

The Productive Life of a Data Center GPU

What realized resale prices, rental rates, and fleet retirements show about useful life and residual value.

Technology leaders are being asked to set useful-life assumptions for GPU fleets, and lenders have begun underwriting against residual value directly. NVIDIA now guarantees up to 25% of residual value to unlock roughly \$500B in AI infrastructure financing. The assumption is no longer an accounting footnote.

Most of the public debate runs on modeled depreciation curves and listing prices. This note sets out four things the observed record shows instead: how long these fleets stay in service, what they earn while they are in it, what they are worth at the end, and whether there is a market to sell into. Every figure here is a realized transaction or a named public source. No modeled curve is used.

Here are the four insights the data shows:

01 Service life runs closer to eight years than to three.

Microsoft ran its K80, P100, and P40 fleets **6.9 to 8.8 years** before retiring them, and its V100 fleet **8.4 years**, with V100 capacity still scheduled in two regions through June 2027. AWS's Matt Garman said in February 2026 that AWS has **never retired an A100** and is sold out. Operator depreciation schedules have moved the same direction, from three years to five and six. The largest single block in Sprout's current inventory is 2017 Volta, which puts the decommissioning wave arriving on our floor today at roughly eight years after launch.

Microsoft Azure VM retirement notices, [learn.microsoft.com](https://learn.microsoft.com/en-us/azure/virtual-machines/retirement-notices). Datacenter Dynamics, 5 Feb 2026. Sprout's inventory system, 24 Aug 2026.

02 Prior generation rental rates rose through 2026 rather than falling.

H100 rental rates moved from roughly **\$2.00 to \$3.00 per GPU-hour between January and August 2026**. A100, H200, and B200 all rose over the same period. Published on-demand rate cards in July 2026 carried **A100 80GB at \$2.21 to \$2.79 per GPU-hour**, six years after that platform shipped. Inference and batch workloads cascade onto older silicon, where the payback arithmetic often beats current-generation hardware, and the binding constraint on new capacity is power and shell rather than chips.

Silicon Data H100 and A100 indices. Lambda and CoreWeave published on-demand rate cards, Jul 2026.

03 Six-year-old systems are clearing at 38% to 43% of launch price.

Sprout is currently clearing 8-GPU **A100 systems at \$75K to \$85K** against the DGX A100's roughly \$199K launch price in May 2020. 8-GPU **H100 systems clear at \$230K to \$260K**. These are settled sale prices rather than listings. Most public GPU price indices are built on asking prices, and asking prices overstate the market. For reference, the residual level now being used to underwrite AI infrastructure lending is 25%.

Sprout realized clearing prices, Aug 2026. DGX A100 launch price, May 2020.

04 There is a working market underneath those prices, not a handful of trades.

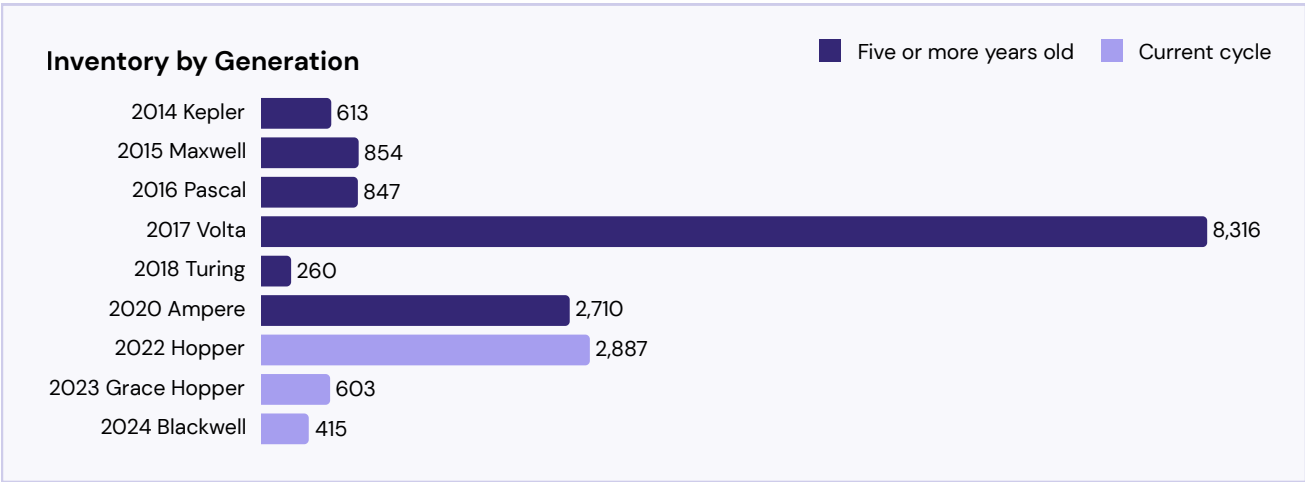
Sprout sold **11,664 GPUs** in the 13 months ending September 2026, with volume in 12 of those 13 months. Nine NVIDIA data center generations sit in live inventory simultaneously, 2014 Kepler through 2024 Blackwell, and every one of them has a bid against it. Liquidity is what makes a residual assumption underwritable, and it is the part of the picture that published price data does not capture at all.

