AWS GPU Instance Pricing 2026: P6-B300, P6-B200, P5 Costs

Published July 18, 2026 34 min read

Executive Summary

Amazon Web Services (AWS) prices its graphics processing unit (GPU) instances for artificial intelligence (AI) workloads through four distinct mechanisms, on-demand hourly billing, Spot discounts, Savings Plans, and EC2 Capacity Blocks for machine learning (ML), and as of July 2026 those prices sit meaningfully higher than they did twelve months earlier. The flagship **Amazon EC2 P6-B300** instance, powered by eight **NVIDIA Blackwell Ultra** GPUs, carries an on-demand list price of $142.42 per hour for the p6-b300.48xlarge size (Source: economize.cloud) (Source: instances.vantage.sh), while the predecessor **P6-B200** (eight NVIDIA Blackwell GPUs) lists at $113.93 per hour (Source: economize.cloud) (Source: instances.vantage.sh). The older **P5** family, built on NVIDIA H100 and H200 Tensor Core GPUs, ranges from $55.04 per hour for the eight-GPU p5.48xlarge on-demand (Source: instances.vantage.sh) up through the p5en.48xlarge, which starts at $63.30 per hour (Source: instances.vantage.sh).

The pricing picture is complicated by AWS's reservation product, EC2 Capacity Blocks for ML, which most large training customers actually use because on-demand capacity for eight-GPU nodes is scarce. AWS raised Capacity Block rates by approximately 15 percent in January 2026 and by a further 20 percent effective July 1, 2026 (Source: theregister.com) (Source: finance.yahoo.com), meaning a customer paying list price today pays roughly a third more per accelerator-hour than they did at the start of the year for identical NVIDIA silicon. Under the current Capacity Block schedule, the P6-B300 bills at $14.04 per accelerator per hour, the P6-B200 at $12.355, the P5 at $5.191 (US regions), the P5e at $5.97, and the P5en at $6.865 (Source: finance.yahoo.com) (Source: economize.cloud).

AWS frames the increases as routine supply-and-demand repricing rather than an admission of scarcity, but the timing follows an unusually aggressive capital program: AWS reported first-quarter 2026 revenue growth of 28 percent year over year to $37.6 billion, its fastest pace in three years, and Amazon has committed roughly $200 billion in 2026 capital expenditure to AI infrastructure while lining up delivery of a million Nvidia GPUs by the end of 2027 (Source: finance.yahoo.com). Just seven months before the January hike, AWS had announced a price cut of up to 45 percent on P4 and P5 on-demand and Savings Plan pricing (Source: aws.amazon.com), a reversal that trade press has called the end of two decades of cloud prices moving only downward (Source: theregister.com).

Positioned against competitors, AWS's P5 on-demand rate of roughly $6.88 per H100 GPU per hour undercuts Microsoft Azure's ND96isr H100 v5 (about $12.29 per GPU) and Google Cloud's a3-highgpu-8g ($11.06 per GPU), but sits well above Oracle Cloud Infrastructure's $10.00 flat rate and far above specialized "neocloud" providers such as CoreWeave ($6.16 per GPU) and Lambda ($3.99 to $6.16 per GPU depending on configuration) (Source: thundercompute.com) (Source: coreweave.com) (Source: lambda.ai). For buyers, the practical decision hinges less on the sticker price than on availability: EC2 Capacity Blocks let customers reserve up to 64 instances (512 GPUs) as far as eight weeks in advance for terms up to six months (Source: aws.amazon.com), a guarantee that on-demand and Spot pricing cannot match given how rarely eight-GPU P5 and P6 capacity appears in the on-demand pool (Source: reddit.com). This report walks through every current AWS GPU instance price, the mechanics of each purchasing option, the twelve-month pricing history, cross-cloud benchmarking, and named customer deployments to give a complete, verifiable answer to what AWS GPU instances cost as of July 2026.

Introduction and Background

Training and serving large language models and other frontier AI systems requires clusters of specialized accelerators, and Amazon Elastic Compute Cloud (EC2), AWS's virtual server product, is one of three hyperscale clouds (alongside Microsoft Azure and Google Cloud) that rents such clusters by the hour. AWS organizes its GPU-accelerated compute under the "P" instance family name, distinguishing it from general-purpose and other accelerator families such as Trainium-based Trn instances. Understanding AWS GPU instance pricing in mid-2026 requires separating two questions that are often conflated: what does an instance cost, and which purchasing mechanism will actually deliver that instance when a customer needs it.

The current P-family lineup spans two silicon generations. The **P5 series** (P5, P5e, P5en) runs on NVIDIA's Hopper-generation H100 and H200 Tensor Core GPUs and became generally available starting in 2023, with P5e and P5en variants adding higher-memory H200 GPUs and improved networking in subsequent quarters (Source: spheron.network). The **P6 series** (P6-B200, P6-B300, and the UltraServer-only P6e-GB200 and P6e-GB300) runs on NVIDIA's newer Blackwell and Blackwell Ultra architecture. P6-B200 reached general availability on May 15, 2025, and P6-B300 followed on November 18, 2025 (Source: aws-news.com) (Source: aws-news.com). The UltraServer-only P6e-GB300, accelerated by NVIDIA's GB300 NVL72 architecture, reached general availability shortly afterward, in December 2025 (Source: aws-news.com). Both generations are sold under four purchasing models: On-Demand (pay per second with no commitment), Spot (spare capacity at a discount, subject to interruption), Savings Plans (a one- or three-year usage commitment in exchange for a lower rate), and EC2 Capacity Blocks for ML (a reservation of specific accelerator capacity for a defined future window, paid up front).

