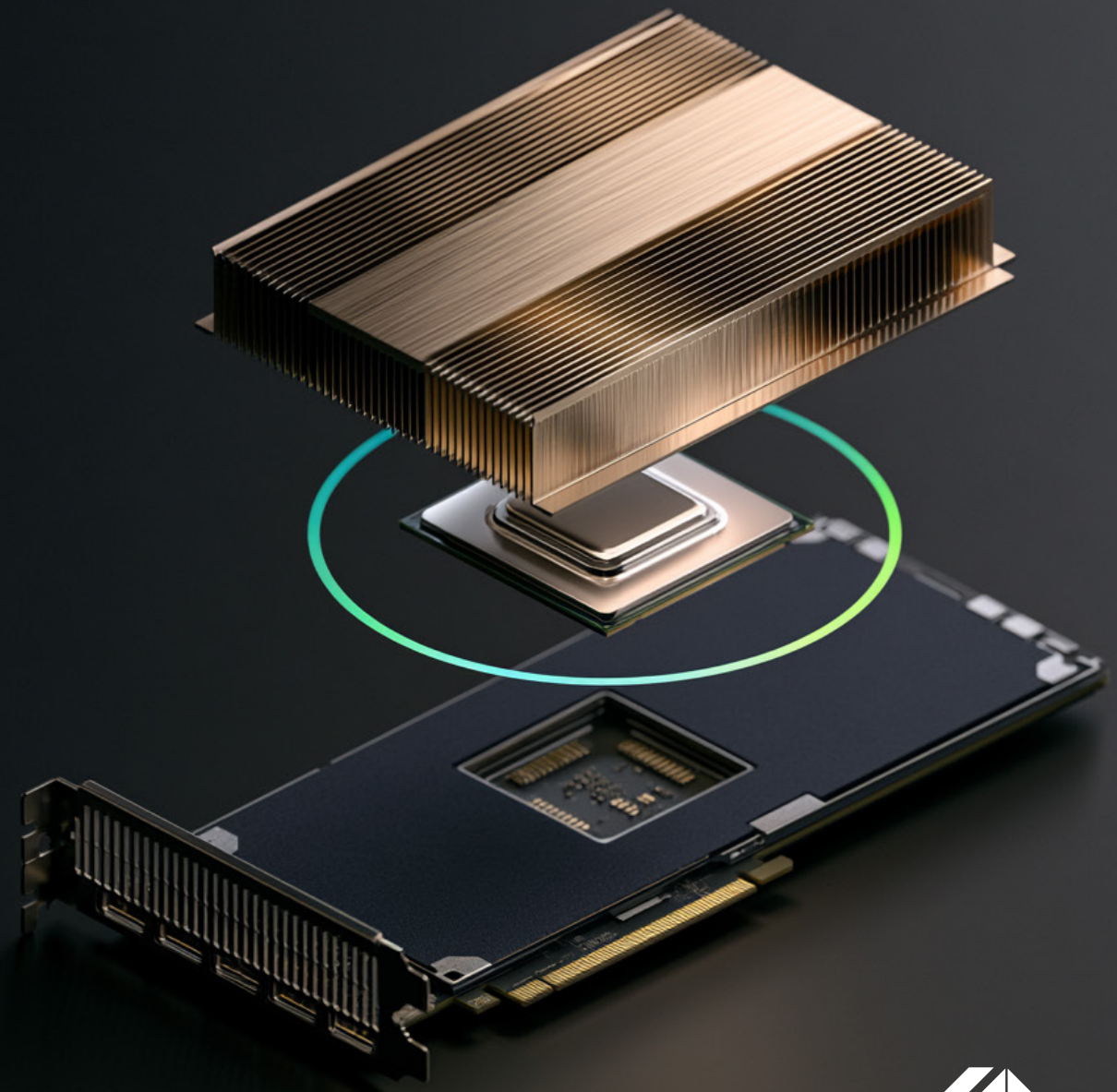
The Second Life of the GPU

What a retired AI fleet is worth, and why most operators recover less



SIMS LIFECYCLE SERVICES

The Second Life of the GPU

Executive Summary

Every data center operator who has built AI capacity in the last three years now faces a second decision that carries almost as much money as the first: what to do with the hardware when it is replaced with the latest generation. The first big fleets of AI accelerators, the GPUs that run AI training and inference, went into production across 2022 and 2023, and those fleets have now hit their refresh cycles. Thousands of GPUs are leaving the floor, and most of them run exactly as well as the day a technician racked them.