Sprout's inventory system, pulled 14 Sept 2026.

Nine generations, all live at once

NVIDIA data center accelerator inventory by generation, GPUs on hand

Hardware that a three year useful-life assumption would treat as written off is still in custody, still being tested, and still being sold.



Method: Sprout's inventory system, non-shipped assets, GPU-holding facilities only, pulled 20 Aug 2026. 17,505 GPUs classified to generation by SKU. Excludes consumer and crypto-era cards.

How the largest operators have revised useful life

Every major operator has extended its server depreciation schedule over the past six years, in most cases more than once. These are disclosed accounting lives, not projections.

Operator	Current	Prior	How it moved
Amazon (AWS)	5–6 years	3 years	3 to 4 in 2020, 4 to 5 in 2022, 5 to 6 in 2024; a subset shortened back to 5 in 2025
Microsoft	6 years	4 years	Extended to 6; Nadella has cited not wanting to be locked to a single generation
Google	6 years	3–4 years	Moved to 6 across 2023 and 2024
Meta	5.5 years	3 years	Extended three times since 2020; reduced Q1 2025 depreciation expense by \$2.9B
CoreWeave	6 years	n/a	6 years from inception

Source: company disclosures and press reporting, compiled by Sprout. Accounting life is a conservative proxy for physical life, not a ceiling on it. Microsoft's V100 fleet ran 8.4 years against a 6-year book life.

About this data, and what it does not tell you

Where it comes from:

Sprout takes physical custody of end-of-life data center hardware across five certified facilities in North America and Europe, and has processed over 22M assets since 2014. Prices quoted here are settled transactions from that ledger. We do not manufacture this hardware; we sell it on behalf of the clients who retire it. Sprout holds inventory whose value this note bears on.

What it covers:

System-level sales in the configurations we receive. Loose module prices behave differently and are not interchangeable with these figures. Sprout transaction and inventory data are presented in aggregate and do not identify any client, client asset, or client transaction. Rental rates cited are published on-demand rates, which typically sit above committed and reserved pricing.

What it does not do:

Nothing here forecasts future residual value. Successor-product timing, power and shell availability, and export controls all move it, and rack-scale systems carry a materially narrower buyer pool than modular node-scale systems. Treat these as observations about the installed base, not as a curve to underwrite against.

Sprout is the intelligent technology-lifecycle platform that turns IT assets, from endpoint devices to AI data center infrastructure, into a strategic driver of security, sustainability, and ROI through real-time visibility, auditable compliance, and measurable value recovery.



Selected sources: CNBC, "NVIDIA, Wall Street asset managers in \$500B AI push," 10 Aug 2026 • CNBC, "How long before a GPU depreciates?", 14 Nov 2025 • Microsoft Azure VM size retirement notices, learn.microsoft.com • Datacenter Dynamics, 5 Feb 2026 • Silicon Data H100 and A100 indices • Lambda and CoreWeave published on-demand rate cards, Jul 2026 • J.P. Morgan Asset Management, Eye on the Market Outlook 2026, p. 16 • arXiv 2503.11901 (SC'25), "Story of Two GPUs" • operator 10-K and 10-Q filings • Sprout's inventory system and realized sales ledger.

This note is for general information only. It is not investment, lending, accounting, or valuation advice, or an appraisal of any asset. Figures are historical, as of the dates stated, and may not represent future results or any particular asset. Sprout makes no representation or warranty as to third-party information and undertakes no obligation to update this note. Do not rely on it for any credit, investment, or accounting decision. Company and product names mentioned are trademarks of their respective owners. Their use here is for identification only and does not imply affiliation or endorsement.