GPU SMITH
COMPUTE ENGINEERING

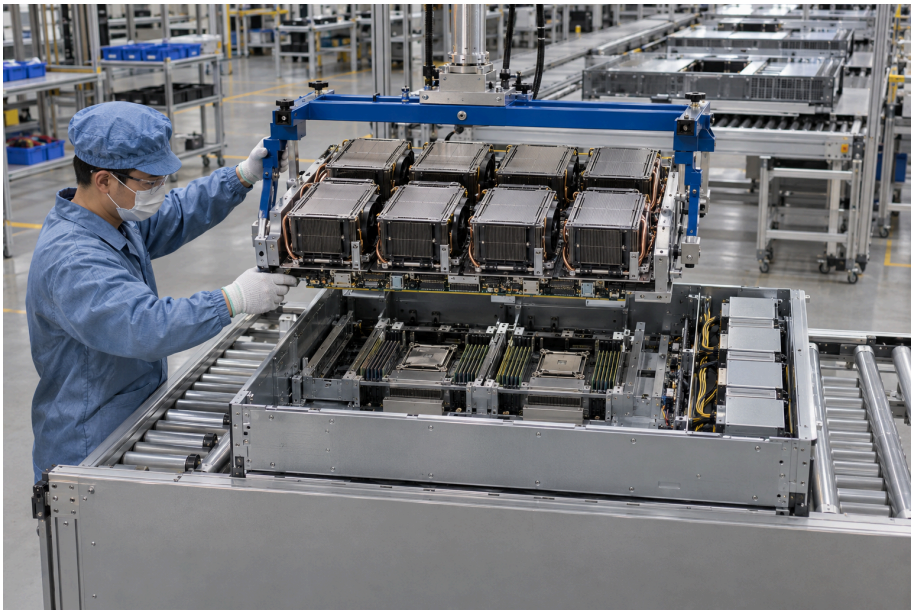
GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

Keep or Replace a GPU Fleet? Carbon Payback Worksheet

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-11-08 · keep or replace gpu fleet · gpu refresh cycle carbon payback · embodied carbon gpu · gpu server useful life · carbon payback period · gpu replacement cost benefit analysis · gpu server tco · nvidia hgx b200 · nvidia hgx h100 · private ai infrastructure

Original article: https://gpusmith.com/articles/en/keep-or-replace-gpu-fleet-carbon-payback



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes in the Decision Frame
4. Implementation Considerations and Process Changes
5. Data Analysis and Evidence
6. Implications and Future Directions
7. Frequently Asked Questions (FAQs)
8. Conclusion

Executive Summary

Deciding whether to keep a serviceable GPU fleet or buy a replacement fleet for an unchanged workload is a question about **incremental future emissions**, not about which chip is greener on a datasheet. The only manufacturing carbon that counts as new is the carbon of the products actually purchased. The carbon already embedded in retained equipment is sunk and is not charged again to the "keep" option. This report supplies a product-scope ledger and a keep-versus-replace worksheet, including an explicit **unknown** state, so a buyer can see whether measured energy savings repay new manufacturing carbon inside the planned operating horizon.

The public manufacturing numbers are narrow. NVIDIA reports **1,312 kg CO2e** for one HGX H100 baseboard (images.nvidia.com) and **2,274 kg CO2e** for one HGX B200 baseboard (images.nvidia.com), both cradle-to-gate (raw materials to factory exit), with use-phase and end-of-life emissions excluded (images.nvidia.com). A baseboard is not a server: it is paired with cooling and additional server components, as the vendor's own summary notes. The widely repeated **24%** improvement is a different kind of number: a vendor-normalized intensity of **0.66 to 0.50 gCO2e per exaflop** (developer.nvidia.com), not a 24% cut in absolute manufacturing carbon or in a buyer's workload emissions (developer.nvidia.com).

Operating-side inputs must be measured by the buyer. MLCommons states that measured power is valid only for the accompanying benchmark (mlcommons.org), and rated maximum power figures are nameplate limits, not consumption. Grid factors are dated and versioned: the latest U.S. EPA edition used here is eGRID2023 (www.epa.gov), while the IEA estimates global intensity at **435 g CO2/kWh** in 2025, falling to **360 g CO2/kWh** by 2030 (www.iea.org).

Using only cited inputs, repaying a baseboard-only **2,274 kg CO2e** requires about **6,500 kWh** of avoided facility electricity at the U.S. eGRID2023 average and about **5,200 kWh** at the IEA global figure (calculated below). Whether a real replacement achieves that depends on equal-service energy, utilization, additional replicas, power usage effectiveness (PUE) and the grid trajectory. Published modeling is mixed: one study finds replacing platforms older than four years could cut projected emissions by at least 75% in many carbon-intensive grids (arxiv.org), while another shows offsetting new-hardware embodied carbon can take months to tens of years (www.vldb.org). Those are model results, not measured fleet outcomes.

