The Economic Life of an Accelerator

Rental Prices, Power and the Six-Year Depreciation of Data-Centre GPUs

Abhay Chakra Sadineni
University of California, Berkeley
abhaychakra@berkeley.edu

November 2026
DOI: 10.5281/zenodo.24210935

Abstract

The largest buyers of data-centre GPUs depreciate them over five to six years. A common critique holds that a new chip every year makes three years more realistic. Both sides argue from the release calendar. This paper argues from prices. It sets public rental indices for the NVIDIA A100, H100, H200 and B200 against each chip's arithmetic throughput, memory bandwidth, memory capacity and system power, and asks when an owner should stop running a chip. Three results follow. First, older chips rent well above arithmetic parity with the newest one. On 26 October 2026 the six-year-old A100 rented for 70 percent more than the B200's price per unit of dense BF16 throughput would imply. Second, once launch scarcity passed, rents fell slowly, by about 13 percent a year for the A100 since April 2023 and 9 percent a year for the H100 since October 2024. Third, the decision to retire a chip turns on power. Where power and space can be bought at market prices, an A100 earns about five times its operating cost before idle hours, and operating cost alone would not retire it before age eight even if rents fell 40 percent a year. Where power is the binding constraint, the answer turns on how fast the replacement's own rent falls. If B200 rents fall 9.4 percent a year, as H100 rents have, the A100 sits at the replacement threshold now, at six and a half years. If they hold flat, both the A100 and the H100 should already be replaced. If they fall 13 percent a year or more, both remain economic. The six-year schedule is therefore plausible for power-limited sites and conservative elsewhere, and its accuracy depends on a rent path no one yet observes. The paper closes with what this means for lenders who take GPUs as collateral.

Keywords: GPU; economic depreciation; useful life; replacement; rental price index; data-centre power; AI infrastructure.
JEL: D24, G31, L63, M41.

1 Introduction

In January 2025 Amazon shortened the estimated useful life of a subset of its servers and networking equipment from six years to five, citing the pace of development in artificial intelligence [1]. In the same month Meta lengthened the life of most of its servers and network assets to five and a half years [2]. Alphabet and Microsoft use six years for servers [3, 4], and so does CoreWeave, the largest of the specialised GPU clouds [5]. Two companies with the same chips, the same supplier and the same year moved in opposite directions.

The disagreement matters beyond earnings per share. Lenders now accept GPUs as collateral for tens of billions of dollars of debt, and the public record of those facilities contains no market

residual value assumption [6]. If a chip's economic life is three years, much of that collateral is worth little before the loans amortise. If it is eight, the collateral is better than its critics think.

The usual argument starts from the release calendar. NVIDIA ships a new data-centre architecture roughly every year, and each one delivers more arithmetic per dollar. It follows, the argument goes, that the previous chip is obsolete. That inference skips a step. A chip is retired when the cash it earns no longer covers what it costs to keep running, or when the space and power it occupies would earn more holding something newer. Neither condition depends on the release calendar directly. Both can be read from prices.

This paper reads them. Section 3 describes the hardware and price data. Section 4 shows how rent varies with capability across generations and within one. Section 5 estimates how fast rents have fallen once launch scarcity passed. Section 6 sets out two exit rules, one for owners with spare power and one for owners without, and applies both. Section 7 draws the consequences for depreciation and for lending.

2 Related work

Economists distinguish physical wear from economic depreciation, the fall in an asset's value with age, which includes obsolescence [7]. For personal computers, resale evidence shows machines losing roughly half their remaining value with each year of use, and part of that loss reflects revaluation as better machines arrive rather than wear [8]. The decision to replace a working machine is a classic problem in its own right, with the replacement point set by the comparison of running costs against the gain from a new unit [9]. Collateral value, in turn, depends on how easily the asset can be redeployed to another user [10].

