



GPU Cloud SLA Decoder: Uptime, Credits and Capacity

Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-10-15 · gpu cloud sla · gpu cloud uptime · service credits · capacity reservations · cloud procurement · gpu infrastructure · availability guarantees · maintenance exclusions · sla comparison

Original article: <https://gpusmith.com/articles/en/gpu-cloud-sla-decoder>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Key Changes in How a GPU SLA Should Be Read
 4. Public SLA Clause Matrix
 5. Credit Mechanics, Maintenance and Claim Evidence
 6. GPU Workload Gaps and Acceptance Criteria
 7. Implementation Considerations and Process Changes
 8. Data Analysis and Evidence
 9. Implications and Future Directions
 10. Conclusion
-

Executive Summary

A **GPU cloud service level agreement (SLA)** should be read as a conditional remedy for a defined service failure. It does not, by itself, establish when a requested graphics processing unit (GPU) cluster will arrive, how quickly a training job will finish, or how much business loss will be reimbursed. The distinction starts with scope: **Amazon EC2** targets **99.99%** for qualifying regional deployments and **99.5%** per instance. (aws.amazon.com) **Google Compute Engine's** Premium Tier targets are **99.99%** for most qualifying multi-zone deployments and **99.9%** for single non-memory-optimized VMs. (cloud.google.com) These are conditional commitments, with different boundaries and eligibility requirements.

GPU eligibility needs a separate check. Google's accelerator documentation explicitly excludes **A4X Max and A4X** from its Compute Engine SLA. (docs.cloud.google.com) **Vultr** explicitly excludes GPU instances from its standard SLA, while describing a separate requested downtime refund for hardware failure. (www.vultr.com) Conversely, **Nebius** includes specified GPU and fabric failures but excludes degradation without connectivity loss. (docs.nebius.com) A percentage alone cannot express these differences.

The recoverable amount depends on **qualifying downtime, eligible charges, the applicable tier, and a valid claim**. AWS excludes specified upfront payments. (aws.amazon.com) **DigitalOcean** requires written notification within **24 hours**, even though its credit-request window extends further. (www.digitalocean.com) **CoreWeave's** public terms require notification within **ten days** of eligibility and cap aggregate credits at **50%** of covered-service monthly charges. (docs.coreweave.com) Maintenance must also be read across documents: CoreWeave's separate policy permits unplanned maintenance without notice. (docs.coreweave.com)

This report therefore supplies a term decoder, a source-linked comparison, and a reproducible worksheet rather than a provider ranking. In the **Hypothetical Example** below, qualifying downtime of **20 minutes** in a **30-day month** produces **99.953704%** uptime. Applying AWS's regional schedule to hypothetical eligible spend of **\$48,000** produces **\$4,800** credit, versus **\$0** under its single-instance threshold. (aws.amazon.com) The arithmetic illustrates scope, not an actual entitlement. Capacity needs its own review: AWS distinguishes

regional discounts from capacity reservations, (docs.aws.amazon.com) and Azure documents a separate capacity-reservation SLA. (learn.microsoft.com) Before signing, procurement should require the workload boundary, allocation obligation, evidence procedure, exclusions, and remedy to agree.

Introduction and Background

Enterprise GPU procurement often begins with an uptime percentage and a reservation proposal. The more useful starting point is a named workload and its consequences: an inference endpoint cannot serve requests, a distributed training run cannot communicate, or a replacement cluster cannot launch. These situations can require different contractual evidence. For example, AWS measures individual-instance external connectivity; (aws.amazon.com) Nebius includes specified GPU and interconnect states. (docs.nebius.com) Neither definition should be silently replaced with the buyer's own description of an unsuccessful job.

This report takes an engineering and procurement perspective. **GPU Smith** describes its scope as specifying, procuring, integrating, and validating private AI infrastructure. That makes written service boundaries and acceptance criteria an appropriate perspective for the analysis; the consultancy is not presented as a competing GPU cloud in the comparison. (gpusmith.com) The objective is to turn an advertised commitment into an auditable decision: what failed, whether that failure counts, which charges qualify, and what must be submitted to recover a remedy.

