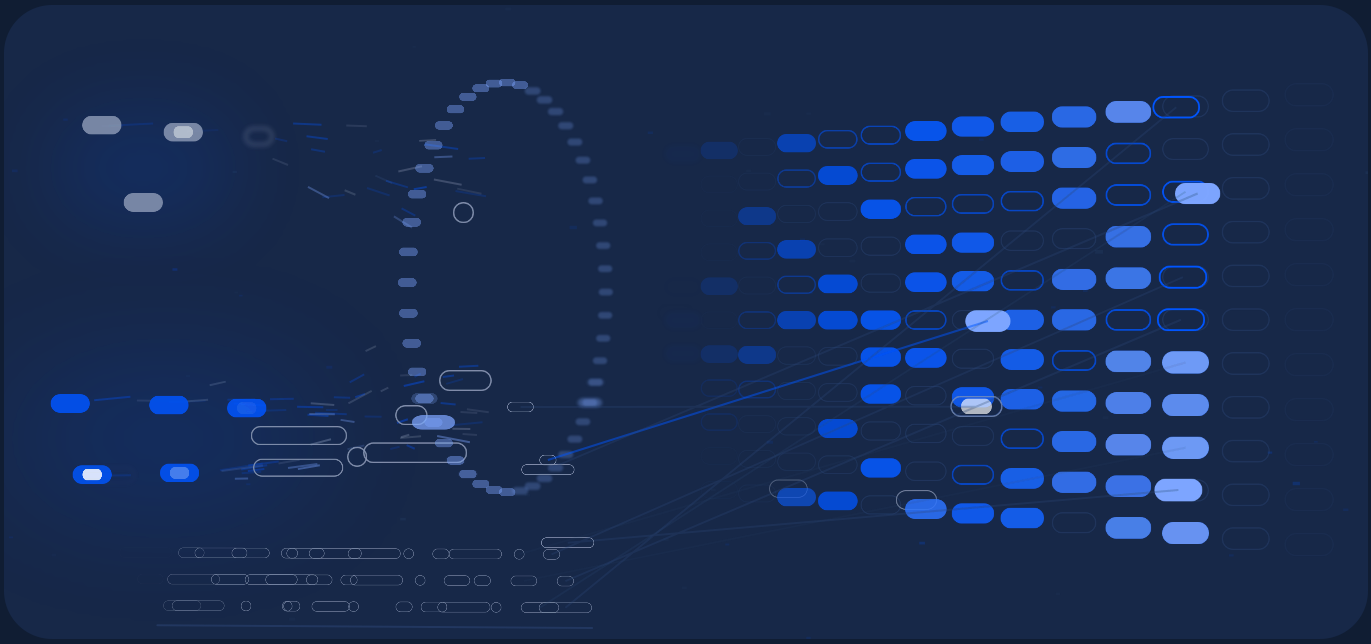
Rillor

EXPLAINER

# Why cloud GPU launches fail and how reservations hold capacity

InsufficientInstanceCapacity and other GPU launch errors come from two separate checks, quota and capacity. Provider documents show what each error means and on what terms reservations hold GPUs.



Rillor

# Contents


01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

# Summary

A cloud GPU launch passes two separate checks, quota and capacity, and the documents of AWS, Google Cloud and Microsoft Azure, as they read on 8 November 2026, show which error each check returns and which products hold GPUs ahead of a run.