Here is the part many operators miss: a retired GPU fleet commands far less on the open market than it should.

The gap does not track the hardware's value. It tracks what a buyer cannot verify. A used accelerator can power on, clear a quick diagnostic, and look pristine, yet still hide memory degradation, a punishing thermal history, or silent computational errors that surface only under sustained load. Buyers know this, so they price in the risk, and that discount lands straight on the seller's recovery.

Sims Lifecycle Services (SLS) built its GPU certification program to close that gap. The program swaps "trust me" for documented evidence: performance validation, silent-data-corruption screening, machine-vision inspection, firmware verification, and certified data sanitization, each tied to a serialized record that follows the unit. Verified inventory clears faster, and it clears at a higher price.

The same substantiation that lifts a seller's recovery lets a buyer deploy on day one. Regional cloud providers, research institutions, neoclouds, and enterprises scaling inference now write second-life GPUs into their capacity plans. They want hardware they can get today, hardware that returns more per dollar, and hardware someone has proven before it ships. This paper traces where the supply comes from, why the value gap opens, and how a lifecycle turns retired accelerators into recovered value for sellers and trusted compute for buyers.

1. The refresh wave is here, and it is larger than the last one

The constraint on AI compute has shifted from budget to hardware availability. New-generation accelerators carry long lead times, and the largest buyers claim the bulk of each production run. The four largest hyperscalers, Amazon, Alphabet, Microsoft, and Meta, guided to combined 2026 capital spending of roughly $725 billion, up about 77 percent from 2025, with most of it aimed at AI infrastructure.[ii] Everyone else waits, and the compute they planned for rarely arrives on schedule.

At the same time, a large new source of capacity is opening up: the GPUs coming out of production. Newer silicon displaces the accelerator fleets that hyperscalers and the largest AI operators bought during the 2022 and 2023 build-out. This follows a planned cycle rather than a wave of failures, and accelerators as a class reach that cycle at scale for the first time.

The first factor is efficiency. Operators do not retire GPUs that stopped working. They retire GPUs that a better performance-per-watt argument has beaten. The next generation earns more from the same power and rack space, so operators pull hardware that still runs perfectly well.

The second factor is heat. An AI accelerator runs flat out around the clock, and that pace ages it faster than a general-purpose server that idles between jobs. Operators plan a two-to-three-year service life for these parts, far shorter than the five to seven years a traditional server earns.

Liquid cooling is the third factor. These systems share cooling across an entire rack, so operators pull whole racks at a time rather than single nodes. Supply arrives as full-rack and full-fleet events instead of a trickle of returns, which raises the bar for any disposition partner.

For most organizations that want this hardware, the previous generation fits the job. Training frontier models demands the newest silicon, but most teams do not train frontier models. They run inference, private AI, HPC, and research, and those workloads reward predictable performance, software compatibility, and fast deployment over raw peak throughput. Inference alone will consume roughly two-thirds of AI compute in 2026, up from about a third in 2023. [iii] More and more buyers want the very hardware now coming out of production.

2. The trust discount: why sellers recover less than a fleet is worth

A buyer can size up a used server quickly. The market spent twenty years building a shared language for the category. Configurations follow standards, condition grades carry roughly the same meaning from one seller to the next, and a buyer prices a unit off a spec line without much anxiety.

A used GPU offers none of that comfort. Its failure modes stay hidden until the right conditions draw them out. A card can power on, clear a basic diagnostic, and look pristine to the eye, then throttle under sustained load, carry high-bandwidth-memory degradation, or reveal a thermal history that surfaces only hours into a real workload. None of that shows up in a five-minute bench check, which leaves the buyer writing a large check on faith.

