RESEARCH FIELD GUIDE / 01

BEYOND THE GPU-HOUR

The 2026 GPU Cloud Pricing Buyer's Guide

What compute actually costs, and how to compare competing offers on a like-for-like basis.



THE PRICE ON A CARD IS NOT THE COST OF A JOB.

15% **illustrative uplift from idle time alone**

230 billable hours / 200 productive hours

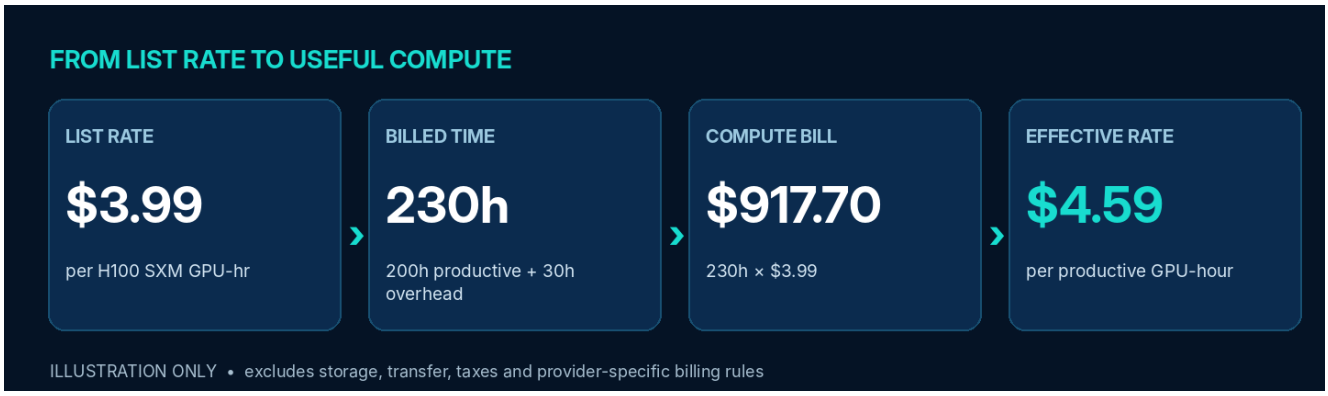
PUBLIC LIST PRICES / WORKLOAD MATH / QUOTE WORKSHEET

01 / THE COST METRIC

The GPU-hour is only the start

Convert every offer into an invoice and a completed result.

A posted rate tells you what is charged for a defined meter. It does not tell you how many hours the workload will occupy that meter, which resources are bundled, or what happens while the GPU is idle. The economic unit is the successful training run, validated batch, accepted inference request or useful output.



The calculation buyers should carry into every quote

Total workload cost = GPU/instance charges + mandatory storage + data movement + managed-service/support fees + minimum commitments + applicable tax. Effective cost per completed workload = total workload cost ÷ successful completed jobs.

For a single GPU, 200 productive hours and 30 additional billable hours at Runpod's published $3.99 H100 SXM Pods rate produce a $917.70 compute-only bill, or $4.59 per productive GPU-hour. This is a transparent arithmetic scenario, not an observed Runpod workload or a claim about its idle behavior. [1]

