BARKR

GPU Resale vs Rental: Robust, Divergent Markets

Barkr Research Series

April 2026 Market Analysis of:

GPU Resale vs. Rental Dynamics and Asset Collateralization

Executive Summary

Barkr solves a specific problem for asset-backed lenders who face two primary risks, the borrower and the asset. Underwriters are talented at tackling borrower risk. Asset value risk is still largely an enigma, especially with hard to price assets. At Barkr we provide loan collateral valuations with a contractual warranty backed by Munich Re. We cover a wide range of assets, from art to private jets and Graphics Processing Units (GPUs). We deliver clear valuations throughout the course of the loan, providing lenders with a high level of confidence should default loan collateral need to be liquidated.

> *"When assets are complex to value, like GPUs, underwriting becomes challenging. Barkr addresses this challenge with real-time, AI-powered valuations, backed by Munich Re's aiSure™. At Munich Re, we support Barkr's commitment to delivering insured and trustworthy AI empowered price transparency to the broader asset-backed lending market."* ~ Michael von Gablenz, Head of Insure AI at Munich Re & HSB.

This broad pricing capability is powered by our domain-specific LLM, a foundational model that helps inform accurate pricing across diverse asset types. One thing we've noticed is that almost ALL loan collateral suffers from often inaccurate valuations and often no recourse when those valuations are wrong. This can be especially damaging when dealing with high value assets or large loans. In addition, these markets often operate in an opaque fashion - art, private jets and GPUs are wildly different assets but they all share limited transparency in their resale markets. At Barkr we are in a unique position to provide insights using data others may not have access to, thereby increasing market transparency.

With that in mind, we've set out to publish a series of reports that provide a layer of transparency that we hope will help clients, lenders and curious minds understand more about these market trends. Starting with GPUs.

It's hard to miss that the global market for high-performance GPUs has undergone a structural shift over the past few years.

> NVIDIA adds key context - *"We are at the beginning of the largest infrastructure buildout in human history. AI factories are being built around the world to produce intelligence, powered by accelerated computing. Long-term demand for GPU infrastructure continues to grow, and NVIDIA's approach to extreme co-design at gigawatt scale and software improvements extend the performance and value of deployed systems over time,"* ~ Nico Caprez, VP AI Infrastructure Growth at NVIDIA.

This new industry is still emerging, and the market is evolving fast. Whereas 2023/24 was defined by scarcity, pushing resale prices higher, by 2025/26 the market has matured and, interestingly, the resale and rental markets have largely bifurcated. We believe the resale market acts as a more stable indicator of long-term asset value than the more volatile rental market. Some propose that the rental market is a leading indicator of where the asset resale market will go, we've yet to see this show up in our data.

This report investigates the thesis that these two markets, while linked by underlying compute demand, move on different trajectories. Furthermore, it examines how the "compute crunch" has effectively immunized the data center sector from shutdowns, as demand continues to outstrip supply.

I. Resale vs. Rental

The Barkr Global GPU Price Index

The Barkr Global GPU Price Index is a quality-adjusted monthly index of secondary market GPU resale prices, normalized to 100 in July 2024. As of March 2026, the index stands at 118.2, its highest recorded value. The index is constructed using a hedonic regression with GPU-model fixed effects, isolating the time trend in prices while holding product quality constant. Data are drawn from public secondary market listings and Barkr's proprietary transaction dataset.

Full period: July 2024 to March 2026

The index opens at 100.0 and closes at 118.2 across 21 months, with three distinct phases. From July to December 2024, the index traded in a narrow band around baseline as H100 supply normalized following the 2023 scarcity peak. From January to June 2025, it rose steadily to 113.9, driven by accelerating inference demand and the memory premium commanded by H200-class hardware. From July 2025 through March 2026, the index remained elevated before a sharp move to 118.2 in March, coinciding with Blackwell-class hardware entering secondary channels and compressing available H100 inventory.