Pricing for these instances has been unusually volatile for a cloud product. AWS cut P4 and P5 on-demand and Savings Plan prices by up to 45 percent in June 2025 to make Hopper-generation capacity more broadly accessible (Source: aws.amazon.com), then raised Capacity Block reservation prices twice within the following twelve months, first by roughly 15 percent in January 2026 and again by roughly 20 percent on July 1, 2026 (Source: theregister.com)(Source: finance.yahoo.com). That reversal, following two decades in which AWS's public messaging emphasized falling prices, is itself a data point worth understanding before comparing sticker prices across instance types or competing clouds. This report catalogs, with sourced figures, what each P5 and P6 instance costs under each purchasing model as of July 2026, how those prices have moved over the past year, how AWS pricing compares to Azure, Google Cloud, Oracle Cloud, and specialized GPU cloud providers, and what the pricing trend implies for organizations planning AI infrastructure budgets.

P6-B300 and P6-B200 Pricing: The Current Blackwell Generation

AWS's newest and most expensive GPU instances are the **P6-B300** and **P6-B200**, both built around NVIDIA's Blackwell architecture and both sold in a single size, the 48xlarge, which packages eight GPUs with 192 virtual CPUs (vCPUs). The P6-B300, accelerated by eight **NVIDIA Blackwell Ultra** GPUs, provides 4,096 GiB of system memory, 6.4 Tbps of Elastic Fabric Adapter (EFA) networking bandwidth, and eight 3.84 TB NVMe storage volumes on 192 vCPUs (Source: instances.vantage.sh). AWS states that P6-B300 delivers "2 times more networking bandwidth, and 1.5 times more GPU memory compared to previous generation instances" (Source: aws.amazon.com), a comparison anchored specifically against the P6-B200. On-demand, the p6-b300.48xlarge lists at $142.42 per hour, corroborated independently by two third-party pricing aggregators that track AWS's published rate card (Source: economize.cloud) (Source: instances.vantage.sh).

The **P6-B200**, accelerated by eight standard NVIDIA Blackwell GPUs, carries 2,048 GiB (2 TiB) of system memory on 192 vCPUs with 3,200 Gbps of aggregate network bandwidth (Source: instances.vantage.sh), and pairs its eight Blackwell GPUs with fifth-generation Intel Xeon Scalable ("Emerald Rapids") host processors. At general availability, AWS quantified P6-B200's improvement over the prior P5en generation precisely: "up to 125 percent improvement in GPU TFLOPs, 27 percent increase in GPU memory size, and 60 percent increase in GPU memory bandwidth compared to P5en instances" (Source: aws.amazon.com). The p6-b200.48xlarge lists on-demand at $113.93 per hour, again cross-confirmed by two independent pricing trackers (Source: economize.cloud) (Source: instances.vantage.sh).

Because eight-GPU on-demand capacity for either instance is difficult to obtain at scale, most P6 buyers transact through EC2 Capacity Blocks for ML instead. Under the rate schedule that took effect July 1, 2026, the p6-b300.48xlarge Capacity Block costs $112.32 per instance-hour in US East (N. Virginia) and US West (Oregon), equivalent to $14.04 per accelerator per hour, and $117.00 ($14.625 per accelerator) in AWS GovCloud (US-East) (Source: aws.amazon.com). The p6-b200.48xlarge Capacity Block costs $98.84 per instance-hour ($12.355 per accelerator) across US East (Ohio, N. Virginia), US West (Oregon), and Asia Pacific (Mumbai), rising to $102.96 ($12.870 per accelerator) in GovCloud regions, per the same schedule (Source: aws.amazon.com). Both instances remain available only in a limited set of regions, having launched first in US West (Oregon) before expanding into additional Availability Zones and AWS GovCloud regions during 2026 (Source: aws.amazon.com). Neither instance is available inside AWS's UltraServer configuration; that architecture is reserved for the separate P6e-GB200 and P6e-GB300 products, which pool up to 72 GPUs in a single NVLink domain for the largest frontier-model training runs.

P5, P5e, and P5en Pricing: The Hopper Generation

The **P5** family remains AWS's most widely deployed GPU instance line and is now the value tier beneath P6. Amazon EC2 P5 instances pair eight NVIDIA H100 GPUs with 192 vCPUs and 2 TiB of instance memory on the flagship p5.48xlarge size, alongside a smaller single-GPU p5.4xlarge configuration for lighter workloads (Source: instances.vantage.sh). The p5.48xlarge lists on-demand at $55.04 per hour, roughly $6.88 per GPU-hour when divided across its eight accelerators (Source: instances.vantage.sh). AWS markets P5, P5e, and P5en collectively as delivering "up to 40% savings on DL training and HPC infrastructure costs compared to previous-generation GPU-based EC2 instances" (Source: aws.amazon.com), a claim measured against the earlier P4d and P4de generation rather than against P6.

P5e and **P5en** upgrade the accelerator to NVIDIA's H200, adding memory capacity over the standard H100-based P5, and P5en further upgrades the host-to-GPU interconnect to Gen5 PCIe with a third-generation EFA implementation, which AWS says "shows up to 35% improvement in latency compared to P5 that uses the previous generation of EFA and Nitro" (Source: aws.amazon.com). The p5en.48xlarge lists on-demand starting at $63.30 per hour (Source: instances.vantage.sh). AWS does not publish a standard on-demand rate for p5e.48xlarge or p5en.48xlarge on its general EC2 pricing pages in the way it does for p5.48xlarge, and third-party pricing trackers note that customers typically need to work directly with an AWS account team to secure these configurations at scale (Source: spheron.network).