The assessment is anchored to **October 15, 2026** and records the dates displayed on the public sources. AWS's policy is labeled **May 25, 2022**; (aws.amazon.com) Google's is labeled **October 14, 2026**. (cloud.google.com) A recent retrieval date does not make an older document a newly negotiated enterprise promise. Microsoft's licensing directory identifies an **October 2026** consolidated SLA edition, but its downloadable contractual text could not be reliably extracted during this research. Consequently, this report does not reproduce Azure VM credit tiers from a marketing headline or an older edition. (www.microsoft.com)

The evidence boundary matters equally for private terms. A public website cannot establish which amendments, service descriptions, or precedence clauses govern a particular account. Buyers should retain the signed agreement and incorporated documents alongside this worksheet. Counsel should assess contract-specific interpretation and enforceability; this is informational engineering and procurement analysis, not legal advice. No professional review or private order-form access is represented.

Key Changes in How a GPU SLA Should Be Read

The changes below are changes to the buyer's review process, rather than claims that every provider recently revised its contract.

Separate the contractual layers

Treat the following as distinct commitments, even when a sales proposal bundles them:

- **Availability:** the service state that qualifies as downtime and the remedy for missing its objective.
- **Support:** the time to acknowledge or initially respond to a case, with a defined severity and coverage schedule.

- **Performance:** the measured throughput, latency, bandwidth, or job-completion condition, if expressly committed.
- **Capacity and delivery:** the specified resources, location, allocation date, duration, and replacement obligation.
- **Maintenance:** permitted intervention, notice, disruption limits, and its treatment in the uptime calculation.

The separation is observable in public documentation. **Lambda's 1-Click Cluster support** page specifies a **four-hour** initial response for severity-one cases. That is a response commitment, not a statement that service will be restored within the same interval. ([docs.lambda.ai](https://docs.aws.amazon.com/lambda/latest/dg/lambda-1-click-cluster-support.html)) AWS separately documents capacity reservations, and CoreWeave publishes maintenance rules apart from its compute credit schedule. ([docs.aws.amazon.com](https://docs.aws.amazon.com/lambda/latest/dg/lambda-1-click-cluster-support.html)) (docs.coreweave.com)

Replace the headline with a service boundary

A **virtual machine (VM)**, a server node, a region, and a distributed job are different measurement objects. Google's qualifying multi-zone definition requires all applicable running instances to be affected; a failed member of a training group need not establish that regional condition. (cloud.google.com) DigitalOcean's GPU SLA instead measures an individual **server node**, explicitly distinguishing that object from a cluster. (www.digitalocean.com) The buyer should identify the smallest workload failure that matters and compare it with the contractual object.

Check the exact accelerator configuration

A GPU product name does not establish coverage under a general compute policy. Google's express A4X exclusions illustrate why the series and attached configuration need checking. (docs.cloud.google.com) Its SLA also excludes pre-general-availability features unless the associated documentation provides otherwise. (cloud.google.com) Vultr's published GPU exclusion likewise prevents applying its standard uptime schedule to a GPU instance without separate authority. (www.vultr.com) Record the stock keeping unit (SKU), region, network tier, storage configuration, purchasing mode, and incorporated service description before selecting a target percentage.

Separate diagnostic evidence from a performance promise

NVIDIA Data Center GPU Manager (DCGM) distinguishes health monitoring based on ordinary telemetry from diagnostics that actively exercise the system. Its multi-node documentation also warns that a successful management connection or distributed launch does not establish a working GPU data path. (docs.nvidia.com) (docs.nvidia.com) These observations support better **acceptance tests**. They do not create a provider obligation to meet a buyer-selected bandwidth target.

Read allocation promises independently

An **AWS Regional Reserved Instance** does not reserve capacity, whereas a zonal reservation has different capacity treatment. (docs.aws.amazon.com) Azure similarly distinguishes reservation discounts from capacity reservations and documents a separate capacity SLA for accepted, supported configurations.