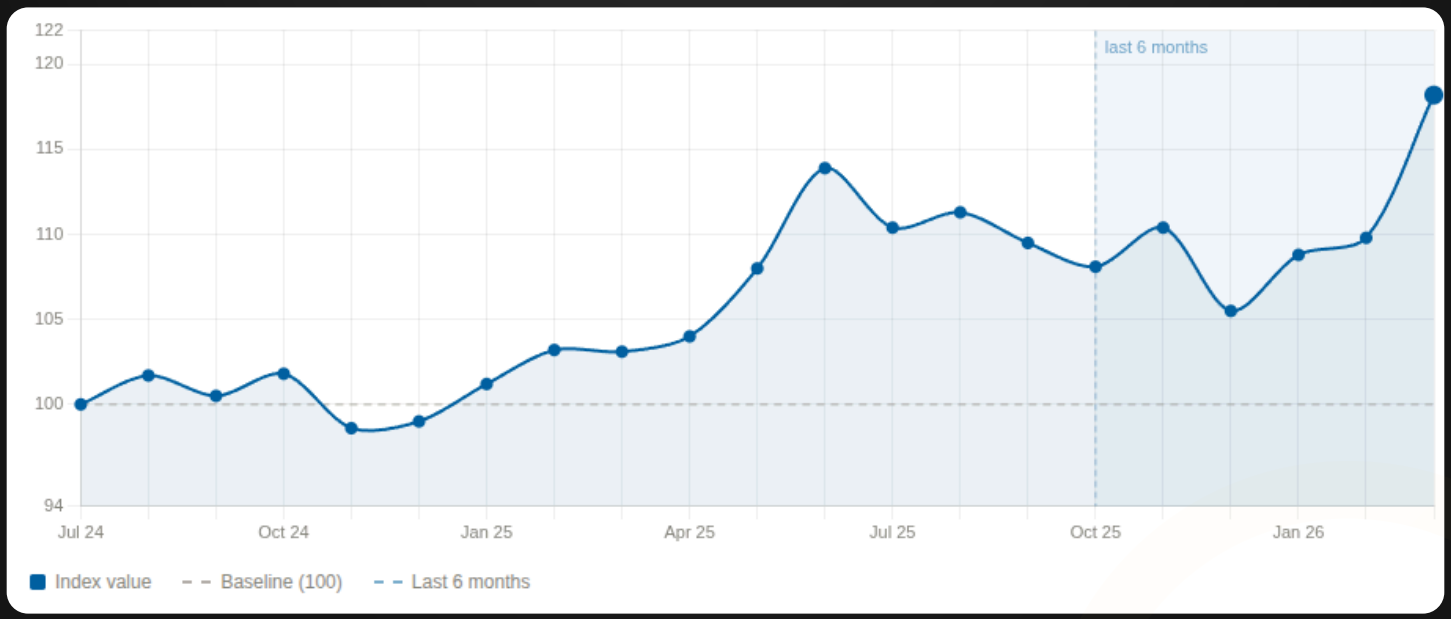

Figure 1: Barkr global GPU price index

Secondary resale market · July 2024 – March 2026 · Base = July 2024 (100.0) · Quality-adjusted hedonic index

Recent period: October 2025 to March 2026.

The index averaged approximately 108 through Q4 2025, with a trough of 105.5 in December before recovering through January and February 2026. March 2026 delivered a 7.65% price increase, the strongest single-month movement in the dataset. The contrast with rental market dynamics over the same period is material. Hourly H100 rental rates declined 60 to 70 percent from their 2024 peak while the Barkr index reached an all-time high in March 2026.