Three numbers to record separately

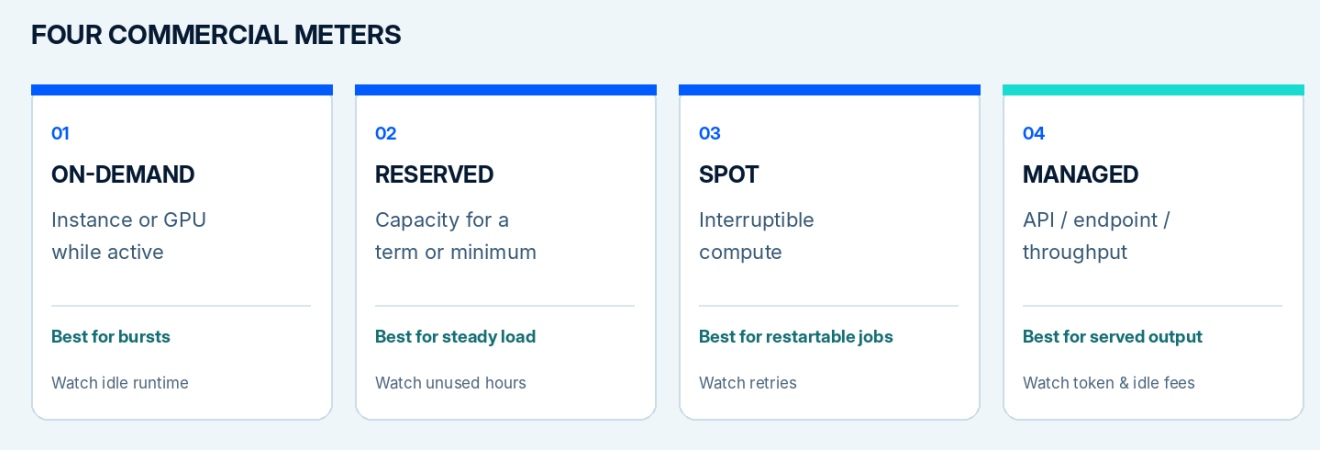
- Quoted rate: the unmodified provider amount, SKU, billing object, native currency, term and source date.
- Billable consumption: time or output actually charged, including setup, queue, pause, retry and idle rules.
- Completed-workload unit: successful runs, accepted tokens, images or jobs at agreed quality and latency.

BUYER RULE If a vendor cannot explain the billable object and the exact start/stop conditions, do not turn its headline into a comparable GPU-hour rate.

02 / COMMERCIAL STRUCTURE

What exactly are you buying?

The service boundary changes the meaning of every price.



A 1-GPU instance, an eight-GPU HGX node, a reserved cluster and a serverless model endpoint are different products. Dividing a node price by eight may be useful for inventory accounting, but only when all eight GPUs, the interconnect, CPU/RAM, storage and the minimum purchase are recorded. A managed inference price per million tokens should not be converted into GPU-hours without workload measurements.

The contract mechanics that change the winner

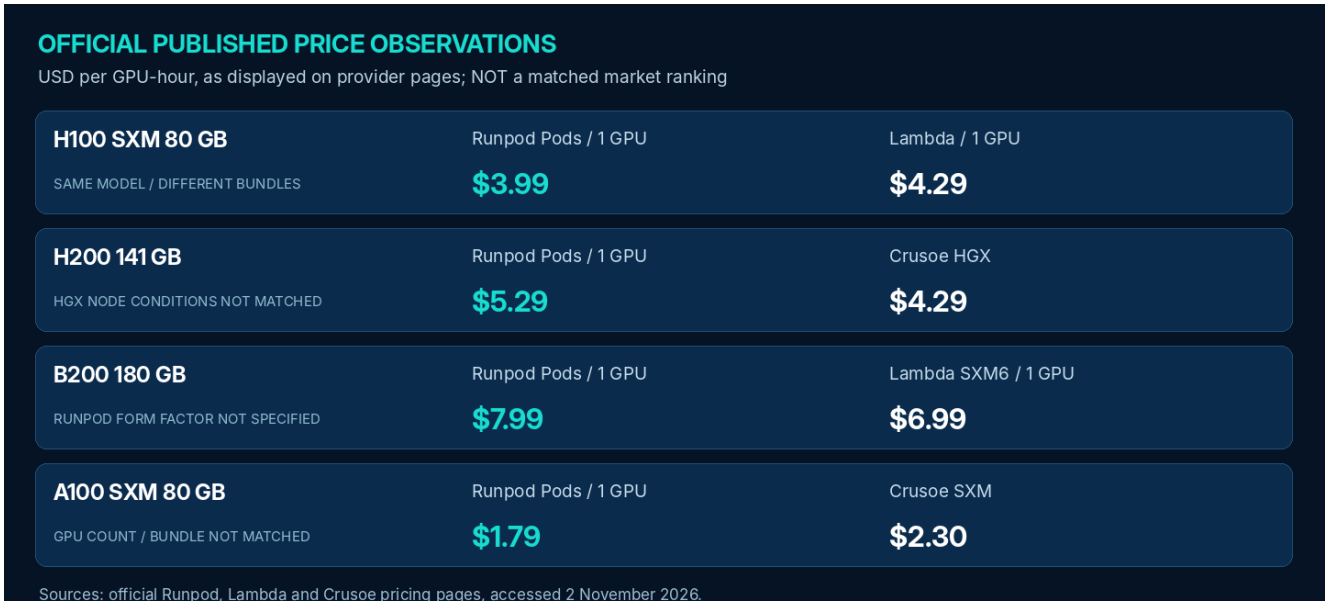
- On-demand: flexible when compute can be stopped. Check launch-to-termination billing, minimum increments and what survives shutdown.
- Reserved / dedicated: compare total obligation, minimum GPU count, prepayment, start date, replacement SLA, cancellation and renewals. A discount is valuable only if capacity is used.
- Spot / preemptible: lower price can be offset by lost work, restarts, checkpoints and missed deadlines. Record interrupted runs and final completed output.
- Managed inference: compare successful output at the same model, quality, context length, latency objective and provisioned-idle policy.