(learn.microsoft.com) Therefore, the question is not simply whether a contract contains the word “reserved.” It is what resource the provider must make available, when, where, and under which acceptance and remedy terms.

Public SLA Clause Matrix

Table 1 summarizes selected public terms. Dates are source labels, not dates of negotiated acceptance. Percentages are service level objectives (SLOs), not observed reliability. Multi-zone and multi-region requirements are preserved as different concepts.

Public source and date	Covered object and stated objective	Downtime or eligibility boundary	Credit basis and published percentages	Claim clock and qualification	Capacity or delivery in this comparison
AWS Compute , May 25, 2022	Qualifying regional deployment 99.99% ; individual instance 99.5% . (aws.amazon.com)	Regional condition requires concurrent connectivity loss across qualifying instances; individual condition is no external connectivity. (aws.amazon.com)	Affected regional or instance monthly EC2 charges, excluding specified upfront payments; 10%, 30%, 100% . (aws.amazon.com)	End of the second billing cycle after the incident; regional and instance claims cannot stack. (aws.amazon.com)	Separate capacity-reservation documentation; verify supported GPU configuration. (docs.aws.amazon.com)
Google Compute Engine , October 14, 2026	Premium multi-zone 99.99% , except Stockholm 99.95% ; single non-memory-optimized instance 99.9% . (cloud.google.com)	Connectivity or persistent-disk access; explicit product exclusions still apply. (docs.cloud.google.com)	Affected covered-service regional bill or individual-instance bill; 10%, 25%, 100% , with scope-specific bands. (cloud.google.com)	Notify support within 60 days of eligibility and provide qualifying downtime logs. (cloud.google.com)	Separate reservation mechanism, subject to creation and matching conditions. (docs.cloud.google.com)
CoreWeave public terms , June 30, 2022	Multiple-region instances 99.9% ; single-region instances 99% ; covered-service and location conditions apply. (docs.coreweave.com)	Public text uses container, disk-access, and connectivity definitions; advance-communicated maintenance is excluded. (docs.coreweave.com)	Affected regional service or single-instance monthly charges; 10%, 25%, 50% . (docs.coreweave.com)	Notify within ten days of eligibility; timestamped logs required; aggregate 50% cap. (docs.coreweave.com)	Named GPU allocation and delivery not established by the cited public SLA.
Nebius Compute , October 6, 2026	Ordinary VMs and rack-scale GPU systems measured separately; 99.50% . (docs.nebius.com)	Specified GPU, boot-disk, networking, and interconnect states; rack-scale maintenance-buffer provisions apply. (docs.nebius.com)	Affected-resource fees; published 10%, 15%, 30% tiers. (docs.nebius.com)	General SLA requires console submission within 14 calendar days after the reporting period. (docs.nebius.com)	Named GPU allocation and delivery not established by the cited public SLA.

Public source and date	Covered object and stated objective	Downtime or eligibility boundary	Credit basis and published percentages	Claim clock and qualification	Capacity or delivery in this comparison
DigitalOcean GPU Droplets , October 21, 2025	Individual GPU server node 99% . (www.digitalocean.com)	Five-minute intervals without external connectivity; partial-month denominator follows node existence. (www.digitalocean.com)	Affected-node monthly charges; 10%, 25%, 50% . (www.digitalocean.com)	Written notice within 24 hours ; credit request within three months after the billing cycle. (www.digitalocean.com)	Named GPU allocation and delivery not established by the cited public SLA.
Scaleway GPU Instances , no version date found	GPU instance monthly objective 99.5% . (www.scaleway.com)	At least four continuous minutes without external connectivity; performance degradation and maintenance excluded. (www.scaleway.com)	Affected-resource invoice voucher; 10%, 25%, 100% , subject to the table's boundary wording. (www.scaleway.com)	Within 30 days following the affected month, with dated and timed unavailability logs. (www.scaleway.com)	Hardware stock shortage is an express exclusion. (www.scaleway.com)

The matrix should be read across each row. It does not normalize different failure definitions into a ranking. CoreWeave's older public text has location exclusions and uneven boundary notation; Scaleway's credit table uses strict inequalities that leave exact equality cases unstated. Buyers should obtain written clarification instead of silently repairing those boundaries. (docs.coreweave.com) (www.scaleway.com) The examples later avoid those ambiguous schedules.