- Quota is an account's permission to deploy resources, and capacity is the hardware available in a region or zone at that moment. Azure calls them separate checks, and Google says a quota increase does not resolve a capacity error [20, 33, 36].
- InsufficientInstanceCapacity is the AWS error for a launch that finds too little available capacity. AWS lists it among its server errors, while the quota errors InstanceLimitExceeded and VcpuLimitExceeded are client errors [1, 2]. Google's ZONE_RESOURCE_POOL_EXHAUSTED, Azure's AllocationFailed and the GPU cloud Lambda's insufficient-capacity code mark the same condition [20, 32, 39].
- As AWS's quota page read on 8 November 2026, the default quotas are 0 vCPUs for Running On-Demand P instances and for All P Spot Instance Requests, and 0 for Concurrent P5 Capacity Blocks, all adjustable [3]. One p5.48xlarge, with eight H100 GPUs and 192 vCPUs, needs 192 vCPUs of P instance quota (computed for this report from [3, 6]).
- On Google Cloud, the A4 (B200), A3 Ultra (H200), A4X (GB200), A4X Max (GB300) and A5X (Vera Rubin) machine series have no standard (on-demand) GPU quota, only quota for committed use and for preemptible VMs such as Spot and Flex-start, as the quota page read on 8 November 2026 [22].
- Reservations that start at once are billed whether they are used or not, at the rates of the instances they hold, at AWS, Google Cloud and Azure [10, 25, 36]. Their GPU coverage differs, and Azure's on-demand capacity reservations exclude ND-series and NCads H100 v5 VMs, as the page read on 8 November 2026 [36].
- Scheduled GPU capacity comes in fixed windows. AWS Capacity Blocks run from one day to 182 days, are paid up front and cannot be cancelled, and Google's calendar mode reserves 1 to 90 days, requested at least 87 hours ahead [12, 13, 14, 26].
- A billing discount holds no capacity unless it carries reservations or reserves capacity itself, as Google's resource-based GPU commitments and AWS zonal Reserved Instances do. AWS regional Reserved Instances and Savings Plans reserve none, Azure Reserved Instances are a billing discount only, and Google's flexible commitments carry no assured capacity [7, 16, 28, 29, 36].
- Availability scores do not guarantee a launch. AWS's Spot placement score and Google's obtainability score both say so, and a 2023 study found that clouds do not publish availability information explicitly [18, 31, 40].

# Background

## Quota and capacity

Microsoft's capacity reservation documentation draws the line in two sentences: "Quota and capacity are separate checks. Quota is your subscription's permission to deploy resources, while capacity is the underlying infrastructure available in a specific region or zone." [36]. Its quota page adds that a deployment can fail with enough quota if the requested VM size lacks capacity in the target region or zone [33].

Google Cloud states the same rule. Its capacity errors indicate "a shortage of physical hardware in a zone rather than a project usage limit", so "a quota increase doesn't resolve them" [20]. Quota "is an administrative limit", and resource availability "reflects the real-time supply of physical hardware in a zone" [20]. Google's example is an a2-highgpu-2g instance, with 24 vCPUs, 170 GB of memory and two A100 GPUs [20]. The request succeeds only if all of them are available in the zone at that time [20].

## What clouds publish about availability

SkyPilot, a broker that places jobs across clouds, reported in 2023 that "clouds do not publish availability information explicitly", so a broker may have to learn it from the preemptions and allocation failures it observes [40]. It found the variations especially acute for scarce resources such as GPUs and spot instances [40]. A 2026 study adds that provider datasets, such as AWS's Spot placement score, indicate relative availability trends and do not guarantee whether a given number of instances can be obtained at a given time [41].

This report is for engineers whose GPU launch has just failed and for planners who need GPUs held for a training run. It answers two questions from provider documents: whether a failed launch is a quota problem or a capacity problem, and which product holds GPUs for which window.

# Method

**What was examined.** The quota rules, launch errors and capacity products of AWS, Google Cloud and Microsoft Azure that apply to GPU instances: default quotas and how they grow, error codes, the fixes each provider lists, reservations, scheduled capacity, billing discounts and availability scores.

**Sources.** AWS documentation [1-19], Google Cloud documentation [20-31], Microsoft Learn [32-38], the Lambda Cloud API reference as one neocloud example [39], a USENIX NSDI 2023 paper [40] and a study accepted at IEEE CLOUD 2026 [41].

**How.** Every page was opened on 8 November 2026, and [20, 23, 25-29, 31] again on 10 November 2026 after their update dates changed. Google and Microsoft Learn pages carry an update date, recorded in the Sources, and AWS pages, which show none, are dated by the day they were viewed. Lists of machine families, quota defaults and limits are given as the pages read on 8 November 2026, and held on the 10 November reading. No instances were launched, no reservations made and no capacity measured. Figures computed for this report are labelled. Providers are named only as publishers of their own documents.

**What is left out.** Prices for reserved and scheduled capacity, Kubernetes-level tools and negotiated terms.

# Findings

### 1. Quota and capacity are separate checks, with separate errors

**AWS.** InstanceLimitExceeded means "you have reached the limit on the number of instances that you can launch in a Region", and an increase can be requested per Region [1]. Creating a Capacity Reservation can return it too, because "Capacity Reservations count towards your On-Demand Instance limits" [2]. VcpuLimitExceeded marks the limit on vCPUs assigned to running instances, and MaxSpotInstanceCountExceeded the limit on Spot Instances [2].

InsufficientInstanceCapacity has a different cause: "AWS does not currently have enough available On-Demand capacity to fulfill your request" [1]. The EC2 API returns it when an instance is launched or restarted, or a Capacity Reservation created or modified [2]. ReservationCapacityExceeded means the reservation a launch targets has too little available capacity [2], and a Spot request that cannot be filled carries the status capacity-not-available [19].

