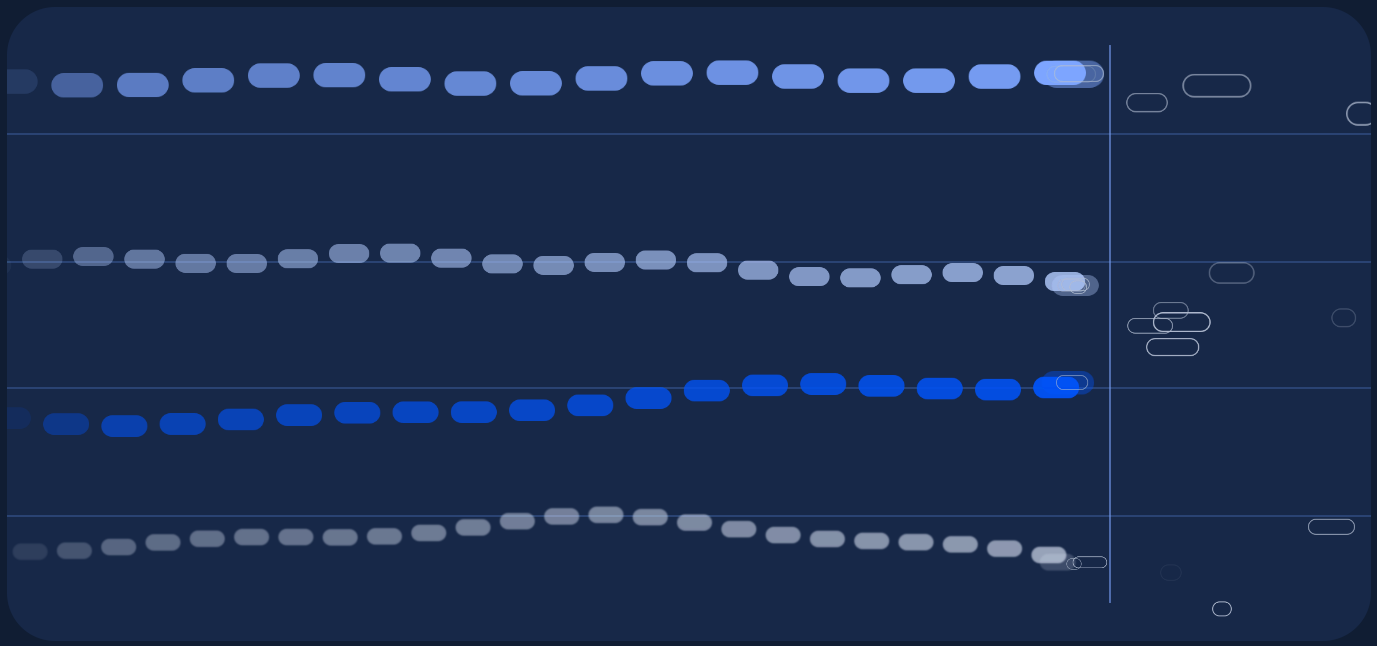
Rillor

METHODS NOTE

How to compare GPU generations on price per FLOP

GPU price performance over time depends on the precision counted, peak or delivered throughput and the price used. This note compares data-center GPU generations on one stated basis.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Price per FLOP compares GPU generations only on a stated basis: the FLOP counted, peak or delivered throughput, and the price and its date.

- Published trends in GPU price performance over time differ by sample, FLOP and price basis. Epoch AI found FLOP/s per dollar doubling every 2.46 years for 470 GPUs released from 2006 to 2021 at FP32, and a 49% yearly gain since 2023 for chips bought each quarter, in constant 2025 dollars [1, 4].
- NVIDIA's H100 page prints its TF32, BF16, FP16, FP8 and INT8 tensor rates with sparsity, which assumes a pruned network, and the dense rate is half. One A100 sparsity test measured a gain of up to 20%, against the 2 times the printed rates imply [9, 14].
- The precision and sparsity quoted can change a generation step fourfold. At dense FP16, H100's datasheet rate is 3.17 times A100's. H100's sparse FP8 figure is 12.7 times A100's dense FP16 rate (this note's arithmetic) [7, 9].
- Delivered throughput, measured on a benchmark, differs from peak. In MLPerf® Inference v5.0, eight B200s produced 1.60 times the text-to-image throughput of eight H200s at FP8, where the dense FP8 peak ratio is 2.27 (this note's arithmetic) [10, 11, 22].
- The price basis can reverse a result. Per dense FP16 FLOP, H100 costs 21% less than A100 at AWS on-demand rates and 14% more at Azure on-demand rates in East US (this note's arithmetic) [7, 9, 26, 27, 28, 29, 30, 31].
- At the same precision, B200 costs 9% less per dense FLOP than H100 at AWS on-demand rates and 26% less at Lambda's rates from 16 November 2026 (this note's arithmetic) [9, 11, 26, 27, 32]. Counting B200 at FP4 against H100 at FP8 shows 54% less at AWS (this note's arithmetic) [9, 11, 26, 27].
- The U.S. Bureau of Labor Statistics separates price change from quality change. Matched-model indexes follow the same item over time, and hedonic methods subtract the estimated value of a quality change. One study's preferred index of microprocessor prices fell 43% a year from 2008 to 2013, while the matched-model producer price index fell 8% [33, 36].
- A matched-model series follows the same item over time, so its changes are price changes [33, 35, 36]. For GPU rents that item is one class and memory variant at one provider and pricing type, and comparing generations then needs a stated measure of quality, such as dense FLOP/s at one precision or measured throughput.