The decision therefore comes down to missing data, not missing arithmetic. No public full-server, network, storage or facility footprint accompanies the baseboard summaries. Until a vendor supplies product-boundary data for the whole purchased configuration, and the buyer supplies equal-service measurements, the defensible answer is a bounded interval or "unknown", never a single payback figure. Boundaries should be confirmed by a qualified lifecycle assessment practitioner before any figure is reported externally.

Introduction and Background

A GPU fleet that still meets its service objective carries a real option value. Replacing it creates new manufacturing emissions on day one, and any benefit arrives later as lower electricity use. For enterprise procurement teams, private AI operators and sustainability engineers, the question is whether that later benefit

exceeds the upfront cost within the horizon the organization will actually operate the hardware. This worksheet frames the problem this way: fix the workload and service objective first, then compare fleets on incremental emissions, and treat missing boundary data as a first-class result.

Two kinds of numbers circulate and are often confused. **Absolute** product carbon footprints (PCFs) state kilograms of CO2 equivalent (CO2e) for a defined product and boundary. **Normalized** claims divide that footprint by a performance measure. NVIDIA's developer article by Zoe Kessler and Aditi Suresh (developer.nvidia.com) is the best-known example for GPUs: it compares HGX H100 and HGX B200 baseboards using summaries that were both based on a cradle-to-gate analysis (developer.nvidia.com). The article has been referenced in other research contexts. For instance, a November 2025 California utility-program report prepared by Energy Solutions (calnext.com) notes in its introduction that hardware manufacturers like NVIDIA have announced a focus on reducing embodied carbon in their products (calnext.com), and lists the NVIDIA article in its bibliography (calnext.com). That report concerns building materials, so it shows how the vendor method is being cited, not that the method has been independently validated for GPU purchasing. NVIDIA's own headline figure is a 24% decrease in embodied carbon emissions between the two generations (developer.nvidia.com).

The scale of the electricity side explains why the question matters. Lawrence Berkeley National Laboratory (LBNL) estimated U.S. data centers consumed **176 TWh** in 2023, **4.4%** of U.S. electricity (escholarship.org), and projects roughly **325 to 580 TWh** by 2028 (escholarship.org). Embodied carbon is a smaller but non-trivial counterweight. An early hyperscale analysis concluded that most emissions of modern data-center equipment come from hardware manufacturing and infrastructure rather than electricity use, in a setting where renewable procurement had already shrunk operational emissions (arxiv.org).

This report does not rank Blackwell against Hopper and does not republish specification tables. Power-cap work is treated as upstream inputs. The added contribution is a boundary-aware ledger, a worked incremental calculation with unit checks, a sweep that exposes no-payback and unknown cases, and an evidence request a buyer can send to vendors. All numeric examples are either calculated from cited inputs or explicitly labeled hypothetical. Figures are as of November 2026.

Key Changes in the Decision Frame

Change 1: Fix the functional requirement before comparing fleets

A keep-or-replace comparison is valid only when both options deliver identical useful work at the same service level. That means the same model set, request mix, latency or throughput objective, availability target and data-residency constraints. A replacement fleet that serves more traffic, adds replicas for redundancy, or is run at higher utilization is a different decision and must be priced as such. The worksheet therefore starts with a service-equivalence statement, and the operating comparison uses energy per unit of delivered service measured under that statement, not peak throughput. The statement should fix at least:

- **Workload mix:** models, context lengths, batch sizes and request distribution, stated as a replayable trace.
- **Service level:** latency percentiles, throughput and availability targets that both fleets must meet.

- **Constraints:** data residency, air-gap requirements and security controls that limit where and how hardware can run.
- **Growth:** whether demand is flat, since added capacity for growth belongs to a separate decision.

Benchmark power data illustrate the caution. MLCommons states that MLPerf Power is only capable of measuring and validating full system power (mlcommons.org), and that the measured power is valid only for the accompanying benchmark (mlcommons.org). The MLPerf Power paper describes a methodology built by a consortium of more than 20 organizations that collected 1,841 reproducible measurements from 60 systems (arxiv.org). These are useful as method templates. They are not a substitute for a buyer's own wall-power measurements on the buyer's own workload.

Change 2: Replace "the GPU's footprint" with a product-boundary ledger

The public PCFs describe an eight-GPU baseboard. NVIDIA states the HGX B200 baseboard result covers raw material extraction through baseboard assembly, and that the baseboard can be paired with advanced air or liquid cooling solutions and additional server components (images.nvidia.com). A purchased server also contains host CPUs, system memory, storage, network adapters, power supplies and a chassis. Third-party work shows how large those pieces can be: a bottom-up model amortizes a 1,780 kg CO2e Lenovo SR675 V3 base module across four accelerator slots (arxiv.org), and a Dell PowerEdge R650 PCF (a general-purpose server, not a GPU server) reports a mean of 9,860 kg CO2e over its stated life (www.delltechnologies.com). Lenovo's own server sheets state that results are for the server only and that data-center values are not reported (static.lenovo.com).

