



Why is NVIDIA so confident in long-term GPU values?

Barkr Research Series

Executive Summary

NVIDIA has announced two key measures to accelerate investment and risk capital into AI infrastructure, both of which expose it to some degree of financial risk. One creates financing platforms with Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs and KKR to mobilize more than \$500 billion of third-party capital.[1] The other helps AI-cloud operators procure NVIDIA infrastructure through revenue-sharing and credit-support arrangements.[2] In selected transactions, the announcement indicates a willingness to support up to 25% of the residual value gap for a given project, which raises two questions. Why is NVIDIA confident in the underlying equipment value? Who can independently determine that value, without bias, so markets can trust the output?

Who determines the supported value?

It's important to note that the 25% referenced is an indication. Clearly, NVIDIA will define the conditions it is willing to support. Lenders and investors will determine advance rates, covenants and source of credit. Each participant therefore answers a different part of the chain. But who can and should value the assets to inform each deal?

"AI infrastructure is a new asset class and the market hasn't built conviction on useful lives or resale values. The problem isn't that these assets depreciate; it's how do they depreciate, and unpriceable risk kills deals in a way that expensive risk doesn't. Independent valuation is what turns that into a measurable exposure that can be priced, insured, and lent against at a coupon that reflects the actual risk."

— Conor Moore, USD.AI

At Barkr, we predict current and forward asset value independently of the financing structure. Our domain-specific models and data pipelines draw on more than 300 million data points, 1.6M GPU-specific. They combine equipment-market evidence with system configuration, workload utility, deployment status and broader market conditions. We then stress-test the prediction against changes in demand, supply, efficiency and infrastructure delivery. Asset status is among the first distinctions. A live server system and an accelerator awaiting deployment are economically different. To date we've provided ~\$3B in GPU valuations, ~\$200m of which are backed by our guarantee, and those valuations indicate that a 25% residual gap support might be tested more than expected without robust independent valuations.

Live status is a material part of the asset

Sometimes, viewing complex markets through the simplest economic principles helps eliminate the noise – there is more demand for compute than supply and the imbalance is projected to get worse. The net effect is that whatever live compute is available retains more value over time until that imbalance is corrected.

An accelerator becomes available compute only after integration into a server and installation in a facility with power, cooling and networking. The completed system must then be accepted for service. A live system has cleared these steps and can generate revenue immediately. Operating readiness does not create an automatic premium, but it reduces the time and execution risk separating the asset from productive use.

U.S. delivery has accelerated rather than slowed. Aterio records 4.14 GW of live additions in H1 2025 and 5.84 GW in H1 2026. That represents an increase of approximately 41%. Active nameplate reached 50.1 GW by June 2026.[3] A larger volume is coming online, but it has not yet created a broad reserve of available capacity.

