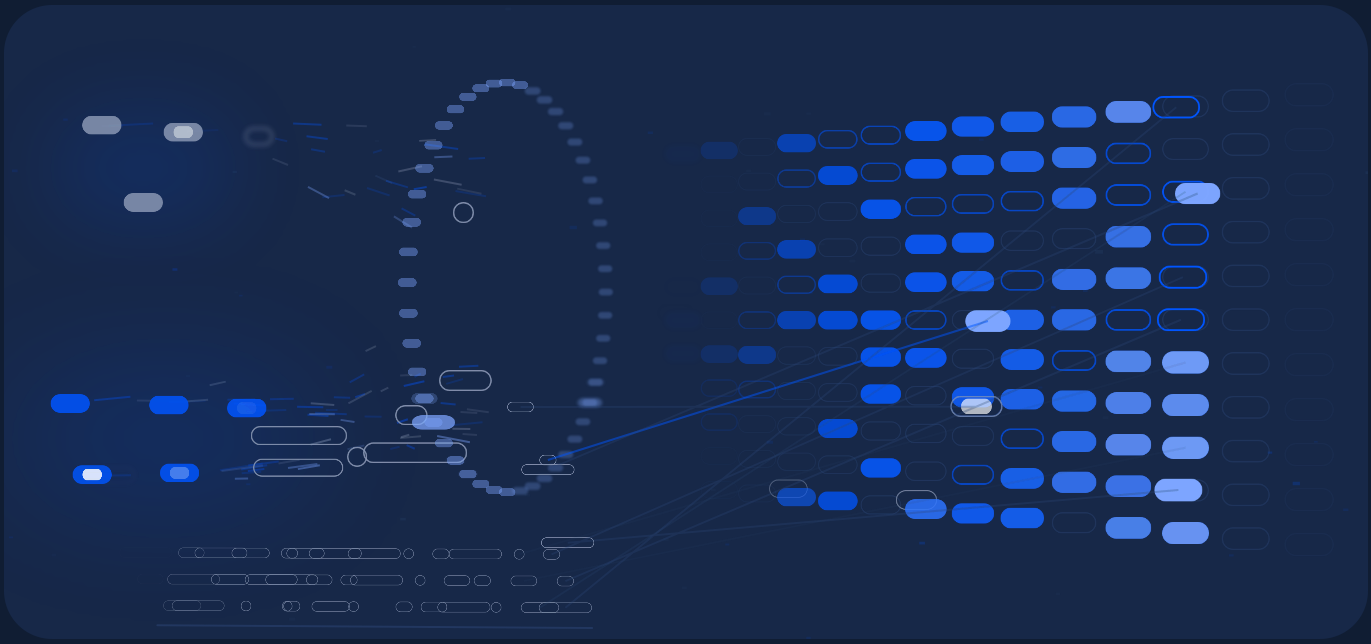
Rillor

EXPLAINER

What to check before signing a GPU cloud contract

A GPU cloud provider checklist built from published service level agreements, reservation terms, filed contracts and acceptance tests: what each term covers and who pays when hardware fails.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

A GPU cloud contract decides who pays when hardware fails, and the service level agreements, filed contracts and reservation terms reviewed here, as published when last viewed on 8 or 14 November 2026, settle that in different ways.

- The Llama 3 Herd of Models (2024), a published account of pre-training on up to 16K H100 GPUs, logged 419 unexpected interruptions in 54 days, about 78% of them from confirmed or suspected hardware issues [1].
- Published service level agreements (SLAs) define downtime differently. AWS, Google Cloud and Oracle write theirs around external connectivity, Google adding disk access, and Nebius also counts a failed GPU or lost NVLink or fabric connectivity [3, 4, 5, 7].
- The six published SLAs reviewed pay service credits, in steps that start at 10% except at Oracle's single-instance level, and the SLA documents of AWS, Google Cloud, Oracle, DigitalOcean and Voltage Park call credits the sole or exclusive remedy [3, 4, 5, 7, 8, 9].
- Rack-scale systems need a rack definition. Nebius's rack SLA covers 16 of the 18 nodes in its NVL72 racks, and SemiAnalysis's ClusterMAX 3.0 rating describes a common pattern where 3 or more failed nodes put the rack down [7, 11].
- Three contracts filed with the U.S. Securities and Exchange Commission (SEC) tie payment to acceptance, with acceptance periods of 5 or 35 days and capacity deemed accepted without a timely rejection [17, 18, 19].
- In the CoreWeave and Microsoft master services agreement, fees for services not performed are cancelled once a hardware failure lasts two consecutive business days [16].
- AWS Capacity Blocks cannot be cancelled and are paid up front, and Google's future reservations in calendar mode can be cancelled only in two cases, both before the start [27, 29, 32].
- The EU Data Act requires switching rights in a written contract, a transition of at most 30 calendar days and, from 12 January 2027, no switching charges [35].