Other public evidence belongs outside a like-for-like numerical comparison. **Crusoe's** July 21, 2026 SLA states **99.5%** per GPU-enabled VM and **10%, 25%, 100%** credit tiers; its replacement-machine and maintenance rules must accompany that percentage. (legal.crusoe.ai) **Runpod** advertises an enterprise **99.99%** SLA, but the fetched product page does not publish a full credit schedule. (www.runpod.io) **OVHcloud's** worldwide GPU page advertises **99.99%**, while the separately fetched US agreement has its own scope and procedures. Incorporation into the buyer's geographic contract must be verified. (www.ovhcloud.com) (us.ovhcloud.com)

Credit Mechanics, Maintenance and Claim Evidence

Calculate the remedy on eligible spend

A credit calculation has distinct inputs: the denominator, qualifying downtime, credit band, eligible charges, and cap. The entire invoice should never be substituted automatically for the affected-service base. AWS excludes specified upfront payments. (aws.amazon.com) A separately billed storage or network service may need its own contractual analysis.

The payment mechanism also affects value. Google describes financial credits against future covered-service use. CoreWeave caps aggregate monthly credits at **50%** and applies them to future usage. (cloud.google.com) (docs.coreweave.com) DigitalOcean's GPU credits expire after **90 days** and provide no cash refund.

(www.digitalocean.com) A buyer planning to migrate should model whether the credit could actually be consumed.

Keep **contractual credit**, **billing adjustment**, and **business impact** in separate fields. AWS generally supplies future credits as its exclusive SLA remedy, subject to the agreement. (aws.amazon.com) By contrast, Vultr describes a requested GPU downtime refund following hardware failure, despite excluding GPUs from its standard SLA. (www.vultr.com) These are different mechanisms, not interchangeable percentages.

Read maintenance permission and credit exclusions together

CoreWeave's maintenance policy allows unplanned work without notice and describes at least **24 hours'** notice for planned critical infrastructure maintenance. Users are responsible for subscribing to its status notifications. (docs.coreweave.com) (docs.coreweave.com) (docs.coreweave.com) The separate public compute terms exclude advance-communicated maintenance. Neither the notice policy alone nor permission to perform emergency work establishes the credit result for every event.

Nebius's general SLA specifies maintenance exclusions and an **eight-hour** per-period node-maintenance allowance, with provisions for excess time and failure recovery. Planned cluster upgrades have a separate notification provision. (docs.nebius.com) DigitalOcean explicitly excludes maintenance and other scheduled interruptions; Scaleway excludes maintenance, performance degradation, short interruptions, and hardware unavailability caused by stock shortage. (www.digitalocean.com) (www.scaleway.com) The relevant question is which exact exclusion applies to the recorded event.

Treat claim preparation as an operational process

Notice deadlines can be shorter than the final credit-request window. Under DigitalOcean's GPU Droplet SLA, downtime begins accruing when the customer reports the outage, subject to DigitalOcean's confirmation, or acknowledges DigitalOcean's proactive outage report. Report an outage immediately or promptly acknowledge DigitalOcean's report; an earlier monitoring timestamp alone does not start the contractual downtime clock. ([DigitalOcean GPU Droplet SLA, clause 3](#)) The separate **24-hour** written notification requirement still applies. (www.digitalocean.com) Crusoe requires notification within **30 days** of eligibility and timestamped logs or monitoring data. (legal.crusoe.ai) OVHcloud US requires requests within **60 calendar days**, and its unavailability clock starts when the customer opens a support case. (us.ovhcloud.com) (us.ovhcloud.com)

A claim packet should contain:

- **Contract identity:** accepted agreement, incorporated SLA, amendments, and service version.
- **Resource identity:** account, region, zone, instance or node identifiers, and configuration.
- **Clock evidence:** synchronized timestamps, time zone, start, end, and monitoring interval.
- **Failure evidence:** raw connectivity, disk, GPU, or fabric observations matching the definition.
- **Exclusion analysis:** maintenance notices, customer changes, and provider explanations.
- **Support record:** severity, case number, first notice, acknowledgments, and updates.
- **Financial base:** affected invoice lines, excluded charges, currency, and calculation.
- **Submission proof:** required form, attachments, submission time, and confirmation.