The API reference sorts errors into client errors, with 400-series codes, usually caused by something the client did, and server errors, with 500-series codes, usually caused by an issue on AWS's side [2]. The three quota errors above are client errors, and InsufficientInstanceCapacity is a server error [2].

**Google Cloud.** A request over a Compute Engine quota typically returns HTTP 403 with QUOTA_EXCEEDED [21]. Capacity shortages return ZONE_RESOURCE_POOL_EXHAUSTED, ZONE_RESOURCE_POOL_EXHAUSTED_WITH_DETAILS, REGION_RESOURCE_POOL_EXHAUSTED_WITH_DETAILS or POOL_CAPACITY_INSUFFICIENT [20]. For multi-node GPU or TPU slices, ACCELERATOR_INTERCONNECT_CONFIG_FAILED comes with the advice to wait a few minutes and retry [20].

**Microsoft Azure.** A quota error carries the code OperationNotAllowed and a message that the operation results in exceeding quota limits of Core [34]. A capacity error carries AllocationFailed or ZonalAllocationFailed and the message "Allocation failed. We do not have sufficient capacity for the requested VM size in this region." [32]. Requests with too many constraints, such as a zone, Accelerated Networking or a proximity placement group, return OverconstrainedZonalAllocationRequest or OverconstrainedAllocationRequest [32].

**A neocloud.** Lambda's launch endpoint lists both kinds among its common 400 error codes [39].

| Provider | Quota errors | Capacity errors |
|---|---|---|
| AWS | InstanceLimitExceeded, VcpuLimitExceeded and MaxSpotInstanceCountExceeded, all client errors [1, 2] | InsufficientInstanceCapacity, a server error, plus ReservationCapacityExceeded and the Spot status capacity-not-available [1, 2, 19] |

| Provider | Quota errors | Capacity errors |
|---|---|---|
| Google Cloud | QUOTA_EXCEEDED, HTTP 403 [21] | ZONE_RESOURCE_POOL_EXHAUSTED and its variants, POOL_CAPACITY_INSUFFICIENT [20] |
| Microsoft Azure | OperationNotAllowed, exceeding quota limits [34] | AllocationFailed, ZonalAllocationFailed and the overconstrained variants [32] |
| Lambda | global/quota-exceeded [39] | instance-operations/launch/insufficient-capacity [39] |

Table 1. Quota and capacity errors as each provider's documents list them, viewed 8 November 2026.

## 2. Default GPU quotas and how they grow

**AWS.** On-Demand Instance quotas are managed "in terms of the number of virtual central processing units (vCPUs)" that running instances use, whatever the instance type [4]. Stopped and hibernated instances do not count, and "Capacity Reservations count toward your On-Demand Instance quotas, even if they are unused" [4].

| Quota | Default | Adjustable |
|---|---|---|
| Running On-Demand P instances | 0 vCPUs | Yes |
| Running On-Demand G and VT instances | 0 vCPUs | Yes |
| Running On-Demand Standard instances | 5 vCPUs | Yes |
| All P Spot Instance Requests | 0 vCPUs | Yes |
| All G and VT Spot Instance Requests | 0 vCPUs | Yes |
| Concurrent Capacity Blocks for P4d, P5, P5e, P5en, Trn1 and Trn2, per account and per organization | 0 vCPUs | Yes |

Table 2. AWS default quotas, viewed 8 November 2026 [3].

The 8-GPU H100, H200, B200 and B300 instances in AWS's specifications have 192 vCPUs each: p5.48xlarge with eight H100 GPUs, p5e.48xlarge and p5en.48xlarge with eight H200, p6-b200.48xlarge with eight B200 and p6-b300.48xlarge with eight B300 [6]. One node of any of them needs at least 192 vCPUs of Running

On-Demand P instances quota, against a default of 0 (computed for this report from [3, 6]).

"Amazon EC2 automatically increases your On-Demand Instance quotas based on your usage", and a customer can also request an increase [3, 4]. Smaller increases are usually approved automatically. Larger requests go to Support, "take time to review, process, approve, and deploy", and can be approved, denied or partially approved [5].

**Google Cloud.** "New accounts and projects have a global GPU quota that applies to all regions", and a region's GPU quota is the total number of GPUs in all VMs there [22]. "Running instances and reservations consume GPU quotas." [22]. Standard, Spot (preemptible), virtual workstation and committed use GPUs have separate quotas, requested per GPU model and region alongside the global "GPUs (all regions)" quota [22].