Background

Price per FLOP divides a price by a throughput figure. Its inverse, FLOP/s per dollar, is the measure trend studies usually report [1, 3, 4]. Analysts, capacity planners and researchers use it to ask whether a newer data-center GPU class costs less per unit of compute, and how fast that changes.

Three choices sit behind any such figure. The FLOP: its precision, dense or sparse, and a spec-sheet peak or a measured rate. The price: a launch price, an estimate of what buyers paid or an hourly rent, from one provider under one pricing type. The time basis: nominal or deflated dollars, and the date of each price. Epoch AI's 2022 report counts FP32 FLOP/s [1]. Its 2026 data insight counts a precision-normalized peak against card or module purchase prices, weighting each chip by its share of quarterly spending [4]. The two measure different things.

Data-center GPU generations often add a lower precision. A100's datasheet has no FP8 row [7], H100 adds FP8 [9] and B200 adds FP4 [11]. Counting each generation at the lower precision it adds multiplies the apparent gain, and NVIDIA's product pages quote tensor rates with sparsity [9, 13].

This note sets out each choice, shows how far it moves the result, works one comparison of three generations from posted prices and explains quality adjustment as the Bureau of Labor Statistics uses it. It makes no forecast and gives no investment advice.

Method

What was examined. Published trend estimates of GPU price-performance and the bases behind them. Tensor rates for A100, H100, H200, B200 and B300 by precision, dense and sparse. Measured throughput from one MLPerf® Inference round and training efficiency from two papers. Purchase-price evidence, posted rents from three providers, and quality-adjustment methods used by the U.S. Bureau of Labor Statistics.

Sources. Thirty-seven sources: Epoch AI and AI Impacts publications [1-6], NVIDIA documents [7-15], an Open Compute Project specification [16], a Federal Register rule [17], two arXiv papers [18, 19], MLCommons pages and results [20-23], a CNBC report [24], AWS, Azure and Lambda price pages, price files and instance documentation [25-32], Bureau of Labor Statistics (BLS) publications [33-35] and two NBER working papers [36, 37], all viewed on 8 November 2026 and [26] again on 11 November 2026.

How. Each FLOP figure carries its precision and whether it assumes sparsity. Dense rates are half the sparse rates where the source says so [11, 12, 13, 14]. Each price carries its source, date and kind: launch price, estimate, list rate or on-demand rent. Instance prices are divided by the GPU count in the provider's documentation [27, 30, 31]. Cost per FLOP is price per GPU-hour divided by dense PFLOP/s, in U.S. dollars per PFLOP/s-hour. "This note's arithmetic" marks calculations on the sources' own numbers. MLPerf® results follow MLCommons' messaging guidelines: round, division, category, scenario and result IDs are stated, ratios and costs are labeled as derived, and MLPerf® results are compared only with each other [23]. The worked comparison is an illustration from posted prices, with no Rillor Compute Index values, Rillor prices or transaction data.

Where sources disagree. Epoch's 2022 report prints the FP16 doubling time as 2.30 years in its table and 2.32 years in its text [1]. NVIDIA gives DGX B200 64 TB/s of memory bandwidth and HGX B200 62 TB/s [11, 13]. Epoch deflates with a producer price index in 2024 and the consumer price index in 2026 [3, 4].

Findings

1. Published trend estimates and the samples behind them

Estimate	Sample and period	FLOP basis	Price basis	Result
Epoch AI 2022, all GPUs	470 GPUs, 2006 to 2021	FP32	Prices in two merged datasets	Doubling every 2.46 years (95% CI 2.24 to 2.72) [1]
Epoch AI 2022, GPUs used in ML	26 GPUs	FP32	Same	Doubling every 2.07 years (1.54 to 3.13) [1]
Epoch AI 2022, top GPUs	57 GPUs	FP32	Same	Doubling every 2.95 years (2.54 to 3.52) [1]
Epoch AI 2023	47 ML accelerators, 2010 to 2023, and 1,948 other GPUs, 2006 to 2021	FP32	Release price, inflation-adjusted, two-year amortization	Doubling every 2.1 years (ML GPUs) and 2.5 years (general) [2]
Epoch AI 2024	P100, V100, A100 and H100 in its table	Any one precision, held fixed	DGX node price divided by eight, 2022 dollars, in its table	30% cheaper each year [3]
Epoch AI 2026	24 chips, Q1 2023 to Q4 2025	Peak, precision-normalized (TPP)	Card or module purchase price, weighted by share of quarterly spending, constant 2025 dollars	49% a year, doubling every 1.7 years [4]
AI Impacts 2020	352 GPUs with prices from Amazon, Newegg and eBay, 2011 to 2020	FP32 and half-precision FMA	Reseller prices and release prices, CPI-adjusted	13% a year (FP32, reseller prices), 46% a year (half-precision FMA, release prices, four years) [6]

In one unit (this note's arithmetic), a doubling every 2.46 years is 32.6% more FLOP/s per dollar a year, every 2.07 years 39.8% and every 2.95 years 26.5% [1]. "30% cheaper each year" is 1.43 times the FLOP/s per dollar each year, a doubling every 1.94 years [3]. A 49% yearly gain doubles every 1.74 years [4].

