GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

GPU Cluster Power Consumption Calculator: PUE and MWh

> 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 · 2026-09-19 · gpu cluster power consumption calculator · gpu cluster energy calculator · annual mwh · power usage effectiveness · pue calculator · data center power · gpu infrastructure · electricity cost · energy metering

Original article: https://gpusmith.com/articles/en/gpu-cluster-energy-calculator

In this article

Executive Summary

A useful **GPU cluster power consumption calculator** is an auditable energy model, not a multiplication of graphics processing unit (GPU) thermal design power by device count. It must distinguish **peak power in kilowatts (kW)** from **energy in kilowatt-hours (kWh) or megawatt-hours (MWh)**. The U.S. Energy Information Administration explains that watts measure power at a moment, whereas watt-hours measure energy over time (eia.gov). For every operating phase, the core equation is **IT MWh = IT kW x hours / 1,000**. A non-leap year has **8,760 hours**, but calendar years may have 365 or 366 days (aa.usno.navy.mil).

The calculator should begin with a dated inventory of servers, accelerators, switches, storage and management equipment. Vendor limits must remain distinct: NVIDIA specifies **10.2 kW maximum system input** for a DGX H100/H200 at 200 to 240 V AC (docs.nvidia.com), while AMD lists **750 W peak Typical Board Power** for one MI300X accelerator (amd.com). Those are different boundaries. Installed redundant power-supply ratings are also not operating demand: the DGX has six supplies in a **4+2** arrangement (docs.nvidia.com).

After inventory, model training, inference, idle, maintenance and ramp phases so their hours equal the analysis period. Apply **Power Usage Effectiveness (PUE)** only after calculating IT energy. The Green Grid defines PUE as facility-entering energy divided by IT-equipment energy (thegreengrid.org). ISO defines annual PUE over a continuous **12-month** period (iso.org), and DOE guidance says PUE does not address IT efficiency (eere.energy.gov). Therefore, annual facility energy is **IT MWh x matched-period annual PUE**, not evidence that one workload or server is efficient.

Cost and emissions remain optional modules. Energy charges use kWh, while demand charges commonly use the maximum kW in a billing period (energy.gov). Emissions should use the facility's applicable grid region, because EPA directs users to the eGRID subregion where the facility is located (epa.gov). The result is a scenario worksheet, not a universal forecast.

Introduction and Background

Facilities teams often receive a proposed GPU count before they receive a defensible energy budget. That ordering encourages three category errors: treating component thermal limits as whole-system input, treating peak kW as annual MWh, and treating PUE as a universal constant. A calculator can prevent all three if its inputs are dated, boundary-defined and assigned to an owner.

The planning question is specific: **How much IT and facility energy will this identified cluster consume during this identified period under this workload schedule?** That is different from asking whether the utility can serve the peak, how much cooling equipment must be installed, or how much electricity all artificial intelligence data centers may consume. Berkeley Lab reported that U.S. data-center electricity use rose from **58 TWh in 2014 to 176 TWh in 2023** (eta.lbl.gov), but that national total is context, not a site input.

In practical terms, the same workbook serves as a **GPU cluster energy consumption calculator**, **GPU cluster annual MWh calculator**, **GPU cluster PUE calculation**, and **GPU server load factor energy-use worksheet**. Those labels describe different views of one governed data model, not separate estimates.

The calculator proposed here uses four evidence labels:

- **Measured:** Interval data from a calibrated meter or device telemetry, with timestamps and coverage recorded.
- **Vendor-specified:** A dated system or component specification linked to the exact model and configuration.
- **Buyer-supplied:** A tariff, deployment date, operating schedule, contract value or site constraint supplied by the decision owner.
- **Hypothetical:** A planning value used solely to explore a scenario, visibly separated from observed data.

GPU Smith is an **adjacent independent engineering advisor**, not a chip manufacturer, utility or cloud provider. Its public scope includes capacity planning and telemetry for existing facilities and colocation (gpusmith.com), and its assessment deliverables include a power, space, cooling and network requirement schedule (gpusmith.com). That perspective belongs in the review process, not as a vendor row in a product comparison.