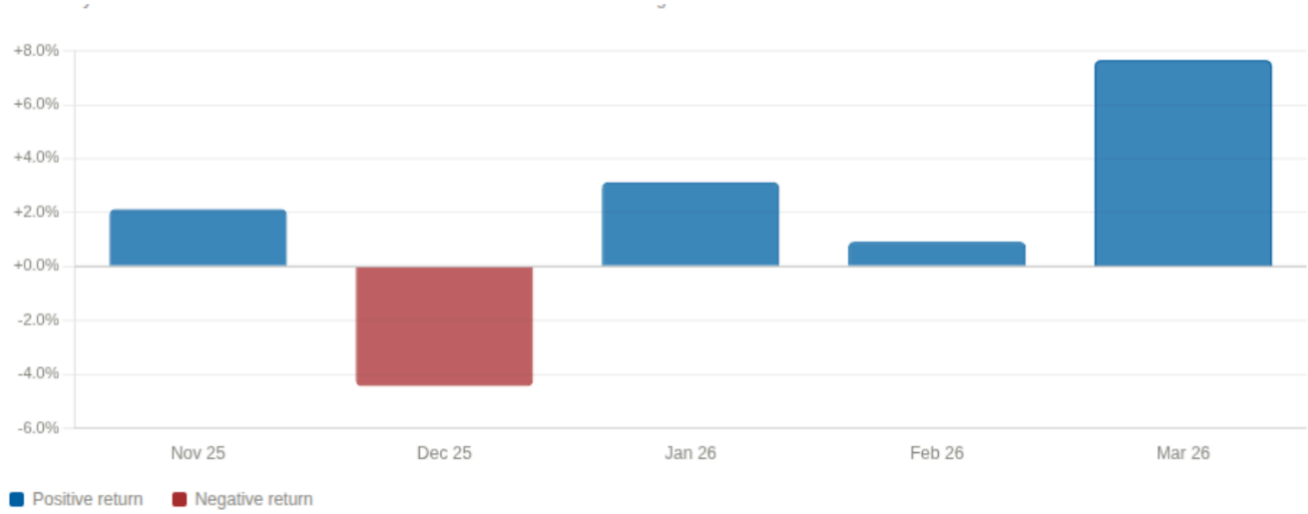

Figure 2: Barkr global GPU price index, monthly returns

Secondary resale market · October 2025 – March 2026 · Month-on-month % change

Rental Market Dynamics

The central finding of this research is that the GPU resale market is fundamentally more stable and less prone to volatility than the rental market.

H100 rental rates peaked near $8.00 per hour in early 2024, reflecting acute supply scarcity as AI laboratories and hyperscalers competed for limited capacity. By Q1 2026, on-demand rates across specialist cloud providers have compressed significantly. Hyperscalers retain a material pricing premium: Azure prices single-GPU H100 instances at $6.98 per hour in East US (October 2025), while specialist providers compete well below $3.00 per hour.

The compression is structural rather than cyclical. The specialist cloud segment, commonly referred to as neoclouds, expanded rapidly through 2025. ABI Research estimates GPU-as-a-Service revenue reached approximately $42 billion in 2025,

with projections exceeding $250 billion by 2030 [12]. Aggressive hyperscaler responses, including a reported 44% H100 price reduction by AWS in June 2025, accelerated the transition toward commodity pricing dynamics: elevated elasticity, spot market volatility, and sustained margin pressure on smaller providers. Rental prices are inherently more elastic and subject to near-term capacity fluctuations, making independent pricing benchmarks such as Ornn's rental index increasingly important reference points for the market.

The rental market prices marginal compute availability: the cost of accessing one GPU-hour at a given moment. This is categorically different from the liquidation value of the underlying physical hardware. The analogy to oil markets is instructive. Crude spot prices move daily on supply signals, geopolitical developments, and inventory data. The value of the drilling rig producing that crude operates on an entirely different cycle, anchored to reserve estimates, replacement cost, and the long-run economics of the field. A lender financing a rig against crude spot prices as a collateral signal is measuring the wrong variable. The same misalignment applies to GPU-backed lending.

Resale Market Dynamics

H100 units trade below their late-2023 secondary market peak of $40,000-plus, but the depreciation trajectory has been gradual and the Barkr index currently stands at a record high. The resale price of an H100 80GB has remained resilient, with units currently trading between $25,000 and $30,000, a considerably slower and more predictable depreciation curve than rental yield compression would imply. This pattern characterizes an asset class with a steep initial discount from Manufacturer Suggested Retail Price (MSRP), followed by a sustained valuation plateau, punctuated by generational transition events.

Secondary demand durability is evident in publicly available data. CoreWeave's Q3 2025 SEC filings show H100 units from 2022-vintage contracts rebooked immediately upon expiration. Its revenue backlog stood at $55.6 billion as of September 30, 2025, growing to $66.8 billion by year-end, driven by contracts with Meta, OpenAI, and others [1][2]. CEO Michael Intrator characterized the business as supply-constrained on the Q3 earnings call. The binding constraint is on the supply side.

Hyperscalers have signaled a consistent view through their capital accounting decisions. Amazon, Microsoft, and Google extended GPU and server depreciation schedules to six years; Meta adopted five and a half years. As Microsoft CEO Satya Nadella stated publicly: "I didn't want to go get stuck with four or five years of depreciation on one generation" [16]. Amazon's Q3 2025 10-Q discloses a partial reversal, reducing useful

life assumptions for a subset of servers to five years, resulting in a $677 million reduction in net income over the first nine months of 2025 [15]. The implications of this ongoing debate for lender underwriting models are addressed in Section IV.

The Divergence

The two markets have decoupled because they price fundamentally different things.

The rental market prices transactional access to compute capacity. Rates respond rapidly to changes in available supply, new provider entry, and near-term demand fluctuations. The resale market prices physical asset control: the premium enterprises place on hardware ownership for reasons including data sovereignty, cost predictability, and independence from third-party availability constraints. As the population of hardware owners grows, so does the aggregate compute demand they generate. Ownership decisions operate on long capital allocation cycles, largely insulated from short-term rental rate movements.