Background

Published accounts of training runs and research clusters put failures on large jobs hours apart [1, 2]. A contract for months of GPU capacity sets who carries that cost: what counts as downtime, what the provider owes, when billing starts and stops, and how either side leaves.

Three layers of public documents show the terms. Online SLAs state uptime commitments and service credits for instances, nodes and, at one provider, racks [3, 4, 5, 6, 7, 8, 9]. Customer agreements set how SLAs change and cap liability, and Google Cloud's terms state that they do not apply to a customer that signed an offline variant of the agreement [14, 15]. Contracts filed as exhibits with the SEC show negotiated terms for acceptance, delivery delays, hardware failure and termination [16, 17, 18, 19].

SemiAnalysis's ClusterMAX, a GPU cloud rating published in a newsletter with paid sections, adds a buyer's yardstick in its free text, including SLAs committed "in written form, with penalties in the form of credits or deductions off a customer's monthly bill" [10]. Its third edition, of 23 September 2026, lists what a standardized SLA document should contain [11]. NIST observes "many variants of a 'service availability percentage' metric across providers" [12], and guidelines presented to the European Commission in 2014 note that stated uptime commonly excludes maintenance periods [13].

This report is research on published terms for buyers and infrastructure leads about to sign reserved or multi-month GPU capacity. It is never legal advice, and any contract is for the buyer's counsel to review.

Method

What was examined. The terms a buyer of reserved or multi-month GPU cloud capacity meets before signing: uptime and remedies, rack-scale definitions, acceptance, billing during provisioning, failure and maintenance, reservations and cancellation, exit, and the price terms of a quote.

Sources. Two published accounts of failures at scale [1, 2]. SLAs from AWS, Google Cloud, Oracle, Nebius, DigitalOcean and Voltage Park, and Microsoft's capacity reservation page [3, 4, 5, 6, 7, 8, 9]. The free sections of ClusterMAX 2.0 and 3.0 [10, 11]. NIST SP 500-307 and the 2014 Cloud SLA Standardisation Guidelines [12, 13]. The AWS and Google Cloud customer terms [14, 15]. Four contracts filed with the SEC [16, 17, 18, 19]. Documentation of test tools [20, 21, 22]. Provider pages on billing, maintenance and reservations [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. The EU Data Act [35].

How. Every source was opened on 8 November 2026, and [5], [24], [26], [32], [33] and [35] again on 14 November 2026, [35] from the Publications Office's copy. Each term is restated or quoted from the section cited, with the document's own date where it gives one. Providers are named only as the publishers of their own documents. ClusterMAX is cited for its criteria and SLA patterns, never for its ratings of named providers. The customer in one filed order form is described as a model developer customer. "Derived" marks arithmetic on a source's numbers, with a 30-day month of 720 hours.

Findings

1. How often hardware fails at scale

The Llama 3 Herd of Models (2024), a published account of pre-training on up to 16K H100 GPUs, reports 466 job interruptions in a 54-day snapshot: 47 planned, for automated maintenance or operator-initiated changes, and 419 unexpected [1]. About 78% of the unexpected ones were attributed to confirmed hardware issues, such as GPU or host component failures, or to suspected ones such as silent data corruption, and GPU issues accounted for 58.7% [1]. That is about 7.8 unexpected interruptions a day (derived) [1]. The account notes that "a single GPU failure may require a restart of the entire job", and the run still reached higher than 90% effective training time, with significant manual intervention needed only three times [1].

Revisiting Reliability in Large-Scale Machine Learning Research Clusters (2024), a published account of 11 months, 4 million jobs and over 150 million A100 GPU hours in two research environments, measured a mean time to failure (MTTF) of 7.9 hours for 1,024-GPU jobs against 47.7 days for 8-GPU jobs [2]. Its authors project 1.8 hours at 16,384 GPUs and 0.23 hours, about 14 minutes (derived), at 131,072 GPUs [2].

Rack-scale systems widen a failure. ClusterMAX 2.0 writes that on GB200 NVL72 "a single faulty component can require draining an entire 72-GPU rack" [10]. ClusterMAX 3.0 adds that a failed tray takes 4 GPUs down, that 8 more GPUs cannot be hot swapped into the NVLink domain, and that the options are to run the rack degraded or to swap the whole rack [11].