The ledger in *Table 1* below lists each scope, its unit, the source status and the state the worksheet must record. Where no public source exists, the entry is **UNKNOWN**, which propagates into the result.

Scope	Quantity and unit	Source status as of November 2026	Worksheet state
Single GPU package	kg CO2e per GPU	No NVIDIA per-GPU PCF in the reviewed sources. An A100 card study gives 127.6 kg CO2e per GPU (arxiv.org). A parametric H100 model gives about 225 kg CO2e per card, against 164 kg from dividing the baseboard value by eight (arxiv.org)	Bounded (third-party, different method)
HGX H100 baseboard (8 GPUs)	kg CO2e per baseboard, cradle-to-gate	**1,312** (images.nvidia.com). Use and end-of-life excluded	Known, baseboard only
HGX B200 baseboard (8 GPUs)	kg CO2e per baseboard, cradle-to-gate	**2,274** (images.nvidia.com). Use and end-of-life excluded	Known, baseboard only
Host server platform (CPU, DRAM, chassis, PSU, NIC)	kg CO2e per server	A 1,780 kg CO2e base module is cited in one model (arxiv.org). The model is an academic estimate, not a vendor PCF for the purchased configuration	UNKNOWN until vendor supplies configured-product PCF

Scope	Quantity and unit	Source status as of November 2026	Worksheet state
Storage (SSD, HDD)	kg CO2e per TB or per drive	Architectural carbon model covers SSD and HDD by capacity but is described by its authors as a lower bound (hsienhsinlee.github.io)	Bounded or UNKNOWN
Network (switches, optics, cables)	kg CO2e per port or per rack	Google's TPU analysis excludes rack, shelf and network equipment from its totals (arxiv.org)	UNKNOWN
Facility changes (power, cooling, construction)	kg CO2e per kW of IT capacity	No product-level public figure in the reviewed sources	UNKNOWN unless a retrofit is required

Table 1 shows that only the two baseboard rows are backed by a vendor-published, reviewed number, and that every other row a procurement team needs is either a third-party estimate with a different boundary or empty. The practical consequence is that any payback computed from baseboard values alone is a **lower bound on upfront manufacturing carbon**, and therefore an **optimistic** (short) payback. A defensible report states that direction explicitly.

Change 3: Separate four kinds of carbon number

Four quantities answer four different questions, and mixing them is the most common source of error. *Table 2* below classifies them, with an example drawn from cited inputs.

Class	Unit	Example from cited inputs	Answers	Cannot answer
Absolute product carbon footprint	kg CO2e per purchased product	2,274 kg CO2e per HGX B200 baseboard (Table 1)	How much manufacturing carbon enters the buyer's Scope 3 on purchase	Whether replacement lowers total future emissions
Vendor-normalized intensity	g CO2e per unit of theoretical computation	0.66 to 0.50 gCO2e per exaflop, FP16, a 24% decrease	How the vendor's footprint compares per theoretical lifetime operation	Absolute savings or buyer workload emissions
Workload-allocated emissions (SCI style)	g CO2e per functional unit R	SCI is a rate: carbon emissions per one unit of R (sci.greensoftware.foundation)	How to apportion fleet emissions to a service or job	Whether to replace, because it spreads sunk emissions
Incremental replacement impact	kg CO2e over a stated horizon	New manufacturing minus cumulative operating difference	Whether replacing reduces future emissions within the horizon	Reporting-year inventory totals

The 24% figure deserves a specific caution. The intensity is expressed per exaflop and is estimated based on FP16 precision (developer.nvidia.com). A cross-check using the datasheet FP16 figures supports the reading that the ratio is driven by peak throughput. The B200 baseboard footprint is 1.73 times the H100 value (2,274 divided by 1,312, calculated), while the datasheet FP16 tensor throughput is 36 PFLOPS for the B200 baseboard (images.nvidia.com) against 16 PFLOPS for the H100 baseboard (images.nvidia.com) (both

datasheet figures state sparsity footnotes), a ratio of 2.25. Dividing 1.73 by 2.25 gives 0.77, close to the 0.50 to 0.66 ratio of 0.76 (calculated). This is an inference from rounded public numbers, not a statement of NVIDIA's method, but it shows why the normalized figure cannot stand in for a buyer's measured energy per request.

Change 4: Keep allocation accounting apart from the incremental decision

The Green Software Foundation's Software Carbon Intensity (SCI) specification, published as ISO/IEC 21031:2024 (webstore.iec.ch), allocates embodied emissions using M = TE * (TiR/EL) * (RR/ToR) (sci.greensoftware.foundation), where EL is the expected lifespan, the anticipated time that the equipment will be installed (sci.greensoftware.foundation). For grid electricity it requires short-run marginal, long-run marginal or average grid intensity (sci.greensoftware.foundation). This is an attributional allocation: it spreads fleet emissions over services. The GHG Protocol Product Standard likewise attributes emissions to the unit of analysis of the studied product (ghgprotocol.org), and an academic review characterizes attributional life cycle assessment as attributing a share of the potential environmental impact of the world to a product (research.vu.nl).