This is a recommended common dossier, not a claim that every provider demands identical fields. AWS requires request logs and validation data. (aws.amazon.com) Scaleway calls for logs identifying unavailability periods with dates and times. (www.scaleway.com)

GPU Workload Gaps and Acceptance Criteria

An SLA can be satisfied while a workload misses its own acceptance criteria. The contract may measure external connectivity, while the application depends on collective communication, storage throughput, checkpoint recovery, and [replacement capacity](#). Nebius makes the distinction explicit by including specified interconnect failures while excluding degraded performance without connectivity loss. (docs.nebius.com) NVIDIA likewise distinguishes a working management path from a working GPU data path. (docs.nvidia.com)

Table 2 maps workload consequences to evidence and proposed terms. The clause requests are analytical recommendations, not representations of a provider's existing commitments.

Workload consequence	Evidence to retain	Clause or acceptance condition to request	Public-document boundary illustrated
Requested GPU configuration cannot launch	Requested SKU, zone, quota, reservation identifier, allocation response	Named quantity, configuration, allocation date, and shortfall remedy	Google checks capacity before creating a reservation; requested future capacity can be only partially provisioned. (docs.cloud.google.com) (docs.cloud.google.com)
Training runs but communication is degraded	Fixed test configuration, topology, collective size, bandwidth and latency results	Reproducible fabric acceptance threshold and retest procedure	NVIDIA warns that successful distributed launch does not prove the GPU data path. (docs.nvidia.com)
GPU or interconnect reports a fault	Raw diagnostic output, target identifiers, debug logs, relevant statistics	Counted hardware states, observation interval, replacement and restart responsibilities	NVIDIA recommends preserving complete diagnostic evidence; diagnostics do not repair faults. (docs.nvidia.com) (docs.nvidia.com)
Job cannot resume from checkpoint	Saved state inventory, object identifiers, restore attempt and result	Checkpoint destination, persistence, restore acceptance, and responsible operator	PyTorch says resuming training requires more than model parameters, including optimizer state. (docs.pytorch.org)
Maintenance interrupts the serving fleet	Notice, affected resources, eviction and application health records	Maintenance notice, permitted disruption, drain procedure, and credited exceptions	Kubernetes disruption budgets cannot prevent involuntary disruptions. (kubernetes.io)

The mapping separates measurement from entitlement. A failed diagnostic can support a hardware investigation without proving that a contractual downtime threshold was met. A completed diagnostic can also leave performance qualification unresolved: NVIDIA's multi-node guidance calls for comparison against a known-good system or site requirement for the relevant tests. (docs.nvidia.com)

Maintenance planning needs workload-specific recovery evidence. Google's GPU documentation states that attached **Local SSD** data is lost when an instance stops for host maintenance. (docs.cloud.google.com) A checkpoint strategy should therefore specify where state is written and demonstrate restoration from that

destination. PyTorch's guidance establishes that a resumable checkpoint needs additional state; it does not establish the cloud storage's durability or a provider's recovery obligation. (docs.pytorch.org)

Likewise, a **Kubernetes PodDisruptionBudget (PDB)** is an orchestration mechanism, not a cloud credit contract. Involuntary disruptions still count against its budget, and direct pod or deployment deletion can bypass it. (kubernetes.io) (kubernetes.io) The maintenance procedure should identify who uses the eviction mechanism, who validates application recovery, and how a missed recovery condition is escalated. Availability, support response, checkpoint recovery, and performance should each retain their own pass condition.