On the hardware side, the roofline model ties attainable performance to arithmetic intensity, the number of operations performed per byte moved from memory [11]. Peak arithmetic on accelerators has grown much faster than memory bandwidth, and many modern workloads are limited by the latter [12]. Large-language-model inference is a prominent case, because generating each token reads the full set of model weights [13]. If the market prices what workloads can actually use, rents should follow bandwidth and capacity more closely than peak arithmetic. Section 4 looks at that idea and finds that three chips cannot settle it.

3 Data

3.1 Hardware

Table 1 lists the four chips for which daily or monthly rental prices are public. Throughput figures are dense, without structured sparsity, because vendors headline sparse figures that are twice as large and rarely realised. The B200 row uses the 1,000 W HGX configuration, which has lower throughput and power than the B200 in NVIDIA's GB200 rack systems. System power is the maximum draw of NVIDIA's own eight-GPU DGX server divided by eight, which includes the host processors, network adapters and fans that a GPU cannot run without.

3.2 Prices

Two sources are used. Silicon Data publishes daily rental price indices for the A100, H100, H200 and B200, standardised to a price per GPU-hour on neo-cloud providers [21–23]. The cross-section uses the values for one date, 26 October 2026, when the indices stood at \$1.56 for the A100, \$2.89 for the H100 and \$6.63 for the B200. The A100 and B200 pages also carry

Chip	First available	BF16 dense (TFLOPS)	Bandwidth (TB/s)	Memory (GB)	GPU power (W)	System kW per GPU
A100 SXM	2020	312	2.04	80	400	0.81
H100 SXM	2022	989	3.35	80	700	1.28
H200 SXM	2024	989	4.80	141	700	1.28
B200 HGX	2025	2,250	7.70	180	1,000	1.79

Table 1. Hardware used in the analysis. Sources are the NVIDIA datasheets for the A100, H100 and H200 [14–16], the Lenovo product guide for the HGX B200 [17], and the DGX A100, DGX H100 and DGX B200 datasheets for system power (6.5, 10.2 and 14.3 kW per eight GPUs) [18–20]. The DGX H200 is assumed to draw the same as the DGX H100.

undated FAQ figures of $1.29 and $5.45, which disagree with their dated series and are not used. Silicon Data also published a twelve-week comparison of its H100 and H200 indices for May to July 2026 [24]. Its longer histories are not public.

For the longer history, the paper uses the monthly series published by Virtualized [25]. Before September 2026 its publisher describes these as monthly approximations of the typical on-demand rate, compiled from several trackers and archived price pages, and as directional rather than tick data. From September 2026 they are computed from 123 tracked rates across 24 providers. The two sources differ in level. Virtualized put the H100 at $2.69 in November 2026, and its own FAQ quotes a late-2026 median nearer $3.35, while Silicon Data's index stood at $3.06 on 12 November. The analysis therefore uses Virtualized only for rates of change and Silicon Data only for levels.

3.3 Costs

Wholesale colocation in the primary North American markets averaged an asking rate of $196.25 per kilowatt-month for 250 to 500 kW requirements in the second half of 2025 [26], or $0.269 per kilowatt-hour of IT load. US industrial electricity averaged 8.62 cents per kilowatt-hour in 2025 [27]. The base case assumes a power usage effectiveness of 1.3, below the 1.56 industry average reported for 2024 [28] because GPU halls are newer and more often liquid-cooled. Together these give an operating cost of $0.38 per kilowatt-hour of system power. Per GPU-hour that is $0.31 for the A100, $0.49 for the H100 and $0.68 for the B200, of which electricity is $0.09, $0.14 and $0.20. Staff, maintenance and network transit are excluded, so these figures are lower bounds.

4 The price schedule across chips

Chip	Rent ($/h)	$ per BF16 PFLOPS-h	$ per TB/s-h	$ per GB-h	$ per system kW-h
A100	1.56	5.00	0.77	0.019	1.92
H100	2.89	2.92	0.86	0.036	2.27
B200	6.63	2.95	0.86	0.037	3.71

Table 2. Rent per unit of capability, Silicon Data indices, 26 October 2026.