A keep-versus-replace choice is a consequential question: it asks what changes if a decision is made. The same review describes consequential assessment as assessing the environmental consequences of a decision, such as an increase of product demand (research.vu.nl), and the GHG Protocol notes consequential approaches include processes in the boundary to the extent they are expected to change (ghgprotocol.org). Practically, retaining equipment adds no new manufacturing emissions, because the manufacturing already happened. Charging an amortized share of it to the "keep" option and comparing against the "replace" option's full new manufacturing double counts the sunk carbon. The worksheet below therefore treats allocation (SCI) and incremental comparison as two separate calculations that share inputs but not totals.

A related reporting point: the GHG Protocol Scope 3 Standard says companies should not depreciate, discount or amortize the emissions from producing capital goods over time (ghgprotocol.org), so a purchase appears in full in the year of acquisition. Under that corporate-inventory convention a replacement year shows a spike, even when the decision view shows a net multi-year reduction. Both views are legitimate. They answer different questions and should be labeled.

Implementation Considerations and Process Changes

Step 1: Collect measured operating inputs for both options

For each option, record facility electricity under the service-equivalence statement, not rated power. Nameplate and design values are limits, not consumption. NVIDIA's DGX H100 guide describes six power supplies configured for 4+2 redundancy (docs.nvidia.com), its DGX B200 guide describes six in a 5+1 configuration (docs.nvidia.com), the H100 product page lists a maximum thermal design power of up to 700 W (configurable) (www.nvidia.com), and the Blackwell datasheet lists configurable up to 1,000 W for HGX B200 (dam-cdn.nvd.orangelogic.com). None of these is measured energy for any buyer's workload, and none should enter the worksheet except as an upper-bound sanity check.

Measure at the wall or rack power distribution unit, over a period that includes realistic idle and reserve time, and record:

- **IT energy** in kWh per defined period for each option, at equal delivered service.
- **Replica count and utilization**, including redundancy that exists only for availability.
- **Idle and standby energy** for hardware that is powered but not serving.
- **PUE** or direct facility energy, because the carbon factor applies to facility kWh. Uptime Institute's 2026 survey reports an industry average PUE of 1.52 (datacenter.uptimeinstitute.com), LBNL estimates a U.S. average of 1.4 in 2023 (escholarship.org), and Google reports a 2025 fleet average annual PUE of 1.09 for its own facilities (datacenters.google). Use the actual site value, since averages vary by facility class.
- **Measurement dates and conditions**, so the energy ledger can be re-run.

Step 2: Choose and date the grid factor

Pick a factor deliberately and record its source, geography, year, unit and method. The GHG Protocol Scope 2 Guidance defines the location-based method as reflecting the average emissions of the grids where consumption occurs (ghgprotocol.org) and the market-based method as reflecting emissions from electricity that companies have chosen (ghgprotocol.org). Companies with operations in markets providing contractual instruments shall report scope 2 emissions in two ways (ghgprotocol.org). A replacement decision should show results under both so the reader can see sensitivity to procurement claims.

Factors are versioned and lag the real grid. The guidance itself flags temporal representativeness problems due to time delays (ghgprotocol.org). EPA's eGRID2023 was revised in June 2025 (www.epa.gov), and EPA's Power Profiler still reports 2023 eGRID data (www.epa.gov). In the UK, the 2026 conversion-factor edition cut the electricity factor, with most of the decrease attributed to grid decarbonization (assets.publishing.service.gov.uk). Globally, Ember reports an average of 458 gCO2e/kWh in 2025 (ember-energy.org) and a 14% decline over the last decade (ember-energy.org), on a CO2-equivalent basis (ember-energy.org), while the IEA's figure is CO2 only, calculated as total CO2 emissions divided by total generation (www.iea.org). The two are not interchangeable.

Step 3: Compute incremental manufacturing and operating differences

Define, with explicit units:

- **ΔM** (kg CO2e) = sum of new purchased product-scope footprints (baseboards, host servers, network, storage, installation, any facility retrofit) minus a documented reuse credit, if and only if the retired equipment has a verified second use. By default the credit is zero and the retained equipment contributes no new manufacturing term.
- **O(option, year)** (kg CO2e) = facility kWh in that year × grid factor (kg CO2e/kWh) for that year.
- **ΔO(year)** = O(keep, year) minus O(replace, year).
- **Cumulative position** C(t) = ΔM minus the sum of ΔO over years 1 to t. Payback occurs in the first year where C(t) is zero or below.
- **Simple payback** T = ΔM divided by annual ΔO, valid only when ΔO is positive and constant. If ΔO is zero or negative, there is **no payback**. If any input is unknown, the result is **unknown**.