Key Changes

Change 1: Separate capacity from energy

Power is a rate. **Energy** is the accumulation of that rate over time. One kWh is one kilowatt supplied for one hour (eia.gov). Consequently:

- **Peak IT kW** informs switchgear, uninterruptible power supply (UPS), power distribution unit (PDU), busway and utility-capacity checks.
- **Average IT kW** describes mean demand during a stated phase or reporting interval.
- **IT kWh or MWh** supports energy budgets after duration is applied.
- **Facility kWh or MWh** adds cooling and power-delivery overhead within a stated boundary.
- **Demand kW** supports tariff calculations and may depend on the utility's billing interval.

Never label a kW result “annual consumption.” Conversely, an annual MWh result does not establish that the facility can serve a short peak. DOE's PUE Category 0 is a demand-based peak calculation during a 12-month measurement period (eere.energy.gov), while Category 1 uses **12 consecutive months of energy data** (eere.energy.gov). These answer different questions.

Change 2: Inventory the whole IT boundary

A GPU-only sum omits CPUs, memory, fans, local storage, network interface cards, switches, management nodes and storage. NVIDIA's scalable-unit design, for example, requires two management racks for fabric, management and storage infrastructure (docs.nvidia.com).

For every item, record both the applicable maximum and the best available measured series. Specifications illustrate why this discipline matters:

- **H200 component limit:** NVIDIA lists configurable maximum thermal design power up to **700 W** for H200 SXM and **600 W** for H200 NVL (nvidia.com).
- **DGX system limit:** NVIDIA lists **10.2 kW maximum** for a complete DGX H100/H200 system (docs.nvidia.com).
- **Switch operating range:** Arista lists **640 W typical and 2,218 W maximum** for two 7060X6 configurations (arista.com).
- **Test condition:** Arista's typical switch power uses 25 C ambient and 50 percent load (arista.com).
- **Another switch boundary:** Cisco lists **2,029 W typical and 2,960 W maximum** for the Nexus 9364E-SG2-Q (cisco.com).

Do not sum the ratings on redundant power supplies. Cisco notes that PSU capability can exceed a switch's maximum requirement (cisco.com).

Change 3: Replace one utilization factor with a phase schedule

A single annual “GPU utilization” percentage hides material operating states. Training, batch inference, latency-sensitive inference, checkpointing, data preparation, idle reservation, maintenance and commissioning can have different system power. SPEC defines **active idle** as an interval with no workload transactions scheduled (spec.org).

The schedule should include:

- **Start and end:** Dates or timestamps for each phase.
- **Hours:** Exact duration, with an automated check that phases sum to the analysis period.
- **Inventory active:** Which systems, switches and storage devices are energized.
- **Power basis:** Measured average kW, modeled kW, or maximum kW multiplied by an explicitly hypothetical load factor.
- **Ramp treatment:** Deployment waves and commissioning loads rather than assuming the full fleet existed all year.

- **Confidence:** High, medium or low, defined by the organization.

NVIDIA-SMI can report a one-second average of entire-board GPU power on supported devices (docs.nvidia.com). That is useful telemetry, but it still omits the host and facility boundary. NVIDIA-SMI can also set a maximum power limit, so the inventory should capture configured caps rather than assume vendor defaults (docs.nvidia.com).

Change 4: Treat PUE as a boundary-defined ratio

For a matched period and boundary:

PUE = total data-center energy / IT-equipment energy

Facility MWh = IT MWh x PUE

Overhead MWh = facility MWh - IT MWh

ISO includes equipment used to store, process and transport data in IT energy (iso.org). EU reporting rules include electricity, fuels and other energy sources used for cooling in total facility energy (eur-lex.europa.eu). Renewable electricity remains energy in the PUE numerator rather than an efficiency credit (eere.energy.gov).

The calculator must reject mismatched periods. ISO calls measurements covering other than continuous 12 months **interim PUE** (iso.org). It also sets no universal PUE target (iso.org). Climate, reliability design, load level and meter location can make cross-site comparisons misleading.

Implementation Considerations and Process Changes

Build the provenance-first input sheet

Table 1 defines the minimum auditable record. Blank is preferable to an invented default.