Break-even test: if a reserved rate is 80% of the on-demand rate, it needs more than 80% paid-hour utilization to beat perfectly stoppable on-demand compute, ignoring other charges. This is a ratio example, not an observed market discount.

03 / PRICING EVIDENCE

Published rates: four GPU families

These are first-party observations, not interchangeable quotes.



H100 has a narrow comparison here: the public 1-GPU H100 SXM listings from Runpod and Lambda are $3.99 and $4.29 per GPU-hour respectively. The GPU model and count align, but attached resources, runtime billing and service terms must still be reconciled. [1][2]

The H200, B200 and A100 rows deliberately retain their provider-specific labels. They expose real public rates for those GPU families while showing where topology or form-factor detail prevents a strict peer comparison. Their numbers must not be turned into a market average. [1][2][3]

Qualification gate before an offer enters a matched cohort

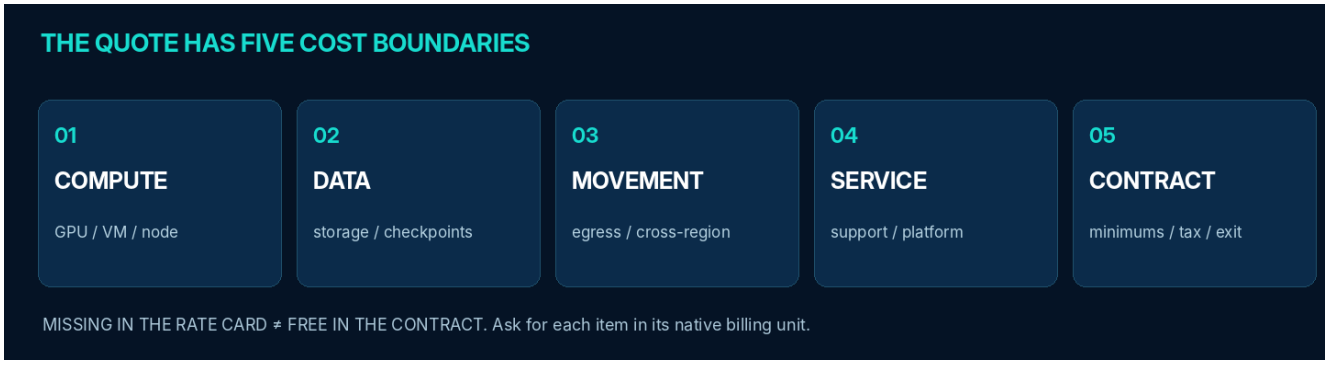
GPU model + memory/variant | physical GPU count or partition | sold object | CPU/RAM/storage and fabric | region | term | tax treatment | billing granularity | official dated source. If any essential field is missing, label the pair “directional” or “not comparable.”