Unit check: kg CO2e divided by (kWh per year × kg CO2e per kWh) leaves years. Facility kWh already includes PUE, so PUE must be applied once.

Step 4: Sweep the assumptions that move the answer

Vary each of the following one at a time, then in combination, and report the range rather than a point:

- **Utilization:** average and peak load, including idle hours.
- **Remaining life:** of the retained fleet, including repair and maintenance exposure.
- **Replacement life:** the planned service period of the new fleet.
- **PUE:** site-specific, before and after any facility change.
- **Grid trajectory:** dated factors for each year, under both Scope 2 methods.
- **Redeployment and end-of-life:** reuse credit at zero by default.

The published literature shows why the sweep matters: modeled outcomes swing widely with grid carbon intensity and lifetime. Four studies illustrate the spread:

- A bottom-up study of GPU platforms finds that, in many medium-to-high carbon-intensity countries, replacing platforms older than four years could reduce projected emissions by at least 75% (arxiv.org), yet it also reports that current platform prices discourage the timely replacement of inefficient hardware (arxiv.org).
- A database-server study finds that replacement can have a longer breakeven period, especially in regions with low-carbon electricity (www.vldb.org), and that break-even extends by 4× from Poland to the UK and by an additional 9× (www.vldb.org).
- A hyperscale systems paper warns that upgrading every cloud server every two years would cause high embodied emissions (www.microsoft.com), and that matching savings by extending lifetimes to 13 years requires a radical data center stack redesign (www.microsoft.com).
- A scheduling study reports that commonly used periodic replacement strategies are often sub-optimal in terms of total carbon emissions (par.nsf.gov).

Step 5: Record redeployment and end-of-life explicitly

If retired GPUs are redeployed internally to displace a purchase that would otherwise occur, that avoided purchase is a legitimate part of ΔM, but it must be documented as a separate line. If they are resold, the effect on other buyers' purchases is uncertain, and the GHG Protocol requires that estimates of avoided emissions be reported separately from the scope 1, 2 and 3 inventory (ghgprotocol.org). The ITU-T L.1410 recommendation, an ICT life cycle assessment framework (www.itu.int), includes application scenarios for goods with extended operating lifetime and multiple life cycles (www.itu.int). In the European Union, the server ecodesign regulation requires that joining techniques not prevent disassembly for repair or reuse (eur-lex.europa.eu) and that a functionality for secure data deletion be made available (eur-lex.europa.eu), which supports a documented reuse path. Default the reuse credit to zero unless evidence is on file.

Step 6: Publish a versioned ledger

Store every input with source, date, unit and owner, and re-run the worksheet whenever a grid factor, PCF revision, utilization measurement or price changes. The *Evidence request* list in the Implications section below gives the minimum vendor documents.

Data Analysis and Evidence

Break-even energy from cited inputs

The following figures are calculated from cited inputs and carry the baseboard-only caveat from *Table 1*. The U.S. factor is EPA's eGRID2023 total output emission rate, published in a summary-tables document (www.epa.gov). Converting its U.S. value of 770.9 lb CO2e/MWh from pounds per megawatt-hour to kilograms per kilowatt-hour gives 0.350 kg CO2e/kWh (calculated). The IEA global estimate is 0.435 and Ember's is 0.458 (both kg per kWh, converted from the g/kWh figures cited above).

- **U.S. eGRID2023 average (0.350 kg/kWh):** 2,274 kg ÷ 0.350 = about **6,500 kWh** of avoided facility electricity to repay one B200 baseboard's manufacturing carbon.
- **IEA global 2025 (0.435 kg/kWh):** about **5,200 kWh**.
- **Ember global 2025 (0.458 kg/kWh):** about **5,000 kWh**.

For scale only: NVIDIA's H100 summary lists a typical baseboard power consumption of 5,600 W as a specification (images.nvidia.com). Running that continuously for a year would use about 49,000 kWh, so 6,500 kWh is roughly 13% of that nameplate figure (calculated). This comparison uses a rated value, not a measurement, and is shown only to illustrate magnitude. A replacement that does not cut facility energy by a double-digit percentage of the old baseboard's actual consumption cannot repay baseboard-only manufacturing carbon in about one year at the U.S. average factor, and a full-server boundary raises the bar further.

Hypothetical payback sweep

Table 3 below is a **hypothetical** sweep, not a finding. It applies T = ΔM / (annual avoided facility kWh × grid factor) with ΔM fixed at 2,274 kg CO2e (baseboard-only lower bound, one B200 baseboard purchased, zero reuse credit). The avoided-energy values are arbitrary placeholders to show sensitivity.