Capacity Block pricing is where P5e cost is most clearly published. Under the July 2026 rate schedule, the p5e.48xlarge Capacity Block costs $47.76 per instance-hour, equivalent to $5.97 per accelerator per hour, uniformly across all listed regions from US East (Ohio) to Asia Pacific (Tokyo) (Source: aws.amazon.com). The p5en.48xlarge Capacity Block costs $54.92 per instance-hour ($6.865 per accelerator) in US regions and $49.928 ($6.241 per accelerator) internationally, while theregister.com's coverage of the January 2026 increase separately documented the prior p5en.48xlarge rate climbing "from $36.18 to $41.61" per hour in that round alone (Source: theregister.com). Standard p5.48xlarge Capacity Blocks cost $41.528 per instance-hour ($5.191 per accelerator) in US regions, dropping to $37.76 ($4.720 per accelerator) in most international regions including Tokyo, Mumbai, London, and São Paulo, and the identical $5.191 per-accelerator rate applies to the single-GPU p5.4xlarge, useful for fine-tuning or inference workloads that do not need a full eight-GPU node (Source: finance.yahoo.com). The same July 2026 schedule also lifted the older P4de family's US rate to $2.214 per accelerator-hour, the sole non-Hopper, non-Blackwell instance swept into the increase (Source: finance.yahoo.com).

Since launch, AWS has repeatedly used P5-family customer testimonials to justify the instance's value proposition. Anthropic cofounder Tom Brown said at P5's 2023 debut that the company was "using Amazon EC2 P4 instances extensively today, and we are excited about the launch of P5 instances. We expect them to deliver substantial price-performance benefits over P4d instances" (Source: aws.amazon.com), while Cohere chief executive Aidan Gomez said P5's H100 GPUs "will unleash the ability of businesses to create, grow, and scale faster with its computing power" (Source: aws.amazon.com).

Purchasing Options Compared: On-Demand, Spot, Savings Plans, and Capacity Blocks

AWS sells every P5 and P6 instance through as many as four purchasing mechanisms, and the mechanism chosen changes the effective price by a factor of two to four. **On-Demand** pricing, billed per second with no upfront commitment, is the reference rate quoted throughout this report: $142.42 per hour for p6-b300.48xlarge, $113.93 for p6-b200.48xlarge, and $55.04 for p5.48xlarge (Source: economize.cloud) (Source: instances.vantage.sh). On-demand is the most expensive per hour but requires no commitment and no advance planning, provided capacity is actually available, which for eight-GPU P5 and P6 nodes is frequently the binding constraint rather than price.

Spot Instances draw on AWS's unused capacity and are priced dynamically, at a discount AWS advertises as "up to 90% off compared to On-Demand pricing" (Source: aws.amazon.com), though that ceiling rarely applies to P5 and P6 given persistent demand. Third-party analysis of AWS's published Spot data puts P5 Spot discounts closer to 44 percent off on-demand, bringing an eight-GPU p5.48xlarge to roughly $30.64 per hour (about $3.83 per GPU) when capacity is available, with the caveat that P5 Spot inventory "shows up infrequently and comes with high interruption rates due

to persistent demand" (Source: spheron.network). Reddit commenters familiar with AWS billing note a structural advantage of Spot over Capacity Blocks in this respect: "With Spot instances, they will never exceed the published On-demand rate. These have no cap," one AWS practitioner observed in a discussion of the July 2026 price increase, contrasting Spot's ceiling with Capacity Blocks' unbounded dynamic repricing (Source: reddit.com). A separate commenter on the same thread pushed back that the underlying mechanism is nothing new, arguing "Capacity Blocks have ALWAYS been Dynamic pricing, i don't know why he is whining about 'spot' price fluctuation" (Source: reddit.com).

Savings Plans, AWS's commitment-based discount, extend to P6-B200 as well as the P5 family, offering EC2 Instance Savings Plans (discounts tied to a specific instance family and region) or the more flexible Compute Savings Plans that apply across families, which AWS's own savings-plans page advertises as offering "savings up to 72% in exchange for commitment to usage of individual instance families in a Region" (Source: aws.amazon.com). AWS's June 2025 repricing quantified P5-specific discounts precisely: a one-year P5 Instance Savings Plan cut 45 percent off the May 31, 2025 baseline and a three-year commitment cut 44 percent, while P5en Instance Savings Plans cut 26 percent (one-year) and 25 percent (three-year) (Source: aws.amazon.com). That same announcement extended Savings Plan support to the then-new P6-B200, and P6-B300 launched with Savings Plans support from day one seven months later (Source: aws-news.com).

EC2 Capacity Blocks for ML is the fourth and, for large training runs, the most commonly used mechanism. As one wire report summarized it, "Capacity Blocks for ML are a reserved-capacity product that lets enterprises secure scarce GPU instances on a future date for time-bound workloads, typically large-scale model training" (Source: finance.yahoo.com). Because eight-GPU P5 and P6 capacity is scarce on the general on-demand market, one Reddit user summarized the practical reality bluntly: "Capacity Blocks are really the only way you can even use these instance types. It's extremely rare that you can ever just spin one of these up on-demand. So in effect, it's a way for them to advertise one price (on-demand) while actually charging more" (Source: reddit.com). AWS's own framing acknowledges the pricing is intentionally dynamic: "Amazon EC2 Capacity Blocks for ML reservation prices are updated periodically based on supply and demand," a company spokesperson told trade press covering the January 2026 increase (Source: theregister.com).

