GPU Residual Value Report: 2026 Outlook

In AI infrastructure, the GPU residual value, or resale price, impacts every lease, loan, and equity return, collectively influencing the flow of trillions of dollars, yet misconceptions are too dangerous to ignore.

Key Takeaways

- GPU financing is a young asset class: investors only began underwriting the GPUs separately from datacenter transactions in 2023. As a result, GPU residual expertise is still developing across asset management teams at private credit, equipment finance, and private equity firms.
- As capital and speculation pour into GPU financing, misconceptions around residual value have become too dangerous to ignore. The closest precedent is IBM mainframes in the 1970s. Lessors who booked aggressive residuals went bankrupt within months of IBM's price cuts.

- ➔ GPUs are a rapidly evolving technology, more volatile than the assets that lenders typically underwrite. The volatility means traditional methods for modeling residuals can fall short. The most tempting, and the most hazardous, is curve-fitting. Fitting a curve to sparse data points invites overbooked residuals and unnecessary losses.
- ➔ Instead of predicted depreciation curves, we are publishing residual bands. Alongside these bands, we rank five potential adverse events: NVIDIA margin compression (high), AI bubble pop (high), bullwhip over-ordering (medium), hyperscaler refresh waves (medium), and technology obsolescence (low).
- ➔ GPU residuals are underwritable. We have transaction data covering over 600,000 secondary-market units and run scenario analysis with actuarial firms and leading reinsurers.
- ➔ A GPU's useful life can reach **eight years**, well beyond the typical financing term. Investors and lenders can also book residuals, even **north of 10%** of equipment cost over five years. But use extreme caution, as the downside from overbooking residuals is severe.

I. Why Residual Value Is the Number That Matters

What Is a Residual?

The residual is most broadly defined as "the resale value of the equipment as a percentage of its original cost." In practice, residuals let lenders take high-value GPU servers as collateral. They put a number on worst-case recovery in a default, and they give equity investors and lessors a path to higher IRR. Typical commercial lending works the other way, resting on the full faith and credit of the borrowing enterprise (or the end equipment user, the "offtaker") for repayment.

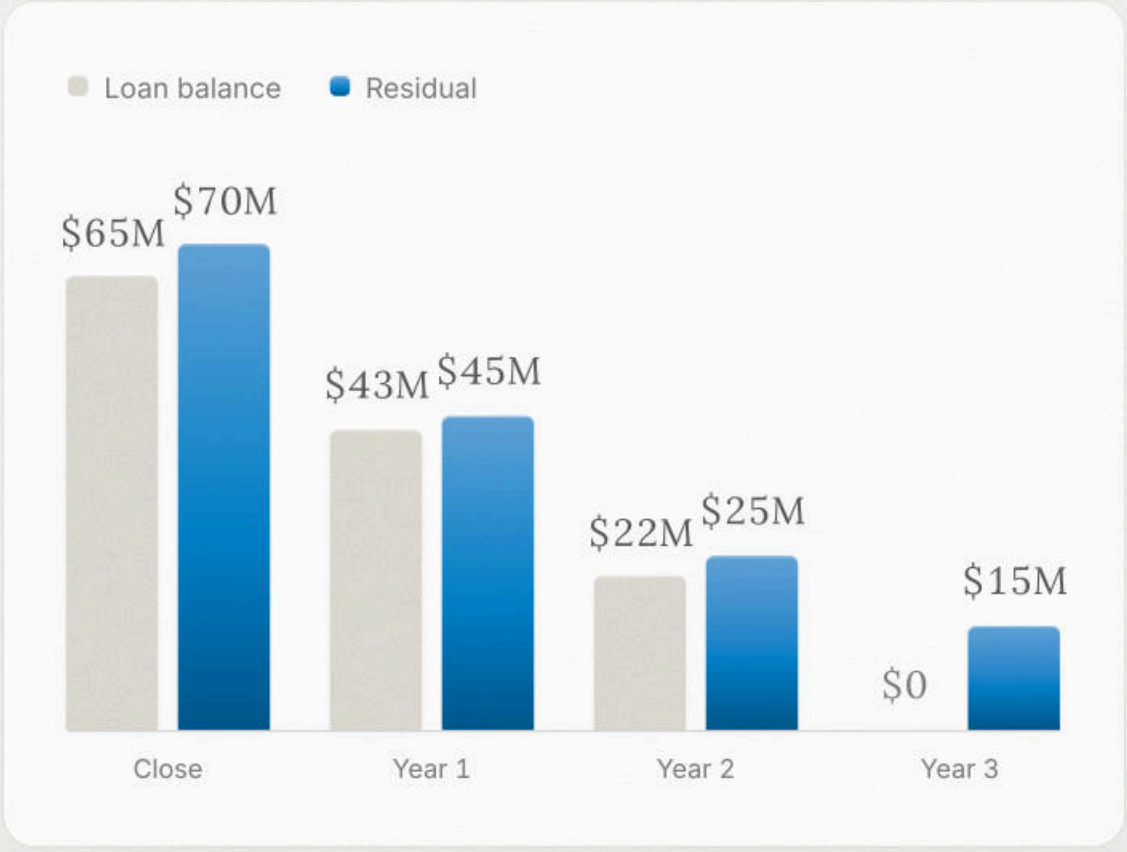
Most lenders agree that over a three-year term a GPU server should fetch more than 5% of original cost. They debate whether the right number is 10% or 60%.

Three Lenses on One Number

For the lender, the residual is used as the worst case: it sets the advance rate and the recovery math. A lender only loses when both a default and a weak resale market arrive together. The residual never comes into play if a default never occurs. But it still needs to be conservatively booked, because defaults and weak resale markets are strongly correlated. And a deal priced off of the equity investor's target residual is mispriced for the lender.