Table 2 divides each chip's rent by each measure of capability. Per unit of arithmetic, the A100 is by far the most expensive, while the H100 and B200 are priced almost identically. From the A100 to the B200, throughput rises 7.2 times while rent rises 4.2 times. A line through the three chips on logarithmic axes has a slope of 0.72. A market that priced chips purely by arithmetic would rent the A100 at $0.92 an hour, the B200's price per PFLOPS-hour times the A100's throughput. It rents at $1.56.

Bandwidth tracks the cross-section at least as closely. Rent rises 4.2 times from the A100 to the B200 while bandwidth rises 3.8 times, and a line through the three chips on logarithmic axes has a slope of 1.08. Priced at the B200's rate per TB/s, the A100 would rent at $1.76, within 13 percent of its actual price. Both lines pass close to all three points. With three chips and one degree of freedom, neither fit supports inference, and the two explanations cannot be separated, because bandwidth and throughput rise together across generations. The ranking of the two fits also changed with the date of the prices used, which is a further reason not to read much into it.

The H100 and H200 provide a cleaner test. The two have identical arithmetic throughput. The H200 has 43 percent more bandwidth and 76 percent more memory. Over May to July 2026 the H200 index ran 0.4 to 16.2 percent above the H100, averaging 5.8 percent in May and 12.1 percent in July [24]. Those averages imply an elasticity of rent to bandwidth between 0.16 and 0.32. Within a generation, then, extra bandwidth and memory earn a modest premium. Across generations, something that moves with both throughput and bandwidth is being paid for. It may be bandwidth, newer interconnect and software support, or simple preference for the current model.

The conclusion that survives is narrower than any of these stories. The oldest chip rents well above arithmetic parity with the newest one, by 70 percent for the A100. An obsolescence model that values chips by peak throughput relative to the frontier will understate the cash an older chip earns.

5 How fast rents fall

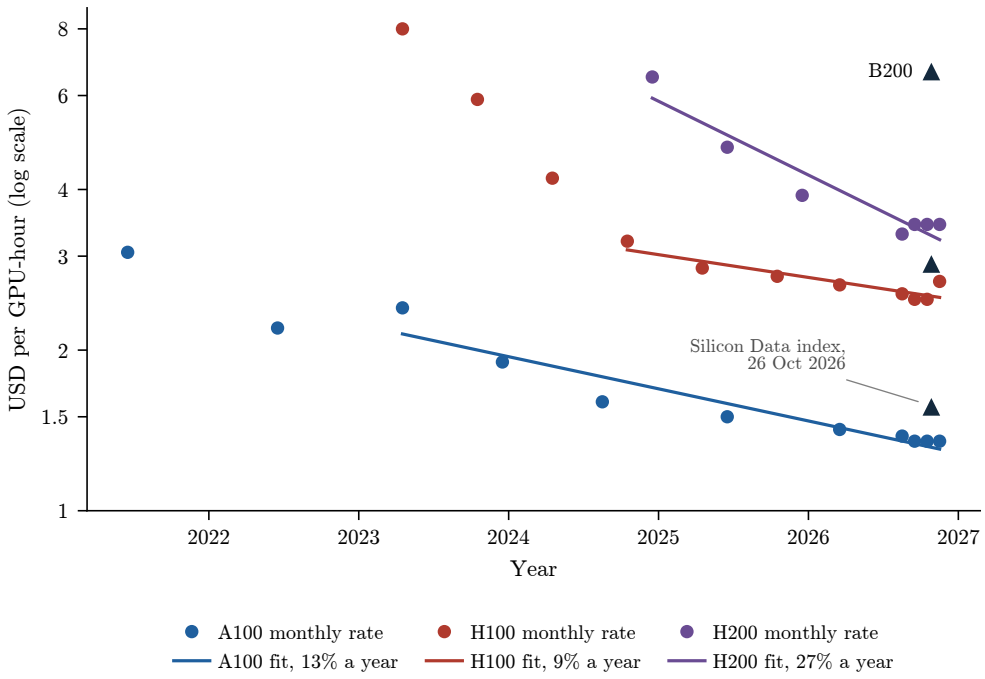


Figure 1. Monthly typical on-demand rents from Virtualized [25], with log-linear fits over the period after launch scarcity, and Silicon Data index levels for 26 October 2026 [21–23]. The H100 fit starts in October 2024, the A100 fit in April 2023 and the H200 fit at its first observation.