Sample composition moves the rate. In Epoch's 2022 data, GPUs used in ML research improved faster than the top GPUs. Epoch suggests "relevant labs spending more resources on procuring top GPUs over time" as one reason [1], a change in the sample. It also notes that "the GPUs for which we have price data tend to be

the ones with higher FLOP/s values" [1]. Its FP16 subsample of 91 GPUs gave 2.30 years [1].

The FLOP basis moves the rate. Epoch's 2023 report found that switching from FP32 to tensor-FP16 gave "a further 10x performance increase", and that the FP32 calculation "introduces bias against specialized ML chips" [2]. AI Impacts found price per FLOPS falling tenfold in about 17 years at single precision and about 5 years for half-precision fused multiply-add [6].

A spending-weighted series measures the mix bought. In Epoch's 2026 insight, "widely sold chips contribute more to each quarter's figure" [4]. The series was nearly flat at 6% a year from 2023 to mid-2024, then "roughly doubled as Blackwell-generation chips grew to a majority of new spending" [4]. In nominal dollars the rate is 45% a year, and the constant-dollar 90% interval runs from 36% to 66% [4]. It keeps one price per chip except H100 and H200 ($30,000 in 2023, $28,000 in 2024, $26,000 in 2025), so it does "not capture any later price cuts or spot prices" [4].

Endpoints and demand set the headline. Epoch puts GB300 at about 5.5 times the P100's 2016 launch price in 2025 dollars, for roughly 200 times its performance [4]. AI Impacts links a 2017 rise in GPU price per FLOPS, with no change in hardware, to demand from cryptocurrency miners [6].

Each estimate answers its own question, and this note does not average them. Epoch's 2022 report "does not attempt to project future GPU price-performance" [1].

2. Which FLOPs: precision, sparsity and peak figures

Dense tensor rates per GPU, from NVIDIA's datasheets and product pages:

GPU	FP16 or BF16, dense	FP8, dense	FP4, dense	As printed	Sources
A100 SXM	312 TFLOPS	No FP8 row	No FP4 row	Dense, with sparse rates starred (624 TFLOPS at FP16)	[7]
H100 SXM	989.5 TFLOPS	1,979 TFLOPS	No FP4 row	Sparse only (1,979 and 3,958 TFLOPS)	[9]
H200 SXM	989.5 TFLOPS	1,979 TFLOPS	No FP4 row	Sparse only, same rates as H100 SXM	[10]
B200 (HGX)	2,250 TFLOPS	4,500 TFLOPS	9,000 TFLOPS	Sparse, "Dense is one-half of the sparse spec shown"	[11]
B300 (HGX)	2,250 TFLOPS	4,500 TFLOPS	14,000 TFLOPS	FP4 printed sparse and dense (18 and 14 PFLOPS), others sparse	[12]

Halved rates are this note's arithmetic: the Blackwell datasheets state that dense is half the sparse figure [11, 12], and the H100 and H200 rates, marked only "with sparsity", are halved per NVIDIA's sparsity guide [9, 10, 14].

Sparsity doubles a rate under one condition. Sparse Tensor Cores accelerate a 2:4 pattern "where two of every four values are zero", completing "the same effective calculation in half the time" [14]. The network has to be pruned first, and one A100 benchmark measured a gain "up to 20% at larger batch sizes" [14]. A sparse rate is a ceiling for pruned networks. NVIDIA's launch release gave DGX A100 5 petaflops for eight GPUs [8], which matches eight times the sparse FP16 rate, 4,992 TFLOPS (this note's arithmetic) [7].

Mixed bases inflate a generation step. At dense FP16, H100 is 3.17 times A100 and B200 is 2.27 times H100 (this note's arithmetic) [7, 9, 11]. H100's sparse FP8 figure against A100's dense FP16 gives 12.7 times, and B200 at dense FP4 against H100 at dense FP8 gives 4.55 times [7, 9, 11]. Each step holds on its own basis, so the basis has to match at both ends.

Halving is not a rule for every chip. The B300 datasheet prints dense FP4 as 14 PFLOPS beside 18 PFLOPS sparse, while its FP8 and FP16 rows are sparse with dense at half [12]. Its dense FP8 and FP16 rates equal B200's, and its FP64 rate falls from 37 TFLOPS on B200 to 1.2 TFLOPS [11, 12]. For FP64 work, price per FLOP moves the other way between those classes.

One GPU name can carry several rates. H100 NVL lists 3,341 TFLOPS at sparse FP8 against 3,958 for H100 SXM [9], and a B200 in GB200 NVL72 lists 10 PFLOPS against 9 in HGX B200 [11]. A class label needs its form factor.