EXHIBIT 1

Residuals cover the loan balance across the term

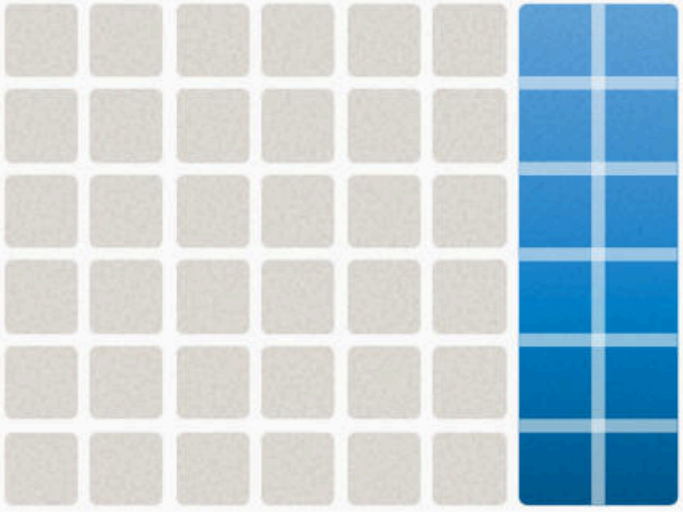


Illustrative: $100M system, 65% advance rate, 36-month amortizing loan.

For the lessor, the residual goes into every transaction. The lessor books a residual into the FMV lease on day one and must turn it into cash at the end of the term, by selling or re-leasing the equipment; a weak market alone produces a loss. That is why overbooked residuals have historically killed lessors first.

EXHIBIT 2

The residual covers part of the cost, so lease payments can be lower



Illustrative: $100M system, 36-month FMV lease, 10% financing, $30M booked residual.

For the equity investor, the residual adds upside: what the hardware could sell for in a strong market boosts IRR. An overbooked residual costs equity a weaker return but leaves no hole in the balance sheet. The residual is optional, but can take an IRR from roughly 20% to 35%.

EXHIBIT 3

Residual sales boost IRR at the end of term



Illustrative: $100M system, $65M loan, $35M equity, 4 year term; cash flow is after servicing debt.

II. The Overbooking Trap

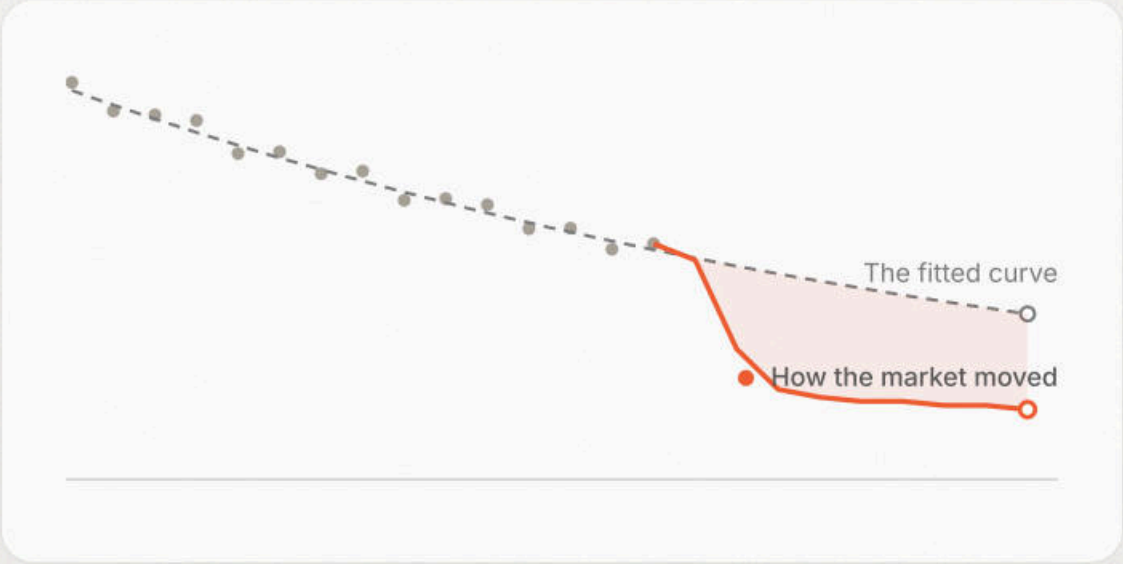
Overbooking, or booking a residual higher than it is worth, is common during speculative periods. A high residual prices the lease lower, lets the loan advance more, and lifts IRR projections. But when the market cools off, resale prices can come in far below the booked residual.

The Curve Illusion

Curves commonly cause overbooked residuals. Curves assume that historical performance predicts future outcomes, which the market routinely proves wrong.

EXHIBIT 4

Curves break when markets change



Illustrative; not market data.

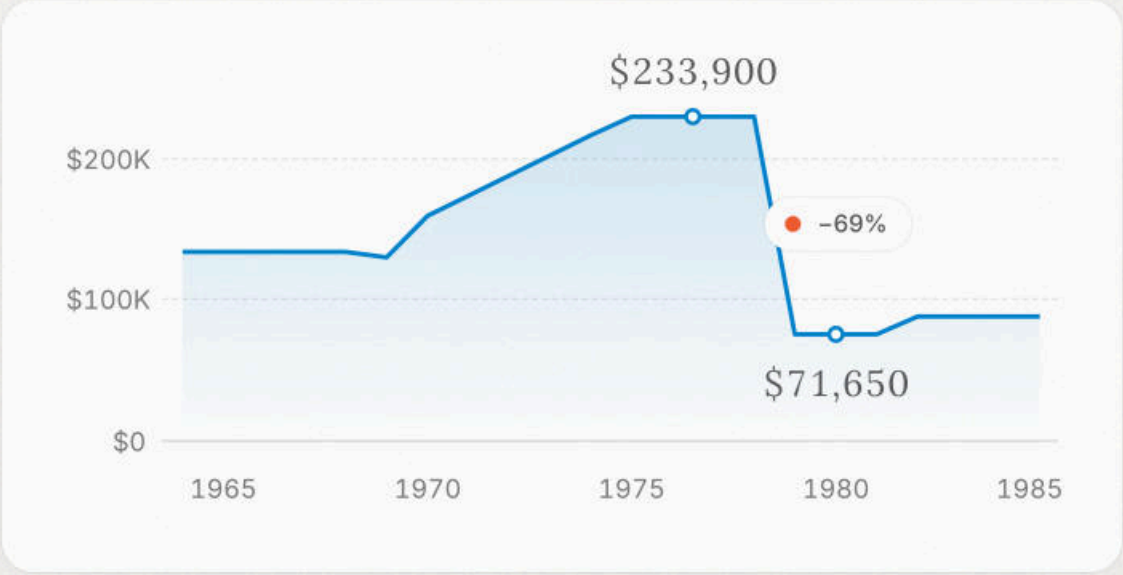
The IBM Precedent

In 1979, IBM cut its own pricing on its newest mainframes due to mounting competition and an “overly healthy” secondary market. It wanted companies to buy new IBM instead of secondhand IBM.