Figure 1 shows the history. Each chip launches into scarcity and its rent falls fast until supply catches up. The H100 fell from about $8 an hour in April 2023 to about $3.20 by October 2024, a decline of about 46 percent a year over those eighteen months. After that the decline slowed sharply. Fitted from October 2024 to November 2026, the H100 rent fell 9.4 percent a

year ($R^2 = 0.83$, eight observations). The A100, fitted from April 2023, fell 12.9 percent a year ($R^2 = 0.90$, nine observations). The H200, still close to its launch, fell 27 percent a year from December 2024.

The fits describe a period of strong demand, and late 2025 to late 2026 included outright increases. Silicon Data reported its H100 index at $2.00 on 9 December 2025 [29], and the index stood at $3.06 on 12 November 2026 [21]. The slow post-scarcity decline should not be read as a forecast. Section 6 therefore tests faster declines of 25 and 40 percent a year alongside the fitted ones.

6 When to stop running a chip

6.1 Two exit rules

Let $r(a)$ be the rent a chip earns per GPU-hour at age a, u the share of hours it is rented, k its system power in kilowatts, and c the operating cost per kilowatt-hour of system power from Section 3. The capital cost of a chip already installed is sunk.

An owner who can buy more power and space at market prices should keep a chip running as long as it covers its own operating cost,

$$u\, r(a) > k\, c. \tag{1}$$

An owner whose site is full has a different problem. Every kilowatt running an old chip is a kilowatt not running a new one. Over a planning horizon of five years, the old chip should stay only while it earns more per kilowatt than its replacement would after paying for itself,

$$\frac{u\,\bar{r}}{k} - c > \frac{u\,\bar{R}}{K} - c - \frac{C\,\phi}{K_s \cdot 8760}. \tag{2}$$

Here $\bar{r}$ and $\bar{R}$ are the levelised rents of the old chip and its replacement over the five years, each falling at its own annual rate and discounted monthly at 10 percent. K is the replacement's system power per GPU, C the price of a replacement server holding K_s kilowatts, and $\phi = 0.264$ the factor that recovers capital over five years at 10 percent. The base case uses the B200 as the replacement, a DGX B200 listed at $515,410 [30], and $K_s = 14.3$ kW, which gives a capital charge of $1.09 per kilowatt-hour. The old chips' rents fall at their fitted rates from Section 5. The B200's rate of decline is unknown, because its post-scarcity history is too short, so it is varied. Utilisation is 70 percent for every chip.

Rule (2) is the one that bites in 2026, because new power connections in the main markets are scarce. CBRE reported a record-low primary-market vacancy of 1.4 percent at the end of 2025 [26].

6.2 Results

Under rule (1), nothing in the observed range retires either chip soon. At its 26 October index the A100 earns 5.0 times its operating cost per hour, the H100 6.0 times and the B200 9.7 times, before allowing for idle hours. At 70 percent utilisation the multiples are 3.5, 4.2 and 6.8. At the fitted declines the A100 would cover its operating cost until about age 15 and the H100 until about age 19. Even at a 40 percent annual decline, the A100 would run until about age 9 and the H100 until about age 7. These are extrapolations far beyond the data and should be read as bounds, not forecasts. They say that operating cost alone does not end a GPU's life inside six years.