A lower-precision FLOP is a different unit of work. NVIDIA reports its NVFP4 format at "1% or less accuracy degradation versus FP8" on one benchmark set for a reasoning model [15]. Its table marks FP4 and MXFP4 with a "Risk of noticeable accuracy drop compared to FP8" [15]. The microscaling specification treats E5M2 and E4M3 as "different formats" under one 8-bit name [16], so a label should name the format where the source gives it.

A precision-normalized metric removes the choice. The Bureau of Industry and Security defines total processing performance (TPP) as "2 x 'MacTOPS' x 'bit length of the operation'", using dense values [17]. Epoch's 2026 insight uses it [4]. On dense rates, H100 gives 15,832 at FP8 and FP16, and B200 36,000 at FP4, FP8 and FP16 (this note's arithmetic) [9, 11]. B300 gives 56,000 at FP4 and 36,000 at FP8 (this note's arithmetic) [12]. The metric holds steady across precisions for a chip whose rate doubles as bit width halves, and it counts B300's extra FP4 rate. This note restates the rule text only and gives no export-control advice.

Vendor multiples carry their own conditions. NVIDIA's DGX B200 page projects 3 times the training and 15 times the inference performance of DGX H100, at 50 ms token-to-token latency and 32,768 input tokens for inference and a 32,768-GPU scale for training [13].

3. Delivered throughput from MLPerf® results

MLPerf® Inference: Datacenter measures "how fast systems can process inputs and produce results using a trained model" [20]. Its Closed division requires the reference model, and its Available category holds components "available for purchase or for rent in the cloud" [20]. Round v5.0 (2 April 2025) held 17,457 performance results from 23 organizations [21].

Entries from one submitter for three NVIDIA systems in the Closed division, Available category, single nodes with eight accelerators, all on TensorRT 10.8 and CUDA 12.8 [22]:

System	Result ID	70-billion-parameter language model, Offline (tokens/s)	Same, Server (tokens/s)	Weights
DGX H100, 8x H100-SXM-80GB	5.0-0057	31,306.8	31,106.3	FP8
H200, 8x H200-SXM-141GB	5.0-0060	34,988.2	33,071.5	FP8
DGX B200, 8x B200-SXM-180GB	5.0-0056	98,858.0	98,443.3	FP4

Source: [22].

Equal peaks, different results. H200 has H100's tensor rates with 1.76 times the memory and 1.43 times the bandwidth [9, 10]. Eight H200s produced 1.12 times eight H100s' Offline tokens per second (this note's arithmetic) [22]. Price per peak FLOP rates the two as the same hardware.

Delivered against peak at one precision. On the text-to-image benchmark at FP8, eight B200s produced 30.3752 samples per second Offline and eight H200s 18.989 [22], a ratio of 1.60 against a dense FP8 peak ratio of 2.27. B200's delivered ratio was about 70% of its peak ratio on that test (this note's arithmetic) [10, 11, 22].

Delivered across precisions. DGX B200 at FP4 weights produced 3.16 times DGX H100's Offline tokens per second at FP8 (this note's arithmetic) [22], between the dense peak ratios of 2.27 at equal precision and 4.55 for B200 at FP4 against H100 at FP8 [9, 11]. In samples per second the ratio is 3.39, because the B200 run produced 271.7 tokens per sample against 291.8 [22].

Submitter and software move results. Four eight-H100 results from three other submitters ranged from 17.9889 to 18.375 samples per second on the text-to-image benchmark, a 2.1% spread on one class (this note's arithmetic) [22]. On the 70-billion-parameter benchmark, the median submitted score doubled from v4.0 [21]. A delivered figure needs its round.

Training efficiency counts dense peaks. A 2022 paper on a 540-billion-parameter model defines model FLOPs utilization (MFU) as observed throughput "relative to the theoretical maximum throughput of a system operating at peak FLOPs" [18]. That model's training reached 46.2% MFU on TPU v4 [18]. A 405-billion-parameter model's pre-training on H100s reported 38% to 43% BF16 MFU, with 430 TFLOPs per GPU at 43% on 8,192 GPUs [19]. 430 is 43.5% of H100's dense BF16 rate and 21.7% of the sparse 1,979 (this note's arithmetic) [9, 19], so the paper's denominator is the dense rate.

MLPerf® notes: MLPerf® Inference: Datacenter v5.0, Closed division, Available category, benchmarks llama2-70b-99.9 and stable-diffusion-xl, Offline and Server scenarios. Result IDs 5.0-0056, 5.0-0057, 5.0-0060, 5.0-0010, 5.0-0039, 5.0-0040 and 5.0-0065, retrieved 8 November 2026 from the MLCommons results repository [22]. Ratios, percentages and the costs in Finding 6 are this note's arithmetic and are not MLPerf® metrics. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

4. Which price: launch price, estimate, listing or rent per GPU-hour

Purchase evidence spreads by kind. NVIDIA announced in May 2020 that DGX A100 systems "start at $199,000" [8], or $24,875 per GPU with the server (this note's arithmetic). Epoch's chip-sales documentation uses $10,000 to $15,000 for an A100 in 2022 and 2023, and $27,000 to $35,000 for H100 and H200 in 2023 and $22,000 to $30,000 in 2025 [5]. It cites an NVIDIA disclosure of $100 billion from 4 million Hopper GPUs through October 2025, outside China, or $25,000 per GPU, and puts H100 server listings at about $35,000 per GPU [5]. CNBC reported NVIDIA's chief executive estimating in March 2024 that one Blackwell chip could cost $30,000 to $40,000 [24]. Epoch, noting he held up a B200 die as he said it, applies the range to standalone B200s [5].

