SK Hynix and Samsung Electronics. If these memory makers experience poor bit efficiency or yield issues on early HBM4 runs, AMD's shippable volume will be instantly capped, regardless of how much raw silicon it prints," said Danish Faruqui, CEO at Fab Economics.

Secondly, there is a severe Chip-on-Wafer-on-Substrate (CoWoS) allocation squeeze as hyperscalers and Tier-1 cloud providers are locking down packaging capacity years in advance.

Faruqui noted that because Nvidia absorbs the lion's share of TSMC's packaging allocation for its Blackwell platforms, AMD's ability to substantially increase its 2027 volumes therefore depends entirely on how many packaging slots it can successfully wrestle away from competitors.

Lastly, while AMD is exploring broader ecosystem investments (such as expanding its multi-billion-dollar presence in Taiwan), it remains 100% tied to TSMC as its primary foundry for advanced nodes, added Faruqui. Efforts to diversify, such as evaluating potential packaging or wafer capacity from US-based fabs (like TSMC's Texas initiatives),

will not yield meaningful volume relief by 2027, leaving its ramp entirely vulnerable to supply chain shocks or capacity rationing in Taiwan, added Faruqui.

AI compute crunch may persist through 2027:

The AI compute supply crunch is expected to remain acute till the first half of 2027, say experts. This is because the shortage extends beyond GPUs to high-bandwidth memory, advanced packaging, data-centre space and reliable power, and demand from AI training and inference workloads continues to outpace the substantial new capacity being added.

"Enterprises can expect moderate relief in 2027 as new chip packaging and data-centre capacity comes online, although meaningful improvement is unlikely before the second half of the year. The outlook is therefore one of easing constraints rather than full normalisation, and premium pricing is likely to persist until supply expands across the entire infrastructure stack," said Devroop Dhar, co-founder and MD at Primus Partners.

Dhar added that most incremental capacity will initially be absorbed by hyperscalers such as AWS, Microsoft, Google and Meta, along with leading AI labs, given their long-term procurement commitments and scale advantages. However, enterprises will benefit largely indirectly, through broader availability of cloud-based AI services and improved access to GPU resources via public-cloud and managed-service providers.

More supply may not mean cheaper AI compute:

With demand for GPUs, CPUs and HBM already outpacing supply, additional capacity should eventually ease some pressure on AI infrastructure costs. But the new capacity will be arriving in a market where AI demand continues to grow rapidly.

"Lead times for premium enterprise data center GPUs range from 36 to 52 weeks in 2026, extending deliveries into 2027, driving GPU prices up by 25-35% above the manufacturer's suggested retail price, and cloud service providers' share increased premium GPU instance pricing (like EC2 Capacity Blocks) by roughly 15-25% already," added Faruqui.

While total global AI compute stock, including AMD's hardware, is projected to scale up through 2027, enterprises can realistically expect only marginal price reduction, and that too in stages.

Dhar noted that in the near term, the most significant effect will be greater negotiating leverage, as a credible alternative supplier gives enterprises and cloud providers more flexibility. Selective discounts may first appear on AMD-based cloud instances and older-generation GPUs.

Prices for frontier compute capacity could begin to moderate in the second half of 2027, once initial large-scale deployments are completed and more supply becomes available to the broader market. A more substantial price correction may therefore not emerge before 2028. He added that, amongst the various products, the MI350P and MI355X may be AMD's trump cards in this market and can give it an advantage in optimizing AI compute requirements and spend.

Compute will become more accessible to corporate buyers over the course of 2027, but will require deliberate planning and secured commitments. As AI compute is becoming strategic infrastructure that needs multiyear planning, Jain cautioned that CIOs should estimate AI demand under different growth scenarios, reserve capacity for critical workloads, use the cloud for unexpected demand, and avoid depending completely on one GPU supplier. The key measure should be cost per successful AI workload, not simply GPU price.