At these rates a large job meets failures within hours or days, and the terms below decide what the buyer pays while they last.

2. Uptime commitments and their remedies

Each online SLA reviewed sets a monthly commitment for a unit it defines, a definition of downtime and a schedule of credits.

Document and date	Unit covered	Commitment	What counts as down	Credit steps
AWS Compute SLA, last updated 25 May 2022	A region, for instances in two or more zones. A single instance	99.99% region. 99.5% instance	Instance: "no external connectivity"	10%, 30%, 100% [3]
Google Compute Engine SLA, last modified 14 October 2026	Instances in multiple zones. A single instance	99.99% multi-zone. 99.95% memory-optimized. 99.9% other families	Loss of external connectivity or persistent disk access, in whole minutes	10%, 25%, 100% [4]

Document and date	Unit covered	Commitment	What counts as down	Credit steps
Oracle pillar document, October 2026	Compute SKUs including H100, H200, B200, GB200 and MI300X	99.9% availability. 99.9% network performance	A problem that prevents external connectivity. Throughput below 90% of the published figure	10%, 25%, 100% in multi-domain regions. 100% for a single instance below 99.9% [5]
Nebius Compute Cloud, effective 6 October 2026	A virtual machine. A rack-scale GPU system	99.50% each	An allocated GPU missing or failed, lost NVLink or fabric connectivity, an inaccessible boot disk or lost external connectivity	10.0%, 15.0%, 30.0% [7]
DigitalOcean GPU Droplets, last updated 21 October 2025	A server node	99%	No external connectivity in a 5-minute interval	10%, 25%, 50% [8]
Voltage Park, last revised 5 January 2025	A GPU virtual machine. A node	99.5% each	Unavailability, for example from lost network connectivity or hardware failure	10%, 25%, 100% [9]

Google's figures are its Premium Tier levels for regions other than Stockholm [4].

What counts as down. The definitions test different things. Google counts downtime in periods of one or more consecutive minutes, and partial minutes do not count [4]. Nebius tests conditions inside the machine, every allocated GPU present and not failed among them, and excludes "distributed training inefficiency or benchmark degradation" and reduced bandwidth that causes no loss of connectivity [7].

When the clock runs. DigitalOcean counts from the customer's report until service is restored or an in-kind replacement node is provided [8]. Voltage Park counts from the customer's notice until a new machine is available [9]. Nebius stops counting when it notifies the customer that a replacement host or rack is available, so relocation time is not counted [7].

What a percentage allows. Over a 30-day month, 99.5% allows 3.6 hours of downtime before a credit, 99.9% allows 43.2 minutes and 99% allows 7.2 hours (derived) [3, 4, 5, 7, 8, 9]. Each allowance applies to one unit, and DigitalOcean measures uptime "at the server node level, rather than at the cluster level" [8]. Microsoft's worked example for capacity reservations shows the arithmetic: 55 minutes of downtime on a reserved unit gives 99.87% uptime and a 10% credit, and 15 minutes gives 99.97% and no credit [6].

What the remedy is. AWS applies credits "only against future payments for Amazon EC2 otherwise due", states that they give no right to a refund, and computes them on the monthly bill excluding one-time payments such as upfront Reserved Instance payments [3]. Google applies a monetary credit to future use, capped at the month's amount due for the covered services [4]. Oracle computes credits on net fees for the quantity of the service actually used, and its Pay as You Go credits expire at the end of the month in which they are granted [5]. DigitalOcean's unused credits expire 90 days after they are provided [8]. AWS, Google Cloud, Oracle, DigitalOcean and Voltage Park each call credits the sole or exclusive remedy [3, 4, 5, 8, 9].

Each sets a deadline for claims: at AWS the end of the second billing cycle after the incident, with request logs [3], at Google and Oracle 60 days, with logs [4, 5], and at Voltage Park 30 days [9]. DigitalOcean requires written notice within 24 hours of the downtime, without which the right to credits is forfeited, and a credit request within three months of the end of the billing cycle in which the issue occurred [8]. AWS's region and instance claims cannot be stacked, and Google pays a machine's downtime as a single instance or as instances in multiple zones, not both [3, 4].

A negotiated order form can differ. In the form of order between a model developer customer and an Nscale subsidiary, filed on 18 September 2026, credits are the "sole and exclusive monetary remedy", but unapplied credits are paid out on request or at termination [19].

How the terms change. The AWS Customer Agreement, last updated 29 October 2026, gives at least 90 days' notice of adverse SLA changes, denies credits during a suspension and caps liability at the amounts paid for the services in the 12 months before the liability arose [14]. Google's terms, last modified 12 October 2026, make material updates to URL Terms, the SLA among them, effective 30 days after posting and cap liability at the fees paid in the prior 12 months [15].