The quota table marks standard quota "Not available" for A5X (Vera Rubin), A4X Max (GB300), A4X (GB200), A4 (B200) and A3 Ultra (H200), which have committed and preemptible quotas only, such as COMMITTED_NVIDIA_B200_GPUS and PREEMPTIBLE_NVIDIA_B200_GPUS [22]. The H100 series (A3 Mega, A3 High and A3 Edge) and A100 keep standard quotas [22].

Preemptible quota applies to Spot VMs, GPU VMs with predefined run times and Flex-start VMs. Without it, these consume standard quota, but once a project requests preemptible quota they "can only consume preemptible quota and can't revert to consuming standard quotas" [22].

**Microsoft Azure.** Azure enforces vCPU quotas per subscription and region at two tiers, regional vCPUs and vCPUs per VM-size family, with separate spot quotas for Spot VMs [35]. "If you exceed either of those quotas, the VM deployment won't be allowed." [33]. Quota counts cores in use, "both allocated and deallocated" [33]. An approved VM-family increase raises the regional vCPU quota automatically, and non-adjustable quotas have a hard limit [35].

## 3. What each capacity error means, and the fixes providers list

**AWS.** For InsufficientInstanceCapacity, AWS lists four steps: wait a few minutes and retry, because "capacity can shift frequently", split a large request (15 instances as three requests of five, or 15 of one), launch without naming an Availability Zone, or choose another instance type [1]. Launches into a cluster placement group can return the error too [1].

**Google Cloud.** Google warns against releasing capacity already held: "Don't stop and restart, or delete and recreate, compute instances to resolve a resource availability error if you lack a reservation." [20]. Stopping or deleting returns the vCPUs, memory, accelerators and disks to the zone [20].

"A gpu_availability error indicates that the zone lacks capacity for your requested GPU model." [20]. The fixes are another zone or region with the same GPU model, or the flex-start, spot or reservation-bound provisioning model, since "Different provisioning models draw from separate capacity pools." [20]. A configuration_availability error marks an unsupported combination, and a later retry still fails [20].

How a request is made changes the result. bulkInsert "checks capacity upfront without queuing or retrying", and since minCount defaults to the full count, a bulk request that cannot create every instance creates none unless a lower minCount is set [20]. A managed instance group (MIG) provisions what current capacity allows and retries the rest in the background, and for GPUs with a flexible start Google points to MIG resize requests, which queue the request, or to Flex-start [20]. When a matching reservation's properties differ from the request, Compute Engine tries on-demand capacity in the zone instead, an attempt that "can fail if the zone lacks capacity" [20].

**Microsoft Azure.** The fixes listed are to retry later, resize to another VM size, change the region or zone, remove the zonal constraint, or relax constraints such as Accelerated Networking, a proximity placement group or Ultra Disk [32]. The page also points to reservations: "To ensure that capacity is always available for your workloads, consider using On-demand Capacity Reservations." [32].

**Measured attempts.** In 2023, SkyPilot reported that allocating 8 V100 GPUs on AWS took 3 to 5 location attempts and allocating 8 T4 GPUs took 2 to 7, depending on when the request was made, and found failover especially valuable for scarce resources such as large CPU or GPU VMs [40].

## 4. Reservations that hold capacity

**AWS On-Demand Capacity Reservations.** They reserve capacity "in a specific Availability Zone for any duration", and a reservation for immediate use has no term commitment and can be cancelled at any time [7]. They carry "No billing discount" of their own, and Savings Plans or regional Reserved Instances can discount them [7, 10]. They are "charged at the equivalent On-Demand rate whether you run instances in reserved capacity or not" [10]. In AWS's example, an account that reserves 20 m4.large instances and runs 15 is charged for 15 active instances and for 5 unused instances in the reservation [10]. Reserved capacity counts against the On-Demand quota [7], and creating a reservation can fail when Amazon EC2 lacks sufficient capacity or the quantity exceeds the account's On-Demand Instance limit [9].

Future-dated reservations are requested 5 to 120 days ahead, and AWS recommends at least 56 days [8]. The assessment is typically completed within 5 days [8]. A future-dated reservation needs at least 32 vCPUs and a commitment of at least 14 days, or 12 weeks for U and X instances, and a cancellation charge may apply during the commitment [8]. They are offered for the families "C, G, I, M, R, T, U, and X" [7]. The P family, which holds the H100, H200, B200 and B300 instances, is not in that list, and the G family of GPU instances is [6, 7]. Interruptible Capacity Reservations let an owner lend unused reserved capacity inside its account or organization, and on reclaim the borrowing instances are terminated with "no fallback to On-Demand or Spot" [11].

