INTELLIGENCE UPDATE

Cloud AI price cuts challenge dedicated deployments

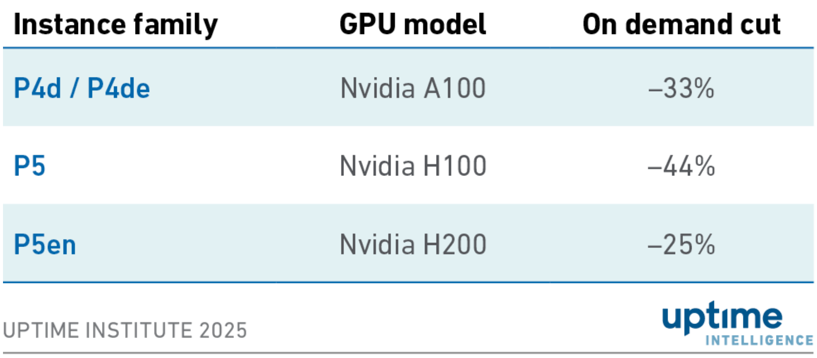
Dr. Owen Rogers

7 Aug 2025

In an Uptime Intelligence report published earlier this year utilization was shown to be the key parameter in determining whether public cloud or dedicated infrastructure was more cost-effective for a given AI workload (see *Neoclouds: a cost-effective AI infrastructure alternative*). With a higher level of utilization (as a proportion of total available capacity), dedicated infrastructure can have lower unit costs than public cloud. With a lower utilization level, public cloud becomes more affordable. The breakeven threshold — the minimum average utilization where dedicated becomes cheaper than cloud — was found to be 33%.

In this Uptime Intelligence report, neoclouds (a new category of cloud providers specializing in AI infrastructure as a service) were typically found to be charging less for GPUs than hyperscalers. In July, however, AWS announced significant price cuts of up to 44%, bringing its GPU-backed instance prices closer to those of neoclouds (**Table 1**).

Table 1 AWS GPU price cuts



Instance family	GPU model	On demand cut
P4d / P4de	Nvidia A100	–33%
P5	Nvidia H100	–44%
P5en	Nvidia H200	–25%

UPTIME INSTITUTE 2025

Why cut prices at all? AWS's explanation is that it now has the economies of scale to reduce costs. Greater demand for services by customers drives the purchase of more servers by AWS, which enables it to negotiate lower prices per server from suppliers such as Nvidia. These savings are passed on to their customers through lower cloud service prices.

However, Uptime Intelligence has identified additional factors influencing these price cuts:

- The new P6 instance, which contains Nvidia B200 Blackwell GPUs, is in high demand,

thereby reducing the demand on the less powerful P4 and P5 instance families. Selling these at a lower price can stimulate demand.
- Amazon is continually learning and developing ways to better balance demand and optimize its infrastructure. The company has developed a centralized GPU orchestration platform that enables customers to share GPU capacity across teams and maximize utilization via a project called "Greenland."
- Greater demand for (and therefore a higher volume of sales of) GPUs allows AWS to reduce its gross margin percentage, while still retaining — and even growing — its profit margins in absolute terms.
- Lower prices from neoclouds are forcing cloud providers to cut prices to remain competitive.

These changing market conditions affect other types of data center operators. Those using, or planning to use, dedicated infrastructure in colocation facilities and private facilities may see capital cost reductions due to the improved availability of GPUs and the lower cost of new hardware — making older hardware less expensive. However, economies of scale and innovative automation give hyperscaler cloud providers a cost advantage that most enterprises are unable to obtain.

AWS price cuts reduce the average price for a Nvidia H100 instance across hyperscalers and neoclouds from $66 to $59 per month. Now, operators need to ensure utilization beats 38% to undercut public cloud.

These reductions make it harder for dedicated infrastructure to beat cloud on price. An operating expenses (OpEx) purchasing model means cloud users typically benefit from price cuts immediately. In a capital purchase, the amount invested remains the same.

Variable cloud prices make it challenging to build a robust comparison between AI hosting models. If dedicated infrastructure has been purchased on the assumption of stable cloud pricing, the return on investment may be less favorable if cloud costs decrease. A dedicated server purchased a few months ago with an expected lifetime utilization of 35% would have been above the 33% threshold, making it more cost-effective than a public cloud instance. Today, at a breakeven point of 38%, the public cloud is the more cost-effective option.

Is there a likelihood that cloud GPU prices will rise? It would be unusual; providers would lose significant trust, as well as sales, if they were to raise prices. A customer may consider it as price gouging if, after committing to a cloud provider for their IT estate, the provider was to increase their monthly prices without offering an easy option to migrate elsewhere. A dedicated GPU server purchase gives stability in costs and protects against unlikely — yet possible — cloud price rises.

Other cloud providers, notably Google Cloud and Microsoft Azure, are likely to follow suit by cutting their prices in response. No provider wants to be seen as more expensive and risk losing new business. These price cuts may shift the breakeven threshold further, making it more difficult for dedicated GPU infrastructure to win on price.

ABOUT THE AUTHOR

Dr. Owen Rogers

Dr. Owen Rogers is Uptime Institute’s Senior Research Director of Cloud Computing. Dr. Rogers has been analyzing the economics of cloud for over a decade as a chartered engineer, product manager and industry analyst. Rogers covers all areas of cloud, including AI, FinOps, sustainability, hybrid infrastructure and quantum computing.

orogers@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.