And faith carries a price. Buyers discount for the risk of the unknown, and that discount comes straight out of the seller's recovery. Yet the hardware itself has lost no value along the way. It has lost only the proof of that value, and that is the whole argument: the lost value points to a solvable problem, not a permanent condition.

> Silent data corruption is the most severe failure mode of all, and the hardest to detect.

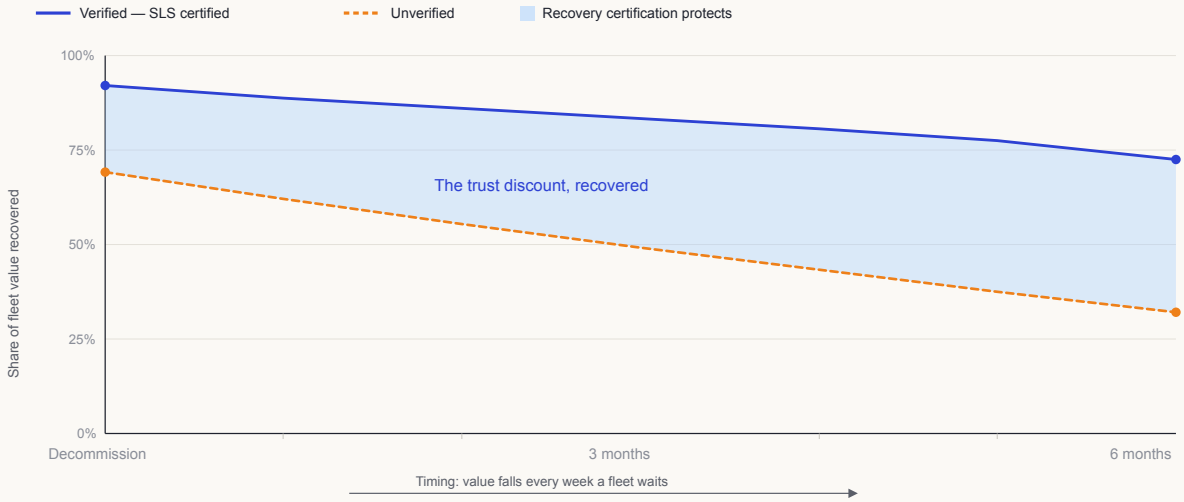
A card with this fault produces wrong results and reports that everything looks fine. It might compute that one plus one equals three and pass the answer downstream without raising a flag. Across a training run or a data pipeline, one faulty card can quietly poison an entire model or dataset. This is a problem at data-center scale. Google engineers documented “mercurial cores” that miscompute in ways manufacturing tests never caught. Meta’s account of training Llama 3 on more than 16,000 GPUs reported silent data corruption among the hardware faults that interrupted the run. These faults dodge quick functional checks by definition, which explains why unverified inventory carries such a steep discount.

3. What determines how much you recover?

FIGURE 01

What a fleet recovers, and what erodes it

Recovery falls with time. Certification sets how high it starts and how slowly it drops.



THE FOUR LEVERS OF RECOVERY

Timing
Value falls every week a fleet sits in the cage. Speed to market protects it.

Credible grading
Verified inventory clears higher. The shaded gap is what certification recovers.

Channel access
A specialist buyer list sets the ceiling the verified line starts from.

Completeness of scope
Cards, racks, network and cooling recovered together, not just the GPUs.

Illustrative. The curves show the shape of the effect, not measured values.

Four variables decide how much value a retiring fleet returns, and each one compounds the others.

- Timing moves the recovery more than anything else, and operators underestimate it most consistently. Accelerator values fall fast. A fleet that sits in a cage for a quarter, waiting on procurement, sheds value every week. Speed to market protects the recovery.

- Credible grading comes next. Verified inventory sells faster and for more than unverified inventory, and that speed feeds back into the timing above. When a buyer trusts the grade, they stop padding the offer with a discount that protects them against what they cannot see.