**Azure on-demand capacity reservation.** It has no one- or three-year term: "Capacity is reserved for you until you delete the reservation." [36]. It is priced at the VM size's rate even when unused. In Microsoft's example, 10 reserved D2s_v3 VMs with 6 deployed are billed as 6 VMs and 4 unused reservations, and

Reserved Instances and savings plans can discount both parts [36]. "Creating capacity reservations requires a quota in the same manner as when you create VMs." [36]. The reservation deployment fails if Azure lacks the capacity or the subscription lacks adequate quota [36].

GPU coverage, as the page read on 8 November 2026 (last updated 16 June 2026): the supported GPU series, all for multi-zone high availability only, are NC v3, NCasT4_v3, NCADSA10_v4, NC_A100_v4, NV v3 and newer, NVadsA10_v5, NVads V710 v5 and NGads V620_v1 [36]. ND-series and NCads H100 v5-series VMs are not supported [36]. Nor are Spot VMs, availability sets, proximity placement groups or Ultra Disk [36].

**Google Cloud on-demand reservations.** They are created "within a few minutes", consumed by standard VMs, and can be modified or deleted at any time. "After Compute Engine provisions your reserved resources, you pay for them whether you use them or not, and for as long as the reservation exists." [24]. They are zone-specific, cannot be used for Flex-start or Spot VMs, and cover "All machine families (except A5X, A4X Max, A4X, A4, A3 Ultra, A3 High with less than 8 GPUs, TPU7x, TPU v6e, and TPU v5p)" [24]. A reservation needs unused quota, and once it is created "quota for those resources is consumed immediately" [25]. It is billed at the reserved resources' rates, "including the same on-demand prices and 1-minute minimum charges", used or not [25].

Future reservations for capacity blocks are arranged through Google's account team, for a contract period, and "Google automatically increases your quota before delivery." [24]. They cover A5X, A4X Max, A4X, A4, A3 Ultra, A3 Mega, A3 High with 8 GPUs, A3 Edge, H4D and TPUs [24]. Reserving for a year or longer requires an attached resource-based commitment [23]. Some A3 Mega and A3 High zones are available with limited capacity, reached through the account team [30].

| Product | How long | Billing | GPU coverage, as read 8 November 2026 |
|---|---|---|---|
| **AWS On-Demand Capacity Reservation** | Any duration, no term for immediate use [7] | On-Demand rate, used or not [10] | Future-dated requests: C, G, I, M, R, T, U and X families [7] |
| **Azure on-demand capacity reservation** | Until deleted [36] | The VM size's rate, used or not [36] | NC, NV and NG series listed, ND-series and NCads H100 v5 not supported [36] |
| **Google on-demand reservation** | Until modified or deleted [24] | The reserved resources' rates, used or not [25] | All families except A5X, A4X Max, A4X, A4, A3 Ultra and A3 High under 8 GPUs [24] |

| Product | How long | Billing | GPU coverage, as read 8 November 2026 |
|---|---|---|---|
| **Google future reservation for capacity blocks** | Contract period, through the account team [24] | Standard instance pricing, discounted with commitments, with a resource-based commitment attached for a year or longer [23, 24] | A5X, A4X Max, A4X, A4, A3 Ultra, A3 Mega, A3 High with 8 GPUs, A3 Edge [24] |

Table 3. Reservations that hold capacity, from the pages cited.

## 5. Scheduled and short-term GPU capacity

**AWS Capacity Blocks for ML.** Capacity Blocks reserve GPU instances on a future date for short-duration machine learning work, placed in EC2 UltraClusters [12]. A block can start up to eight weeks ahead, or "in as soon as 30 minutes", and holds up to 64 instances, with up to 256 across an account's blocks [12]. "Capacity Block cancellations aren't allowed." [12]. Instances in a block do not count against On-Demand Instance limits, and a launch must target the reservation ID [12]. Durations run from one to 14 days, or in multiples of 7 days up to 182 days [13]. The price "depends on available supply and demand at the time the offering was delivered" and does not change after purchase [13].

The price is charged up front, "Savings Plans and Reserved Instance discounts don't apply to Capacity Blocks", and "There is no additional charge for unused time" [14]. Supported types include p6-b300.48xlarge, p6-b200.48xlarge, p5.48xlarge, p5e.48xlarge, p5en.48xlarge and p4d.24xlarge, and P6e-GB200 UltraServers, in listed Regions only [12]. Extensions are requested 1 hour to 57 days before a block ends, up to 182 days in total, and "Capacity Blocks can only be extended if there is sufficient capacity available to support them, which is not guaranteed." [15]. The default quota for Concurrent P5 Capacity Blocks is 0 [3].