Hypothetical annual avoided facility energy per replaced baseboard	U.S. eGRID2023 (0.350 kg/kWh)	IEA global 2025 (0.435 kg/kWh)	Ember global 2025 (0.458 kg/kWh)
2,000 kWh	3.25 years	2.61 years	2.48 years
5,000 kWh	1.30 years	1.05 years	0.99 years
10,000 kWh	0.65 years	0.52 years	0.50 years
20,000 kWh	0.32 years	0.26 years	0.25 years
Zero or negative (more replicas, higher utilization, unmeasured)	**No payback**	**No payback**	**No payback**
Any component of ΔM or ΔE unknown	**Unknown**	**Unknown**	**Unknown**

Table 3 shows that the grid factor changes the answer by roughly 30% across the three cited values, while the avoided-energy input changes it by an order of magnitude. That asymmetry is the argument for measuring equal-service energy before debating factors. The last two rows are not edge cases: if the replacement fleet is

sized with additional replicas or runs at higher utilization to meet the same objective, avoided energy can be zero or negative. Because the table uses baseboard-only ΔM, every payback shown is shorter than a full-server calculation would give.

Grid trajectory example (hypothetical)

Constant factors overstate savings when the grid is decarbonizing. The IEA projects global intensity falling from 435 g CO2/kWh in 2025 to 360 g CO2/kWh in 2030 (www.iea.org). As a **hypothetical** illustration, interpolate linearly (0.435, 0.420, 0.405, 0.390, 0.375 kg/kWh for years 1 to 5) and assume 2,000 kWh avoided per year. Annual avoided emissions are 870, 840, 810, 780 and 750 kg CO2e, cumulative 870, 1,710, 2,520, 3,300 and 4,050. The cumulative position crosses ΔM = 2,274 kg during year 3, at about 2.7 years, against 2.61 years with a constant factor (calculated). The effect is small here but grows with longer paybacks.

Allocation example (hypothetical)

An SCI-style allocation for a workload that reserves 10% of a baseboard's resources for one year of a hypothetical 4-year expected lifespan gives M = 2,274 × (1/4) × 0.10 = 56.85 kg CO2e (calculated, baseboard-only). This number belongs in a service-level carbon report. It should not be subtracted from, or added to, the incremental replacement comparison, because the retained baseboard's manufacturing carbon is already sunk.

Lifetime and economic context

Lifetime assumptions vary by source and purpose, and accounting lives are not engineering lifetimes. Company filings give financial depreciation estimates for servers:

- **Alphabet:** changed to six years effective fiscal 2023 (www.sec.gov).
- **Microsoft:** completed a July 2022 assessment of server and network equipment lives, raising them from four to six years effective fiscal 2023 (www.sec.gov).
- **Amazon:** shortened the life of a subset of servers and networking equipment effective January 1, 2025 (www.sec.gov), citing the increased pace of technology development (www.sec.gov).
- **Meta:** extended certain servers and network assets to 5.5 years (www.sec.gov).
- **Oracle:** increased the estimate of server and networking lives from five years to six years (www.sec.gov).
- **CoreWeave:** a change from five to six years for computing equipment is recorded in its accounting policies (www.sec.gov).
- **LBNL:** models average server lifetime reaching 5.7 years by 2023 (escholarship.org).

Research assumptions are shorter: an architectural carbon model cites typical server-processor lifetimes of 3 to 5 years (hsienhsinlee.github.io), and an early hyperscale paper notes data centers typically maintain server-class CPUs for three to four years (arxiv.org). Google's TPU life cycle assessment finds operational emissions dominate AI hardware lifetime at 70% or more (arxiv.org), a result specific to its hardware, location and electricity procurement.

On cost, the U.S. Energy Information Administration reports average commercial electricity at 14.47 cents/kWh and industrial at 9.64 cents/kWh in August 2026 (preliminary) (www.eia.gov). In the bottom-up GPU platform study, a five-year upgrade cycle would need an additional carbon incentive of about 87 USD/tCO2e on average to be justified on carbon grounds (arxiv.org), and the same study puts the global average carbon price at 32 USD/tCO2e. For scale, valuing the 2.274 tonnes of one B200 baseboard at 32 USD/tCO2e gives about $73 (calculated), small next to hardware prices, which is one reason economic and carbon logic can diverge.

A total cost of ownership (TCO) calculator for the same decision uses the same ledger inputs:

- **Capital:** purchase price of the replacement configuration minus resale value of retired units.
- **Energy cost:** facility kWh per year × tariff, using the buyer's actual contract rather than a national average.
- **Operations:** support, spares, space and staff time, which differ for older and newer fleets.
- **Carbon term (optional):** ΔM and ΔO multiplied by an internal carbon price, reported separately from cash cost.

