Special Analysis

The Varying Cost of Nvidia's B200 Access in the Cloud

Tom Sorensen and Mark Nossokoff
March 2026

HYPERION RESEARCH OPINION

Hyperion Research's continued tracking of GPU market offerings reveals an emerging yet corroborated trend in the widely variable pricing of public cloud instances offering Nvidia's B200 GPU. The B200, the latest widely available addition to the Blackwell family of high-end GPUs, brings with it a significant premium over its predecessor. However, its publicly stated price per GPU/hour varies considerably across the popular public cloud services (AWS, Azure, and Google Cloud), as well as for some second-tier providers. Although GPU usage prices from cloud providers can change rapidly due to factors such as regional and reservation specifics, with all else equal, as seen in Table 1, there appears to be a wide spread of cost when it comes to leveraging the B200.

Table 1

B200 and Other GPU Dollar/Hour Breakdown

GPU Model	AWS (us-east-1/us-west-2)	Azure (East US/West US 2)	Google Cloud (us-central1/us-west1)	Lambda Labs (US East/West)
B100	~$5.00–$6.00	~$4.80–$5.80	~$4.50–$5.50	*Custom Pricing*
B200	~$8.00–$10.00	~$7.50–$9.50	~$7.00–$9.00	~$4.80–$6.00
B300	*Not yet available*	*Not yet available*	*Not yet available*	*Not yet available*

Note: Estimates assume 1-year reserved instances. 3-year commitments typically offer 15-25% discount. Pricing is per GPU hour, though actual instance pricing bundles GPU with other resources. Data primarily sourced direct from CSP/Hyperscaler portals and corroborated in other research. All values are in USD.

Source: Hyperion Research, 2026

ANALYST COMMENTARY

Lower prices for GPU/hour come with trade-offs. Oftentimes smaller cloud providers have more limited regional availability, fewer features such as auto-scaling or networking options, and perhaps most important, less mindshare as a reliable pillar of the cloud marketplace. These limitations highlight the differences with the major, fully-featured cloud providers that offer wide global infrastructure, more complex integrated ecosystems of storage and networking, as well as support perks such as dedicated account managers.

But these benefits often come at a premium, especially when the jobs being run within the cloud are highly specialized, such as those with a focus on raw GPU computing power. This pricing diversity further highlights the ongoing and rising importance of parametrizing and rigorously scoping the computational needs of a job's lifecycle before committing to a single or even multiple cloud provider. Leveraging the resources of a large, exhaustively diverse, and highly-serviced cloud environment in cases where many of these tools go unused can be significantly cost-ineffective.

Organizational leaders looking to define their cloud needs for optimal cost-effectiveness should ask questions including:

- What comprises my unique computational, networking, and storage needs?
- What specific software, storage, file systems are required?
- Do I have any obligations or preference to regional zones?
- How much ongoing support do I expect to need during the lifecycle of my cloud usage?
- What kind of Service Level Agreement (SLA) will satisfy my performance metrics, uptime, and responsiveness needs?
- How flexible is my budget? Am I prepared for billing complexities regarding data egress, storage, or additional services?
- What time window of commitment is my organization willing to make, if any?

While certainly not exhaustive, questions like these must be answered within the context of a wide exploration of cloud offerings when looking to maximize the cost-effectiveness of cloud resources in a time when compute costs, especially from advanced accelerators, continue to rise across the industry.

FUTURE OUTLOOK

As GPU prices rise and access to cutting-edge accelerators and hardware stays at a premium, organizations that focus on a complete knowledge of their needs and goals when purchasing cycles in a cloud for their cloud-only or hybrid computing systems stand to thrive. Cloud providers offering a narrower, less one-stop-shop set of resources could excel from their ability to limit computing costs as a trade-off for expansive, diverse offerings, and organizations that properly define their own needs and identify providers that suit them can realize higher computing yield per cost.