**Google Cloud calendar mode.** A future reservation request in calendar mode reserves capacity for 1 to 90 days [26]. For GPU VMs, availability can be viewed up to 60 days in advance, the start has to be at least 87 hours (three days and 15 hours) away, and Google reviews a request within a minute [26]. A request covers 1 to 80 GPU VMs of the A4, A3 Ultra, A3 Mega, A3 High with 8 GPUs or A3 Edge series, or 1 to 256 H4D VMs [26]. It can be cancelled only while the start is more than 30 days away, or within 24 hours of approval when the start is at least four days away, and it cannot be modified [26].

Charges start when the request is fulfilled, at what Google calls DWS pricing, a discount to standard pricing [26], given as up to 53% in Google's comparison table [24]. For GPU and TPU machine types, "no quota is required to reserve capacity" [23, 26]. At the end time the reservation is deleted, and the VMs that consume it are stopped or deleted [26].

**Google Cloud Flex-start VMs.** Flex-start VMs run on Google's Dynamic Workload Scheduler for "a maximum of seven days" [27]. A standalone request waits 0 seconds or 90 to 7,200 seconds, and the VM starts within two minutes of capacity becoming available if it is allocated in that time [27]. Otherwise the request fails [27].

A MIG resize request adds all its VMs at once when all the resources are available [27]. Flex-start consumes preemptible quota, is paid as you go at DWS pricing and "can't use reservations" [27]. The provisioning models page lists a 10-minute minimum run, a discount of up to 53% for A4, A3, A2 and G4, and the A4, A3, A2, G4 and G2 series plus N1 VMs with GPUs attached [23]. The Flex-start page excludes A4X Max and A4X [27].

**Which provisioning models each Google GPU series allows.** The standard model covers "any machine series, except A5X, A4X Max, A4X, A4, and A3 Ultra", and Spot any series except A5X, A4X Max, A4X and bare metal [23]. Reservation-bound VMs run on capacity blocks or calendar mode [23]. As the pages read on 8 November 2026, B200 (A4) and H200 (A3 Ultra) VMs run only as Spot, Flex-start or reserved capacity, and GB200 (A4X) and GB300 (A4X Max) VMs only on capacity reserved through the account team [22, 23, 24, 27]. The provisioning models page puts Vera Rubin (A5X) VMs in the second group, while the Flex-start page excludes only A4X Max and A4X [23, 27].

| | AWS Capacity Blocks | Google calendar mode | Google Flex-start |
|---|---|---|---|
| **Length** | 1 to 14 days, or multiples of 7 days up to 182 [13] | 1 to 90 days [26] | 10 minutes to 7 days [23, 27] |
| **Lead time** | Up to 8 weeks ahead, or as soon as 30 minutes [12] | At least 87 hours ahead for GPU VMs [26] | A standalone VM waits 0 seconds or 90 to 7,200 seconds for capacity [27] |
| **Payment** | Up front, no cancellation [12, 14] | From fulfilment, at DWS pricing [26] | Pay as you go, at DWS pricing [27] |
| **Quota** | Concurrent P5 Capacity Blocks default to 0 [3] | None required for GPU machine types [26] | Preemptible quota [27] |

Table 4. Scheduled and short-term GPU capacity, from the pages cited.

## 6. Discounts that hold no capacity

**AWS.** The comparison table marks regional Reserved Instances and Savings Plans "No capacity reserved" [7]. A regional Reserved Instance "does not reserve capacity", while "A zonal Reserved Instance reserves capacity in the specified Availability Zone." [16]. Savings Plans are "a commitment of using a specified amount of compute power (measured per hour) for a one or three year period" [17].

**Microsoft Azure.** "Reserved Instances provide a billing discount only and don't guarantee capacity." [36]. With Reserved Instances, customers can choose Capacity priority to gain better access, an option that carries no SLA [36]. Savings plans are billing offers that do not affect the runtime state of resources, and their purchases cannot be canceled or refunded [37].

**Google Cloud.** Resource-based commitments for GPUs carry capacity: "When you purchase resource-based commitments for GPUs, you must reserve those GPUs and attach those reservations to your commitment." [28]. The attached reservations cannot be deleted, resized or modified during the commitment [28]. Compute flexible commitments, which cover G2 and G4 GPUs only, "don't support attached reservations and don't offer the assured capacity that comes with attached reservations" [29].