- Channel access sets the ceiling. A broker with three buyers cannot reach the same market as a specialist with a deep list of buyers, and that difference shows up in the final price.

- Completeness of scope captures the rest. GPUs are never removed on their own. Trays, rack frames, network components, cabling, and cooling infrastructure come out with them, and each piece holds real value in the right hands. A partner that grabs only the cards strands the seller with everything else.

4. Certified, not assumed: replacing uncertainty with evidence

The SLS certification program removes the doubt that drives the trust discount. Instead of a visual check and a short diagnostic, every accelerator moves through a documented certification process that measures how the hardware behaves under real operating conditions.

Functional validation measures the envelope rather than the center of it. Diagnostics report a clean pass or fail on the card's health. Stress testing then drives the card for hours under sustained and stepped load, logging junction, hotspot, and memory temperatures and recording the point where throttling begins, the same point that creeps earlier as a card ages. Benchmarking measures output against the manufacturer's baseline for that model, using workloads that mimic large language

model training and inference, so a card that has quietly lost performance shows the gap. The process screens for silent data corruption at this stage as well, matching each unit against a known-good model and flagging any variance. Silent errors surface at the voltage, frequency, and temperature corners, and only after a part has aged in service, which describes every card in a refresh fleet. A card that looks healthy but cannot compute reliably never slips through.

> Physical condition matters as much as performance for hardware that ships as individual cards or as complete racks.

A machine-vision system inspects connectors, pins, and cooling surfaces at a precision no manual check can match. It compares each image against a reference for a perfect unit and flags oxidation, corrosion, or bent pins for an operator to confirm. The manufacturer's own tools then verify the firmware, which proves the card matches its label and nobody has altered it. That step guards against the counterfeit and misrepresented accelerators that still circulate in the secondary market.

Not every deployment needs the same depth of testing. A standard tier applies to high-volume inventory where reliable throughput and consistent performance matter most, verified through burn-in testing. A premium tier runs the full test stack on premium SXM inventory, where higher-value deployments justify the added testing. Every certified GPU carries a documented grade and a digital certificate linked to its serial number. That record travels with the hardware the way a vehicle history report travels with a car, so a buyer sees the testing behind any unit instead of taking a seller's word for it. The buyer receives the thermal curve, the throttle point, and the error history.

5. Security, custody, and compliance by design

Performance settles only half the question, for the buyer's trust and the seller's position alike. The other half rests on proof: every card carries a recorded history, has moved securely since it left the rack, and meets the rules that govern advanced computing hardware.

The security question always reduces to the same two parts: what sits on the card, and how you know it is gone. This concern runs deeper than it does with drives. People understand storage media. They trust accelerators less, and they wonder what persists in high-bandwidth memory, in firmware, or in the residue a model checkpoint leaves behind. Often the workload that ran on the card represents the most valuable intellectual property the seller owns.

SLS answers in three parts. High-bandwidth memory holds nothing once power drops, and the small block of non-volatile storage that carries system-level configuration gets sanitized to NIST 800-88 Rev. 1 standards, with a certificate of destruction for every serial number. [vi] Custody stays continuous: serialized manifests, GPS-tracked transport, and a chain of custody that runs from the client's loading dock through final disposition. Intake reconciles every unit against the client's manifest, so discrepancies surface as corrections rather than later disputes. The Nashville Circular Center invites a walk-through, and clients who press hardest on security tend to relax once they have watched the floor.

None of it escapes export control. Advanced accelerators fall under ECCN 3A090.a, which governs where a card can go and who can receive it. The regime is in flux: a 2025 federal framework that would have imposed worldwide licensing was rescinded before it took effect, licensing reverted to the pre-existing country-group controls, and the government has signaled a replacement rule to come. A GPU disposition amounts to a trade-compliance event that happens to involve hardware. It demands a screen on every buyer, a clearance on every destination, and a record on every transaction that holds up when a regulator asks two years later. SLS runs all of it, holds C-TPAT and TAPA FSR-A, tracks each shift in the rules, and hands the seller a documented trail instead of an exposure. Most vendors are not experts in movement of GPU export and the risk exposure carries fines and/or jail time.