Implementation Considerations and Process Changes

Procurement should start with an acceptance record that engineering can reproduce and finance can reconcile. GPU Smith's public site describes its position as independent engineering without a cloud of its own. (gpusmith.com) In that advisory framing, the useful deliverable is a traceable decision about the agreement and workload, rather than a universal score for the provider.

A practical review sequence is:

- **Name the workload:** specify the training run, batch pipeline, or inference endpoint being protected.
- **Map dependencies:** identify compute, fabric, persistent storage, control plane, and external access.
- **Select the object:** distinguish instance, node, zone, region, rack, and application measurement.
- **Verify eligibility:** match GPU series, location, network tier, storage, and purchase mode.
- **Fix the denominator:** document calendar month, billing cycle, or resource-existence period.
- **Translate exclusions:** list which planned work, short events, and customer actions will not count.
- **Define acceptance:** attach executable checks with configuration, baseline, and output format.
- **Allocate responsibility:** assign monitoring, notice subscriptions, case creation, and claim ownership.
- **Model the remedy:** calculate eligible charges, tier, cap, expiry, and usable future credit.
- **Confirm precedence:** obtain the controlling order form and all incorporated service descriptions.

These steps should produce an agreed evidence vocabulary. NVIDIA recommends retaining the diagnostic command, selected targets, complete output, debug log, and relevant plugin statistics when escalating failures. (docs.nvidia.com) The buyer can extend that pattern to application probes and restore tests, while respecting the provider's actual claim requirements.

Negotiation should also specify **when downtime ends**. Crusoe's public hardware-failure measurement can end when a replacement machine becomes available; Nebius likewise has a replacement-host notification boundary, with a continued-unavailability exception for provider-attributable restart failure. (legal.crusoe.ai) (docs.nebius.com) A workload may still need provisioning, data restoration, process startup, and validation afterward. A proposed recovery commitment should state which of those activities is included, who performs it, and what evidence closes the event.

Support severity must be actionable. Lambda distinguishes infrastructure and networking support from customer code, later-installed third-party applications, and customer network or virtual private network connections. (docs.lambda.ai) Procurement should map those boundaries to the internal on-call team. A fast

provider acknowledgment is of limited operational use if no owner can supply the required logs or authorize recovery.

Finally, keep an operational calendar alongside the contract. Record notice obligations immediately, prepare the claim before the deadline, and preserve an immutable copy of submitted evidence. OVHcloud US states that provider records are its sole validation source. (us.ovhcloud.com) Where customer and provider monitoring differ, the agreement should explain reconciliation and access to supporting records. This is a negotiation request; the report does not assume that every provider offers it.

Data Analysis and Evidence

Reproduce uptime before selecting a credit band

For a simple minute-based contract, use:

$$\text{Uptime} = 1 - \text{qualifying downtime minutes} / \text{denominator minutes}$$

“Qualifying” is essential. Google’s formula uses calendar-month minutes, and partial or intermittent downtime below **one minute** does not count toward a downtime period. (cloud.google.com) (cloud.google.com)

Scaleway assumes availability during the part of a month when resources are not running, while DigitalOcean uses the node’s existence period for partial months. (www.scaleway.com) (www.digitalocean.com) These denominator differences can change results even when observed events are identical.

The following **Hypothetical Example** assumes a complete **30-day month**, hence **43,200 minutes**, an established covered deployment, no disputed observations, and hypothetical eligible charges of **\$48,000**. The regional rows use AWS’s published credit schedule. The single-instance row uses its separate threshold. No actual customer event or entitlement is asserted. (aws.amazon.com) (aws.amazon.com)

Table 3 shows the worksheet outputs. “Excluded” means observed minutes removed under an assumed applicable exclusion; it does not reduce this example’s full-month denominator.