Its new mainframe delivered four times the hardware performance per dollar of the machines it replaced, and IBM's published entry price, the price of the entry level mainframe from IBM, fell from $233,900 to $71,650, a 69% reset in a single announcement. The used market repriced immediately. Losses ran into the billions in today's dollars.

EXHIBIT 5

IBM cut entry prices 69%



Sources: Datapro reports; IBM announcement archives. Public entry-configuration purchase prices, nominal USD.

In 1979, IBM held roughly 70% of the computer market, with margins wide enough to fund a price cut whenever competition demanded one. In 2026 NVIDIA ships roughly 90% of datacenter GPUs, at above 70% gross margin. One IBM decision repriced every used mainframe on earth. NVIDIA could decide to do the same if pressured.

EXHIBIT 6

IBM's price cuts repriced the installed base overnight

1974



Many U.S. companies were leasing IBM System/370s.

JULY 1979



IBM released the more powerful 4300, 30% cheaper.

SEPTEMBER 1979



The Washington Post

Leasing Firm Fights for Life

COMPUTER MIDDLEMEN STRUGGLING

By John F. Berry · September 29, 1979

If ever there was a classic American success story, William T. Evans Jr. was it. He joined the Navy in 1965 and

Leases canceled; $2B in losses (inflation-adjusted).

Sources: The Washington Post, IBM press photos.

III. The Residual Bands

We publish GPU residuals as three bands rather than a single curve.

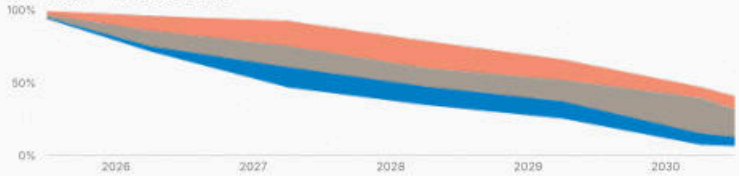
- **The conservative band** covers the safe scenario.
- **The moderate band** books aggressive but potentially defensible residuals.
- **The dangerous band** holds only when everything goes right.

EXHIBIT 7

Our outlook on GPU residuals as of June 2026

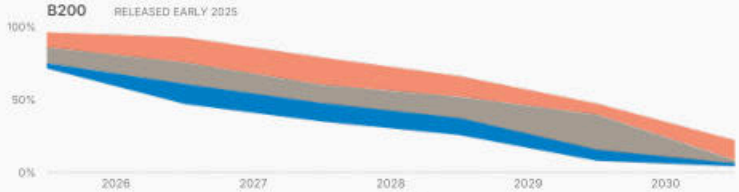
CONSERVATIVE MODERATE DANGEROUS

B300 RELEASED LATE 2025



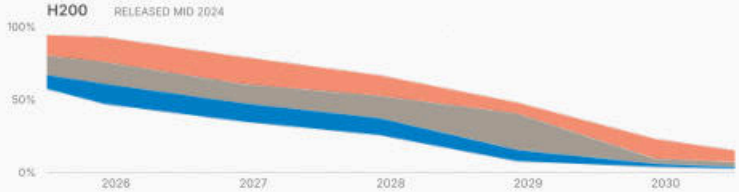
YEAR	CONS.	MOD.	DANGER
2026	78–81	81–89	>89
2027	52–64	64–78	>78
2028	37–50	50–64	>64
2029	28–39	39–54	>54
2030	12–21	21–43	>43

B200 RELEASED EARLY 2025



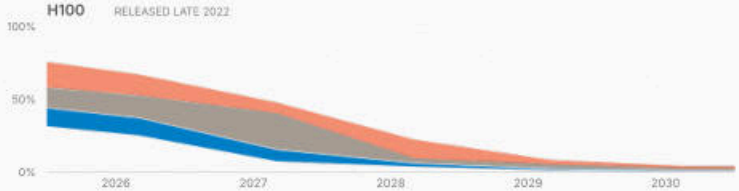
YEAR	CONS.	MOD.	DANGER
2026	59–67	67–80	>80
2027	40–54	54–68	>68
2028	30–42	42–56	>56
2029	16–26	26–46	>46
2030	6–11	11–24	>24

H200 RELEASED MID 2024



YEAR	CONS.	MOD.	DANGER
2026	45–59	59–74	>74
2027	34–46	46–59	>59
2028	24–35	35–51	>51
2029	7–15	15–37	>37
2030	4–6	6–9	>9

H100 RELEASED LATE 2022



YEAR	CONS.	MOD.	DANGER
2026	27–39	39–53	>53
2027	11–19	19–42	>42
2028	5–7	7–14	>14
2029	2–3	3–6	>6
2030	1	1–3	>3

Source: The bands are built from a subset of American Compute's proprietary dataset. The full dataset includes 76,775 completed secondary-market transactions, covering 622,098 units across 679 components, collected continuously since 2023. The GPU bands draw on datacenter training GPUs, Pascal through Hopper; consumer cards are excluded. We use arm's-length sales only: no asking prices, no rental rates, no appraisal indices. The residual is measured against original MSRP, after outliers are trimmed. We use these sales to project out distributions from the GPU release date, then adjust them further for specific technology considerations such as memory, power efficiency, and more.

IV. Modeling Five Adverse Events

We rank five adverse events: NVIDIA margin compression (high), AI bubble pop (high), bullwhip over-ordering (medium), hyperscaler refresh waves (medium), and technology obsolescence (low). Two of them can break the moderate band, and potentially even the conservative band.

For each event, we name the mechanism that impacts used prices and the leading indicators we watch.

1. NVIDIA margin compression (HIGH): NVIDIA is the most valuable public company in the world today, and its gross margin runs above 70%. Customers are recognizing the dependency on NVIDIA as a risk and funding billions into alternatives. That margin, combined with mounting pressure, sets up NVIDIA to take the same steps that IBM did in 1979: drop prices to retain market share.