6. Choosing a partner: SLS against the alternatives

An operator weighing a retired fleet has four realistic options, and three of them fall short for accelerators.

- Handling it internally rarely works. A client's own staff are already fully committed to other job tasks, and GPU testing and channel development is a specialist capability they would build from scratch for something that happens once a cycle.

- A general ITAD vendor absorbs most of this volume today, and it fits the job poorly. These firms handle servers and laptops well. They lack GPU-specific test capability, and they sell into a generalist channel that prices accelerators like ordinary server parts.

- A broker moves product, sometimes fast, but brings none of the compliance stack. No R2v3, no SOC 2, no auditable chain of custody, no downstream vendor controls. For a client with an ESG commitment and an audit committee, the broker route opens a risk that dwarfs whatever spread the broker offered.

- Sitting on the assets costs the most and happens the most, since every week of delay drains value from a fast-moving market.

SLS holds the position none of the others can claim: specialist GPU capability inside a compliance and reporting infrastructure that survives an audit. One partner handles the GPUs, the racks, the network gear, the cooling infrastructure, the media, and the recycling tail.

7. From bench to rack, and into the reporting

A certified GPU follows a defined path back into service. It arrives under chain of custody, serialized and GPS-tracked, and intake reconciles its manifest against what the client shipped. Data sanitization closes out before anything reaches the test bench. Certification then runs the protocol for the unit's tier, and the card leaves the bench with a grade and a certificate tied to its serial number. Sell-side work runs alongside processing, so SLS usually knows a lot's destination by the time it certifies. SLS builds lots in coherent quantities of consistent grade, since buyers want production-ready inventory rather than a pile of mixed condition. Settlement and revenue share close the event in the client's reporting.

The buyers on the other side of that transaction keep multiplying. Regional cloud providers build capacity where the unit economics rule out new silicon. Grant-funded research and academic institutions, precise about their needs, put prior-generation hardware straight to work. Neoclouds and independent AI builders need real compute

that allocation never reaches. Enterprises running production inference grow fastest of all, and they reward the closest watch, since inference now holds far more of AI's lifetime cost and lifetime value than training does.

What keeps clients coming back is visibility. SLS reporting portal shows asset-level status, chain-of-custody records, data-destruction certificates, grading outcomes, settlement detail, and the landfill-diversion and ESG figures a sustainability team needs for its own disclosures. Most clients want that data in their own systems rather than a vendor portal, so SLS builds API connections that drop disposition records where their asset management already lives. That is what turns a one-time transaction into a standing relationship: a client who can pull an audit trail on demand and hand clean numbers to an ESG team has no reason to re-bid.

What keeps clients coming back is visibility. SLS reporting portal shows asset-level status, chain-of-custody records, data-destruction certificates, grading outcomes, settlement detail, and the landfill-diversion and ESG figures a sustainability team needs for its own disclosures. Most clients want that data in their own systems rather than a vendor portal, so SLS builds API connections that drop disposition records where their asset management already lives. That is what turns a one-time transaction into a standing relationship: a client who can pull an audit trail on demand and hand clean numbers to an ESG team has no reason to re-bid.

When an asset holds more value whole than in parts, RackRenew remanufactures and remarkets it as an integrated rack, which shortens the buyer's deployment and simplifies planning. SLS picks the highest-value path for every asset in a decommission, and the whole-rack route counts as one of them. Some racks return to the client for redeployment. Some move out through RackRenew. Some come apart, since the value sits in the components. Judging which path fits which asset is the service.

When a card reaches the end of its usefulness, value recovery continues. A unit that no longer runs at peak may still suit a buyer with lighter requirements, and SLS hands that buyer the performance data to decide. Whatever fails resale goes to the shredder for densification and on to a precious-metal smelter, which recovers gold, copper, and other materials. This recovery anchors the Sims circular-economy model, and it

counts for more as embodied carbon climbs. Building a high-performance accelerator carries a heavy manufacturing footprint, and for data-center hardware the carbon embodied in manufacturing can rival the carbon from years of operation. [vii] A second life preserves the carbon already spent to build the card and cuts demand for new manufacturing, which turns a sustainability target into a measurable result.

