

Read a GPU cloud price before you compare it

Compare billing units, GPU variants and extra charges.

Updated October 29, 2026

A GPU price is useful only when its unit and conditions are clear. Start with the exact device, memory variant, region and host size. A per-GPU rate can still require you to rent a full host.

A sourced example

Modal lists H100 SXM5 GPU compute at USD 0.001097 per second. This equals USD 3.9492 per GPU-hour. It lists A100 with 80 GB at USD 0.000694 per second, or USD 2.4984 per GPU-hour. These are conversions of the rates on Modal's pricing page fetched 2026-10-29. They exclude separately billed CPU and memory. They are not a quote for a full workload.

Lambda's pricing page fetched 2026-10-29 lists an H100 SXM rate of USD 3.99 per GPU-hour in its eight-GPU instance table. The host costs more than that per hour because the rate is per GPU. Do not treat this as a single-device offer.

Build the comparison

Record the quoted rate and billing unit. Record whether a stopped machine still has billable storage. Add network transfer, attached volumes, images and support. Ask whether the provider rounds time at the task, machine or invoice level.

Use measured runtime from your workload. A cheaper device can cost more per completed job if it runs longer. Keep dated quotes with the benchmark output. Recheck both before a purchase.

Sources

- [Modal pricing](#) — fetched 2026-10-29.
- [Lambda GPU cloud pricing](#) — fetched 2026-10-29.

Frequently asked questions

Is this a vendor quote?

No. Confirm current terms and capacity with the provider before you rent.