EXHIBIT 8

The accelerator competition is arriving from every direction

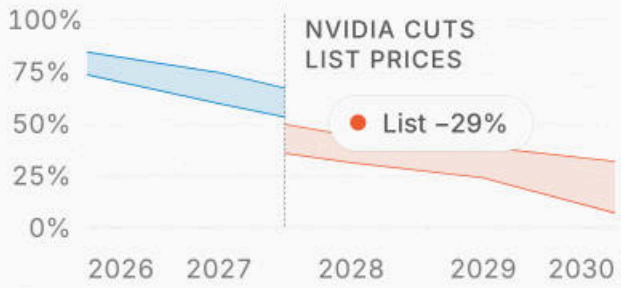
VENDOR & ACCELERATOR	LATEST	CAPABILITIES
Google Trillium · Ironwood (TPU v7)	2025	Ironwood matches Blackwell per chip ar scales to 9,216-chip pods
AWS Trainium2 · Trainium3	2026	Trainium3 now generally available; AWS claims ~50% lower cost than comparab NVIDIA
AMD MI355X · MI400 (MI455X)	2026	MI400 series and Helios rack land in 20 432 GB HBM4, OpenAI and Meta comm
Microsoft Maia 200	2026	Maia 200 live on Azure; FP4/FP8 inferer serving GPT-5.2, ~30% better perf per
Huawei Ascend 910C · 950	2026	Ascend 950 now sits between H100 anc H200; default training silicon inside Chi
Intel Gaudi3 · Crescent Island	2026	Gaudi line winding down; Crescent Islar inference GPU (160 GB) in trials H2 202
Cerebras WSE-3	2024	Fastest published inference tokens-per-second; serves Mistral and Perplexity

VENDOR & ACCELERATOR	LATEST	CAPABILITIES
SambaNova SN40L	2023	High-token-rate inference cloud on Llan class models

Based on media reports as of June 2026.

EXHIBIT 9

If NVIDIA cut gross margins from 75% to 65%, residuals could slip 29%



% MSRP	
2026	74–85
2027	60–75
2028	31–44
2029	24–39
2030	7–32

Illustrative scenario, not a forecast. An estimated −29% residual depression would happen if NVIDIA decided to cut prices to roughly a 10% premium over AMD's MI-series, taking NVIDIA's gross margin from 75% to roughly 65%.

2. Technology obsolescence (LOW): Prior-generation GPUs keep producing economic value through a use-case ladder: training, inference, fine-tuning, edge, etc. At the same time, new generations have power, cooling, and software friction that slow down adoption. New technologies like photonics could leapfrog current silicon, but they are unlikely to be developed, scaled, and adopted quickly enough within the next few years to depress residuals. New generations always arrive late. Announcing a chip is easy, but ramping a new design's supply chain is hard.

EXHIBIT 10

The GPU use case ladder

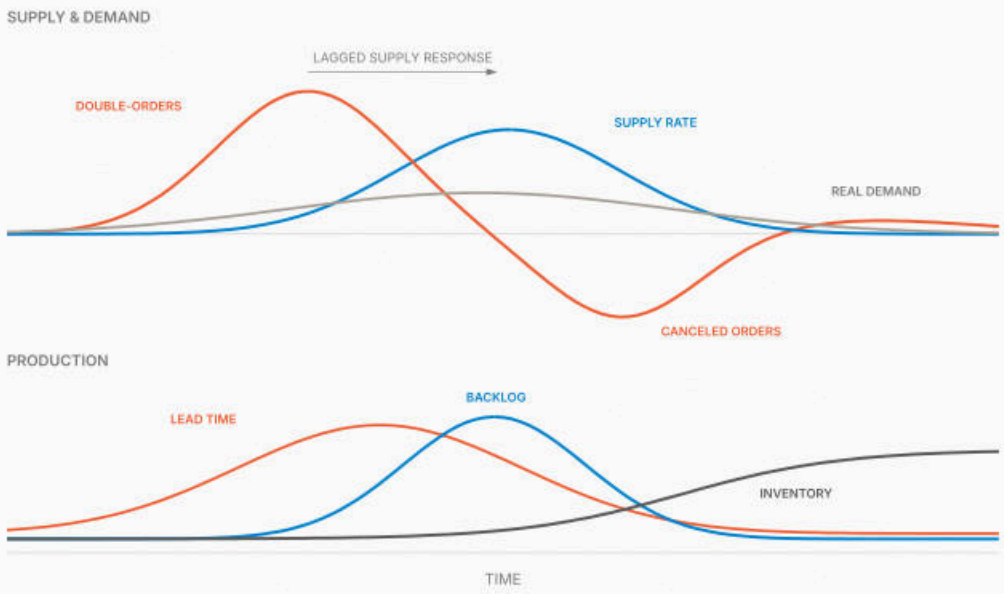


3. Bullwhip over-ordering (MEDIUM): The industry ordered through a shortage, and likely over-ordered. Double-orders placed to secure allocation sit in order books and channel inventory, and they surface as supply at exactly the moment demand normalizes (Exhibit 11).

> *“When suppliers are unable to fill orders, delivery delays increase and customers receive less than they desire. Customers often respond by seeking larger safety stocks (hoarding) and by ordering more than they need to meet demand (phantom ordering).” (Sterman & Dogan)*

EXHIBIT 11

Initial shortages in supply can lead to extreme overordering



How to read: both panels share the same timeline and are read together. The top panel reflects compute **orders** against **real demand**: when supply cannot keep up, buyers pile on extra orders to secure allocation. Orders spike far above real demand, but once supply catches up, the excess orders are canceled and orders crash below real demand. The bottom panel tracks the **physical pipeline** those extra orders create: lead times stretch first, the order backlog peaks next, and finished inventory arrives last.

Illustrative; redrawn after Sterman & Dogan, "Behavioral causes of phantom ordering in supply chains."

4. Hyperscaler refresh (MEDIUM): AWS, Google, and Microsoft hoard the largest GPU fleets ever assembled, and they refresh on their own cadence. A single refresh event could release billions of dollars of used hardware into the secondary market at once, depressing prices. A default event at the same time as a hyperscaler refresh would lead to a lower residual than expected. We rank it medium because it would at most cause a temporary drawdown, but nobody can predict when it could happen.