**Spot capacity.** Spot is a discount on capacity the provider can take back. Azure's Spot VMs carry no SLA and are evicted with 30 seconds of notice [38]. Google keeps Spot in its own quota and capacity pool [20, 22], and AWS reports a Spot request it cannot fill as capacity-not-available [19]. The cost of interruptions to training runs is the subject of What research shows about training on spot GPUs (https://rillor.com/insights/spot-gpu-training-research).

| Product | Holds capacity | Source |
|---|---|---|
| **AWS regional Reserved Instance, Savings Plans** | No | [7, 16] |
| **AWS zonal Reserved Instance** | Yes, in one Availability Zone | [16] |
| **AWS On-Demand Capacity Reservation** | Yes, with no billing discount of its own | [7] |
| **Azure Reserved Instance, savings plan** | No | [36, 37] |
| **Azure on-demand capacity reservation** | Yes | [36] |
| **Google resource-based GPU commitment** | Yes, through its attached reservations | [28] |
| **Google compute flexible commitment** | No | [29] |

Table 5. Which commitments hold capacity, from the pages cited.

## 7. What reported availability shows

**Scores.** AWS's Spot placement score indicates how likely a Spot request is to succeed in a Region or Availability Zone, and it "does not provide any guarantees in terms of available capacity or risk of interruption" [18]. Google's obtainability score, in Preview, is a value from 0.0 to 1.0 based on "real-time availability of your requested resources and the success rate of recent creation requests", and "Obtainability scores don't guarantee capacity." [31].

**Forward views.** Calendar mode shows GPU availability up to 60 days ahead [26], and Capacity Block offerings show GPU capacity on future dates [12].

**Region lists.** Lambda's instance-types response carries regions_with_capacity_available, "A list of the regions in which this instance type is available." [39].

**Research.** Ding-Dong Ditch probed 68 instance types in 15 regions on AWS and Azure with 336,033 spot requests and analyzed 2,635 interruption events [41]. It covers spot instances in general. It found that provider datasets show relative trends and leave open whether a specific number of instances can be obtained at a given time [41].

# Implications

**For an engineer whose GPU launch just failed.** The error code says which check failed. A quota error (InstanceLimitExceeded, VcpuLimitExceeded, QUOTA_EXCEEDED, OperationNotAllowed or global/quota-exceeded) calls for a quota request. A capacity error (InsufficientInstanceCapacity, ZONE_RESOURCE_POOL_EXHAUSTED, AllocationFailed or insufficient-capacity) does not go away with more quota [1, 2, 20, 21, 32, 34, 39]. The listed fixes change the zone, size, type or provisioning model of the request, and Google advises keeping running instances that have no reservation behind them [1, 20, 32].

**For a planner holding GPUs for a run.** Quota comes first: AWS's P instance quotas start at 0 vCPUs [3], larger increases go to Support review [5], and Google's B200, H200, GB200, GB300 and Vera Rubin series have no standard quota [22]. The window then picks the product: Capacity Blocks for one day to 182 days with no cancellation [12, 13], calendar mode for 1 to 90 days booked at least 87 hours ahead [26], Flex-start for up to seven days, after a wait of up to two hours for a standalone VM [27], and on-demand reservations for open-ended holds [7, 24, 36].

**For finance and procurement.** Reserved capacity is billed while idle. On-demand reservations bill whether used or not [10, 25, 36], and Capacity Blocks are paid up front for the whole block [14]. A billing discount holds no capacity unless it carries reservations, as Google's GPU resource-based commitments do [28], or reserves capacity itself, as AWS's zonal Reserved Instances do [16].

**For anyone reading availability data.** In Rillor's GPU and neocloud pricing method, availability means what a provider reports as available to rent, recorded as published beside the price at the moment of capture, and it describes rentable offers at that moment. Whether a launch of a given size succeeds is a separate question, and the provider scores built for it carry no guarantee [18, 31, 41].

# Limits of this analysis

- Provider pages change often. The Google pages read were last updated between 26 October and 9 November 2026, the Azure capacity reservation page on 16 June 2026, and the AWS pages carry no date. Every list here is as the pages read on 8 November 2026.
- No source read gives a launch failure rate for H100, H200, B200 or B300 instances at any provider. The only measured location attempts are SkyPilot's 2023 figures for V100 and T4 GPUs [40].
- No source read gives a turnaround time for quota increases. AWS says only that larger requests take time [5].
- The Azure quota pages read give example output, not default quotas for GPU families [33].
- Whether AWS sells zonal Reserved Instances for P instances was not checked, and Azure products for scheduled GPU blocks were not researched.
- Neocloud coverage is one API reference, Lambda's [39]. No instances were launched and no reservations made for this report.

