

Compare GPU cloud approaches

Match the service model to control, capacity and operations needs.

Updated October 29, 2026

Start with the service boundary rather than a provider label. A single company can offer more than one approach. Check the terms of the specific product you plan to buy.

Infrastructure choices

A broad cloud platform can put GPU compute near existing storage and identity services. A GPU-focused cloud can offer a simpler path to accelerator capacity. In either case, compare the host specification, region, network and support terms.

A marketplace introduces a host selection step. Check who supplies the hardware, who handles failures and which party owes you a refund. A decentralized network also needs a clear review of supplier trust and data access. Do not upload sensitive data until that review is complete.

Execution choices

Bare-metal clusters give control over hosts and their connections. They also require an operations plan. Serverless GPU execution moves allocation and scaling into the service. Check startup behaviour, timeout rules and charges for supporting compute.

Managed inference moves the boundary further toward a model endpoint. Check supported weights, model versions and response behaviour under load.

Lambda's pricing page distinguishes instances and clusters. Modal publishes serverless compute prices. These sources show why a headline GPU rate is not enough to compare service models. Use a workload trial and a written quote to decide. Regional or sovereign services need an additional check of legal jurisdiction, storage location and support access. A local marketing presence is not sufficient evidence of data control.

Sources

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