Support and monitoring. ClusterMAX 2.0 expects 24x7 support with a 15-minute response SLA, and health checks that detect and drain nodes on signals such as GPUs falling off the bus, ECC errors and NCCL stalls [10]. The 2014 EU guidelines list support hours, support responsiveness and resolution time as SLA terms [13].

3. Rack-scale systems and failure domains

An NVL72 rack joins 72 GPUs in 18 nodes on one NVLink domain, so a contract needs to say when the rack, and not only a node, is down [7, 10, 11]. ClusterMAX 2.0 reports that "top-tier providers are moving beyond node-level uptime and introducing rack-level SLAs", and describes three approaches [10]:

- A node SLA of 99% plus a rack SLA of 95%, a working rack being "16 of the 18 nodes, or 64 of 72 GPUs", with access to all nodes.
- A node SLA of 99% only, with access to 16 of the 18 nodes.
- A rack SLA only, where "all 18 of 18 nodes must be up 95% of the time".

ClusterMAX 3.0 calls a common pattern "NVL64+": an SLA on each node, with the whole rack considered down when 3 or more of its 18 nodes fail at a given time [11]. On HGX systems it expects "replacements in under an hour end-to-end and detection in around 2 minutes" [11]. It notes that a provider in colocation can depend on the landlord's remote hands and their schedule, and that "Both show up in how fast capacity comes back, and that is what an SLA has to measure" [11].

Nebius's rack-scale SLA of 99.50% covers GB300 NVL72 and Vera Rubin NVL72 racks with 18 nodes each, 16 of them SLA-covered and 2 kept as a maintenance buffer [7]. A rack is unavailable while fewer nodes are functional and usable than the covered count, impairments within the buffer do not count, and buffer nodes may be used but do not count toward uptime [7]. Virtual machines inside a rack-scale system are outside the virtual machine SLA [7].

AWS notifies the customer when it detects impairment on an UltraServer Capacity Block, "but generally your instances are not terminated", and remediation means terminating all instances on the block and opening a support case [27].

A rack-scale contract can therefore define the node count that puts a rack down, the buffer, the billing of a degraded rack and how replacement time is measured [7, 10, 11, 27].

4. Acceptance testing and burn-in

In the filed contracts, acceptance decides when payment starts.

Filed contract and date	Validation before delivery	Acceptance period	Without a timely rejection	Late delivery
Statement of work between an IREN subsidiary and Microsoft, filed 5 February 2026 [17]	GPU diagnostics, memory error detection, HPL-MxP, HPCG, InfiniBand topology and NCCL validation	5 days, then 15 business days to cure a rejection	Deemed accepted	Daily delay credits for shortfall GPUs, amount redacted
Form of statement of work between Microsoft and an Nscale subsidiary, filed 18 September 2026 [18]	GPU diagnostics, memory tests, HPL-MxP, HPCG and NCCL tests	35 days for the first GPU service or the first after a technology change, 5 days for the rest	Deemed accepted	100% of a daily bill for each day of late handoff

Filed contract and date	Validation before delivery	Acceptance period	Without a timely rejection	Late delivery
Form of order between a model developer customer and an Nscale subsidiary, filed 18 September 2026 [19]	Acceptance criteria in a redacted schedule	Written notice of deficiencies within the period, then re-delivery and retest	Deemed accepted	No payment during a further delay and a right to terminate, or a late delivery discount

The Nscale order form makes no fees payable for a tranche until written acceptance of all of it, with retesting until the customer accepts or terminates [19]. The CoreWeave and Microsoft agreement, filed on 3 March 2025, has the customer pay after it accepts services and receives a proper and undisputed invoice [16]. The acceptance criteria of the IREN statement of work and the Nscale order form are redacted [17, 19].

The tests these contracts name are public tools. NVIDIA's DCGM Diagnostics, documentation last updated 17 June 2026, run at four levels, the longest under 2.25 hours on an 8-GPU Hopper system, so a contract can name the level and the time it takes [20].

NVIDIA states that DCGM Diagnostics are not designed to provide comprehensive hardware diagnostics or to facilitate any RMA process [20]. NCCL Tests "check both the performance and the correctness of NCCL operations" [21]. AWS's SageMaker HyperPod runs deep health checks when a cluster is created or updated, including DCGM diagnostics level 4, an NCCL test and multi-node Megatron training checks for "NaN or infinite loss, loss divergence, and training time" [22]. ClusterMAX 2.0 describes NVL72 burn-in as "running tests such as dcgmi diag -r 3, nccl-tests, multinode ubergemm, and a sample" multi-node training job [10]. ClusterMAX 3.0's SLA template includes "Definitions of acceptance testing" and termination rights for the buyer when an agreed acceptance date is missed [11].