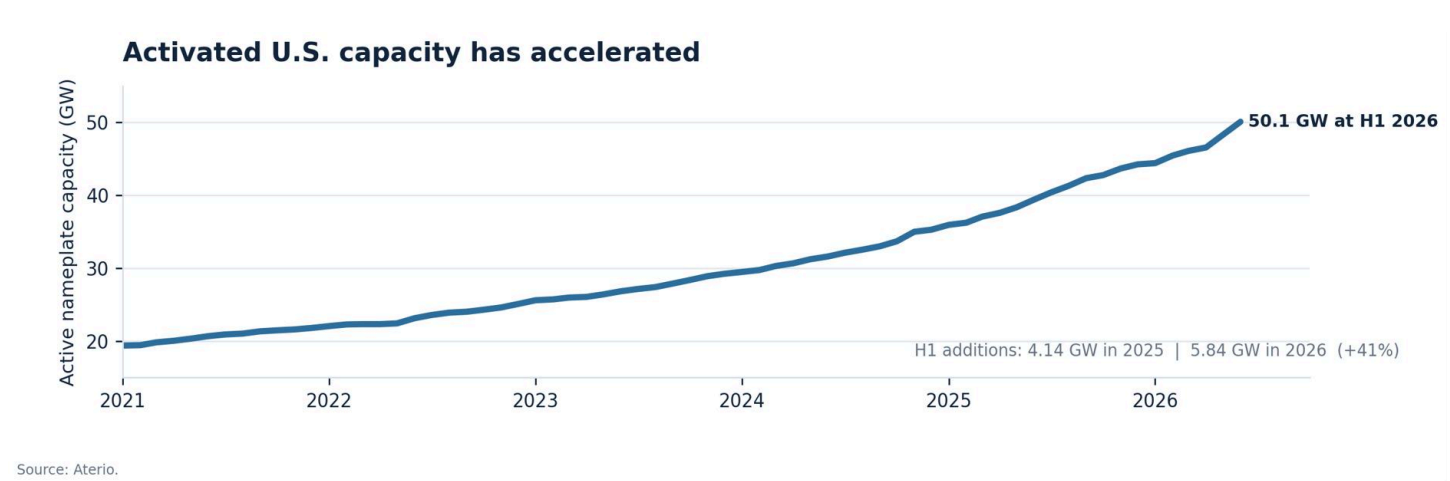


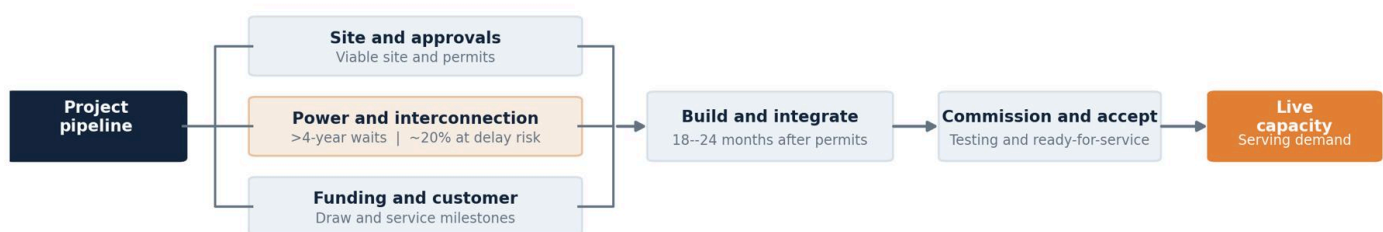
Figure 1. Monthly activated U.S. nameplate capacity through H1 2026.

Faster delivery matters, but it does not remove that distinction. The valuation question depends on whether the project pipeline can become commissioned supply quickly enough to create spare operating capacity.

Why scheduled capacity does not equal live supply

Published pipeline totals measure scheduled completion, not live capacity. Goldman Sachs Research finds that about 72% of facilities scheduled for the next four quarters activated on time.[4] The remaining 28% missed the stated delivery window. Figure 2 shows the development and commissioning stages separating an announcement from usable compute.

Parallel workstreams converge before capacity can go live



Sources: Goldman Sachs Research; IEA; JLL; Carbon Direct; U.S. DOE; CoreWeave and Galaxy SEC filings.

Figure 2. Parallel prerequisites converging into construction, commissioning and live capacity.

Before construction can begin, the site and power decisions must align. Developers often apply in several regions before choosing one location. The early pipeline can therefore include alternatives rather than independent projects, and withdrawals or denials reduce the number that advance.[4,6] Even a selected site cannot proceed without an executable interconnection. The IEA estimates that grid constraints could delay around 20% of global capacity planned through 2030. JLL reports connection waits exceeding four years in primary markets.[7,5] Until that power path is secured, the project remains prospective supply.