Rate-card capture: 2 November 2026. Provider web pages can change and displayed prices may depend on page selectors, region or availability. Confirm the live order configuration before signing.

04 / TOTAL BILL

What the actual invoice includes

Unknown must remain unknown. It must not silently become zero.



Work backwards from the path of a single successful job

COST LINE	BUYER INPUT	QUESTION TO CLOSE
Compute	GPU-hours / instance-hours × price	When does the meter start and stop?
Storage	GiB-months, IOPS, snapshots, retention	Does storage persist after VM deletion?
Movement	GB/TiB by source, path and destination	Which egress, cross-zone and interconnect rules apply?
Software + support	Platform, endpoint, scheduler, support tier	Bundled, separately billed or pass-through?
Contract + tax	Minimum term, deposit, credit, tax	What is owed if utilization is zero?

Provider billing rules vary. Lambda documents that running on-demand instances are billed even when unused, while separate filesystems continue billing while they exist, including after an instance is terminated. Lambda also states no charges for filesystem ingress or egress. This is a documented exception to blanket claims that all clouds charge transfer fees. [4]

Useful sensitivity check: every hypothetical $0.01/GiB of charge on 2 TiB of transferred data adds $20.48. Do not treat the $0.01 as a provider tariff. Ask for the actual rate, allowances, tiers and destination before using it in a quotation.

05 / WORKED BUY

When the cheaper GPU-hour loses

A controlled comparison using two genuine provider list offers.

Offer A is Runpod's published 1-GPU H100 SXM Pods rate of $3.99/hour. Offer B is Lambda's published 1-GPU H100 SXM on-demand instance at $4.29/hour. These are actual public list offers, not negotiated quotations. We do not claim identical bundles, regions or performance. [1][2]

WHEN A HIGHER RATE CAN WIN

Same 1-GPU H100 SXM. Equal quality, no extra fees. Times are scenarios, not measured speeds.

RUNPOD	200 h	$3.99/h	$798.00
LAMBDA	186 h	$4.29/h	$797.94

Threshold: $3.99 / $4.29 = 0.9301. Lambda must use >6.99% fewer billed hours to undercut Runpod.

The decision is conditional, not a provider verdict

If both require 200 billable GPU-hours and incur no other differences, A costs $798 and B costs $858 in compute charges. If B requires 186 hours for the identical accepted workload, it costs $797.94, marginally below A's $798. No supplier benchmark supports the 186-hour premise; it is a break-even illustration. A buyer must measure or contract the workload result before using this outcome.

A three-case bid test

TEST	OFFER A	OFFER B
Same billed time	200 h × $3.99 = $798	200 h × $4.29 = $858
Break-even runtime	200 h / agreed quality	Less than 186.01 h at same quality
Unpriced obligations	Storage / network / tax: TBD	Storage / network / tax: TBD

Procurement question: "For this pinned model, dataset, image, region and completion standard, show the exact offer configuration, billable time, storage and data-movement line items, then reconcile the result with the usage export."

Also test a failed or interrupted job. Inference should use cost per successful request or accepted million tokens at the same quality and latency target, not the underlying GPU list rate.

06 / BUYER WORKSHEET

Make two quotations comparable

Copy these fields into an RFP or require both suppliers to complete them.

REQUIRED FIELD	SUPPLIER A	SUPPLIER B
Seller / actual infrastructure operator	_____	_____
Official quote ID / issue & expiry dates	_____	_____
GPU model, variant, VRAM, count / MIG	_____	_____
Sold object: GPU / VM / node / endpoint	_____	_____
Topology, network fabric, CPU/RAM	_____	_____
Included / persistent storage	_____	_____
Region, residency, start date / SLA	_____	_____
Native price, currency, unit, tax status	_____	_____
On-demand / spot / reserved; duration	_____	_____
Billing increment, idle, pause, retry rules	_____	_____
Minimums, deposits, exit, renewal	_____	_____
Egress / cross-zone / data-movement fees	_____	_____
Support, platform and extra service charges	_____	_____
Proven billable hours / accepted output	_____	_____
All-in cost / completed workload	_____	_____