The container shipping industry offers a useful structural parallel. Freight rates on a major transoceanic route can fall by half within a single season when new tonnage enters service. The resale value of the vessel operating that route does not move in proportion. It is anchored to steel replacement cost, yard availability, charter market depth, and the long-run economics of global trade. The rate and the asset respond to the same underlying demand, across entirely different timescales and with different structural sensitivities. GPU resale and rental markets exhibit the same pattern.

This ownership premium establishes a price floor for resale assets that is structurally independent of hourly cloud pricing. Across the asset classes where Barkr observes both a rental and a resale market for the same underlying asset, the rental-resale divergence in GPUs is among the most pronounced we have documented.

II. Component Analysis: From Hopper to Blackwell

The GPU market is currently transitioning across three distinct hardware tiers, each with differentiated secondary market characteristics and collateral implications. NVIDIA's bench of GPUs provides an excellent example and reflects the majority of the sector.

NVIDIA H100 and H200 (Hopper Architecture)

The H100 remains the liquidity benchmark for the GPU secondary market. Every major language model deployed at scale between 2023 and 2025 was trained on H100 infrastructure, per NVIDIA's official MLPerf documentation [3][4], establishing a broad and well-understood secondary buyer base.

The H200, which retains the Hopper architecture while upgrading memory to 141GB HBM3e from the H100's 80GB, has emerged as the preferred asset for large-scale inference deployments. Memory capacity has become the primary technical bottleneck for production LLM workloads: greater addressable memory allows a GPU to serve larger model contexts without partitioning across multiple devices, reducing both latency and operational complexity. The H200's memory premium translates directly into a durable secondary market valuation differential relative to the H100.

Secondary market pricing for H100-class hardware is age-stratified and condition-dependent. Based on Barkr's collateral assessment experience, the single most consequential variable in realized secondary prices is the operational context of the hardware, specifically whether it is installed within a power-secured, operational data center or sitting undeployed. This point is addressed in Section III.

NVIDIA B100 and B200 (Blackwell Architecture)

Blackwell-class GPUs entered secondary channels in meaningful volume in early 2026, primarily as cancelled-order inventory and short-term lease reassignments. Their arrival has repositioned H100s within a clearly defined mid-tier, analogous to the role A100s occupied following the H100's initial deployment at scale.

The intergenerational performance differential is documented in NVIDIA's official benchmark publications. Per NVIDIA's MLPerf Training results [4][5], the B200

delivers approximately 2.2x higher fine-tuning performance on Llama 2 70B and approximately 2x the pre-training throughput on GPT-3 175B versus the H100. MLPerf Inference v5.0 results [3] show the GB200 NVL72 system achieving up to 3.4x higher per-GPU performance on Llama 3.1 405B versus an H200 eight-GPU system. B200 inference performance remains partially software-constrained as frameworks including vLLM continue to mature for the Blackwell architecture; realized gains are expected to increase as the software ecosystem develops.

For lenders with multi-year facilities, NVIDIA's Rubin architecture is broadly anticipated in the 2026 to 2027 timeframe. That transition horizon warrants explicit modeling in any facility extending beyond 18 months.

A100 and Older Generations

A functioning secondary market persists for A100 hardware among academic institutions, regional cloud providers, and enterprises running inference on models that do not require frontier memory capacity. Units in this category continue to transact at $10,000 to $15,000, sufficient for fine-tuning smaller, specialized models. Tier 2 demand of this kind is likely to represent a considerably larger market segment than Tier 1 over time, as small and medium enterprises gain access to high-performance compute at price points that match their scale requirements. As successive generations enter the market, prior-generation hardware cascades into this Tier 2 layer, often with substantial useful life remaining.

The A100's trajectory illustrates the value cascade model: hardware that served frontier training use cases in 2021 and 2022 migrated to inference in 2023 and 2024, and now serves batch and lower-priority workloads. This dynamic closely mirrors the commercial aviation aftermarket. A widebody aircraft retired from a major carrier's long-haul network transitions to a regional operator, then a cargo carrier, then a wet-lease fleet, extracting economic value at each stage before ultimately being parted out. Depreciation occurs across a longer arc and through more value-generating stages than a straight-line accounting schedule implies.