5. AI demand pop (HIGH): Demand temporarily breaks. A 2026 enterprise survey from Cast AI put GPU utilization at 5% across thousands of enterprises. This signals that enterprises haven't fully adopted their own AI, and are still relying on AI assistants from frontier labs like Claude, ChatGPT, and similar subscriptions. If compute pauses because of the lack of utilization, the residuals for unwanted equipment could be very low.

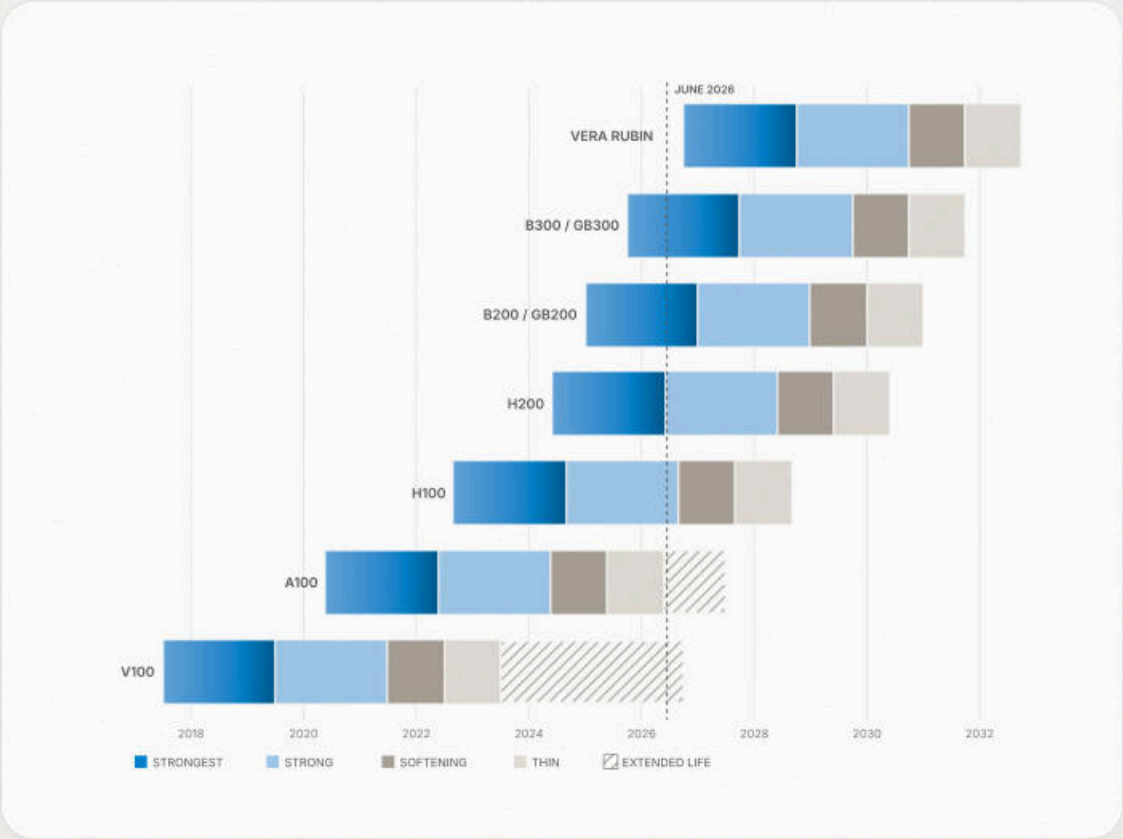
V. The Market Today

The GPU shortage is over. Lead times are back to normal for new orders. What is scarce in mid-2026 is everything *around* the GPU: grid interconnection, datacenter shells, memory, and storage. This has led to a temporary increase in residuals of older generation equipment like the H100s. These GPUs can still be run on older data centers without liquid cooling or higher power. Because of this, it's easier to buy new B200s than used H100s in 2026.

Most GPU financing is done over 36 months, with some deal terms as short as 24 months and others up to 60 months. The asset's economic life should exceed any of these periods, as long as the equipment was first deployed early into the release cycle. The graph below shows NVIDIA speeding up its release cycle and potentially shortening asset life.

EXHIBIT 12

A GPU generation can survive in the market for eight years

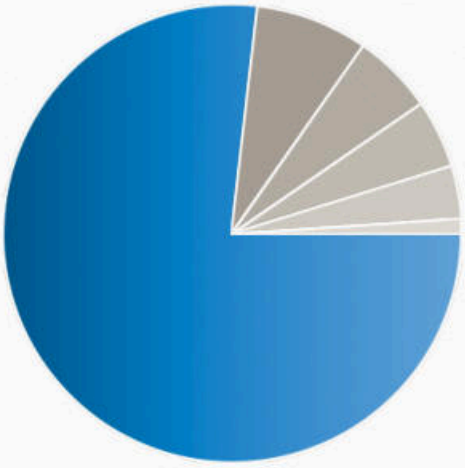


Source: NVIDIA announcement dates; demand phases are American Compute's house view keyed to technology age (0–2, 2–4, 4–5, 5–6 years), consistent with Exhibit 7. The hatched extended-life segments on V100 and A100 reflect demand running past expected life on current AI demand.

The collateral is the whole server around the chip. GPUs are roughly three-quarters of a flagship system's bill of materials. Memory retains surprisingly well; CPUs, storage, and networking fall into the teens or lower. Underwriting the GPU residual comes first, as it is both the bulk of the bill and the bulk of the residual.

EXHIBIT 13

GPUs are ~75% of the bill of materials



COMPONENT	SHARE	Y5 RESIDUAL
GPU	76.8%	8–15%
SSD	7.9%	14–17%
NIC	5.7%	13–18%
IB / Switches	4.8%	7–12%
CPU	3.8%	5–7%
RAM	1.0%	16–20%

Source: American Compute reference BOM weights.

The Power and Cooling Wall

GPU systems are becoming more efficient. SemiAnalysis's InferenceX benchmarks put a B200 at roughly three times the tokens per watt of an H200. But each GPU generation requires far more power and cooling than the last (Exhibit 14). An H100 rack runs near 40 kilowatts, a GB200 NVL72 draws roughly 120, and a 2027 Rubin rack is designed for as much as 600.