Table 1 below summarizes current EC2 Capacity Block rates across the full P5 and P6 lineup as of the July 1, 2026 schedule.

INSTANCE	GPU (COUNT X MODEL)	PER-INSTANCE RATE (US REGIONS)	PER-ACCELERATOR RATE (US)	PER-ACCELERATOR RATE (INTERNATIONAL, WHERE LOWER)
p6-b300.48xlarge	8 x NVIDIA Blackwell Ultra (B300)	$112.32/hr	$14.04/hr	Not offered outside US/GovCloud
p6-b200.48xlarge	8 x NVIDIA Blackwell (B200)	$98.84/hr	$12.355/hr	$12.355/hr (Mumbai, same as US)
p5en.48xlarge	8 x NVIDIA H200	$54.92/hr	$6.865/hr	$6.241/hr (Europe, Asia Pacific)
p5e.48xlarge	8 x NVIDIA H200	$47.76/hr	$5.97/hr	$5.97/hr (uniform globally)
p5.48xlarge	8 x NVIDIA H100	$41.528/hr	$5.191/hr	$4.720/hr (Tokyo, Mumbai, London, Sao Paulo)
p5.4xlarge	1 x NVIDIA H100	$5.191/hr	$5.191/hr	$4.720/hr (same regions)

Source: AWS EC2 Capacity Blocks for ML pricing page and cross-checked wire coverage, effective July 1, 2026 (Source: finance.yahoo.com) (Source: economize.cloud).

The table illustrates two patterns worth calling out in prose. First, the per-accelerator rate does not track memory or compute capability in a simple linear way: P5e and P5en both use H200 silicon, yet P5en's faster networking and PCIe Gen5 host interconnect command roughly a 15 percent premium per accelerator over P5e. Second, the newest Blackwell Ultra silicon in P6-B300 costs more than double the per-accelerator rate of the three-generation-older P5, a gap that reflects both raw performance improvement and the acute early-life scarcity premium that new GPU architectures typically command in cloud capacity markets.

Comparative Context and Market Positioning

AWS GPU pricing sits in the middle of a wide field once compared against Microsoft Azure, Google Cloud Platform (GCP), Oracle Cloud Infrastructure (OCI), and specialized "neocloud" GPU providers such as CoreWeave and Lambda. On raw per-GPU on-demand pricing for H100-class silicon, AWS's $6.88 per GPU-hour on p5.48xlarge undercuts both of its hyperscale rivals. Microsoft's ND96isr H100 v5, an eight-GPU instance whose GPUs each connect through "a dedicated, topology-agnostic 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand connection" (Source: learn.microsoft.com), lists at approximately $98.32 per hour, or roughly $12.29 per GPU (Source: thundercompute.com). Azure's list rate also varies substantially by region: independent analysis of Azure's own calculator found "West US 2 runs roughly 7% higher for H100 instances, Europe West adds about 19%, Southeast Asia around 30%, and Australia East around 33%" relative to the East US baseline used throughout this report (Source: thundercompute.com).

Google Cloud's a3-highgpu-8g, an eight-GPU H100 configuration, is priced at $88.49 per hour on Google's own accelerator-optimized pricing page (Source: cloud.google.com), while the higher-bandwidth a3-megagpu-8g variant runs $93.40 per hour (Source: cloud.google.com) and the H200-based a3-ultragpu-8g runs $84.81 per hour (Source: cloud.google.com). At the smaller end, GCP's a2-ultragpu-1g offers a single NVIDIA A100 GPU at $5.06879789 per hour, a useful reference point for older Ampere-generation comparisons (Source: cloud.google.com). GCP's newest Blackwell-based a4-highgpu-8g (NVIDIA B200) has no standard on-demand rate published and is sold only through Dynamic Workload Scheduler flex-start pricing at $64.44 per hour (Source: cloud.google.com), a pricing structure notably different from AWS's straightforward hourly rate for P6. GCP's committed-use discounting is also steeper than AWS's on paper: Google states a three-year Compute Flexible commitment yields "a 46% discount over your committed hourly spend amount," and separately, its Sustained Use Discounts let a workload that "run[s] the entire month" earn "up to a 30% net discount off of the resource cost" with no upfront commitment at all, a structure AWS does not offer on P5 or P6 (Source: cloud.google.com) (Source: cloud.google.com).