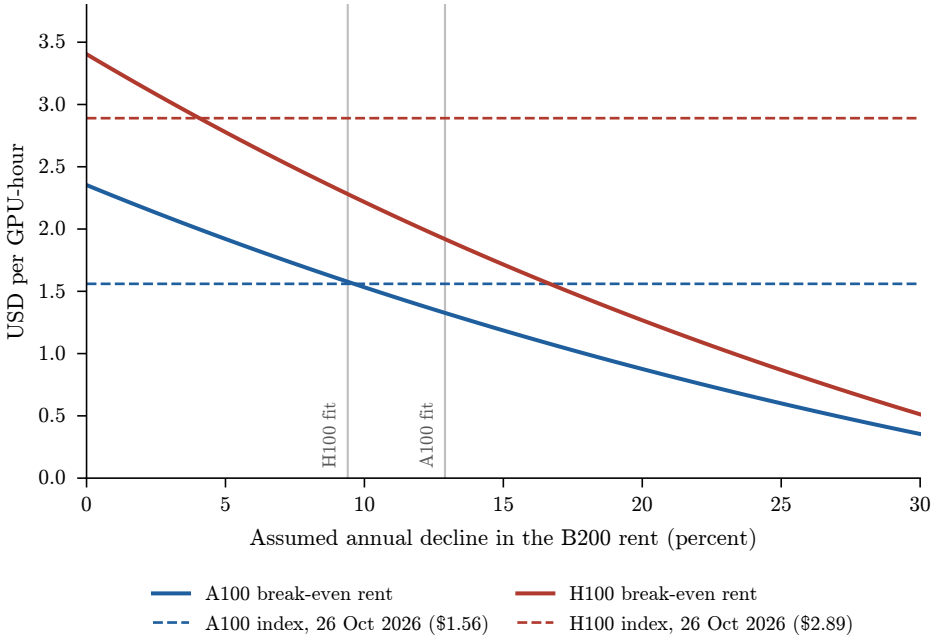


Figure 2. Current rent at which keeping each chip and replacing it with a B200 break even in a power-limited site, rule (2), as a function of how fast the B200's own rent falls. Dashed lines show the 26 October 2026 index. Above its break-even line a chip should be kept. Vertical lines mark the fitted declines of the H100 and A100.

Rule (2) gives an answer that depends almost entirely on the replacement's rent path, as Figure 2 shows. If the B200's rent held at $6.63 for five years, it would net $1.13 per kilowatt-hour after capital, and both older chips should be replaced now. The A100 would need to rent at $2.35 to stay, against an index of $1.56, and the H100 at $3.40, against $2.89. If the B200's rent falls 9.4 percent a year, as the H100's has since late 2024, the B200 nets $0.63, the A100's break-even falls to $1.58, and the A100 sits on the line, 1 percent below it, at six and a half years. The H100 clears its break-even of $2.28 by 27 percent. If the B200's rent falls 13 percent a year or faster, both older chips should stay. At 25 percent a year the B200 does not cover its own capital charge at 70 percent utilisation.

	A100 break-even rent ($/h)			H100 break-even rent ($/h)		
Utilisation	$450k	$515k	$600k	$450k	$515k	$600k
0.5	1.20	0.90	0.51	1.74	1.30	0.74
0.7	1.79	1.58	1.30	2.59	2.28	1.88
0.9	2.12	1.95	1.74	3.07	2.82	2.51

Table 3. Break-even current rent under rule (2), with the B200 rent falling 9.4 percent a year, by utilisation and by the price of a DGX B200. Index values on 26 October 2026 were $1.56 for the A100 and $2.89 for the H100. Reseller listings for eight-GPU B200 systems span roughly $450,000 to $600,000 [30, 31].

Table 3 holds the B200's decline at the H100's observed rate and varies the other two inputs. Higher utilisation makes the new chip more valuable and raises the bar for the old one. A dearer replacement lowers the bar. Across the table the A100 sits on either side of its threshold, which is the honest summary of its position. The H100 clears its threshold in eight of the nine cells.

The rule also omits two forces that shorten life. The comparison chip will not stay the B200. Each new generation raises the rent a kilowatt can earn, and with it the bar an old chip must clear. And chips fail. During a 54-day period of training on 16,384 H100s, Meta recorded 148 interruptions caused by faulty GPUs and 72 by faulty high-bandwidth memory [32]. Scaled to a

year, that is an interruption for roughly 9 percent of GPUs, although an interruption is not a retirement and most faults are repaired.