Power density decides which datacenters can host a system. Past roughly 40 kilowatts per rack, air cooling stops working and direct-to-chip liquid cooling becomes mandatory. That confines the newest systems to a small set of purpose-built, liquid-cooled datacenters with the grid capacity, reinforced floors and plumbing. The far larger installed base of

conventional, air-cooled datacenters would need to be retrofitted to host the latest systems.

So even as faster silicon arrives, the power and cooling wall keeps the prior generation deployable for years, supporting the long economic life behind our residual bands.

Only a limited number of sites will be built on time to host the newest generation of GPUs, and only a handful can adopt Rubins. Very few operators will accept running new silicon at half power or under thermal throttling.

Because of the power and cooling friction, many enterprises will adapt their workloads to continue running on older hardware rather than chasing the newest generation. The software layer of the stack shows the shift. A large body of open-source work is focused on optimizing models to run efficiently on older, and even consumer-grade, GPUs.

Liquid-Cooled Capacity

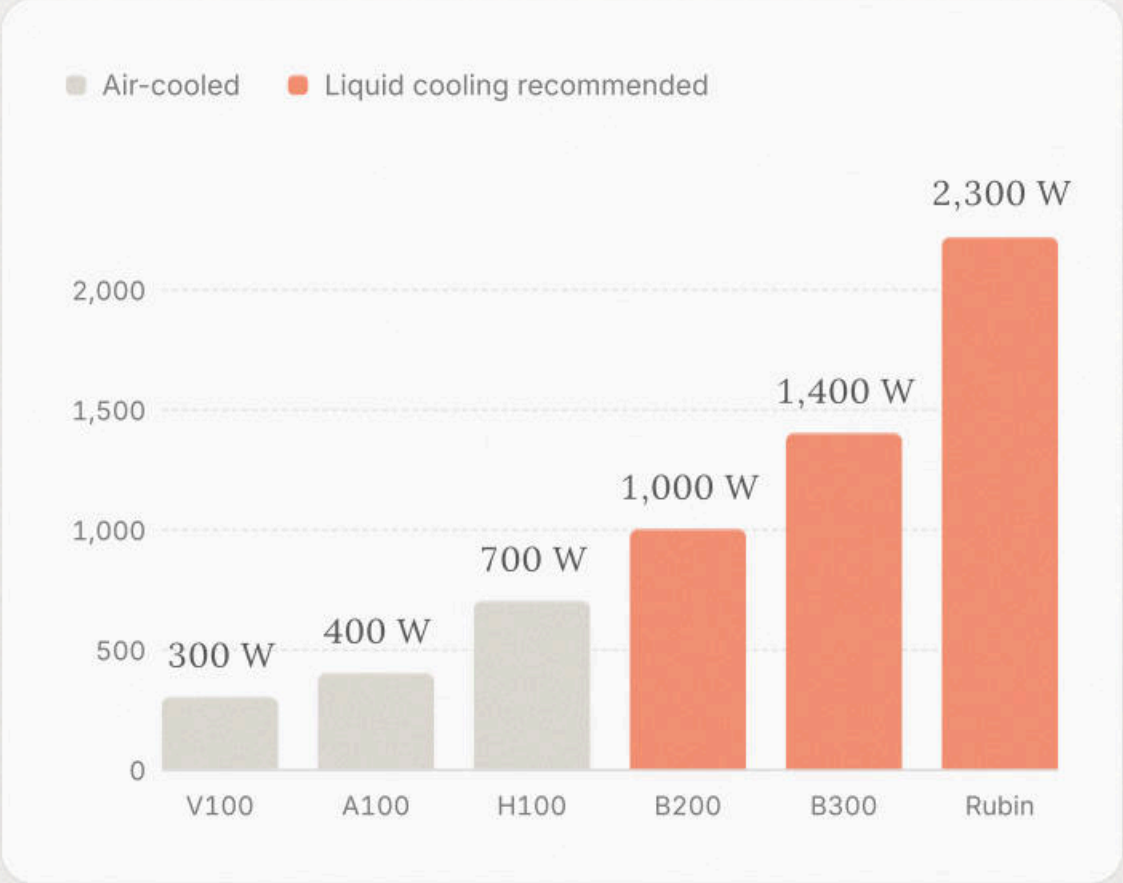
Outside of the five ranked events, one upcoming systemic shock will be the liquid-cooled, high-density facility buildout. Right now, the shortage of qualified space is holding up air-cooled B200s and H100/H200s. New GPUs are not being bought, because the datacenter facilities cannot support them. Once more datacenter capacity is built out, Hopper residuals will drop much quicker, because Blackwell runs inference far more efficiently per watt.

The current GPU residuals therefore depend less on technology obsolescence than on datacenter project delays. Projects are running behind due to grid, transformer, turbine, and switchgear lead times, though that might change within the next 12 months.

Hoppers are one of the last widely adopted air-cooled, lower-power GPUs. The residual curve for Hopper will therefore not translate to Blackwells, as Blackwells will not have a datacenter shortage propping up their residuals the way Hoppers do today.

EXHIBIT 14

Power per GPU keeps climbing, pushing the newest systems toward liquid cooling



Power and cooling requirements create adoption friction for future GPU generations, and most data centers don't support liquid cooling yet. Most B200 deployments are still air-cooled. Sources: NVIDIA product specifications; SemiAnalysis.

VI. Modeling Pitfalls

1. Indices and rental rates: Rental indices are not good indicators of resale value; backtesting shows only a weak correlation between the two. Only an arm's-length secondary transaction prices the asset itself. As Exhibit 15 shows, secondary transactions refuse to follow the rental line.

2. DCF as the basis: An extension of the rental-rate approach is a discounted cash flow model of what the asset is worth. But DCF models lean heavily on rental rates, which are volatile: market regimes shift and pricing competition grinds on, and that volatility bleeds straight into the analysis. The result predicts residuals poorly.

3. Overthinking physical wear: Silicon does not wear out the way cars and other hard assets do, and both the primary and secondary markets are protected by warranties and third-party maintenance contracts. Most equipment resellers run diagnostics, refurbish, and back second-hand gear with 30-day performance guarantees. Second-hand buyers don't care about a single GPU's thermal history, so neither should primary buyers.

4. Curves instead of distributions: A single curve is a single guess, and it hides the only thing underwriting cares about: dispersion. Our data shows GPUs selling at very different prices in the same period, driven by market conditions, reseller networks, and other variables a curve cannot carry.