Implications and Future Directions

What a defensible answer requires

A defensible keep-or-replace carbon statement needs four things the public record does not currently provide together: a configured-product footprint for the whole purchased bill of materials, a stated reuse and end-of-life treatment, measured equal-service energy for both fleets, and dated grid factors under both location-based and market-based methods. The baseboard summaries are rigorous within their scope. NVIDIA reports primary supplier data for components representing over 92% of the H100 baseboard by weight (images.nvidia.com), and a similar coverage statement for the B200 baseboard. The scope is nonetheless a baseboard. Outside the vendor documents, Computer Weekly's 2026 feature on measuring AI's carbon footprint reports that Gartner recommends accounting for the life cycle of datacentre equipment (computerweekly.com), consistent with the ledger approach here, though the feature gives no embodied-carbon values.

Evidence request for vendors and integrators

Before approving a replacement purchase, request in writing:

- **Configured-product PCF** for the exact server SKU, with CPUs, memory, storage, NICs, power supplies and chassis itemized, plus revision date and reviewer.
- **Boundary statement** naming what is excluded (thermal solutions, interconnects, packaging, cables, optics).
- **Method and databases** used, with the standard followed. Dell describes a streamlined PAIA method developed by MIT for its server footprints (www.delltechnologies.com) and notes its use phase assumes a PUE of 1.0 (www.delltechnologies.com), so use-phase values must be rescaled.
- **Uncertainty range**, such as percentile bounds, rather than a single point value.
- **Network, storage and rack bill of materials** with footprints, or a statement that none exist.
- **Reuse, refurbishment and take-back terms** for the retired fleet, including data-deletion support.

- **Energy per unit of service** measured on the buyer's workload under a witnessed protocol, not peak or rated figures.

Versioned calculation ledger

Record each run in a structure such as:

```
run_id: 2026-11-07-r1
service_statement: <workload, SLO, availability>
horizon_years: <min of remaining life, replacement life, plan>
grid_factor: {source, geography, year, unit, method}
deltaM: {baseboard: <kg>, host: <kg|UNKNOWN>, network: <kg|UNKNOWN>, storage: <kg|UNKNOWN>, facility:
<kg|UNKNOWN>, reuse_credit: 0}
energy_keep_kwh_per_year: <measured>
energy_replace_kwh_per_year: <measured>
pue: <site value>
result: payback_years | NO_PAYBACK | UNKNOWN
```

Where the field is heading

Three developments will change the worksheet. First, Scope 2 and product-carbon guidance are evolving, and grid datasets are re-issued with a lag, so ledgers must be re-run annually. Second, as grids decarbonize, operational savings shrink while embodied carbon stays fixed, tilting the balance toward retaining hardware. A model of server replacement finds that in high-carbon grids faster replacement is better as energy efficiency reduces emissions (arxiv.org), which also implies the opposite in low-carbon grids. Third, vendors may publish configured-system footprints, which would turn many UNKNOWN cells into bounded ones. Teams that want support assembling the evidence request or the measurement plan can consult an advisor such as GPU Smith, the publisher of this report. This report is not lifecycle-specialist reviewed, and a qualified practitioner should confirm boundaries and factor choices before external disclosure.

Frequently Asked Questions (FAQs)

When should GPU servers be replaced? When the incremental calculation shows cumulative operating savings exceed ΔM inside the shortest of the old fleet's remaining life, the replacement's life and the planning horizon, and when equal-service energy has been measured. Otherwise the answer is keep, or unknown.

How does the GPU refresh cycle affect carbon payback? Shorter cycles add manufacturing carbon more often. A hyperscale analysis warns that replacing every two years would cause high embodied emissions (www.microsoft.com). Longer cycles delay efficiency gains. The optimum depends on grid intensity and measured efficiency gains.

Is embodied carbon larger than operational carbon for GPUs? It depends on boundary, grid and life. Google's TPU analysis finds operational emissions dominate at 70% or more (arxiv.org), while an early hyperscale analysis finds manufacturing dominates where electricity is largely renewable (arxiv.org).

What is a useful life for a GPU server? Financial lives in company filings cluster at five to six years (www.sec.gov), and research models use three to five (hsienhsinlee.github.io). Treat the planning horizon as an input, and test sensitivity.

Can the 24% figure be used as the carbon saving of replacing H100 with B200? No. It is a vendor-normalized intensity per exaflop, not a measure of a buyer's energy or absolute emissions.

How is a total cost of ownership calculator related? Total cost of ownership adds capital, energy cost (kWh × tariff), maintenance and residual value. The carbon worksheet uses the same energy and horizon inputs, so run both from one ledger.

Conclusion

Keep-or-replace decisions for working GPU fleets turn on a small set of auditable quantities: the manufacturing carbon of the exact configuration to be purchased, the measured difference in facility energy at equal service, the dated grid factor, and the planning horizon. Published baseboard footprints of 1,312 and 2,274 kg CO2e are rigorous inputs for what they cover, and they make clear what is missing: host servers, networking, storage and facilities. Normalized intensity claims and rated power figures cannot fill those gaps.