Oracle Cloud Infrastructure undercuts all three hyperscalers on raw hardware price. OCI's published price list sets its bare-metal eight-GPU H100 and H200 shapes (BM.GPU.H100.8 and BM.GPU.H200.8) at a flat $10.00 per GPU per hour (Source: oracle.com), or $80 per hour for the full eight-GPU node, its Blackwell-generation BM.GPU.B200.8 shape at $14.00 per GPU per hour, and its newest BM.GPU.B300.8 shape, based on NVIDIA Blackwell Ultra silicon, at $15.00 per GPU per hour (Source: oracle.com) (Source: oracle.com). Oracle's UltraServer-class shapes price even higher: the four-GPU BM.GPU.GB200.41 lists at $16.00 per GPU per hour and the four-GPU BM.GPU.GB300.42 at $18.00 per GPU per hour (Source: oracle.com) (Source: oracle.com). Oracle also discounts unused-capacity and preemptible instances directly off its own list price: "preemptible instances are priced at 50% of the price of regular instances," while "Unused Capacity reservations are priced at 85% of the price of regular instances" (Source: oracle.com) (Source: oracle.com), a simpler discount structure than AWS's four-mechanism system. Specialized GPU cloud providers price lower still: CoreWeave lists its eight-GPU HGX H100 node at $49.24 per hour on-demand (about $6.16 per GPU) with Spot pricing around $19.51 to $19.71 per hour, and its HGX B200 node at $68.80 per hour (Source: coreweave.com) (Source: coreweave.com). Lambda prices single H100 SXM GPUs from $3.99 per GPU-hour in its largest eight-GPU configuration and NVIDIA B200 SXM6 GPUs from $6.69 per GPU-hour in the same configuration, with its managed 1-Click Cluster product pricing sixteen-GPU H100 clusters at $6.16 per GPU-hour and discounting B200 clusters of 256 GPUs or more to $8.87 per GPU-hour; for older Ampere-generation hardware, Lambda's smallest on-demand configuration prices NVIDIA A100 SXM GPUs at $1.99 per GPU-hour (Source: lambda.ai) (Source: lambda.ai) (Source: lambda.ai).

Spot and preemptible discounting varies widely by provider and is a second axis worth comparing alongside headline on-demand rates. Google Cloud states its "Spot prices are variable and can change up to once every day, but provide discounts of up to 91% off of the corresponding default price" (Source: cloud.google.com), a deeper advertised discount ceiling than AWS's 90 percent Spot maximum, though as with AWS's own Spot pool, GCP's steepest discounts are least available on the newest, most contested GPU shapes. Azure's spot pricing, per independent analysis, "can reduce costs by 70 to 82%," but carries only "30 seconds of eviction notice" compared to the longer interruption warnings AWS and GCP typically provide (Source: thundercompute.com).

Table 2 collates these figures into a single per-GPU comparison for H100-class, on-demand, eight-GPU configurations, the most directly comparable slice across providers.

PROVIDER	INSTANCE / SHAPE	ON-DEMAND RATE (8-GPU NODE)	PER-GPU RATE
CoreWeave	HGX H100	$49.24/hr	~$6.16/hr
AWS	p5.48xlarge (H100)	$55.04/hr	~$6.88/hr
Oracle Cloud	BM.GPU.H100.8	$80.00/hr	$10.00/hr
Google Cloud	a3-highgpu-8g (H100)	$88.49/hr	~$11.06/hr
Azure	ND96isr H100 v5	~$98.32/hr	~$12.29/hr

Sources: CoreWeave, AWS (via aggregator), Oracle, Google Cloud, and Azure (via aggregator) pricing pages, as cited above (Source: coreweave.com) (Source: instances.vantage.sh) (Source: oracle.com) (Source: cloud.google.com) (Source: learn.microsoft.com).

The pattern in Table 2 places AWS second-cheapest among the four hyperscale-class providers for H100 capacity, ahead of Google Cloud and Azure but behind Oracle's flat-rate bare-metal pricing, and roughly 12 percent more expensive than CoreWeave's specialized offering. That gap narrows further once Capacity Block and Savings Plan discounting are applied on the AWS side, but it also understates the value many enterprise buyers place on AWS's broader managed-service ecosystem, including SageMaker HyperPod for training-cluster orchestration, deep integration with Amazon S3 and other storage, and the AWS Nitro System's hardware-isolated security architecture. One competitor blog framed the tradeoff starkly, arguing "AWS is charging 2.5x to 3.5x more than independent providers for the same NVIDIA GPUs. On an H100, that is roughly $2,300 more per month, per GPU" (Source: servergurus.com), a claim that should be read as marketing from a rival GPU host but that nonetheless reflects the directionally real premium AWS commands versus bare-metal specialists.

Data Analysis and Evidence

The clearest quantitative story in AWS GPU pricing over the past thirteen months is a round trip: a large cut followed by two increases that, compounded, roughly offset it for reservation-based buyers. In June 2025, AWS announced price reductions of up to 45 percent on P4d, P4de, P5, and P5en instances, applying to both On-Demand and Savings Plan purchases beginning June 1 and June 4, 2025 respectively (Source: aws.amazon.com). The reduction table AWS published showed P5 (H100) getting a 44 percent cut to On-Demand pricing and a 45 percent cut to one-year EC2 Instance Savings Plans, while P4d and P4de (A100) received 33 percent On-Demand cuts and P5en (H200) received a smaller 25 percent On-Demand cut (Source: aws.amazon.com).

Roughly seven months later, on Saturday, January 4, 2026, AWS reversed direction and raised EC2 Capacity Block prices for ML by approximately 15 percent across the board, without a formal press release, a move trade press flagged specifically for its quiet timing (Source: theregister.com). That increase moved the p5e.48xlarge Capacity Block rate from $34.61 to $39.80 per hour across most regions, with steeper increases in US West (N. California), where p5e rates rose from $43.26 to $49.75 (Source: theregister.com). AWS's stated rationale, delivered to trade press by an Amazon spokesperson, was that "EC2 Capacity Blocks for ML pricing vary based on supply and demand patterns, as described on the product detail page. This price adjustment reflects the supply/demand patterns we expect this quarter" (Source: theregister.com).