A contract can therefore name the tests and their pass thresholds, who runs them, the acceptance period, deemed acceptance, the cure period, whether billing starts at handoff or at acceptance, and the remedy for a missed date [11, 17, 18, 19].

5. Billing during provisioning and downtime

Online billing follows the instance's state. EC2 bills a `running` instance per second with a one-minute minimum, even when it is idle, and does not bill one that is `pending`, stopped or terminated [23]. Azure bills Starting, Running, Stopping and Stopped (allocated), and does not bill Creating or Deallocated [25]. Google's lifecycle page states that CPU is charged in `RUNNING` and `PENDING_STOP`, and it names CPU, memory and attached resources only [24]. At all three, disks and IP addresses or networking keep charging whatever the instance's state [23, 24, 25].

Maintenance stops GPU instances on a schedule. Compute Engine instances with GPUs cannot be live migrated and stop for host maintenance events, which "typically occur once every two weeks, but might occasionally run more frequently", with 1 hour of notice [26]. They can take up to one hour to terminate after failures or host errors, and Local SSD data is unrecoverable after a maintenance stop [26].

AWS's SLA adds a failure rule: no charge for any clock hour in which a single instance is unavailable for more than six minutes, applied automatically [3].

Filed contracts go further. In the CoreWeave and Microsoft agreement, once a hardware failure continues for two consecutive business days, payment for services not performed is cancelled and does not recommence until the failure is resolved [16]. CoreWeave must keep spares in all locations for nodes and switches that cannot be remediated within that time, and planned maintenance needs at least five business days' notice, planned critical maintenance at least 24 hours' [16]. The Nscale order form subtracts service credits and late delivery discounts from a monthly payment of hours times GPUs times the price per GPU-hour [19]. It allows suspension for scheduled or emergency maintenance and lets the provider substitute servers "of the same or better type, generation and specification" on notice [19].

ClusterMAX 2.0 counts "charging users for GPU hours while a cluster is being created or while servers are down due to hardware failure" among the critical failures of its underperforming tier [10]. ClusterMAX 3.0 lists charging during cluster creation or hardware downtime among the critical issues of its bottom tier [11].

6. Capacity, cancellation and end-of-term terms

Reservations hold capacity on terms that differ in length, lead time, cancellation and billing.

Product, as documented	Term and start	Cancellation	Billing
AWS Capacity Blocks for ML [27, 28, 29]	1 to 14 days, or multiples of 7 days up to 182, starting up to 8 weeks ahead	Not allowed, and a block cannot be moved or split	Charged up front at reservation, no extra charge for unused time
AWS future-dated Capacity Reservations [30, 31]	A minimum commitment of 14 days for most instance types	A cancellation charge might apply, depending on when	The On-Demand rate, used or not
AWS immediate Capacity Reservations [30, 31]	No term commitment	Any time, which stops the charges	The On-Demand rate, used or not

Product, as documented	Term and start	Cancellation	Billing
Google future reservations in calendar mode [32]	1 to 90 days, starting at least 87 hours after the request for GPU VMs	Only while procuring, and only if the start is more than 30 days away, or within 24 hours of approval with the start at least four days away	From the start time to the end time
Google resource-based committed use discounts [33]	The commitment term bought	Not after purchase	Monthly to the end of the term, used or not
Azure on-demand capacity reservation [6]	No term commitment	Delete at any time	The rate of the VM size, used or not

A Capacity Block's price "depends on available supply and demand at the time the offering was delivered" and does not change after reservation [28]. A payment that cannot be processed in time releases the block, and Savings Plans and Reserved Instance discounts do not apply [29]. A block ends at 11:30 UTC, and its instances begin terminating at 11:00 UTC on the final day [27]. AWS future-dated reservations cannot drop below the committed instance count during the commitment [30]. Google's calendar-mode charges stop at the end time, when Compute Engine stops or deletes the VMs using the reservation [32].

Microsoft states that "Reserved Instances provide a billing discount only and don't guarantee capacity" [6]. Its refunds of Azure reservations are capped at USD 50,000 of cancelled commitments in a rolling 12-month window [34]. It states that, starting 1 February 2027, reservations bought after that date are not eligible for exchange where savings plans support the service, Azure Virtual Machines included [34].