Microsoft Azure's hardware retirement history provides relevant empirical grounding. Azure's V100-based VM series, launched in 2016 and 2017, was retired in September 2025, implying approximately seven and a half years of commercial service life [18]. For lenders, A100-class assets represent a lower per-unit value but a more liquid collateral tier: the buyer pool is broader, and forced liquidation timelines are shorter than for H100-class assets.

III. Validation of Demand and Data Center Uptime

The Shift from Component Scarcity to Megawatt Scarcity

Current market data reveals a structural paradox: chip production has scaled materially, yet effective GPU data center capacity remains constrained. The binding bottleneck has migrated. Between 2021 and 2024, the primary challenge was hardware lead times, with component delivery windows reaching 52 weeks at peak. From 2025 onward, the constraint has shifted from the server rack to the electrical substation.

GPU clusters require 50 to 120 kilowatts per rack. Large-scale AI facilities consume hundreds of megawatts, a power density that exceeds the design parameters of conventional hyperscale data centers. Electrical grid infrastructure cannot be upgraded on the same timeline as GPU procurement. Bain & Company's 2030 Global Data Center Forecast [6] projects global capacity demand reaching 163 gigawatts by 2030, double current installed capacity. Gartner [10] projects data center electricity consumption grew 16% in 2025 alone, with a further doubling expected by 2030. CBRE's North America Data Center Trends H1 2025 [7] recorded vacancy rates in primary US markets at a historic low of 1.6%, with large-scale deployment costs (10MW-plus) rising up to 19% year-on-year.

S&P Global's power analysis [8] projects the supply constraint will be most acute through 2028 and 2029. Lawrence Berkeley National Laboratory's interconnection queue data [9] shows median wait times of five to seven years in constrained markets. Primary markets including Northern Virginia, Dublin, Singapore, and Frankfurt face multi-year delays for new grid connections.

Zero closures

No reported cases exist of a modern GPU-centric data center shutting down for lack of tenant demand. This is a critical data point. Projections of GPU values collapsing to zero within 18 months would require widespread data center asset replacement or facility closures. Neither is occurring.

Supply-constrained expansion

The sector is capacity-constrained on the build side. Approximately 50% of planned 2026 data center facilities have been delayed due to insufficient grid capacity or non-GPU supply constraints exacerbated by the war in Iran. This constraint on available rack space maintains the value of existing installed GPUs: the hardware generates no revenue without the supporting infrastructure to power it. The combination of supply constraint and sustained demand growth indicates that GPU economic life cycles are considerably longer than near-term depreciation narratives suggest.

Implications for Collateral Valuation

An H100 server operating within a power-secured, operational data center under a multi-year power purchase agreement is a materially different collateral asset from the same server sitting uninstalled in a warehouse or inside a facility with expiring lease terms. The data center context, comprising secured power supply, cooling infrastructure, network interconnects, and remaining facility tenure, is an increasingly scarce prerequisite for GPU value realization. The North Sea oil platform analogy applies directly: drilling equipment bolted to a producing platform in a licensed block commands a substantially different valuation from identical equipment staged in an Aberdeen yard. Location, infrastructure, and operational status are inseparable from asset value.

At Barkr, facility context is a material input in every GPU collateral assessment. Valuations that treat GPU hardware in isolation systematically misstate liquidation value, understating it for hardware embedded in secured long-tenure facilities and overstating it for undeployed inventory.

IV. GPU Market Stability Stats

The Emergence of GPU-Backed Lending

GPU-backed lending has transitioned from experimental to operational. CoreWeave's $2.3 billion debt facility in 2023, secured by first-lien interests over an NVIDIA H100 fleet, established the structural template. The announcement from CoreWeave on their DDTL 4.0 Facility on March 31, 2026 shows the market has since deepened and matured considerably. Per PitchBook's reporting [13][14], Lambda Labs raised a $500 million special-purpose GPU financing vehicle collateralized by owned chips; Crusoe raised a $750 million credit facility from Brookfield Asset Management to finance additional GPU inventory. AI companies accounted for approximately 21.8% of all venture debt deal count in 2025, consuming nearly one in four venture debt dollars.

Institutional private credit, including Blackstone and Brookfield, has entered the space alongside selective bank participation through structured vehicles, ranging from direct secured loans to SPV-based lease structures with insolvency-remote title segregation. NVIDIA itself has deployed capital through SPV structures, including a reported $2 billion investment in xAI executed through a dedicated GPU-SPV that acquires hardware and leases it back to the operating company.