Evidence, kind and date	Price per GPU	Per dense FP16 TFLOP/s	Sources
A100 in DGX A100, launch price with server, May 2020	$24,875	About $80	[7, 8]
A100, estimate, 2022 to 2023	$10,000 to $15,000	$32 to $48	[5, 7]
H100 or H200, estimate, 2023	$27,000 to $35,000	$27 to $35	[5, 9]
H100 or H200, estimate, 2025	$22,000 to $30,000	$22 to $30	[5, 9]
B200, chief executive's estimated range, March 2024	$30,000 to $40,000	$13 to $18	[5, 11, 24]

Per-FLOP figures are this note's arithmetic.

On these rows A100 and H100 cost about the same per dense FP16 FLOP to buy, and B200 about half. The rows mix a launch price with a server, estimates and a range given at announcement. None is a transaction price.

Rents are posted by provider and pricing type. AWS's on-demand price file for US East (N. Virginia), published 10 November 2026, lists p4d.24xlarge at $21.957642 an hour, p5.48xlarge at $55.04 and p6-b200.48xlarge at $113.9328 [26], each with eight GPUs: A100 40 GiB, H100 80 GiB and B200 179 GiB [27]. That is $2.74, $6.88 and $14.24 per GPU-hour (this note's arithmetic). Azure's Retail Prices API lists Standard_ND96isr_H100_v5 in East US at $98.32 an hour on Linux from 1 December 2023, spot at $18.169536 from 1 January 2026 [28], and Standard_ND96asr_v4 at $27.197 from 1 March 2021 [29], each with eight GPUs, H100 80 GB and A100 40 GB [30, 31]. That is $12.29 on-demand and $2.27 spot per H100-hour, and $3.40 per A100-hour (this note's arithmetic). Lambda's 8x plan lists per GPU-hour rates that take effect on 16 November 2026: B200 180 GB $7.69, H100 80 GB $4.59, A100 80 GB $3.19 and A100 40 GB $1.99 [32].

The same class varies by source. Azure's on-demand H100 rate is 1.79 times AWS's and 5.4 times its own spot rate, and Lambda's A100 80 GB rate is 1.60 times its A100 40 GB rate (this note's arithmetic) [26, 28, 32]. An instance price also bundles more than GPUs: p5.48xlarge carries 192 vCPUs, 2,048 GiB of memory and 3,200 Gigabit networking [27].

Rents move without a change in hardware. From 1 June 2025, AWS cut on-demand prices by 33% for P4d and P4de (A100), 44% for P5 (H100) and 25% for P5en (H200) [25]. Rent per FLOP for each class at AWS fell by that share on one date. Epoch notes that pricing for cluster hardware "is often negotiated privately rather than publicly listed" [2].

5. Quality change in price indexes: matched-model and hedonic methods

Matched models. A price index follows the same item over time, so a change in its price is a pure price change. When the item leaves the market, the BLS chooses a replacement, and a less comparable one can bring "quality change and an associated price differential into the index" [33]. The Producer Price Index (PPI) "tracks only 'pure' price change, which is price change that excludes changes in quality" [35].

Hedonic adjustment. The hedonic method "removes any price differential attributed to a change in quality by adding or subtracting the estimated value of that change from the price of the old item" [33]. In January 2018 the BLS began estimating quality-adjusted price change in the PPI for microprocessors each quarter with a time dummy approach [34]. In a time dummy model, characteristics enter as explanatory variables, and, as the BLS cloud computing study puts it, "the time dummy variable gives the quality adjusted price change between the two quarters" [35]. Personal computers in the CPI use cost-based adjustments instead [33].

What the choice does to a chip index. Byrne, Oliner and Sichel argue that the matched-model method in the PPI for microprocessors "likely started to be biased in the mid-2000s" [36]. From 2008 to 2013 their preferred index fell at an average 43% a year, while the PPI fell 8% [36].

Cloud services. Byrne, Corrado and Sichel built hedonic price indexes for AWS compute, database and storage services from 2009 to 2016, with double-digit yearly declines for all three from early 2014 [37]. A 2023 BLS study found that the PPI's approach to a changed cloud service "does not specifically account for changes to characteristics such as microprocessor, memory, and data storage", and tested a time dummy model as an exploratory alternative for CPU instances [35].

Price per FLOP is a one-characteristic adjustment. Dividing a price by peak FLOP/s treats FLOP/s as the only quality that counts. A hedonic model would also value memory and bandwidth. H200 has H100's tensor rates with more memory and bandwidth [9, 10] and produced 1.12 times the tokens per second [22]. Price per peak FLOP counts it as an H100, so any premium for its extra memory reads as a higher price per FLOP.