Delivery delays. The IREN statement of work gives daily delay credits for shortfall GPUs, a right to terminate with a refund of any upfront payment if the minimum GPU quantity misses the delivery delay window, and extensions for documented delays in NVIDIA supply [17]. The Nscale statement of work credits 100% of a daily bill for each day of late handoff [18]. The Nscale order form offers no payment during a further delay with a right to terminate the tranche, or a late delivery discount [19].

Termination. In the CoreWeave and Microsoft agreement, either party may terminate for cause on 45 days' written notice of an uncured material breach, and a customer that terminates for cause has its prepaid fees for the rest of the term refunded [16]. If CoreWeave terminates for certain customer breaches (a material violation of law that cannot be cured, uncured non-payment of undisputed fees, or gross negligence, willful misconduct or fraud) and chooses not to continue the services, the customer pays the lesser of the remaining fees minus the proceeds of transferring or selling the contract servers, or the difference, from the start of the agreement to termination, between the fees paid and the fees at CoreWeave's on-demand rates [16]. ClusterMAX 3.0's template adds termination rights for downtime below recommended thresholds and two-way commitments for measuring downtime and providing credits [11].

Exit and switching. The EU Data Act requires that switching rights be "clearly set out in a written contract" made available before signing, with a notice period of at most two months, a transitional period of at most 30 calendar days and data retrieval of at least 30 calendar days [35]. From 12 January 2027, providers "shall not impose any switching charges", which include data egress charges, and until then charges may not exceed the costs directly linked to the switch [35]. Services mostly custom-built for one customer and not offered at broad commercial scale fall outside that rule [35]. The 2014 EU guidelines name data retrieval, data retention and residual data retention periods among termination terms [13].

7. Price terms for comparing quotes

The Nscale order form writes the monthly payment as the hours in the month times the GPUs in the tranche times the price per GPU-hour, minus service credits, late delivery discounts and other adjustments [19]. For example, a 1,024-GPU tranche in a 30-day month is billed for 737,280 GPU-hours before credits (derived) [19]. Other terms move that figure:

- **Use.** Azure and AWS capacity reservations and Google resource-based commitments bill whether the capacity is used or not [6, 31, 33].
- **Price fixing.** A Capacity Block is paid up front, at a price fixed at reservation, with no Savings Plans or Reserved Instance discounts [28, 29].
- **Credit base.** AWS computes credits on the monthly bill without one-time payments such as upfront Reserved Instance payments [3], and Oracle on net fees for the quantity actually used [5].
- **Bundle and term.** ClusterMAX 2.0's pricing criteria cover consumption models from 1 month to 3 years, separate or bundled charges for storage, compute nodes and network, and the extension of existing contracts [10].
- **Charges outside the GPU-hour.** Disks and IP addresses keep charging whatever the instance's state [23, 24, 25].

Prices for the same GPUs differ by pricing type, term and the date a price was offered [6, 28, 31, 33], so two quotes compare on a like basis only with those held fixed, along with the GPU configuration and region.

Implications

The findings give a buyer's checklist, each item tied to the documents that show how the term is written. It is research on published terms, never legal advice, and a contract is for counsel to review.

- **Uptime.** The unit covered (instance, node, rack or region), the monthly percentage and its downtime allowance [3, 4, 5, 7, 8, 9].
- **Definition of down.** Whether it names GPU, NVLink and fabric conditions or only external connectivity and disk access [3, 4, 5, 7].
- **The clock.** When downtime starts and stops counting, and whether a replacement notice ends it [7, 8, 9].
- **Credits.** The steps, the base they apply to, caps, expiry, payout, claim deadlines and evidence [3, 4, 5, 8, 9, 19].
- **Changes and caps.** Notice of SLA changes, exclusions and the liability cap [3, 4, 14, 15].
- **Support.** Support hours, response time and resolution time [10, 13].
- **Racks.** The node count that puts a rack down, the buffer, billing of a degraded rack, and detection and replacement times [7, 10, 11, 27].
- **Acceptance.** The named tests and thresholds, who runs them, the acceptance and cure periods, deemed acceptance, and whether billing starts at acceptance [16, 17, 18, 19, 20, 21, 22].
- **Delays.** Credits and termination rights for late delivery or a missed acceptance date [11, 17, 18, 19].
- **Billing during failure.** Billing during creation and failure, and when failed hardware stops being billed [3, 10, 11, 16, 23, 25].
- **Maintenance.** Notice, frequency, substitution of servers and effect on local data [16, 19, 26].
- **Reservations.** Term, start, lead time, cancellation terms and charges, and the end time [27, 28, 30, 32, 33, 34].
- **Termination.** Termination for cause, refunds of prepaid fees and the buyer's payments on early termination [16].
- **Exit.** Switching notice, transition and data retrieval periods, and switching or exit charges [13, 35].
- **Price.** The payment formula, billing whether used or not, upfront payments, discounts and charges outside the GPU-hour [6, 19, 23, 28, 29, 31, 33].