Underwriting standards remain nascent. PitchBook quotes Keri Findley, CEO of Tacora Capital: "You just have to come up with a proper depreciation curve. We've struggled a little bit with GPUs, depending on what contracts look like, because they depreciate really quickly, and the technology is moving much more quickly than a lease can" [13]. Bob Curley, CEO of Bridge Bank, told PitchBook: "In terms of GPUs, I don't think anyone really knows the economic shelf life" [13]. These assessments reflect the current state of practice and define the core underwriting challenge with precision.

Loan-to-Value Ratios and Structural Terms

Economic depreciation front-loading

Current market data indicates GPUs lose approximately 30 to 40% of MSRP within the first 12 months of deployment, after which values tend to plateau for 24 to 48 months. Identifying the precise plateau point is critical for calibrating Loan-to-Value ratios, a variable Barkr monitors through regular pricing updates across its transaction dataset. Applying MSRP as a valuation baseline overstates collateral quality from the inception of the facility.

The HBM premium

The price differential between GPU models tracks High Bandwidth Memory (HBM) capacity more reliably than raw compute teraflops. As foundation model sizes continue to grow, a GPU with 141GB of memory (H200) may command twice the resale value of an 80GB unit (H100), despite similar clock-speed characteristics, because memory capacity is the binding constraint for inference workloads. The parallel to upstream energy assets is apt: an H100 rack retains meaningful value, much like a mature oil field continues to produce, but commands a discount to a newer asset with greater throughput capacity.

Secondary market liquidity cycles

Eight-GPU server nodes transact materially faster in the secondary market than individual PCIe cards. Enterprise and data center buyers procure in node-level units; a single card requires the buyer to supply a compatible host system. Barkr's initial transaction data indicates 8-GPU nodes sell approximately four times faster than individual PCIe cards. The node format represents a significantly more liquid collateral structure, and this distinction should be reflected in facility-level LTV policy.

Terms and amortization

GPU-backed facilities are typically structured as short-dated instruments, three to five years, with front-loaded amortization designed to keep the outstanding loan balance below the asset's declining market value throughout the facility's life. The structural discipline required is that principal repayment outpaces value decay. At Barkr, we have observed the consequences of this misalignment directly in aircraft lending. Airfares on a given route can fall 40% within a single season when a low-cost carrier adds capacity. The resale value of the narrowbody operating that route moves considerably less, anchored to airframe maintenance status, remaining engine life, and the global market for that aircraft type. Lenders who structured facilities against ticket revenue projections rather than appraised aircraft value learned that distinction under adverse conditions. GPU lending sits earlier on the same learning curve. Repayment schedules must be calibrated to projected asset resale value, with accelerated paydown provisions triggered if collateral underperforms its appraisal.

Pricing

GPU-backed loans carry wider spreads than comparable equipment finance in established asset classes, reflecting the risk premium lenders assign to residual value uncertainty in a market where new architectures are introduced on an annual cycle.

Covenants

More sophisticated facilities incorporate real-time telemetry covenants enabling remote confirmation of GPU operational status, periodic revaluation triggers linked to secondary market price movements, and cash sweep mechanisms tied to rental revenue performance.

The Depreciation Curve: Key Underwriting Variables

Hyperscaler depreciation as a reference framework

Amazon, Microsoft, and Google adopted six-year GPU depreciation schedules; Meta adopted five and a half years. These decisions carry material P&L consequences and reflect institutional conviction regarding productive asset life. Amazon's Q3 2025 10-Q discloses a partial reversal for a subset of servers, a reduction to five years that produced a $677 million net income impact over the first nine months of 2025 [15]. Satya Nadella's public acknowledgment that NVIDIA's architectural release cadence exceeded Microsoft's planning assumptions introduces additional conservatism [16].

The value cascade

Prior-generation GPUs migrate through workload tiers as each successive architecture releases. A100-class hardware transitioned from frontier training to inference to batch workloads across a four-year span. Azure's V100 service life of approximately seven and a half years provides empirical grounding for extended economic life assumptions [18]. The cascade model breaks down if purpose-built inference silicon, such as AWS Inferentia or Google TPU, displaces GPU-based inference at scale, removing the middle rung of the value ladder.

Export control impact on orderly liquidation value

US export controls have eliminated China as a legitimate buyer in the advanced GPU secondary market, materially reducing the addressable buyer pool in forced liquidation scenarios. This warrants a downward adjustment to orderly liquidation value (OLV) assumptions. In our experience, this adjustment is frequently omitted from standard collateral assessments.

V. Demand Validation and Market Durability

GPU demand fundamentals remain robust across multiple institutional data sources.