# Sources


1. Troubleshoot Amazon EC2 instance launch issues. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/troubleshooting-launch.html

2. Error codes for the Amazon EC2 API. Amazon EC2 Developer Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/ec2/latest/devguide/errors-overview.html

3. Amazon EC2 instance type quotas. Amazon EC2 Instance Types. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/ec2/latest/instancetypes/ec2-instance-quotas.html

4. Purchasing On-Demand Instances for Amazon EC2. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-on-demand-instances.html

5. Requesting a quota increase. Service Quotas User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/servicequotas/latest/userguide/request-quota-increase.html

6. Specifications for Amazon EC2 accelerated computing instances. Amazon EC2 Instance Types. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/ec2/latest/instancetypes/ac.html

7. Reserve compute capacity with EC2 On-Demand Capacity Reservations. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-reservations.html

8. 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

9. Create a Capacity Reservation. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/capacity-reservations-create.html

10. 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

11. Interruptible Capacity Reservations. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/interruptible-capacity-reservations.html

12. 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

13. 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

14. 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

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

16. Regional and zonal Reserved Instances (scope). Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/reserved-instances-scope.html

17. What are Savings Plans? Savings Plans User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/savingsplans/latest/userguide/what-is-savings-plans.html

18. Spot placement score. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-placement-score.html

19. Get the status of a Spot Instance request. Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-request-status.html

20. Troubleshoot resource availability errors. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/troubleshooting/troubleshooting-resource-availability

21. Troubleshoot quota errors. Cloud Quotas documentation. Google Cloud, last updated 26 October 2026, viewed 8 November 2026. https://docs.cloud.google.com/docs/quotas/troubleshoot

22. Allocation quotas. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/compute/resource-usage

23. Compute Engine instances provisioning models. Compute Engine documentation. Google Cloud, last updated 9 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/provisioning-models

24. Choose a reservation type. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/compute/docs/instances/choose-reservation-type

25. About reservations. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/reservations-overview

26. About future reservation requests in calendar mode. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/future-reservations-calendar-mode-overview

27. About Flex-start VMs. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/about-flex-start-vms

28. Resource-based committed use discounts. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/signing-up-committed-use-discounts

29. Committed use discounts (CUDs) for Compute Engine. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/committed-use-discounts-overview

30. GPU locations. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/compute/docs/regions-zones/gpu-regions-zones

31. View the availability of Spot VMs. Compute Engine documentation. Google Cloud, last updated 6 November 2026, viewed 10 November 2026. https://docs.cloud.google.com/compute/docs/instances/view-vm-availability

32. Troubleshoot allocation failures when you create or resize VMs in Azure. Microsoft Learn. Microsoft, updated 16 April 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/troubleshoot/azure/virtual-machines/windows/allocation-failure

33. Check vCPU quotas. Microsoft Learn. Microsoft, updated 10 April 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/quotas

34. Resolve errors for resource quotas. Microsoft Learn. Microsoft, updated 12 October 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/azure-resource-manager/troubleshooting/error-resource-quota

35. Increase VM-family vCPU quotas. Microsoft Learn. Microsoft, updated 22 January 2025, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/quotas/per-vm-quota-requests

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

37. What are savings plans? Microsoft Learn. Microsoft, updated 29 October 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/cost-management-billing/savings-plan/savings-plan-overview

38. Azure Spot Virtual Machines. Microsoft Learn. Microsoft, updated 6 February 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/spot-vms

39. Lambda Cloud API reference. Lambda, viewed 8 November 2026. https://docs.lambda.com/api/cloud

40. Yang, Wu, Luo and others. SkyPilot: An Intercloud Broker for Sky Computing. 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 437-455. USENIX, April 2023, viewed 8 November 2026. https://www.usenix.org/system/files/nsdi23-yang-zongheng.pdf

41. Kim, Song, Kim and Lee. Ding-Dong Ditch: Peeking Into Spot Instance Availability. Accepted at IEEE CLOUD 2026, arXiv 2604.16457 v2. arXiv, 27 May 2026 (first version 8 April 2026), viewed 8 November 2026. https://arxiv.org/abs/2604.16457

## About this report

Published by Rillor on 11 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. Why cloud GPU launches fail and how reservations hold capacity. 11 November 2026. https://rillor.com/insights/gpu-capacity-errors-quotas-reservations

## 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