Limits of this analysis

- **Published terms only.** Online SLAs describe standard offers, which a signed offline agreement replaces at Google [15]. The report reads no negotiated contract beyond the four filed ones.
- **Redactions.** The CoreWeave and Microsoft agreement places its service level objective in an exhibit that is not in the filed text [16]. The Nscale order form's SLA exhibit and acceptance schedule, and the IREN acceptance criteria and delay credit amounts, are redacted [17, 19].
- **No survey.** The report does not state how many providers offer any term. ClusterMAX's descriptions of common patterns are that publisher's account [10, 11].
- **Coverage.** Azure virtual machine SLA percentages are outside the report, which uses Microsoft's capacity reservation page only [6]. It infers nothing about GPU billing in Compute Engine's provisioning states [24], or about whether a failed GPU on a reachable instance counts as downtime under a definition that names only connectivity or disk access [3, 4, 5].
- **Failure data.** The failure figures come from one account of H100 pre-training and one study of A100 research clusters [1, 2], and the report has no primary failure data for NVL72 racks.
- **Dates.** Each term is as published when last viewed, on 14 November 2026 for [5], [24], [26], [32], [33] and [35] and on 8 November 2026 for every other source. Providers change their terms. The EU Data Act's 12 January 2027 date and Microsoft's 1 February 2027 change are stated as their sources state them [34, 35].
- **No legal analysis.** The report does not assess whether any clause is enforceable or advisable.

Sources

1. The Llama 3 Herd of Models. arXiv 2407.21783 v3, last revised 23 November 2024 (first version 31 July 2024), viewed 8 November 2026. https://arxiv.org/abs/2407.21783

2. Revisiting Reliability in Large-Scale Machine Learning Research Clusters. arXiv 2410.21680 v2, last revised 6 February 2025 (first version 29 October 2024), viewed 8 November 2026. https://arxiv.org/abs/2410.21680

3. Amazon Compute Service Level Agreement. Amazon Web Services, last updated 25 May 2022, viewed 8 November 2026. https://aws.amazon.com/compute/sla/

4. Compute Engine Service Level Agreement (SLA). Google Cloud, last modified 14 October 2026, viewed 8 November 2026. https://cloud.google.com/compute/sla

5. Oracle PaaS and IaaS Public Cloud Services Pillar Document. Oracle, October 2026, viewed 14 November 2026. https://www.oracle.com/contracts/docs/paas_iaas_pub_cld_srvs_pillar_4021422.pdf

6. On-demand capacity reservation in Azure. Microsoft Learn, last updated 16 June 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/capacity-reservation-overview

7. Service Level for Compute Cloud. Nebius AI Cloud, effective 6 October 2026, viewed 8 November 2026. https://docs.nebius.com/legal/sla-levels/compute

8. DigitalOcean GPU Droplet SLA. DigitalOcean, last updated 21 October 2025, viewed 8 November 2026. https://www.digitalocean.com/sla/gpu-droplets

9. Service Level Agreements (Virtual Machine SLA and Node SLA). Voltage Park, last revised 5 January 2025, viewed 8 November 2026. https://voltagepark.com/virtual-machine-sla

10. ClusterMAX 2.0: The Industry Standard GPU Cloud Rating System. SemiAnalysis, 6 November 2025, viewed 8 November 2026. https://newsletter.semianalysis.com/p/clustermax-20-the-industry-standard

11. ClusterMAX 3.0: The Industry Standard GPU Cloud Rating System Returns. SemiAnalysis, 23 September 2026, viewed 8 November 2026. https://newsletter.semianalysis.com/p/clustermax-30-the-industry-standard

12. Cloud Computing Service Metrics Description, NIST Special Publication 500-307. National Institute of Standards and Technology, April 2018, viewed 8 November 2026. https://doi.org/10.6028/NIST.SP.500-307

13. Cloud Service Level Agreement Standardisation Guidelines. Cloud Select Industry Group, SLA subgroup, for the European Commission, Brussels, 24 June 2014, viewed 8 November 2026. https://ec.europa.eu/newsroom/dae/redirection/document/6138 (linked from the Commission's news item of 26 June 2014)