A second increase of approximately 20 percent followed on July 1, 2026, again attributed by AWS to "supply and demand dynamics," and applying uniformly to P6-B300, P6-B200, P5, P5e, P5en, and P4de Capacity Blocks while leaving all other EC2 prices unchanged (Source: finance.yahoo.com). AWS's own pricing page confirms this rate schedule is not final: it states that "the current prices are scheduled to be updated next in October, 2026" (Source: aws.amazon.com), signaling a quarterly repricing cadence rather than a one-off correction. Compounding the two 2026 increases (15 percent then 20 percent) on a representative baseline produces an increase of roughly 38 percent from January 1, 2026 levels, a figure one competitor blog rounded to "roughly 35% more than you were on January 1 for the exact same hardware" (Source: servergurus.com).

The macro backdrop helps explain AWS's pricing leverage. AWS reported first-quarter 2026 revenue growth of 28 percent year over year to $37.6 billion, its fastest growth rate in more than three years (Source: finance.yahoo.com), while Amazon has earmarked approximately $200 billion of 2026 capital expenditure for AI infrastructure and is contracted to receive one million Nvidia GPU chips by the end of 2027 under a supply agreement first reported by Reuters (Source: finance.yahoo.com). On the supply side, cost pressure is traceable to the memory components inside the GPUs themselves: one industry analysis notes that "NVIDIA cannot make enough GPUs. HBM3e memory is the bottleneck. Each stack costs $300, and an

H200 needs four of them," and separately claims "enterprise AI spending hit $401 billion this year and keeps climbing" (Source: servergurus.com) (Source: servergurus.com), pointing to component scarcity as a structural driver behind repeated GPU price hikes across the industry, not an AWS-specific decision alone.

On the technical performance side, AWS has published specific, falsifiable improvement figures rather than vague marketing multipliers. Third-generation EC2 UltraClusters, which host P6e-GB200 UltraServers, deliver substantial power and cabling efficiency gains over prior generations according to AWS Vice President of Compute and ML Services David Brown, while the fourth-generation EFA networking used in P6e-GB200 and P6-B200 delivers "up to 18% faster collective communications in distributed training compared to P5en instances that use EFAv3" (Source: aws.amazon.com). These efficiency gains are a partial counterweight to the higher headline hourly rate of newer instances: a workload that completes 18 percent faster on P6-B200 than on P5en offsets part of P6-B200's higher per-hour cost when measured on a cost-per-completed-job basis rather than cost-per-hour.

Case Studies and Real-World Examples

Anthropic and Cohere: Foundation Model Labs on P5 at Launch

When AWS launched P5 instances in 2023, it did so with public commitments from foundation model developers who had already been running earlier P4 generations at scale, as documented in the customer testimonials cited in the P5 pricing section above. Anthropic and Cohere both credited P5's H100 GPUs with meaningful price-performance gains over the prior P4d generation at the scale required for training next-generation large language models. Both testimonials predate the July 2026 pricing changes but establish the baseline value proposition, faster time-to-train at a lower cost per unit of compute, that subsequent price increases have partially eroded for reservation-dependent customers.

Canva: Scaling Multi-Modal Training with Capacity Block Reservations

Design platform Canva, which AWS says empowers "over 150M monthly active users," used P4de instances to train the multi-modal models behind its generative AI creative tools before moving to reserve P5 capacity through EC2 Capacity Blocks. Canva's Head of Data Platforms said the launch of Capacity Block support for P5 let the company "get predictable access to up to 512 NVIDIA H100 GPUs in low-latency EC2 UltraClusters to train even larger models than before" (Source: aws.amazon.com). The 512-GPU figure matches the maximum Capacity Block cluster size of 64 eight-GPU instances that AWS documents on its Capacity Blocks landing page, indicating Canva reserves at or near the platform's ceiling for its largest training runs.

Dashtoon: Quantified 3x Performance Gain Moving from P4d to P5

AI-driven digital comics platform Dashtoon, which its co-founder says serves more than 80,000 monthly active users and generates over 100,000 images per day through its Dashtoon Studio product, provides one of the more specific quantified performance claims in AWS's published customer material. Dashtoon's Chief Technology Officer stated that the company uses "Amazon EC2 P5 instances to train and fine-tune multi-modal models including Stable Diffusion XL, GroundingDINO, and Segment Anything," and that Dashtoon has "seen performance improve by 3x while using P5 instances, powered by NVIDIA H100 GPUs, compared to using equivalent P4d instances, powered by NVIDIA A100 GPUs" (Source: aws.amazon.com). The same executive credited Capacity Blocks specifically for enabling predictable, low lead times, as soon as next-day, when scaling GPU capacity up and down around release cycles.

AON: Actuarial Simulation Beyond Frontier-Model Training

Not every named P5 customer is training frontier AI models. Insurance and risk-advisory firm AON uses a single-GPU p5.4xlarge instance to accelerate actuarial and economic-forecasting workloads. AON's Global Head of Life Solutions said the P5 family "has been a game-changer for us," explaining that the company can "now run machine learning models and economic forecasts that used to take days in just a matter of hours" using "a single H100 GPU instance (p5.4xlarge)" (Source: aws.amazon.com). This example illustrates that P5's per-accelerator Capacity Block pricing of $5.191 per hour, priced identically whether a customer reserves one GPU (p5.4xlarge) or eight (p5.48xlarge), makes single-GPU HPC and financial-modeling workloads economically viable at a much smaller footprint than the eight-GPU clusters foundation-model labs typically reserve.