6. A worked comparison of three generations from posted prices

This is an illustration from dated posted prices, computed by this note. It is not a Rillor Compute Index value, a quote or a transaction price.

US dollars per dense PFLOP/s-hour (price per GPU-hour divided by dense PFLOP/s):

Price source	A100 FP16	H100 FP16	H100 FP8	B200 FP16	B200 FP8	B200 FP4
AWS on-demand, US East, price file of 10 November 2026	$8.80	$6.95	$3.48	$6.33	$3.16	$1.58
Lambda 8x plan, rates from 16 November 2026 (A100 40 GB)	$6.38	$4.64	$2.32	$3.42	$1.71	$0.85

Sources: [7, 9, 11, 26, 27, 32]. The arithmetic is this note's.

Same precision, same provider. B200 costs 9% less per dense FP16 or FP8 FLOP than H100 at AWS and 26% less at Lambda [9, 11, 26, 27, 32]. H100 costs 21% less per dense FP16 FLOP than A100 at AWS and 27% less against Lambda's A100 40 GB rate, or 55% less against its A100 80 GB rate of $3.19 [7, 9, 26, 27, 32]. Per GPU-hour, B200 costs 2.07 times H100 at AWS and 1.68 times at Lambda (this note's arithmetic) [26, 27, 32].

Mixing formats across generations. B200 at dense FP4 against H100 at dense FP8 shows 54% less per FLOP at AWS and 63% less at Lambda (this note's arithmetic) [9, 11, 26, 27, 32]. That gap counts a precision change as a price change.

A third provider reverses one step. At Azure's East US on-demand rates, A100 costs $10.90 per dense FP16 PFLOP/s-hour and H100 $12.42, so H100 costs 14% more per FLOP (this note's arithmetic) [7, 9, 28, 29, 30, 31]. Azure's spot H100 rate gives $2.30 (this note's arithmetic) [9, 28, 30]. Azure's A100 rate dates from 2021 and its H100 rate from 2023, while AWS cut both of its classes in 2025 [25, 28, 29].

Delivered cost on one benchmark. This derived cost metric combines MLPerf® throughput from NVIDIA's DGX systems with AWS on-demand rates for instances with the same GPU count, so it assumes the instance matches the DGX result. At 31,306.8 tokens per second and $55.04 an hour, eight H100s give 2.05 million tokens per dollar, or $0.49 per million [22, 26]. At 98,858.0 tokens per second and $113.9328 an hour, eight B200s give 3.12 million, or $0.32 per million [22, 26]. The B200 run costs 34% less per token, with FP4 weights against FP8 (this note's arithmetic), between the 9% per-FLOP gap at equal precision and the 54% gap across precisions. MLPerf® notes as in Finding 3. Cost per token is this note's derived metric and is not an MLPerf® metric.

Within each provider the region and pricing type are fixed. The bundle is not: CPU, memory and network differ by instance [27].

7. What a price series kept by GPU class holds fixed

The Rillor Compute Index page describes RCI Rental as published rental prices for GPU capacity in USD per GPU-hour, broken out by GPU class, region, term and pricing type (on-demand, spot or preemptible, reserved), with hyperscaler and neocloud segments shown separately. RCI Hardware covers GPU system prices by comparable grade, with air-cooled and liquid-cooled builds as separate grades. Methodology version 1.1 defines a GPU class as "a GPU model as the index groups it" and a variant as "the form factor and memory configuration a GPU class ships in."

What follows is this note's reading of a class-level series. A series kept by GPU class and variant, at one provider and pricing type, compares like hours, the logic of a matched-model index [33, 35, 36]. In Finding 4, provider, pricing type and memory variant move one class's rent by 1.6 to 5.4 times [26, 28, 32]. In Finding 6, B200's step over H100 at equal precision is 9% to 26% per FLOP [9, 11, 26, 27, 32]. A cross-generation figure that mixed providers or pricing types would carry those effects into the comparison. Class series kept apart let a reader add quality as a second, stated step.

Implications

For analysts quoting a trend. Quote each estimate with its sample, period, FLOP basis, price basis and deflator [1, 3, 4, 6]. The 2.46-year doubling and the 49% yearly gain come from different samples and prices [1, 4], and a trend spliced from them has no basis.

For capacity planners comparing classes. Count FLOPs at the precision the workload runs: dense BF16 for training, as the 405-billion-parameter training report does [19], and FP8 or FP4 for inference only where the model's accuracy at that format has been checked [15]. Use dense rates unless the network is pruned to the 2:4 pattern [14]. Where an MLPerf® result matches the workload, price it as a derived cost metric and say so [23]. Compare classes within one provider and pricing type first.

For researchers building price series. Keep each class and variant separate, record form factor and memory, and date every price [9, 11, 32]. Treat quality as an explicit step, with one characteristic such as dense FLOP/s or a hedonic model on the BLS time dummy pattern [35]. The chip study shows the method can change a measured decline several times over [36].

A checklist for one comparison.