Once site and power are secured, the schedule moves from feasibility to execution. Financing agreements link capital draws and project economics to readiness, reporting and service milestones.[8,9] Construction takes 18–24 months and remains exposed to labor, supply-chain and critical-equipment constraints.[4,5] A completed system must still pass testing, commissioning and acceptance before it can serve workloads.[10,9] Delays can therefore carry from one stage into the next. JLL's finding that 57% of projects experienced delays of at least three months reflects the combined result of this sequence.[5]

These constraints do not imply that every delayed project will fail. They show why scheduled capacity must be weighted by its probability of arriving on time. Goldman expects approximately 60% of next-year capacity and 50% of capacity scheduled two years out to meet the stated date.[4] Those estimates define the 50–60% realization range applied to the 2027 pipeline in the next section. Current live systems retain an execution advantage while the remaining capacity works through that process.

Demand leaves limited room for delivery slippage

H1 2026 additions represent approximately 43% of the 13.6 GW scheduled for the full year. This is close to the 39–46% first-half share observed from 2022 through 2025. Annualizing the current pace produces 12.7–14.9 GW for 2026 and 56.9–59.1 GW of total live capacity.[3,4] Delivery is broadly on its near-term schedule. The pressure emerges as demand and the forward project slate grow.

Goldman expects 50–60% of the 36.3 GW scheduled for 2027 to arrive on time. Applying that range adds 18.2–21.8 GW and produces 75.1–80.9 GW of total live capacity. Another 14.5–18.2 GW remains outside the 2027 delivery window.[4]

The demand case makes the cost of that slippage visible. Goldman's 31 GW of operating demand against Aterio's 44.2 GW of active nameplate in 2025 implies 70.1% utilization, closely matching Goldman's 70% assumption.[3,4] Preserving that operating cushion as demand rises to 41 GW in 2026 and 66 GW in 2027 requires 58.6 GW and 94.3 GW of live capacity. The raw schedule reaches 94.1 GW, but the expected delivery range reaches only 75.1–80.9 GW, leaving a 13.4–19.2 GW gap. A straight extension of the 2026–2027 changes widens that gap to 27.3–36.8 GW in 2028.



Figure 3. Demand–required and expected live capacity through 2027, with a 2028 extension.

The raw development schedule is large enough on paper, but expected delivery is insufficient to preserve the operating reserve embedded in the demand case. Existing systems can carry more load above 70% utilization, but the result is less available headroom rather than surplus capacity.

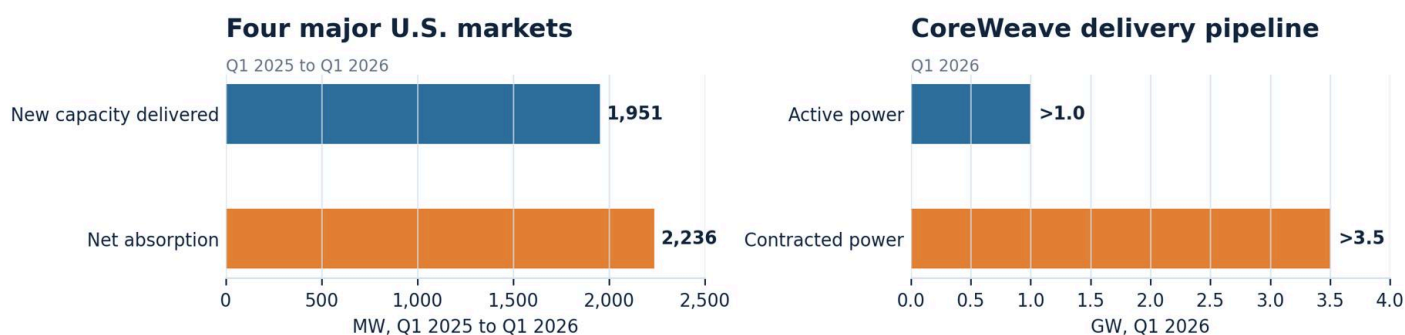
The conclusion survives a material downside test. Across the full expected supply range, the 2027 headroom gap remains when operating demand falls 14% below the base case. At the low end of expected supply, the gap remains after a reduction of approximately 20%. Demand does not need to exceed installed nameplate for this pressure to occur. Realized supply remains below the capacity required to preserve the operating cushion observed in 2025.