14. AWS Customer Agreement. Amazon Web Services, last updated 29 October 2026, viewed 8 November 2026. https://aws.amazon.com/agreement/

15. Google Cloud Terms of Service. Google Cloud, last modified 12 October 2026, viewed 8 November 2026.

https://cloud.google.com/terms

16. Master Services Agreement between CoreWeave, Inc. and Microsoft Corporation (Exhibit 10.23 to CoreWeave Form S-1). U.S. Securities and Exchange Commission, filed 3 March 2025, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1769628/000119312525044231/d899798dex1023.htm

17. Partner Statement of Work between IE US Hardware 3 Inc. (an IREN subsidiary) and Microsoft (Exhibit 10.3 to IREN Limited Form 10-Q). U.S. Securities and Exchange Commission, filed 5 February 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1878848/000187884826000015/exhibit103-partnerstatem.htm

18. Form of Partner Statement of Work between Microsoft and an Nscale subsidiary (Exhibit 10.19 to Nscale Limited Form S-1). U.S. Securities and Exchange Commission, filed 18 September 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/2110365/000119312526395475/ck0002110365-ex10_19.htm

19. Form of Order for GPU Services between a model developer customer and an Nscale subsidiary (Exhibit 10.25 to Nscale Limited Form S-1). U.S. Securities and Exchange Commission, filed 18 September 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/2110365/000119312526395475/ck0002110365-ex10_25.htm

20. DCGM Diagnostics. NVIDIA DCGM Documentation, NVIDIA, last updated 17 June 2026, viewed 8 November 2026. https://docs.nvidia.com/datacenter/dcgm/latest/user-guide/dcgm-diagnostics.html

21. NCCL Tests. NVIDIA, GitHub, viewed 8 November 2026. https://github.com/NVIDIA/nccl-tests

22. Deep health checks. Amazon SageMaker HyperPod documentation, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-hyperpod-eks-resiliency-deep-health-checks.html

23. Amazon EC2 instance state changes. Amazon EC2 User Guide, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instance-lifecycle.html

24. Compute Engine instance lifecycle. Compute Engine documentation, Google Cloud, last updated 10 November 2026, viewed 14 November 2026. https://cloud.google.com/compute/docs/instances/instance-lifecycle

25. States and billing status of Azure Virtual Machines. Microsoft Learn, last updated 22 August 2024, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/states-billing

26. Handle GPU host maintenance events. Compute Engine documentation, Google Cloud, last updated 10 November 2026, viewed 14 November 2026. https://cloud.google.com/compute/docs/gpus/gpu-host-maintenance

27. Capacity Blocks for ML. Amazon EC2 User Guide, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-blocks.html

28. How Amazon EC2 Capacity Blocks work. Amazon EC2 User Guide, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/capacity-blocks-how.html

29. Capacity Blocks pricing and billing. Amazon EC2 User Guide, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/capacity-blocks-pricing-billing.html

30. Concepts for Amazon EC2 Capacity Reservations. Amazon EC2 User Guide, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/cr-concepts.html

31. Capacity Reservation pricing and billing. Amazon EC2 User Guide, Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/capacity-reservations-pricing-billing.html

32. About future reservation requests in calendar mode. Compute Engine documentation, Google Cloud, last updated 10 November 2026, viewed 14 November 2026. https://cloud.google.com/compute/docs/instances/future-reservations-calendar-mode-overview

33. Committed use discounts (CUDs) for Compute Engine. Compute Engine documentation, Google Cloud, last updated 12 November 2026, viewed 14 November 2026. https://cloud.google.com/compute/docs/instances/committed-use-discounts-overview

34. Self-service exchanges and refunds for Azure Reservations. Microsoft Learn, updated 29 July 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/cost-management-billing/reservations/exchange-and-refund-azure-reservations

35. Regulation (EU) 2023/2854 of 13 December 2023 (Data Act). Official Journal of the European Union, OJ L, 2023/2854, 22 December 2023, Publications Office of the European Union, viewed 14 November 2026. https://publications.europa.eu/resource/celex/32023R2854

About this report

Published by Rillor on 16 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. What to check before signing a GPU cloud contract. 16 November 2026. https://rillor.com/insights/gpu-cloud-contract-checklist

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

NVIDIA, AMD, Ceph, Kubernetes and related product names are trademarks of their owners and are used only to identify products. No affiliation or endorsement is implied.

Work with Rillor

Send the GPU classes, regions and pricing types you need, and how often you need them. Contact us at rillor.com/contact?type=index or contact@rillor.com. More on GPU and neocloud pricing: rillor.com/markets/gpu-pricing