Reject, qualify, then select

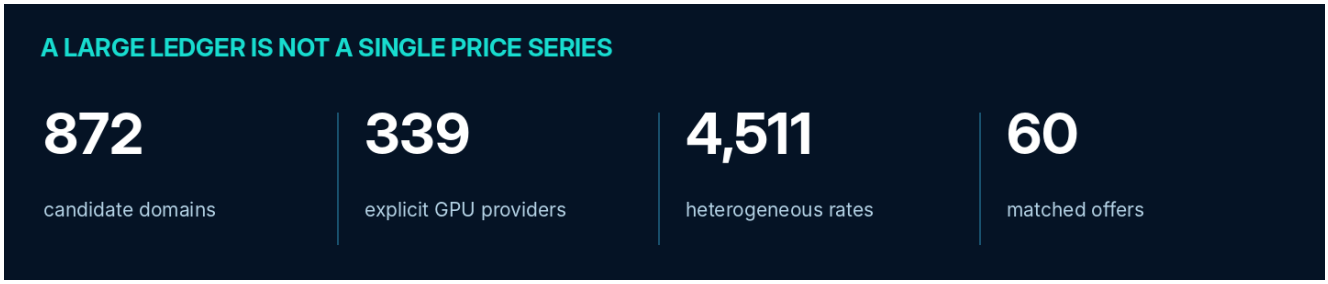
PASS / FAIL Validate mandatory model, region, security, capacity date, tenancy and deadline first. COMPARABILITY Mark exact match / directional / not comparable. ECONOMICS Compare the same low/base/high workload case and fixed contract obligation. ACCEPTANCE Reconcile an independent proof-of-value run against the usage report and first invoice.

Attach the supplier's original quotation, source capture and all commercial schedules. Do not overwrite the native billing object when making a normalized calculation.

07 / FINDINGS & METHOD

What the evidence can support

Independent buyer intelligence with explicit limits.



ViaBandwidth's October checkpoint reviewed 872 candidate domains, retained 722 active candidates and explicitly labeled 339 GPU service-provider domain rows. The private ledger contains 4,511 heterogeneous price representations across 221 domains; 216 retained domains had numeric or indicative public pricing. None of those denominators represents one comparable global price sample. [5]

Five findings for buyers

- The purchase unit is the first filter. GPU, pod, whole node, reservation and API output are not equivalent rate objects.
- Hourly price is a starting point. Completion time, utilization and failure/retry behavior determine effective compute cost.
- Mandatory storage, transfer, software, support, taxes and fixed obligations decide invoice-level economics.
- No price number proves stock, guaranteed start date or supplier-controlled infrastructure.
- Public prices are leads. A quote is decision-grade only when the complete configuration and terms are contractable.

Evidence and limitations

Historical context only: 60 strictly matched Q2-to-October offers from 10 providers included 41 unchanged, 15 higher and 4 lower. This narrow matched set is not a claim of market-wide inflation. No benchmark or completed-workload timings were measured for this paper. All example job hours are expressly illustrative. [5]

SOURCES [1] Runpod official pricing, runpod.io/pricing (28 Oct page update; accessed 2 Nov 2026). [2] Lambda official pricing, lambda.ai/pricing (accessed 2 Nov 2026). [3] Crusoe official pricing, crusoe.ai/cloud/pricing (accessed 2 Nov 2026). [4] Lambda billing documentation, docs.lambda.ai/public-cloud/billing/ (accessed 2 Nov 2026). [5] ViaBandwidth October checkpoint-162 internal aggregate and State of GPU Services Q4 2026 early-quarter edition.

Method: use original official list-rate values as printed; retain native object, GPU quantity, form factor, time unit, date and unknowns; never infer a synthetic price for unqualified configurations. The underlying proprietary domain and pricing ledgers are not published.