5. Consumer vs. enterprise: Review data sources for consumer bias. A single PCIe card on a retail listing site is valuable to a consumer but not to an enterprise. Consumer comps carry gamer demand, hobbyist condition tiers, and retail spreads that have nothing to do with how enterprise systems clear.

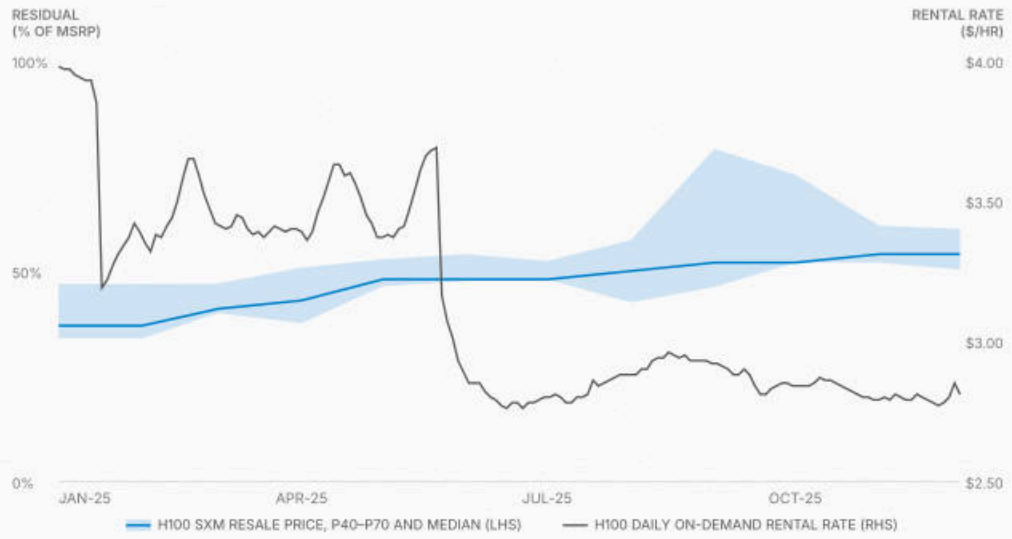
6. Depreciation schedule as residuals: Depreciation schedules are tax and reporting conventions, and they fail to predict residuals. A 3-year straight line writes equipment to zero while the market still pays for it; a 6-year schedule can carry book value the market floor no longer supports.

7. Asset age instead of technology age: Residuals should be booked on technology age, measured from the day a GPU generation first reaches buyers, when the first units ship and can be deployed, rather than from the day a particular server was purchased or from the day NVIDIA announced the chip (announcements often precede shipments by many months). For example, an H200 bought brand-new today is already more than two years into its technology life, because that generation first shipped in 2024, even though the server itself has never been racked.

8. Overlooking delays: Most smaller operators do not get allocation for flagship GPUs at launch, and projects get delayed for all kinds of logistical reasons; a Blackwell project signed today might only be racked and installed in six months. In our deal flow, most financing on smaller clusters closes on GPUs that are already a year past release.

EXHIBIT 15

H100 rental indices and secondary sales in 2025 tell different stories



Rentals fell as compute came online, but median H100 residuals rose because many datacenters could not support the newer Blackwells yet. Sources: SemiAnalysis GPU Pricing Index; American Compute proprietary resale data as of June 2026.

VII. What Holds Value

Certain qualities can help equipment maintain higher residuals. In general, anything that makes the equipment appeal to a larger pool of buyers will help.

1. OEM brands. Despite containing the same NVIDIA GPU, a server from Dell and one from a Taiwanese OEM like Gigabyte can have very different resale values. In general, servers from major OEMs like Dell resell for more than those from smaller builders. For instance, our data shows Dell-built servers reselling about 18% above less-known OEMs.

2. Whole vs. parted-out. Second-hand buyers have little interest in piecing together several orders to build a cluster. A whole cluster that has already been tested, configured, and proven to work will sell better than the same nodes sold individually.

3. Vanilla vs. exotic. Standard configurations clear faster and higher than bespoke ones. A vanilla configuration taps into a wider buyer pool and is more likely to find a buyer willing to pay a good price. A highly specific cluster with unusual pooling, networking, or heterogeneous silicon has a much smaller pool.

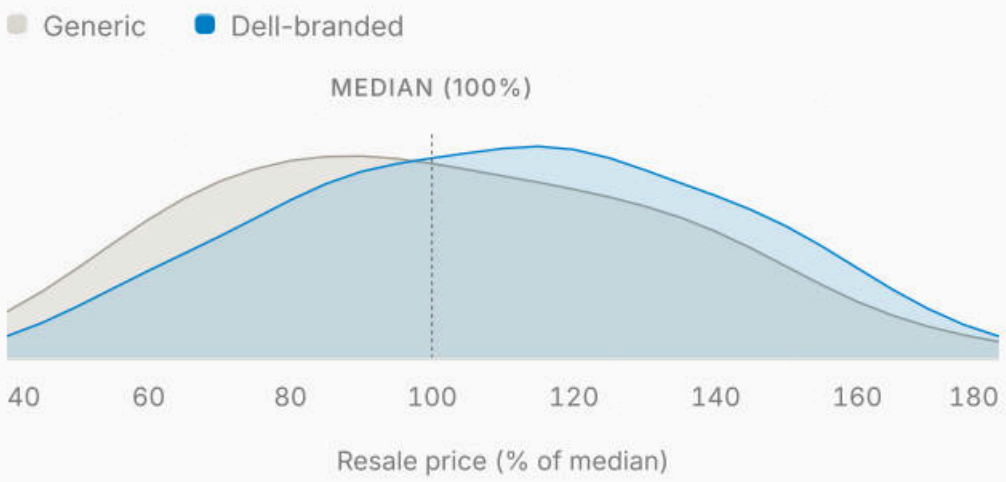
4. Geography. GPUs deployed in the U.S. resell better than identical units deployed abroad.

5. Documentation. Service records and original configuration documents help. Buyers value more information, though they ultimately lean on the reseller's guarantee that everything works and on the third-party maintenance contract.