1. Name the precision and format, and mark dense or sparse.
2. Use the same precision at both ends, or a TPP-style measure.
3. Use the form factor's own rate, and keep memory variants apart.
4. State peak or delivered, and for delivered give the round, result IDs and scenario.
5. Give each price's source, date, region and kind.
6. Divide instance prices by the documented GPU count, and note the bundle.
7. Label the result as arithmetic from posted prices.

Limits of this analysis

- Peak rates come from vendor datasheets and product pages. Delivered rates come from one MLPerf® Inference round, two benchmarks and single eight-GPU nodes, with no A100 result in that round [22]. MLPerf® Training was not examined.
- The delivered cost pairs NVIDIA's systems with AWS instance prices and assumes the two match.
- Prices are posted rates, estimates, a launch price and a range given at announcement, never transaction prices. Lambda's rates take effect on 16 November 2026 [32]. Azure's on-demand rates carry effective dates from 2021 and 2023, and its spot H100 rate from 1 January 2026 [28, 29].
- Rents are nominal 2026 dollars with no deflator applied. Each provider's figures cover one region, and the per-GPU split assigns the whole instance price to its GPUs.
- Every purchase estimate is Epoch's reading of its own sources [5].
- No source reviewed builds a hedonic index for GPUs or GPU rentals. The cloud studies cover CPU-era services [35, 37].
- The note makes no forecast. The worked figures are illustrations, not Rillor Compute Index values.

Sources

1. Trends in GPU price-performance. Marius Hobbhahn and Tamay Besiroglu. Epoch AI, 27 June 2022, viewed 8 November 2026. https://epoch.ai/publications/trends-in-gpu-price-performance

2. Trends in Machine Learning Hardware. Marius Hobbhahn, Lennart Heim and Gökçe Aydos. Epoch AI, 9 November 2023, viewed 8 November 2026. https://epoch.ai/publications/trends-in-machine-learning-hardware

3. Performance per dollar improves around 30% each year. Robi Rahman. Epoch AI, 23 October 2024, viewed 8 November 2026. https://epoch.ai/data-insights/price-performance-hardware

4. The performance per dollar of AI chips purchased each quarter has grown by an average of 49% per year. Venkat Somala. Epoch AI, 13 August 2026, viewed 8 November 2026. https://epoch.ai/data-insights/chip-performance-per-dollar

5. AI chip sales documentation: Methodology. Epoch AI, viewed 8 November 2026. https://epoch.ai/data/ai-chip-sales-documentation/methodology

6. 2019 recent trends in GPU price per FLOPS. AI Impacts, 25 March 2020, viewed 8 November 2026. https://aiimpacts.org/2019-recent-trends-in-gpu-price-per-flops/

7. NVIDIA A100 Tensor Core GPU datasheet. NVIDIA, June 2021, viewed 8 November 2026. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf

8. NVIDIA Ships World's Most Advanced AI System, NVIDIA DGX A100, to Fight COVID-19. NVIDIA Newsroom, 14 May 2020, viewed 8 November 2026. https://nvidianews.nvidia.com/news/nvidia-ships-worlds-most-advanced-ai-system-nvidia-dgx-a100-to-fight-covid-19-third-generation-dgx-packs-record-5-petaflops-of-ai-performance

9. NVIDIA H100 GPU. NVIDIA, viewed 8 November 2026. https://www.nvidia.com/en-us/data-center/h100/

10. NVIDIA H200 GPU. NVIDIA, viewed 8 November 2026. https://www.nvidia.com/en-us/data-center/h200/

11. Datasheet for NVIDIA Blackwell Architecture. NVIDIA, October 2025, viewed 8 November 2026. https://dam-cdn.nvd.orangelogic.com/AssetLink/y441155802qub41q118b2852i557jem5.pdf

12. Datasheet for NVIDIA Blackwell Ultra. NVIDIA, October 2025, viewed 8 November 2026. https://dam-cdn.nvd.orangelogic.com/AssetLink/1k0p832eq8r5ca0u5383ie5o4tp3bst1.pdf

13. NVIDIA DGX B200. NVIDIA, viewed 8 November 2026. https://www.nvidia.com/en-us/data-center/dgx-b200/

14. Accelerating Inference with Sparsity Using the NVIDIA Ampere Architecture and NVIDIA TensorRT. Jeff Pool, Abhishek Sawarkar and Jay Rodge. NVIDIA Technical Blog, 20 July 2021, viewed 8 November 2026. https://developer.nvidia.com/blog/accelerating-inference-with-sparsity-using-ampere-and-tensorrt/

15. Introducing NVFP4 for Efficient and Accurate Low-Precision Inference. Eduardo Alvarez and others. NVIDIA Technical Blog, 24 June 2025, viewed 8 November 2026. https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-

accurate-low-precision-inference/

16. OCP Microscaling Formats (MX) Specification, Version 1.0. Open Compute Project, September 2023, viewed 8 November 2026. https://www.opencompute.org/documents/ocp-microscaling-formats-mx-v1-0-spec-final-pdf

17. Implementation of Additional Export Controls: Certain Advanced Computing Items, 88 FR 73458, document 2023-23055. Bureau of Industry and Security, Federal Register, 25 October 2023, viewed 8 November 2026. https://www.federalregister.gov/documents/2023/10/25/2023-23055/implementation-of-additional-export-controls-certain-advanced-computing-items-supercomputer-and