This squeeze in supply has a net positive impact on existing live GPUs. As evidenced by the recent Coreweave announcement contracting A100's first installed in 2020 being rebooked through 2029, demonstrating a 9-year useful life.[18]

Observed absorption confirms the pressure

Market data already show this pressure. Across four major U.S. markets, net absorption of 2,236 MW exceeded the 1,951 MW delivered between Q1 2025 and Q1 2026. Across the 16 largest markets, supply rose 25% while vacancy fell from 8.3% to 6.7%. In the four largest markets, 80% of capacity under construction was already preleased.[11] New capacity entered the market, yet available capacity tightened. Customers also committed to much of the next wave before completion.

Observed delivery, absorption and operator capacity



Sources: CBRE Global Data Center Trends 2026; CoreWeave Q1 2026 results.

Figure 4. Observed absorption versus delivery in four major U.S. markets and CoreWeave's active-versus-contracted power position.

CoreWeave's Q1 2026 position shows the same timing distinction at the operator level. More than 3.5 GW was contracted, while slightly more than 1 GW was active.[12] Commercial commitments can therefore precede live supply by a wide margin. That spread is not a valuation by itself. It shows that customers are reserving capacity before it becomes operational. This also sets the context for efficiency. Lower compute per task must be weighed against the additional use that lower cost makes possible.

Efficiency can accelerate absorption

Efficiency affects value through two opposing channels. Better software and more efficient models reduce the compute required for an existing task. This can pressure hardware when workloads migrate elsewhere. At the same time, the improvement lowers the cost of using AI. More applications become economical, and users can run inference more frequently. Task-level efficiency can therefore rise while total electricity and infrastructure demand continue to grow.[13] For valuation, the relevant question is whether new use absorbs the capacity released by each improvement.

Improved software, broader workload compatibility and slower replacement cycles can also keep deployed systems useful after newer generations arrive. Barkr's prior market work found that physical equipment values did not mechanically follow falling access prices.[14] Efficiency may change which systems serve which workloads without eliminating the economic use of existing systems.

Kimi K3 provides a current test of that mechanism. It reports approximately 2.5 times the overall scaling efficiency of Kimi K2.[15] An upper-bound stress treats that improvement as a 60% compute reduction for every affected workload. Even then, it closes the high-supply gap only if Kimi reaches roughly one-quarter of all modeled 2027 workload immediately. The calculation also assumes no rebound in use. The launch instead pushed Kimi's existing capacity close to its limit within 48 hours. Kimi then paused new subscriptions while adding capacity.[16]

The launch does not establish a market-wide demand response, but it illustrates a familiar economic mechanism. When efficiency lowers the cost of a service, more uses become economical. Economists call this a rebound effect. If total consumption rises despite the efficiency gain, the stronger outcome is known as Jevons' paradox. The mechanism has a long history in energy economics and has been studied across general-purpose technologies. Its magnitude varies by market.[17] Kimi's launch is consistent with this mechanism because reported demand absorbed the available launch capacity despite a large task-level efficiency gain. Efficiency therefore belongs in the valuation stress through both compute savings and demand expansion. It should not be

treated as an automatic residual-value haircut. The final supported percentage remains an asset- and transaction-level decision.

These capacity conditions do not directly determine the value of an individual GPU system. They establish the market context within which generation, configuration, operating status, workload utility, liquidity and redeployment risk should be assessed.

Why live systems should retain value

While there are other downstream capital considerations to consider that may impact the broader AI sector – for example, many AI-native providers still operate at negative unit economics. A significant drop in revenue (either to reduced VC capital forcing AI labs to charge higher prices or cutting fees to win market share) from consumers or enterprises will potentially have knock-on effects impacting compute providers and data centers as loan payments become harder to fulfill. Something that might bite during a market downturn. However, as long as the demand-supply imbalance persists, the underlying asset value is well positioned to absorb market shocks, which will not impact all parties equally.