Hyperscaler capital expenditure

Goldman Sachs [11] projects aggregate hyperscaler capital expenditure from 2025 to 2027 at $1.15 trillion, more than double the $477 billion deployed from 2022 to 2024, with the top five hyperscalers expected to direct approximately 75% of 2026 spending toward AI infrastructure.

Neocloud contract backlogs

CoreWeave's SEC filings [1][2] show a contracted revenue backlog of $55.6 billion as of September 30, 2025, growing to $66.8 billion by year-end, a figure representing approximately a decade of forward revenue visibility at current run rates.

The workload transition

AI infrastructure demand has shifted from training-dominant to inference-dominant as organizations transition from building foundation models to deploying them at production scale. This shift sustains demand for installed prior-generation hardware. Inference workloads are less sensitive to peak compute throughput and more sensitive to memory capacity, per-unit economics, and deployment reliability. This is the primary structural reason H100-class resale demand has remained durable while rental rates have declined.

VI. Potential Developments

Algorithmic Efficiency

A potential tail risk to GPU collateral values is an algorithmic or architectural breakthrough that materially reduces the accelerator-hours required to train or serve models at a fixed level of capability. A parallel can be drawn to optical fiber and internet bandwidth in the early 2000s: as Daubechie's wavelet compression and related technologies increased the effective capacity of installed networks, the economics of new infrastructure became more sensitive to whether traffic growth could keep pace with rapidly expanding supply [19]. There are precedents within AI itself: the compute required to reach a fixed language-modeling performance threshold has been estimated to halve roughly every 8–9 months since 2012 [20], the compute required to reach AlexNet-level ImageNet performance fell about 44x between 2012 and 2019 [21], and compute-optimal training has already shown that a model such as Chinchilla can outperform a larger predecessor at the same training-compute budget while requiring less downstream fine-tuning and inference compute [22]. Inference-side methods can also reduce hardware intensity: SmoothQuant reported up to 1.56x faster inference and 2x memory reduction with negligible loss in accuracy [23]. Under Barkr's base-case assumptions, aggregate demand may still exceed supply even after substantial efficiency gains, but lenders with multi-year GPU-backed facilities should nevertheless stress-test scenarios in which algorithmic progress materially lowers required accelerator-hours per workload and slows growth in installed-capacity demand, even with demand for AI services remaining strong.

New, Cheaper Hardware

Several new entrants have emerged in the accelerator chip market, and major players like NVIDIA have made defensive acquisitions in response. The most consequential near-term disruption may originate from Apple, which is focused on deploying personal AI capabilities at consumer scale. NVIDIA and others are accelerating development of SME and consumer-oriented compute solutions in parallel. The competitive dynamic may ultimately resemble the historical Microsoft-Apple pattern, with NVIDIA occupying the enterprise incumbent position and Apple establishing dominance in the consumer segment.

At the data center level, hyperscalers are deploying proprietary inference silicon at scale: AWS Inferentia and Trainium, Google TPU, Microsoft Maia, and Meta MTIA. These architectures can offer superior total cost of ownership for inference workloads relative

to prior-generation GPUs. If purpose-built inference silicon displaces GPU-based inference systematically, the value cascade model described in Section IV starts to break down: H100-class hardware lose some of the inference workload tier and face accelerated depreciation.

Macroeconomic Sensitivity

The impact of a global recession on GPU demand depends materially on whether AI-driven productivity gains are sufficient to sustain corporate technology investment through an economic downturn. Barkr's position is that AI is a meaningful productivity multiplier, currently displacing employment only at the margin, and that the larger structural displacement risk to labor lies in robotics rather than software AI. On that basis, demand for GPU infrastructure should remain relatively resilient through moderate macroeconomic stress.

VII. Conclusion

The Barkr Global GPU Price Index reached 118.2 in March 2026, an all-time high. Hourly H100 rental rates are 60 to 70 percent below their 2024 peak. Both observations apply to the same hardware at the same point in time.

The GPU market has ceased to function as a single market. The rental segment exhibits commodity-like price dynamics and declining margins. The secondary resale market is demonstrating resilience and predictability, driven by ownership demand, compute demand growth, physical infrastructure constraints, and data center supply limitations. For lending institutions, GPUs represent a collateral asset with a largely predictable depreciation profile.

The practical implications for lenders are as follows:

> **The resale market requires a dedicated valuation framework**
>
> Rental rate movements are a poor proxy for secondary market prices. The two markets are driven by different buyer populations, operating on different time horizons, responding to different structural forces. A lender calibrating collateral value to cloud pricing is measuring the wrong market.