18. PaLM: Scaling Language Modeling with Pathways. Aakanksha Chowdhery and others. arXiv 2204.02311, version 5, 5 October 2022, viewed 8 November 2026. https://arxiv.org/abs/2204.02311

19. The Llama 3 Herd of Models. Aaron Grattafiori and others. arXiv 2407.21783, version 3, 23 November 2024, viewed 8 November 2026. https://arxiv.org/html/2407.21783v3

20. Benchmark MLPerf Inference: Datacenter. MLCommons, viewed 8 November 2026. https://mlcommons.org/benchmarks/inference-datacenter/

21. MLCommons Releases New MLPerf Inference v5.0 Benchmark Results. MLCommons, 2 April 2025, viewed 8 November 2026. https://mlcommons.org/2025/04/mlperf-inference-v5-0-results/

22. MLPerf Inference v5.0 results, summary_results.json. MLCommons, 2 April 2025, viewed 8 November 2026. https://raw.githubusercontent.com/mlcommons/inference_results_v5.0/main/summary_results.json

23. MLPerf Results Messaging Guidelines. MLCommons, viewed 8 November 2026. https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc

24. Nvidia shares close up after company unveils latest AI chips. Alex Koller. CNBC, 19 March 2024, viewed 8 November 2026. https://www.cnbc.com/2024/03/19/nvidia-stock-dips-after-company-unveils-latest-ai-chips.html

25. Announcing up to 45% price reduction for Amazon EC2 NVIDIA GPU-accelerated instances. Channy Yun. AWS News Blog, 5 June 2025, viewed 8 November 2026. https://aws.amazon.com/blogs/aws/announcing-up-to-45-price-reduction-for-amazon-ec2-nvidia-gpu-accelerated-instances/

26. Amazon EC2 On-Demand price data, US East (N. Virginia), Linux. Amazon Web Services, published 10 November 2026, viewed 11 November 2026. https://b0.p.awsstatic.com/pricing/2.0/meteredUnitMaps/ec2/USD/current/ec2-ondemand-without-sec-sel/US%20East%20%28N.%20Virginia%29/Linux/index.json

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

28. Azure Retail Prices API, Standard_ND96isr_H100_v5 in East US. Microsoft, viewed 8 November 2026. https://prices.azure.com/api/retail/prices?$filter=armSkuName%20eq%20%27Standard_ND96isr_H100_v5%27%20and%20armRegionName%20eq%20%27eastus%27%20and%20priceType%20eq%20%27Consumption%27

29. Azure Retail Prices API, Standard_ND96asr_v4 in East US. Microsoft, viewed 8 November 2026. https://prices.azure.com/api/retail/prices?$filter=armSkuName%20eq%20%27Standard_ND96asr_v4%27%20and%20armRegionName%20eq%20%27eastus%27%20and%20priceType%20eq%20%27Consumption%27

30. ND H100 v5 sizes series. Microsoft Learn, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/ndh100v5-series

31. NDasrA100_v4 sizes series. Microsoft Learn, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/ndasra100v4-series

32. GPU cloud pricing. Lambda, viewed 8 November 2026. https://lambda.com/pricing

33. Quality Adjustment in the CPI. U.S. Bureau of Labor Statistics, last modified 28 September 2026, viewed 8 November 2026. https://www.bls.gov/cpi/quality-adjustment/

34. A Review of Hedonic Price Adjustment Techniques for Products Experiencing Rapid and Complex Quality Change. U.S. Bureau of Labor Statistics, last modified 15 September 2022, viewed 8 November 2026. https://www.bls.gov/cpi/quality-adjustment/hedonic-price-adjustment-techniques.htm

35. Exploring quality adjustment in PPI cloud computing. Steven D. Sawyer and Claire O'Bryan. Monthly Labor Review, U.S. Bureau of Labor Statistics, February 2023, viewed 8 November 2026. https://www.bls.gov/opub/mlr/2023/article/exploring-quality-adjustment-in-ppi-cloud-computing.htm

36. How Fast are Semiconductor Prices Falling? NBER Working Paper 21074. David M. Byrne, Stephen D. Oliner and Daniel E. Sichel. National Bureau of Economic Research, April 2015, viewed 8 November 2026. https://www.nber.org/papers/w21074

37. The Rise of Cloud Computing: Minding Your P's, Q's and K's. NBER Working Paper 25188. David Byrne, Carol Corrado and Daniel E. Sichel. National Bureau of Economic Research, October 2018, viewed 8 November 2026. https://www.nber.org/papers/w25188

About this report

Published by Rillor on 12 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. How to compare GPU generations on price per FLOP. 12 November 2026. https://rillor.com/insights/gpu-price-per-flop-across-generations

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.

Rillor Compute Index series are for research and planning use. They are not offered for settlement of any contract.

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 series, GPU classes and history you need, and how you plan to use them. Contact us at rillor.com/contact?type=index or contact@rillor.com. More on the Rillor Compute Index: rillor.com/markets/compute-index