Arcee and OctoAI: Flexible Reservation Without Long-Term Commitment

Two additional AWS customers illustrate the reservation-flexibility argument for Capacity Blocks over Savings Plans. Arcee, which builds what its CEO describes as "small, specialized, secure, and scalable language models" (SLMs), said Capacity Blocks are "an important part of our ML compute landscape for training SLMs on AWS because they provide us with reliable access to GPU capacity when we need it," adding that "knowing we can get a cluster of GPUs within a couple days and without a long-term commitment has been game changing for us" (Source: aws.amazon.com). OctoAI's chief executive made a similar point about matching reserved capacity to customer-driven demand spikes, saying Capacity Blocks let the company "predictably spin up different sizes of GPU clusters that match our customers' planned scale-ups, while offering potential cost savings as compared to long-term capacity commits or deploying on-prem" (Source: aws.amazon.com). Both testimonials underscore that for many mid-size AI companies, the value of AWS GPU pricing cannot be assessed on hourly rate alone; the ability to reserve short-duration capacity without a one- or three-year Savings Plan commitment is itself a priced feature.

Implications and Future Directions

The most direct implication of AWS's 2026 pricing pattern is that GPU capacity planning now requires the same discipline organizations apply to commodity or energy procurement: locking in rates ahead of quarterly repricing windows, rather than assuming stable list prices, given that AWS has already signaled a further update to Capacity Block rates for October 2026. For organizations with negotiated Enterprise Discount Programs, this dynamic creates a complication that Reddit commenters tracking the January 2026 increase quantified from two angles. One user worked the arithmetic directly: "If it was $100, your discount is $10. You pay $90. Then the price goes up to $115. Your discount is now $11.50. You are paying $103.50. An increase in public pricing of 15% (at $100) results in an increase of $13.50" (Source: reddit.com). Another commenter distilled the same point more concisely: "a discount plan is like a percentage off. An increase is still an increase" (Source: reddit.com), meaning negotiated percentage discounts do not insulate large customers from list-price increases in absolute dollar terms.

The competitive implication is a live question rather than a settled one. AWS's pricing increases coincide with Azure and Google Cloud actively marketing themselves as alternatives, and coverage of the July 2026 increase noted that "rising reservation costs could prompt some AWS customers to evaluate alternatives, including Nvidia-powered offerings on rival clouds or Google Cloud's TPU-based instances" (Source: finance.yahoo.com). The Register made a similar point about the January increase's optics, noting it "hands Azure and GCP a talking point on a silver platter" for enterprise sales teams courting cost-sensitive AI workloads (Source: theregister.com). Whether that dynamic actually shifts workloads is constrained by the same capacity scarcity that makes AWS's own Capacity Blocks necessary in the first place: Azure and Google Cloud face comparable NVIDIA supply constraints, and migrating a multi-week distributed training job between clouds carries its own engineering and data-egress cost that can exceed the pricing delta being chased.

A structural implication concerns whether AWS's pricing pattern becomes a template other hyperscalers replicate. The Register's analysis of the January 2026 increase framed the event as a precedent-setting break from two decades of AWS pricing behavior: "AWS has long benefited from the assumption that cloud pricing only trends in one direction. That assumption died on a Saturday in January, with all the fanfare of a Terms of Service update" (Source: theregister.com). The same analysis speculated about contagion into adjacent constrained resources: "Keep an eye on services where AWS faces genuine supply constraints or where their costs have materially increased," citing Graviton chip supply and data transfer costs as plausible next candidates for repricing (Source: theregister.com). If that pattern extends to Azure and GCP, the effective floor for hyperscale GPU pricing across all three major clouds could rise in tandem over the second half of 2026, narrowing the gap to specialized neocloud providers rather than widening it.

Finally, the efficiency gains embedded in each new instance generation, the 18 percent faster collective communications on EFAv4-equipped P6 instances and the up to 125 percent GPU TFLOPs improvement of P6-B200 over P5en discussed earlier in this report, mean that organizations evaluating "is AWS more expensive" should model cost per completed training run or cost per million tokens served, not cost per GPU-hour in isolation. A newer, nominally pricier instance that finishes the same job in materially less wall-clock time can still be the cheaper option end to end, a calculation that becomes more consequential the more frequently AWS reprices its underlying hourly rates.

Frequently Asked Questions (FAQs)

How much does AWS charge for GPU instances in 2026? Rates vary by instance and purchasing model. On-demand, the p6-b300.48xlarge (eight NVIDIA Blackwell Ultra GPUs) lists at $142.42 per hour, the p6-b200.48xlarge (eight NVIDIA Blackwell GPUs) at $113.93 per hour, and the p5.48xlarge (eight NVIDIA H100 GPUs) at $55.04 per hour (Source: economize.cloud) (Source: instances.vantage.sh) (Source: instances.vantage.sh). Reserved through EC2 Capacity Blocks for ML, the per-accelerator rates are lower: $14.04 for P6-B300, $12.355 for P6-B200, and $5.191 for P5 (Source: finance.yahoo.com).

What is the difference between AWS P5 and P6 instance pricing? P6 instances run on newer NVIDIA Blackwell and Blackwell Ultra GPUs and cost roughly double to nearly triple the P5 family's per-accelerator rate: $12.355 (P6-B200) to $14.04 (P6-B300) per accelerator-hour under Capacity Blocks versus $5.191 for standard P5 (Source: finance.yahoo.com). In exchange, AWS reports P6-B200 delivers "up to 125 percent improvement in GPU TFLOPs" and 60 percent higher memory bandwidth versus the P5en generation (Source: aws.amazon.com).

How does P6-B300 pricing compare to P6-B200? P6-B300's Capacity Block rate of $14.04 per accelerator-hour runs about 14 percent above P6-B200's $12.355 (Source: economize.cloud), while offering "2 times more networking bandwidth, and 1.5 times more GPU memory" (Source: aws.amazon.com). P6-B300 is currently available only in US West (Oregon), US East (N. Virginia), and AWS GovCloud (US-East), a narrower footprint than P6-B200.