Depreciation is front-loaded and non-linear

The initial discount from MSRP is immediate and material. The subsequent trajectory is more gradual, interrupted by generational transition events and workload displacement risk. A scenario-based depreciation model with explicit generational transition triggers is more defensible than either a six-year straight-line or a two-year cliff assumption.

Facility context is a material collateral variable

An H100 cluster inside a power-secured operational data center with substantial remaining tenure commands a higher liquidation value than equivalent hardware held undeployed. Valuations conducted in isolation from facility context are structurally incomplete.

GPU-backed lending is a viable and growing asset class

The underwriting discipline required to execute it well is specific and has yet to standardize across the market. Lenders who develop that discipline, and who assess hardware in its operational context rather than against list price, are pricing a risk that the broader lending market is still in the process of learning to measure.

Bibliography

[1] CoreWeave, Inc. Form 8-K: Third Quarter 2025 Financial Results. U.S. Securities and Exchange Commission, November 10, 2025.

[2] CoreWeave, Inc. Annual Report and SEC Filings, 2024–2025. U.S. Securities and Exchange Commission, 2025.

[3] NVIDIA Corporation. "NVIDIA Blackwell Delivers Massive Performance Leaps in MLPerf Inference v5.0." NVIDIA Technical Blog, January 8, 2026.

[4] NVIDIA Corporation. "Nvidia's B200 Boasts 2.2x Gain Over H100 in MLPerf Training." NVIDIA Technical Blog, November 2024.

[5] NVIDIA Corporation. "NVIDIA Blackwell Enables 3x Faster Training and Nearly 2x Training Performance Per Dollar than Previous-Gen Architecture." NVIDIA Technical Blog, December 12, 2025.

[6] Bain & Company. 2030 Global Data Center Forecast. October 2025.

[7] CBRE Research. North America Data Center Trends: H1 2025. CBRE Group, Inc., 2025.

[8] S&P Global. Navigating the US Data Center Power Crunch: On-Site Solutions Offer a Faster Path to Power. December 2025.

[9] Lawrence Berkeley National Laboratory. Queued Up: Characteristics of Power Plants Seeking Transmission Interconnection. August 2025 update.

[10] Gartner, Inc. Data Center Electricity Consumption Forecast, 2025–2030. 2025.

[11] Goldman Sachs Global Investment Research. Hyperscaler Capital Expenditure Projections, 2025–2027. 2025.

[12] ABI Research. Profiling Seven Leading Neocloud Companies. October 2025.

[13] PitchBook. "As Venture Debt Gambles on GPUs, Not All Are Sold on Silicon-Backed Loans." August 27, 2025.

[14] PitchBook. "AI Startups Gobbling More Than a Third of Venture Debt Dollars This Year." July 2025.

[15] Amazon.com, Inc. Form 10-Q: Third Quarter 2025. U.S. Securities and Exchange Commission, 2025.

[16] Microsoft Corporation. Satya Nadella, quoted in "The Question Everyone in AI is Asking: How Long Before a GPU Depreciates?" CNBC, November 14, 2025.

[17] Yahoo Finance. "How Fast Does an AI Chip Depreciate, and Why Does It Matter for Nvidia Stock?" December 11, 2025.

[18] Microsoft Azure. Virtual Machine Retirement Notices: NCv3-Series (V100). Microsoft Corporation, 2025.

[19] Couper, E.A., Hejkal, J.P., and Wolman, A.L. "Boom and Bust in Telecommunications." Federal Reserve Bank of Richmond Economic Quarterly, Vol. 89, Fall 2003, pp. 1–24.

[20] Erdil, E. and Besiroglu, T. "Algorithmic Progress in Computer Vision." arXiv preprint arXiv:2212.05153, 2022. As summarized in: Pilz, K.F., Heim, L., and Brown, N. "Increased Compute Efficiency and the Diffusion of AI Capabilities." AAAI, 2024.

[21] Hernandez, D. and Brown, T.B. "Measuring the Algorithmic Efficiency of Neural Networks." arXiv preprint arXiv:2005.04305, OpenAI, 2020.

[22] Hoffmann, J. et al. "Training Compute-Optimal Large Language Models (Chinchilla)." arXiv preprint arXiv:2203.15556, DeepMind, 2022.

[23] Xiao, G. et al. "SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models." ICML 2023. arXiv preprint arXiv:2211.10438, 2022.