The bottom line

The refresh wave has arrived, and it will not slow. Each cycle sends more functional accelerators out of production, and their owners recover a fraction of what the hardware is worth. The cause is consistent: a buyer will not pay for what a seller cannot prove. That gap is solvable: the SLS certification program closes it, replacing assumption with evidence and turning a retired fleet into certified, deployment-ready compute. A decommission becomes a recovery, not a write-down. The hardware still holds its value. Certification is how you get it back, and the longer a fleet sits unverified, the less of it you recover.

About Sims Lifecycle Services

Sims Lifecycle Services is the lifecycle partner for data center and AI infrastructure, managing the full scope of decommissioning, from the loading dock through data destruction, testing, certification, remarketing, settlement, and reporting. Part of Sims Limited, a global leader in the circular economy, SLS pairs specialist GPU capability, including its GPU testing and certification program, RackRenew whole-rack remanufacturing, and the SLS client portal, with an audited compliance infrastructure trusted by the world's most demanding technology organizations. SLS gives retired infrastructure a verified second life: recovering more value for sellers, delivering trusted compute to buyers, and keeping high-performance hardware working for the AI economy.

[i] Combined 2026 capital-expenditure guidance for Amazon, Alphabet, Microsoft, and Meta, based on the companies' initial fiscal-2026 guidance following Q4 2025 earnings, as compiled by CNBC, "Big Tech's AI spending approaches $700 billion in 2026." The four guided to roughly $725 billion for 2026, up about 77 percent from $410 billion in 2025. CreditSights, "Hyperscaler Capex 2026 Estimates," puts roughly three-quarters of that spending toward AI infrastructure. Figures reflect early-2026 guidance; mid-2026 updates raised the combined total toward $750 billion.

CNBC: https://www.cnbc.com/2026/02/06/google-microsoft-meta-amazon-ai-cash.html

CreditSights: https://know.creditsights.com/insights/technology-hyperscaler-capex-2026-estimates/

Corroborating: Sherwood News, https://sherwood.news/tech/alphabet-amazon-microsoft-meta-plan-more-than-700-billion-on-capex-this-year/

[i]Dell'Oro Group, "AI Infrastructure Buildouts and Memory Cost Inflation Drove Data Center Capex Higher in 1Q 2026," 2026; Synergy Research Group, hyperscale capex forecast, 2026.
https://www.delloro.com/news/ai-infrastructure-buildouts-and-memory-cost-inflation-drove-data-center-capex-higher-in-1q-2026/

https://www.srgresearch.com/articles

[ii]Deloitte, "More compute for AI, not less," Technology, Media & Telecom Predictions 2026.

https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/compute-power-ai.html

[iii]P. Hochschild et al., "Cores That Don't Count," Workshop on Hot Topics in Operating Systems (HotOS '21), Google, 2021.

https://sigops.org/s/conferences/hotos/2021/papers/hotos21-s01-hochschild.pdf

[iv]Meta, "The Llama 3 Herd of Models," 2024.

https://arxiv.org/abs/2407.21783

[v]NIST Special Publication 800-88 Revision 1, "Guidelines for Media Sanitization."

https://csrc.nist.gov/pubs/sp/800/88/r1/final

[vi]U.S. Bureau of Industry and Security, rescission of the Framework for Artificial Intelligence Diffusion, May 2025 (ECCN 3A090.a):

https://www.akingump.com/en/insights/alerts/bis-rescinds-its-ai-diffusion-rule-and-issues-compliance-guidance-regarding-advanced-computing-items

[vii]U. Gupta et al., "Chasing Carbon: The Elusive Environmental Footprint of Computing," IEEE Micro / HPCA, 2021.

https://arxiv.org/abs/2011.02839