The worksheet treats allocation and incremental decisions separately, never charges sunk manufacturing to the retained fleet, and returns "no payback" or "unknown" when the evidence does not support a number. Published models point in both directions, with large reductions in carbon-intensive grids and long break-even in low-carbon ones. A buyer who obtains configured-product footprints and witnessed equal-service energy measurements can turn this worksheet into a decision. Until then, the honest result is a bounded interval with the missing inputs named.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://images.nvidia.com/aem-dam/Solutions/documents/HGX-H100-PCF-Summary.pdf
2. https://images.nvidia.com/aem-dam/Solutions/documents/HGX-B200-PCF-Summary.pdf
3. https://developer.nvidia.com/blog/nvidia-hgx-b200-reduces-embodied-carbon-emissions-intensity
4. https://mlcommons.org/benchmarks/inference-datacenter
5. https://www.epa.gov/egrid/power-profiler
6. https://www.iea.org/reports/electricity-2026/emissions
7. https://arxiv.org/pdf/2507.08923
8. https://www.vldb.org/pvldb/vol18/p5223-rabl.pdf
9. https://calnext.com/wp-content/uploads/2026/01/ET25SWE0031_Embodied-Carbon-One-for-One-Material-Substitution-Market-Characterization_Final-Report.pdf
10. https://escholarship.org/content/qt32d6m0d1/qt32d6m0d1.pdf
11. https://arxiv.org/pdf/2011.02839
12. https://arxiv.org/html/2410.12032v2

13. https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-r650.pdf
14. https://static.lenovo.com/ww/docs/regulatory/eco-declaration/2022/pcf-thinksystem-st50.pdf
15. https://arxiv.org/html/2509.00093v3
16. https://arxiv.org/pdf/2607.01258
17. https://hsienhsinlee.github.io/MARS/pub/isca2022.pdf
18. https://arxiv.org/pdf/2502.01671
19. https://sci.greensoftware.foundation
20. https://webstore.iec.ch/publication/93436
21. https://ghgprotocol.org/sites/default/files/standards/Product-Life-Cycle-Accounting-Reporting-Standard_041613.pdf
22. https://research.vu.nl/en/publications/attributional-amp-consequential-life-cycle-assessment-definitions
23. https://ghgprotocol.org/sites/default/files/standards/Corporate-Value-Chain-Accounting-Reporing-Standard_041613_2.pdf
24. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html
25. https://docs.nvidia.com/dgx/dgxb200-user-guide/introduction-to-dgxb200.html
26. https://www.nvidia.com/en-us/data-center/h100
27. https://dam-cdn.nvd.orangelogic.com/AssetLink/y441155802qub41q118b2852i557jem5.pdf
28. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/UptimeInstitute.GlobalDataCenterSurvey.2026.pdf
29. https://datacenters.google/efficiency
30. https://ghgprotocol.org/sites/default/files/standards/Scope%202%20Guidance_Final_Sept26.pdf
31. https://www.epa.gov/egrid/known-issues-data-notes-and-revisions
32. https://assets.publishing.service.gov.uk/media/6a2940653b15d05a7ce3202f/2026-GHG-conversion-factors-major-changes-report.pdf
33. https://ember-energy.org/latest-insights/global-electricity-review-2026/2025-in-review
34. https://ember-energy.org/latest-insights/global-electricity-review-2026/electricity-demand-and-supply-trends
35. https://ember-energy.org/latest-insights/global-electricity-review-2026/supporting-materials
36. https://www.microsoft.com/en-us/research/wp-content/uploads/2024/03/2024-GreenSKU-ISCA2024.pdf
37. https://par.nsf.gov//servlets/purl/10616737
38. https://www.itu.int/dms_pubrec/itu-t/rec/l/T-REC-L.1410-202411-I!!SUM-HTM-E.htm
39. https://www.itu.int/dms_pubrec/itu-t/rec/l/T-REC-L.1410-202411-I!!TOC-HTM-E.htm
40. https://eur-lex.europa.eu/legal-content/EN/TXT?uri=CELEX:32019R0424
41. https://www.epa.gov/system/files/documents/2025-06/summary_tables_rev2.pdf
42. https://www.sec.gov/Archives/edgar/data/1652044/000165204424000022/goog-20231231.htm
43. https://www.sec.gov/Archives/edgar/data/789019/000156459022026876/R27.htm
44. https://www.sec.gov/Archives/edgar/data/1018724/000101872426000004/amzn-20251231.htm
45. https://www.sec.gov/Archives/edgar/data/1326801/000132680125000017/meta-20241231.htm
46. https://www.sec.gov/Archives/edgar/data/1341439/000095017025087926/R50.htm
47. https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/R9.htm
48. https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a

49. https://computerweekly.com/feature/Getting-started-with-measuring-AIs-carbon-footprint

50. https://arxiv.org/pdf/2506.14365

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.