Why did AWS raise GPU instance prices? AWS attributes both the January 2026 (15 percent) and July 2026 (20 percent) increases to "supply and demand dynamics" in its official documentation (Source: finance.yahoo.com). Analysts point to NVIDIA GPU and HBM3e memory supply constraints, sustained AI training demand, and AWS's own $200 billion 2026 infrastructure buildout as underlying drivers (Source: servergurus.com) (Source: theregister.com).

What does a P5e instance cost? The p5e.48xlarge (eight NVIDIA H200 GPUs) is priced at $47.76 per instance-hour, or $5.97 per accelerator-hour, under EC2 Capacity Blocks, applied uniformly across all available regions (Source: finance.yahoo.com). AWS does not publish a standard on-demand list rate for p5e.48xlarge on its general pricing pages (Source: spheron.network).

How does AWS GPU instance pricing compare to other clouds? For eight-GPU H100 nodes, AWS's $55.04 per hour ($6.88 per GPU) is cheaper than Azure's roughly $98.32 per hour ($12.29 per GPU) (Source: thundercompute.com) and Google Cloud's $88.49 per hour ($11.06 per GPU) (Source: cloud.google.com), but more expensive than Oracle Cloud's $80.00 flat rate ($10.00 per GPU) (Source: oracle.com) and CoreWeave's $49.24 per hour ($6.16 per GPU) (Source: coreweave.com).

Is there a difference between P5 and P6 on-demand availability? Yes. Eight-GPU on-demand capacity for both generations is scarce enough that AWS built EC2 Capacity Blocks for ML specifically to guarantee access; one AWS customer described the on-demand market for these instance types as a situation where you can "extremely rare[ly]... ever just spin one of these up on-demand" (Source: reddit.com).

Is AWS GPU pricing negotiable for large customers? Enterprise Discount Programs apply percentage discounts off AWS's public list price, but because those discounts are proportional, a Reddit thread analyzing the July 2026 increase noted that when public pricing goes up, the discounted rate becomes more expensive in absolute terms even though the percentage discount holds steady (Source: reddit.com), meaning negotiated agreements reduce but do not eliminate exposure to list-price increases.

Do AWS GPU instances get cheaper with Reserved Instances the way general-purpose EC2 does? Not in the traditional sense. P5 and P6 pricing flexibility comes primarily from Savings Plans and Capacity Blocks rather than classic Reserved Instances, and third-party trackers list three-year reserved-equivalent pricing for p6-b300.48xlarge alongside on-demand and Spot rates on the same rate card (Source: instances.vantage.sh), underscoring that commitment-based discounting, not classic reservations, is the primary lever for P-family cost reduction.

Conclusion

AWS's GPU instance pricing in July 2026 reflects a market defined by scarcity rather than the steady downward drift customers came to expect over the prior two decades of cloud computing. The current lineup spans from the p5.48xlarge's $55.04 hourly on-demand rate for eight NVIDIA H100 GPUs up to the p6-b300.48xlarge's $142.42 hourly rate for eight NVIDIA Blackwell Ultra GPUs, with EC2 Capacity Blocks offering lower, but now twice-raised, per-accelerator rates that most large training customers rely on because eight-GPU on-demand capacity remains difficult to obtain. Two rate increases within a single twelve-month window, 15 percent in January and 20 percent in July 2026, compound to roughly a 38 percent rise from year-start levels, a reversal from the up to 45 percent price cut AWS had announced just seven months before the first increase.

The pattern is best understood as AWS repricing a genuinely scarce resource rather than an arbitrary margin grab: NVIDIA GPU and HBM3e memory supply constraints, sustained frontier-model training demand, and AWS's own unprecedented infrastructure capital program are all documented, independently sourced contributors to the pricing trend. Competitively, AWS remains positioned between the premium-priced Azure and Google Cloud on one side and lower-cost Oracle Cloud and specialized neocloud providers such as CoreWeave and Lambda on the other, a middle position that reflects AWS's scale, managed-service ecosystem, and security architecture as much as it reflects raw hardware cost. Named customers from Anthropic and Cohere at the frontier-model end to AON's single-GPU actuarial workloads at the applied end illustrate that the P5 and P6 families serve a genuinely broad range of budgets and use cases, not solely the largest AI labs.

For organizations budgeting AI infrastructure spend for the remainder of 2026 and into 2027, the most actionable takeaway from this analysis is that AWS has explicitly signaled its next Capacity Block repricing for October 2026, meaning current rates should be treated as temporary rather than stable. Buyers with flexibility in region, instance generation, or purchasing mechanism, favoring P5 over P6 where H100-class performance suffices, favoring international regions where Capacity Block rates run lower, or evaluating Oracle Cloud and neocloud alternatives for workloads that do not require AWS-specific managed services, retain meaningfully more control over total cost than those who default to on-demand P6-B300 pricing without comparison. Whichever mechanism a given workload ultimately uses, the underlying lesson of the past thirteen months is that AWS GPU pricing now moves in both directions, and infrastructure budgets built on the older assumption of only-falling cloud prices will need to be revisited at least quarterly.

Tags: aws gpu instance pricing, aws p6 instance pricing, aws p5 instance pricing, p6-b300, p6-b200, ec2 capacity blocks, nvidia blackwell aws, aws ec2 gpu price increase

---