STRATEGIC INTELLIGENCE DIVISION

Six Nvidia rates rose 38.0% and Trainium has not moved since 22 April

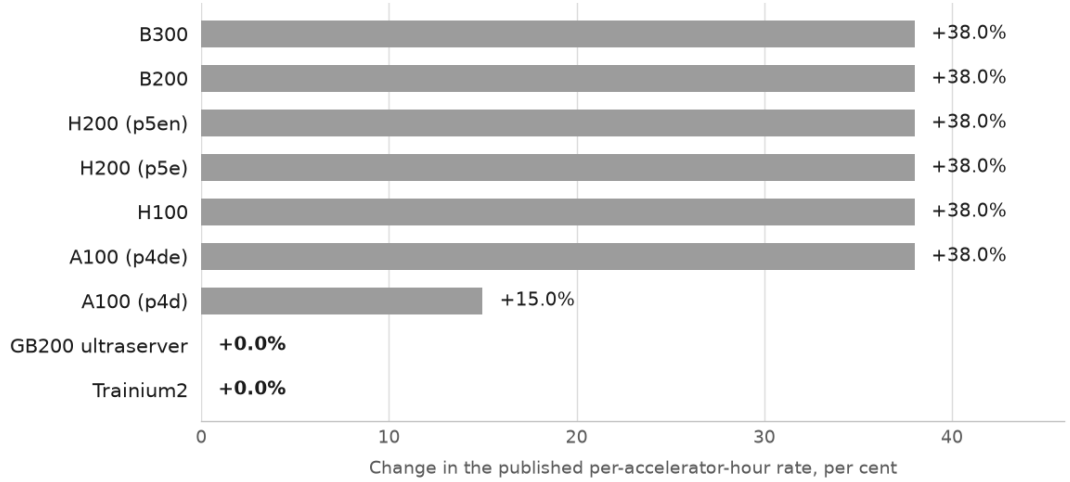
Every Nvidia instance family on AWS's capacity-block price list costs more than it did in April 2026: six by 38.0%, through a 20.0% step on 7 July and 15.0% on 7 October. Trainium2 has held at $2.235 since at least 22 April, and the page names January 2027 for the next update.

Annika Sorensen, Strategic Intelligence Division · 3 November 2026 · 9 minute read

In brief

- The two steps AWS published, 20.0% on 7 July 2026 and 15.0% on 7 October 2026, compound to 38.0% on the B300, B200, H200 and H100 instances, so a 2027 GPU line built on the April rates is 38.0% short before any volume growth.
- Outside the United States the step is 51.8%, because October replaced the regional discount with one world rate: an H100 accelerator-hour in Tokyo, London or Stockholm went from $3.933 to $5.970, so the estate that was cheapest in April took the largest increase.
- Price certainty is a purchase date rather than a contract term. A block is charged at the rate in force when it is booked, must start within 8 weeks, and cannot be cancelled, so a booking made on 3 November 2026 bought a start no later than 29 December 2026, at the price then printed.
- The cheaper rate on the same page is the silicon AWS designs: Trainium2 has held at $2.235 since at least 22 April 2026, and the 16-chip block is listed in two regions, US East (Ohio) and Asia Pacific (Hyderabad), with none in Europe.
- Both announcements state that on-demand and Savings Plan prices were unchanged, so this is a repricing of scheduled capacity: the 20.0% step on 7 July 2026 fell inside the quarter ended 30 September 2026, which AWS reported at a 38.7% segment operating margin against 34.6% a year earlier, and the 15.0% step on 7 October 2026 has not been reported at all.

Change in the per-accelerator-hour rate, 22 April 2026 to 3 November 2026



Units: per cent change. Source: Amazon EC2 Capacity Blocks for ML Pricing, Internet Archive capture of 22 April 2026 against the page as read on 3 November 2026. US regions; Oregon for B300, Ohio for B200, H200 and H100, N. Virginia for p4de, US East (Dallas) Local Zone for the GB200 ultraserver. Mosaic Singularity analysis of the two dated readings, November 2026.

The bill for scheduled GPU capacity on AWS is now set by a rate card that moves on a date the vendor announces in advance, and the last move, on 7 October 2026, put 15.0% on every Nvidia instance type it lists. A B200 block in US East (Ohio) cost $10.296 an accelerator-hour on 22 April 2026 and costs $14.208 on 3 November 2026. The same page prices Trainium2, AWS's own accelerator, at $2.235, unchanged since at least 22 April, and states that the next update is January 2027. The decision that follows sits with the CIO

or chief AI officer who signs the 2027 capacity commitment and the CFO who approves it: book blocks at the price in force before the reset, move a defined share of the estate to Trainium2, or budget a GPU line that is 38.0% higher than the one in the current plan.