7 What this means for depreciation and credit

The debate over useful life has treated obsolescence as a property of the chip. In this simple model it depends on the site as much as on the chip. In a facility with spare power, a GPU keeps paying its way long after a newer chip has arrived, and a six-year schedule is conservative. In a facility that cannot add power, every chip competes for its kilowatts with the newest chip on the market. On the assumptions above, the A100 is close to losing that contest at about the age the accounting assumed, and whether it has lost depends on a rent path for the B200 that has not yet been observed.

Two consequences follow for lenders. First, the residual value of GPU collateral depends on where the collateral sits. Chips in a power-limited building can be displaced on roughly the schedule above, while chips that could be moved to a site with spare power keep more value. The earlier review of GPU-backed facilities did not record relocation rights, and none of the facilities' public summaries mention them [6]. Second, valuing GPU collateral by peak arithmetic relative to the newest chip understates it, at least for the A100, which rents for 70 percent more than that parity implies. A residual value curve should be built from rents per kilowatt and from the replacement threshold, not from throughput ratios alone.

The same logic marks out a segment that ages differently. NVIDIA's Blackwell Ultra generation cut native double-precision throughput to about 1.4 TFLOPS per package, from 40 on the B200 in NVIDIA's GB200 rack systems and 37 on the 1,000 W HGX B200 [17, 33]. For scientific users who need native FP64, the H100 and H200, at 67 TFLOPS and 700 W, remain NVIDIA's strongest parts per watt, and their economic life in that market will be set by demand from science, not by the AI release calendar.

8 Limits

The price data are thin. The cross-section has three chips. The H100 and H200 comparison covers twelve weeks. The monthly history before September 2026 is approximate by its publisher's own description, and Silicon Data's longer history is private. Rental indices measure what renters pay, not what owners receive after discounts, and they say nothing about resale prices, which would be the direct measure of collateral value. The rental indices also publish figures that disagree with one another on the same page, which is why one dated set is used throughout. Utilisation is assumed, not observed. Operating cost excludes staff and maintenance. The replacement rule depends on a rate of decline for the B200 rent that has not been observed, and it ignores the arrival of the next generation. Each of these points could move the exit ages by a year or more. Two findings do not depend on them. The A100 rents well above arithmetic parity, and the exit decision is governed by power rather than by the release calendar.

Data and code

The analysis script and the figure script reproduce every number in this paper from the values in Section 3. Both are published with the paper on the author's research page.

Funding and interests

No external funding was received. The author has no financial interest in any company named.

References

[1] Amazon.com, Inc. (2025). Annual Report on Form 10-K for the fiscal year ended 31 December 2024. Note 1, change in estimated useful lives of servers and networking equipment.

[2] Meta Platforms, Inc. (2025). Annual Report on Form 10-K for the fiscal year ended 31 December 2024. Change in estimated useful lives of servers and network assets.

[3] Alphabet Inc. (2024). Annual Report on Form 10-K for the fiscal year ended 31 December 2023. Change in estimated useful lives of servers and certain network equipment.

[4] Microsoft Corporation (2022). Annual Report on Form 10-K for the fiscal year ended 30 June 2022. Change in accounting estimate for servers and network equipment.

[5] CoreWeave, Inc. (2025). Registration Statement on Form S-1, March 2025. https://www.sec.gov/Archives/edgar/data/1769628/000119312525044231/d899798ds1.htm

[6] Sadineni, A. C. (2026). Lending against machines nobody appraises: What GPU-backed credit facilities disclose, 2025 to 2026. Preprint, November 2026. https://doi.org/10.5281/zenodo.24210925

[7] Hulten, C. R., and Wykoff, F. C. (1981). The measurement of economic depreciation. In C. R. Hulten (Ed.), *Depreciation, Inflation, and the Taxation of Income from Capital* (pp. 81-125). Washington, DC: Urban Institute Press.

