

GPU cloud terms for buyers

Read quotes with clear definitions of capacity and billing.

Updated October 29, 2026

Use these terms to clarify a quote. Ask the provider to confirm its own definitions in the order form.

Hardware and scheduling

GPU memory is the device memory available to the workload. Host memory is separate. A device model alone does not specify the memory variant or the host configuration. Interconnect describes the links between devices or hosts. It can affect a distributed job and must be tested with the intended workload.

A quota is permission to request a resource. It is not necessarily reserved capacity. A reservation may involve a capacity promise, a spending commitment or both. Ask which promise is in your contract.

Billing and execution

On-demand capacity is requested as needed under the provider's terms. Spot capacity can be interrupted; AWS's documentation describes its own Spot model. Do not apply its exact rules to another service.

Serverless describes an execution interface where the service manages allocation. It does not imply that startup is instant or that every supporting resource is included in the GPU rate. Modal's pricing page separates GPU and other compute charges.

A checkpoint is saved state used to resume work. Test that the state can be loaded from durable storage after the original worker is gone. Kubernetes GPU resource requests tell the scheduler which devices a workload needs. They do not state application throughput. Keep measured performance separate from scheduling and billing terms.

Sources

- [AWS EC2 Spot Instances](#) — fetched 2026-10-29.
- [Modal pricing](#) — fetched 2026-10-29.
- [Kubernetes GPU scheduling](#) — fetched 2026-10-29.

Frequently asked questions

Is this a vendor quote?

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