Seven Nvidia families rose and two prices on the page did not

AWS announced both steps itself, in the banner at the top of its own price page. The page as it stood on 4 July 2026 carried an announcement that effective 7 July 2026 the per-accelerator rate would be $14.04 for P6-B300 and $12.355 for P6-B200. The page as it stood on 1 October 2026 carried the next one: effective 7 October 2026, $16.146 and $14.208, with P5en at $7.895, P5e at $6.866 and P5 at $5.970. The table on 22 April 2026 priced the same B300 accelerator-hour at $11.70. 11.70 became 14.04 became 16.146, which is 20.0% then 15.0%, and 1.20 times 1.15 is 1.38. Six of the seven Nvidia instance families on the page compounded to exactly 38.0%; the seventh, the A100 in p4d, was not in the July announcement and rose 15.0% in October alone.

One page, four dated rates, per accelerator-hour

Instance or UltraServer (accelerators)	22 April 2026	From 7 July 2026	From 7 October 2026	Change
p6-b300.48xlarge (8 x B300), US West (Oregon)	$11.70	$14.04	$16.146	+38.0%
p6-b200.48xlarge (8 x B200), US East (Ohio)	$10.296	$12.355	$14.208	+38.0%
p5en.48xlarge (8 x H200), US East (Ohio)	$5.721	$6.865	$7.895	+38.0%
p5e.48xlarge (8 x H200), US East (Ohio)	$4.975	$5.97	$6.866	+38.0%
p5.48xlarge (8 x H100), US East (Ohio)	$4.326	$5.191	$5.970	+38.0%
p4de.24xlarge (8 x A100), US East (N. Virginia)	$1.845	$2.214	$2.546	+38.0%
p4d.24xlarge (8 x A100), US East (Ohio)	$1.475	$1.475	$1.696	+15.0%
trn2.48xlarge (16 x Trainium2), US East (Ohio)	$2.235	$2.235	$2.235	0.0%
u-p6e-gb200x72 ultraserver (72 x B200), US East (Dallas)	$10.582	$10.582	$10.582	0.0%

Source: Amazon EC2 Capacity Blocks for ML Pricing: table as captured 22 April 2026 and 1 October 2026; rates for 7 July 2026 and 7 October 2026 as stated in the announcements on those dates; all figures confirmed on the page as read on 3 November 2026. Mosaic Singularity analysis of the change, November 2026.

Neither announcement lists a Trainium instance, and each closes with the same sentence: "All other EC2 Capacity Block prices, as well as prices on fixed pricing models like On-Demand and Savings Plans, remain unchanged." Trainium2 was $2.235 on the 22 April reading, on 4 July, on 1 October and on the page as read on 3 November 2026. The GB200 ultraserver, the rack-scale Nvidia product sold as 72 accelerators in a Dallas local zone, has been $10.582 an accelerator-hour across the seven readings between 22 April 2026 and 3 November 2026. The increases are attached to the per-instance shapes and to nothing else on the page.

The horizon is eight weeks and the reset is in January

The page states that prices "are scheduled to be updated next in January, 2027". A block is charged "at the prevailing rate at the time of purchase", it can be reserved with "a start time up to 8 weeks in the future", and "Capacity Block cancellations aren't allowed". Those three terms are the whole hedge. On 3 November 2026 a buyer could push a start date no further than 29 December 2026, and paid the November rate for it. Capacity that runs in 2027 is bought at whatever the January update prints, and the page dates that update only to a month. Source: Amazon EC2 Capacity Blocks for ML Pricing and the Amazon EC2 Capacity Blocks user guide, both read on 3 November 2026.

Outside the United States the increase is 51.8%

In April the price list carried a regional discount on Nvidia capacity: an H100 accelerator-hour was $3.933 in Jakarta, Mumbai, Tokyo, Sydney, London, Stockholm and Sao Paulo against $4.326 in the United States, and an H200 on p5en was $5.201 against $5.721. The July announcement kept the split, at $4.72 and $6.241 for those regions. The October announcement removed it and set one rate, $5.970 and $7.895, for every region on the page. An estate in Stockholm therefore absorbed 51.8% between 22 April and 3 November 2026 on the same silicon that cost its Ohio counterpart 38.0%, and the discount that made a European region cheaper is gone rather than reduced.

The only price protection is the day you buy

A capacity block is not a hedge a buyer can hold for a year. It must start within 8 weeks of booking, it is charged at the rate in force when the booking is made rather than the rate when it runs, and it cannot be cancelled, so a wrong forecast is paid for in full. AWS caps a block at 64 instances and the whole organisation at 256 instances on any one date. What the block does give is certainty at a known price for its term: a training run scheduled for the first week of February 2027 and booked in December 2026 is priced on the December card and holds that price through the reset. A run scheduled for March cannot be booked before January, the month the page names for the update.