[8] Doms, M. E., Dunn, W. F., Oliner, S. D., and Sichel, D. E. (2004). How fast do personal computers depreciate? Concepts and new estimates. *Tax Policy and the Economy*, 18, 37-80.

[9] Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5), 999-1033.

[10] Benmelech, E., and Bergman, N. K. (2009). Collateral pricing. *Journal of Financial Economics*, 91(3), 339-360.

[11] Williams, S., Waterman, A., and Patterson, D. (2009). Roofline: An insightful visual performance model for multicore architectures. *Communications of the ACM*, 52(4), 65-76.

[12] Gholami, A., Yao, Z., Kim, S., Hooper, C., Mahoney, M. W., and Keutzer, K. (2024). AI and memory wall. *IEEE Micro*, 44(3), 33-39.

[13] Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., Xiao, K., Agrawal, S., and Dean, J. (2023). Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5.

[14] NVIDIA Corporation (2021). NVIDIA A100 Tensor Core GPU datasheet. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf

[15] NVIDIA Corporation (2023). NVIDIA H100 Tensor Core GPU datasheet. https://www.nvidia.com/en-us/data-center/h100/

[16] NVIDIA Corporation (2024). NVIDIA H200 Tensor Core GPU. https://www.nvidia.com/en-us/data-center/h200/

[17] Lenovo Press (2025). ThinkSystem NVIDIA HGX B200 180GB 1000W GPU, product guide LP2226. https://lenovopress.lenovo.com/lp2226-thinksystem-nvidia-b200-180gb-1000w-gpu

[18] NVIDIA Corporation (2020). NVIDIA DGX A100 datasheet. System power usage 6.5 kW maximum.

[19] NVIDIA Corporation (2022). NVIDIA DGX H100 datasheet. System power usage 10.2 kW maximum.

[20] NVIDIA Corporation (2024). NVIDIA DGX B200 datasheet. System power usage 14.3 kW maximum. https://www.nvidia.com/en-us/data-center/dgx-b200/

[21] Silicon Data (2026). H100 Rental Price Index, SDH100RT. Accessed 13 November 2026. https://www.silicondata.com/products/silicon-index/h100

[22] Silicon Data (2026). A100 Rental Price Index, SDA100RT. Accessed 13 November 2026. https://www.silicondata.com/products/silicon-index/a100

[23] Silicon Data (2026). B200 Rental Price Index, SDB200RT. Accessed 13 November 2026. https://www.silicondata.com/products/silicon-index/b200

[24] Silicon Data (2026). H200 vs H100 rental prices, May to July 2026. 4 August 2026. https://www.silicondata.com/blog/h200-vs-h100-rental-prices-may-july-2026

[25] Virtualized (2026). GPU price trends. Accessed 13 November 2026. https://virtualized.ai/gpu-price-trends

[26] CBRE Research (2026). North America Data Center Trends H2 2025. https://www.cbre.com/insights/books/north-america-data-center-trends-h2-2025

[27] US Energy Information Administration (2026). Electricity explained: Prices and factors affecting prices. https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php

[28] Uptime Institute (2024). Global Data Center Survey 2024. New York: Uptime Institute.

[29] Silicon Data (2026). H100 price spike: Understanding the 10% surge in GPU rental costs. January 2026. https://www.silicondata.com/blog/h100-price-spike

[30] Zuhair, M. (2024). NVIDIA Blackwell DGX B200 listed, prices start at a half a million dollars. Wccftech. https://wccftech.com/nvidia-blackwell-dgx-b200-price-half-a-million-dollars-top-of-the-line-ai-hardware/

[31] GPU per Hour (2026). NVIDIA GPU prices. https://gpuperhour.com/blog/nvidia-gpu-prices

[32] Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The Llama 3 herd of models. arXiv:2407.21783.

[33] Tom's Hardware (2025). Nvidia shares Blackwell Ultra's secrets: NVFP4 boost detailed and PCIe 6.0 support. https://www.tomshardware.com/pc-components/gpus/nvidia-shares-blackwell-ultras-secrets-nvfp4-boost-detailed-and-pcie-6-0-support