Hypothetical scenario	Denominator minutes	Observed downtime	Excluded minutes	Qualifying minutes	Uptime	Eligible spend	Credit rate	Credit
Regional, qualifying deployment	43,200	20	0	20	99.953704%	\$48,000	10%	\$4,800
Single instance, same event	43,200	20	0	20	99.953704%	\$48,000	0%	\$0
Regional, longer event	43,200	500	0	500	98.842593%	\$48,000	30%	\$14,400
Regional, assumed excluded portion	43,200	500	100	400	99.074074%	\$48,000	10%	\$4,800

The first two rows isolate **scope**. The last two isolate **excluded downtime**. They do not imply that AWS automatically excludes the assumed minutes: the buyer must establish the actual exclusion. The examples also assume that every billed amount shown is eligible, the applicable cap does not reduce the result, and the

claim is submitted correctly.

Spreadsheet-ready worksheet

Copy this tab-separated block into a spreadsheet starting at cell A1. Inputs are deliberately hypothetical. Format F and H as percentages and G and J as currency. The formula in H implements the AWS regional bands; it is not a universal provider schedule. (aws.amazon.com)

Scenario	DenominatorMinutes	ObservedMinutes	ExcludedMinutes	QualifyingMinutes	Uptime	
EligibleSpend	CreditRate	CreditCap	Credit			
Hypothetical	regional	43200	500	100	=MAX(0,C2-D2)	=1-E2/B2
					48000	48000
						=MIN(G2*H2,I2)

The expected result is **400 qualifying minutes**, **99.074074% uptime**, and **\$4,800 credit**. For the AWS individual-instance schedule, replace the highest threshold in H2 with **0.995**; preserve the lower bands. (aws.amazon.com) Calculate using unrounded values, then format the displayed uptime.

Before using the worksheet:

- **Deduplicate events:** overlapping observations should not create duplicate downtime minutes.
- **Apply the event rule:** minimum continuous duration and measurement granularity come before aggregation.
- **Verify the base:** use eligible affected charges rather than automatically using the entire invoice.
- **Set the cap:** enter the applicable monetary cap and consider other credited events.
- **Track the deadline:** the sheet calculates an amount, not a valid submission.
- **Check precision:** compare unrounded uptime with the exact contractual tier boundary.
- **Retain provenance:** attach the source version, evidence locations, and exclusion rationale.

The spreadsheet is intentionally transparent. Partial-month operation, multiple affected services, ambiguous tier boundaries, and negotiated definitions require different inputs or formulas. Those changes should be explicit and independently checked rather than hidden inside a single uptime percentage.

Implications and Future Directions

The strongest procurement distinction is between **available service**, **reserved capacity**, and **delivered capacity**. AWS documents reservations for capacity in a specified availability zone and separately explains that regional reservation discounts do not reserve capacity. (docs.aws.amazon.com) (docs.aws.amazon.com)

A buyer should therefore request the actual reservation mechanism, supported configuration, quantity, and start condition, rather than infer delivery from a commercial commitment.

Google verifies capacity before creating an on-demand reservation, and consumption requires matching properties, including GPU type and count. (docs.cloud.google.com) (docs.cloud.google.com) Its future-reservation documentation distinguishes review and approval from provisioning and describes partial provisioning. (docs.cloud.google.com) The order form should state whether the obligation is to consider a request, approve it, allocate resources, or deliver an accepted cluster by a specified date. Each is a different milestone.

Azure's capacity-reservation overview documents a separate capacity SLA but limits supported configurations, identifying **ND-series and NCads H100 v5-series** among unsupported series. (learn.microsoft.com) Its VM reliability guidance separately associates higher uptime commitments with qualifying multi-VM deployment patterns. (learn.microsoft.com) A capacity claim and an availability claim therefore require separate configuration checks. No Azure credit percentage is inferred here from the overview.

Product descriptions are useful discovery evidence, but they leave procurement work unfinished. Runpod's enterprise page combines reserved baseline capacity, committed-use pricing, uptime, and support in one offering description. (www.runpod.io) The buyer still needs the credit denominator, exclusions, delivery obligation, evidence requirements, and governing documents. Similarly, OVHcloud's worldwide GPU availability statement should not be treated as proof that its separately published US procedures govern every geography. (www.ovhcloud.com) (us.ovhcloud.com)

Future negotiations should make these layers easier to audit: machine-readable resource inventories, agreed event timestamps, retained diagnostics, and explicit acceptance records. This is a proposed direction, not a claim of industry-wide adoption. The practical objective is to let infrastructure, finance, procurement, and counsel reproduce the same conclusion from the same contract and event record.