Illustrative photograph, not evidence, and not a depiction of any company's facility or of the pricing described in this article.

The alternative is sold in Ohio or Hyderabad

The cheaper rate is real and its footprint is not. On 3 November 2026 the price list offers the 16-chip Trainium2 block, trn2.48xlarge, in two regions: US East (Ohio) and Asia Pacific (Hyderabad). The single-chip block, trn2.3xlarge, is listed in Hyderabad, Melbourne and Sao Paulo. No European region appears for either, while the B300 block is offered in seven commercial regions and GovCloud. An estate in Frankfurt or Dublin that wants the $2.235 rate is moving work to Ohio or to Hyderabad, because a capacity block is bought in a region.

The gap has widened with the rate. In April 2026 a p5en accelerator-hour cost 2.56 times a Trainium2 accelerator-hour; on 3 November 2026 it cost 3.53 times. For B200 against Trainium2 the multiple went

from 4.61 to 6.36. Those are ratios of published rates per accelerator-hour, not of value: the page prices accelerators, and a B200 and a Trainium2 are not the same machine. What a buyer can take from them is direction and size, and the direction is one way. AWS spent part of the same quarter removing the obvious objection: "TorchNeuron lets developers run existing PyTorch code (which most already use) on Trainium without rewriting a thing."

The counterparty's margin rose with the rate

AWS segment operating income, Q3 2026	**$17.8bn**
AWS net sales growth, year over year	**39.3%**
AWS segment operating margin, Q3 2026	**38.7%**
Purchases of property and equipment, twelve months to 30 September 2026	**$210.7bn**

Source: Amazon.com Form 10-Q for the quarter ended 30 September 2026, segment table and free cash flow table; the margin and the growth rate are Mosaic Singularity analysis of those figures, November 2026.

AWS reported $17.8 billion of segment operating income on $45,984 million of net sales in the quarter to 30 September 2026, a 38.7% segment margin against 34.6% a year earlier, and that quarter closed a week before the 7 October reset. The same filing shows the size of the machine the rate card is paying for: purchases of property and equipment, net of proceeds, of $210.7 billion in the twelve months to 30 September 2026 against $115.9 billion a year earlier, and free cash flow of negative $42.4 billion against positive $14.8 billion. The 38.0% therefore arrives on top of a margin that has already widened 4.0 points, and the reason the page gives for the price is "trends in supply and demand".

Four decisions before 31 December 2026

- CFO: re-base the 2027 accelerated-compute line on the 7 October 2026 rates and show the board the delta separately from volume growth. For a US estate on B200 that delta is 38.0%; for an estate in Stockholm, London or Tokyo on H100 or H200 it is 51.8%. By 31 December 2026, before the January 2027 update and before the 2027 plan is signed.
- CIO or chief AI officer: take one named training or inference workload and price it on Trainium2 in writing, in Ohio or Hyderabad, including the engineering cost of the move. TorchNeuron removes the rewriting objection; it does not remove latency, data residency, or the fact that the block is sold in two regions. By 31 December 2026.
- General counsel: read the rate clause. If the contract sets price by reference to the published rate card, the only protection is the purchase date, and the question for the supplier is whether it will fix a rate for a defined volume across a reset. Get the answer in writing, and date it before the January 2027 update.
- Procurement: book February 2027 capacity in December 2026. A block bought then starts within 8 weeks at the December price and carries that price through the reset, which is the one hedge on the page that costs nothing but a scheduling commitment. Do it before 31 December 2026.

What the board should ask

- What share of the 2027 AI compute commitment is priced off a published rate card rather than a negotiated fixed term, and who signed it?

- If January 2027 repeats the July and October steps, which line in the 2027 plan absorbs the increase, and what is cut to fund it?
- Which workload has been priced on Trainium2 in writing, in which region, and what does the move cost in engineering months?
- What does our contract say about a mid-term rate change, and has anyone read that clause since 7 October 2026?

What to watch, and when

- January 2027: the update the page names. If Trainium2 moves off $2.235, or the GB200 ultraserver moves off $10.582, this reading fails and the increases are a cost pass-through rather than a price structure. If the Nvidia instance rates do not move, the cadence reading fails and the two observed steps were one-off adjustments.
- Early February 2027, reporting the quarter ended 31 December 2026: whether the AWS segment margin holds at or above 38.7%. A fall would be the first evidence that the October rate met demand that could wait.
- By 31 March 2027: whether a Trainium2 capacity block appears in any European region. Until one does, the alternative on this page is a domestic option and a European estate's only decision is when to buy.