6. Timing. Give the resale process room to breathe, ideally 30 to 60 days. A rushed sale fetches a lower residual. It usually takes about 30 days, but can take longer to find the right buyer after preparing remarketing materials, running diagnostics, and refurbishing the equipment.

EXHIBIT 16

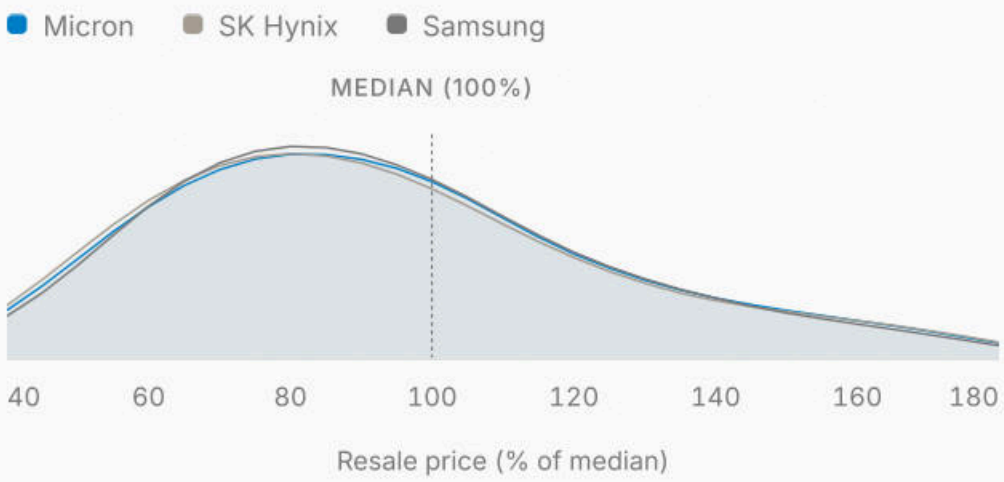
The same equipment resells higher under a name-brand OEM



Source: American Compute secondary-transaction data.

EXHIBIT 17

Some types of equipment are commodities with little resale difference



Source: American Compute secondary-transaction data.

VIII. Conclusion

GPU residuals will not always follow a curve. The market can change suddenly, with booms and busts that catch the unprepared. And not everyone quoting an overly high residual is being honest with themselves.

We leave the reader with three recommendations:

1. Model in bands, not a curve. Pick a defensible band instead of a point. Compare your financial performance at the top and bottom of the range.

2. Pick the band that fits your risk appetite. Lessors should sit at the most conservative end, lenders can be slightly more aggressive, and equity investors can be the most aggressive of all. The biggest danger is every party using the same band. Compare T-bills

and stocks: each suits a different investor's risk-return profile. No single band is right for an operator, a lender, and an equity investor on the same project.

3. Add confidence where it matters. Residual value insurance solutions, such as those Matcha Specialty Insurance Services places under the American Compute program, and a few other market solutions can backstop your underwritten residual. Residual value guarantees from IT asset disposal firms can also help, provided you are comfortable with the counterparty risk. And read the fine print: some products exclude the very event you are worried about, for example by lapsing coverage if an overall GPU index falls past a set threshold.

A condensed version of this report is available as the GPU Residual Value 2026 one-pager (PDF), free to download and share.

A limited sample of the dataset behind this paper is available to download from this page.

If you would like to discuss structuring residual value insurance solutions for a transaction, need a valuation, or want a second opinion on a residual assumption, contact us at deals@amcompute.com.

ABOUT THE AUTHORS

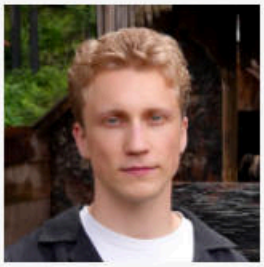


Bernie Margulies
CEO, Chief Underwriting Officer

Bernie leads American Compute's program business end to end: underwriting, reinsurance capacity, deal structuring, and the carrier and client relationships behind them. Prior to founding American Compute, he was chief of staff at a $2.6B insurance carrier and spent four years at Wavemaker Partners in Singapore investing in technology companies from seed through Series B. He also operates an approved on-shore U.S. captive manager.

Bernie holds the Chartered Property Casualty Underwriter (CPCU) designation, the most prestigious credential in property and casualty insurance. He graduated from

Ateneo de Manila University with a BS in Management Engineering.



Illia Kasian
Chief Technology Officer

Illia leads engineering at American Compute and built the data platform behind the firm's underwriting: the collection and processing of the secondary-market transaction data, the pipelines that keep it current, and, together with the firm's actuaries, the rate-making models built on it. He also leads the firm's hardware teardowns and analysis of the ML architectures that drive GPU demand.

Previously, he was a founding engineer at a multi-billion-dollar insurance carrier. Before that, Illia built large-scale fraud detection and content moderation models at Quarks Tech and ML perception systems for defense drones in support of Ukraine. He graduated from Taras Shevchenko National University of Kyiv with a bachelor's degree in Applied Mathematics.

American Compute

American Compute publishes GPU residual value data and research. Residual value insurance solutions under the American Compute program, for transactions up to $500M, are placed by Matcha Specialty Insurance Services (Matcha Specialty LLC), a licensed surplus lines broker, CA License No. 6018236, with insurers supported by A-rated reinsurers.

American Compute, Inc. is not an insurer and is not a licensed insurance producer.

Where needed, we also consult and provide appraisals built on our proprietary datasets.

Contact us at **deals@amcompute.com**.

DISCLOSURE

This document is provided for informational and educational purposes only and does not constitute an offer of insurance, a quotation, or investment, tax, accounting, or legal advice; it should not be relied upon in making any investment, financing, or underwriting decision. The views and estimates expressed are those of the authors and American Compute Research as of June 2026, are based on sources believed to be reliable but not guaranteed, and are subject to change without notice. Residual bands, scenarios, and other forward-looking statements are estimates, involve known and unknown risks and uncertainties, and may differ materially from actual outcomes; past secondary-market performance does not guarantee future results. Transaction data referenced herein is American Compute proprietary data unless otherwise noted; third-party data is attributed to its source and remains the property of its owners. Insurance products are available only where licensed and approved, and are subject to the terms, conditions, and exclusions of the policies as issued. No part of this document may be reproduced without the prior written consent of American Compute, Inc.