Conclusion

A useful GPU cloud SLA review ends with a reproducible answer to a specific workload question. The buyer should be able to identify the covered configuration, point to the qualifying failure definition, calculate the correct denominator, explain every exclusion, identify the affected charges, and prepare a claim that follows the governing procedure.

The public documents examined here show why a headline percentage cannot perform all of those functions. Availability objects differ, GPU eligibility can be explicit, maintenance can be governed across separate documents, and a response commitment is distinct from service restoration. Capacity reservation, delivery acceptance, performance qualification, and checkpoint recovery each need their own treatment.

The purchasing decision should therefore follow a compact test: **does the agreement cover the failure that matters, and is its remedy usable?** If either answer is unclear, procurement has a concrete clarification to request before signing. A named allocation milestone, a measurable fabric condition, a defined replacement boundary, or an agreed evidence procedure is more decision-useful than an additional unexplained uptime digit.

The worksheet provides the financial bridge. Its hypothetical calculations show how scope, qualifying minutes, and affected spend change the recoverable amount without pretending to measure business loss. Preserve those inputs, the applicable document version, and the claim evidence together. The result is an SLA assessment that engineering can reproduce, finance can reconcile, and procurement can compare with the obligations actually offered in the signed agreement.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://aws.amazon.com/compute/sla/>
2. <https://cloud.google.com/compute/sla>
3. <https://docs.cloud.google.com/compute/docs/accelerator-optimized-machines>
4. <https://www.vultr.com/legal/sla/>
5. <https://docs.nebius.com/legal/sla-levels/compute>
6. <https://www.digitalocean.com/sla/gpu-droplets>
7. <https://docs.coreweave.com/policies/terms-of-service>
8. <https://docs.coreweave.com/policies/terms-of-service/maintenance-policy>
9. <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/reserved-instances-scope.html>
10. <https://learn.microsoft.com/en-us/azure/virtual-machines/capacity-reservation-overview>
11. <https://gpusmith.com/>
12. <https://www.microsoft.com/licensing/docs/view/Service-Level-Agreements-SLA-for-Online-Services?lang=1>
13. <https://docs.lambda.ai/public-cloud/1-click-clusters/support/>
14. <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-reservations.html>
15. <https://docs.nvidia.com/datacenter/dcgm/latest/learn/modules/dcgm-diagnostics.html>
16. <https://docs.nvidia.com/datacenter/dcgm/latest/learn/modules/dcgm-multinode-diagnostics.html>
17. <https://gpusmith.com/articles/en/gpu-cluster-acceptance-testing-runbook>
18. <https://docs.cloud.google.com/compute/docs/instances/reservations-overview>
19. <https://docs.nebius.com/legal/sla>
20. <https://www.scaleway.com/en/virtual-instances/sla/>
21. <https://legal.crusoe.ai/>
22. <https://www.runpod.io/enterprise>
23. <https://www.ovhcloud.com/en/public-cloud/gpu/>
24. <https://us.ovhcloud.com/legal/sla/public-cloud/>
25. <https://gpusmith.com/articles/en/gpu-cluster-reliability-budget>
26. <https://docs.cloud.google.com/compute/docs/instances/future-reservations-overview>
27. https://docs.pytorch.org/tutorials/beginner/saving_loading_models.html
28. <https://kubernetes.io/docs/concepts/workloads/pods/disruptions/>
29. <https://docs.cloud.google.com/compute/docs/gpus/about-gpus>
30. <https://learn.microsoft.com/en-us/azure/reliability/reliability-virtual-machines>

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the [hardware library](#), [GPU references](#), [networking references](#) and [vendor research](#). [Contact GPU Smith](#) with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: <https://gpusmith.com>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.