As long as supply remains constrained, live GPUs will retain more value for longer. Generation, configuration, workload utility, operating status, contracted cash flows, redeployment cost, buyer depth and the terms of any external support still matter. The evidence here establishes the market condition against which those asset-specific inputs should be judged.

That condition remains constructive for live systems. Commissioned additions are being absorbed, scheduled projects face repeated conversion constraints, and the modeled headroom gap persists across the full expected supply range even after a 14% reduction in 2027 demand. Efficiency may change which systems serve which workloads, but Kimi shows how lower compute cost can enlarge use before it creates surplus. The evidence is therefore consistent with support reaching 25% in selected transactions, not with a universal valuation at that level. The strongest assets are configured systems that can serve demand now and remain economically useful while new capacity works through the pipeline. In either case, independent valuations are vital to establishing a source of truth when transactions involve values larger than some countries' GDP.

References

- [1] **NVIDIA.** NVIDIA Partners With Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs and KKR to Establish AI Compute Infrastructure Financing Platforms to Mobilize Over \$500 Billion of Third-Party Capital. NVIDIA Newsroom, August 2026. [Announcement.](#)
- [2] **Colette Kress and Raj Mirpuri.** NVIDIA Unlocks AI Compute at Scale, Inviting Partners to Power the AI Infrastructure Buildout. NVIDIA Blog, July 2026. [Announcement.](#)
- [3] **Aterio.** US Data Centers Dashboard. Aterio Monthly, July 2026. [Download page.](#)
- [4] **Wei, Hongcen, Daan Struyven and Samantha Dart.** US Data Center Power Demand Projected to Double by 2027. Goldman Sachs Research, May 2026. [Source page.](#)
- [5] **JLL Research.** 2026 Global Data Center Outlook: Navigating AI Demand, Power Constraints and Global Opportunities in 2026. Jones Lang LaSalle, January 2026. [Report PDF.](#)
- [6] **Carbon Direct.** Community Opposition to AI Data Centers: Lessons Learned. June 2026. [Report page.](#)
- [7] **International Energy Agency.** Energy and AI. Paris, April 2025. [Report page.](#)
- [8] **CoreWeave, Inc.** Credit Agreement Dated March 30, 2026. SEC Exhibit 10.11, March 2026. [Exhibit 10.11.](#)
- [9] **Galaxy Helios I LLC.** Credit Agreement Dated August 15, 2025. SEC Exhibit 10.23, August 2025. [Exhibit 10.23.](#)
- [10] **U.S. Department of Energy.** Federal Distributed Energy Project Implementation Process: Construction and Performance. Federal Energy Management Program, 2026. [Guidance page.](#)
- [11] **CBRE Research.** Global Data Center Trends 2026. CBRE, June 2026. [Report summary.](#)
- [12] **CoreWeave, Inc.** CoreWeave Reports Strong First Quarter 2026 Results. SEC Exhibit 99.1, May 2026. [Results release.](#)
- [13] **International Energy Agency.** Key Questions on Energy and AI. Paris, April 2026. [Report page.](#)
- [14] **Barkr Research.** GPU Resale vs Rental: Robust, Divergent Markets. April 2026.
- [15] **Kimi Team.** Kimi K3: Open Frontier Intelligence. arXiv:2607.24653, July 2026. [Paper.](#)
- [16] **Chan Ho-him.** China's New AI Model Halts New Subscriptions as Demand Swamps Capacity. Associated Press, July 2026. [Report.](#)
- [17] **Steve Sorrell.** Jevons' Paradox Revisited: The Evidence for Backfire from Improved Energy Efficiency. Energy Policy 37(4), April 2009, pp. 1456–1469. [Paper.](#)
- [18] **Alistair Barr .** This 2020 Nvidia chip is still going strong. That matters for the AI boom. Businessinsider. August 12, 2026. [Source Page](#)