Field	Value and unit	Required provenance	Label and owner	Quality check
Server quantity and model	Count, exact configuration	Purchase bill of materials and dated vendor document	Vendor-specified, procurement	Match installed serial inventory
Maximum system input	kW per server	Whole-system electrical specification	Vendor-specified, facilities	Do not substitute GPU TDP
Phase average power	kW by server group	Interval meter or telemetry series	Measured, operations	State interval and coverage
Network, storage, management	kW by device group	Device meter or dated specification	Measured or vendor-specified, network owner	Include optics and active redundancy
Schedule	Start, end and hours	Deployment plan and workload calendar	Buyer-supplied, workload owner	Sum to exact period
PUE	Ratio and period	Facility meters with boundary diagram	Measured or buyer-supplied, facilities	Numerator and denominator periods match
Tariff	$/kWh, $/kW and clauses	Executed tariff or utility schedule	Buyer-supplied, finance	Capture time-of-use and ratchet rules
Emissions factor	kg CO2e/MWh and region	Dated grid-factor source	Buyer-supplied, sustainability	State location-based or market-based method
Confidence	Defined category	Review policy	Assumption, model owner	Record reason and next evidence action

The table makes review ownership explicit. A value can be numerically precise yet still weak if its configuration, date or boundary is unknown. It also prevents a vendor maximum from silently becoming a measured average.

Establish meter boundaries before collecting data

At facility level, diagram the utility handoff, generators, on-site energy, UPS inputs and outputs, PDUs, rack PDUs and device telemetry. EU rules place total-energy measurement before supply transfer switchgear (eur-lex.europa.eu) and require annual IT energy measurement at every UPS connected to IT equipment (eur-lex.europa.eu). DOE Category 2 places IT measurement at the output of the supporting PDU (eere.energy.gov).

Mixed-use sites require allocation rules. Green Grid guidance locates the control-volume boundary at the landlord or shared-facilities handoff (datacenters.lbl.gov). Additional monitoring points may be required even though the measurement method remains the same (datacenters.lbl.gov).

Commission and reconcile the measurement chain

The measurement plan should specify:

- **Meter ID and location:** Physical point, circuit, device group and expected direction of flow.
- **Interval:** Consistent timestamps and time zone across IT and facility systems.
- **Accuracy:** Meter class, current-transformer ratio, calibration status and expected uncertainty.
- **Completeness:** Expected intervals, missing intervals, estimated intervals and exclusion rules.
- **Reconciliation:** Parent meter compared with the sum of child meters within a documented tolerance.
- **Change control:** Firmware, configuration, power-cap and inventory changes recorded with effective timestamps.

DOE describes commissioning as an end-to-end check from sensors through visualization (betterbuildingssolutioncenter.energy.gov). It recommends portable metering in parallel with installed metering (betterbuildingssolutioncenter.energy.gov). A spot check alone is insufficient because accumulation errors converting kW to kWh may remain hidden (betterbuildingssolutioncenter.energy.gov).

EPA allows monthly IT energy to be derived from readings every **15 minutes or more frequently** (energystar.gov). ENERGY STAR defines a data gap as a date not covered by a bill (portfoliomanager.energystar.gov) and explicitly flags meters containing estimated usage (portfoliomanager.energystar.gov). The calculator should do the same rather than silently interpolate.

Calculator Architecture and Formulas

Step-by-step calculation

1. **Set the period.** Store inclusive start, exclusive end, time zone and expected hours.
2. **Freeze an inventory version.** Record deployment and removal timestamps for every device group.
3. **Calculate phase IT kW.** Sum measured group averages, or sum maximum system input multiplied by buyer-supplied phase factors.
4. **Calculate phase IT energy.** `IT_MWh_phase = IT_kW_phase x hours_phase / 1,000`.
5. **Aggregate IT energy.** `IT_MWh = sum(IT_MWh_phase)`.
6. **Apply matched PUE.** `Facility_MWh = IT_MWh x annual_PUE` only when boundary and period align.
7. **Calculate overhead.** `Overhead_MWh = Facility_MWh - IT_MWh`.
8. **Run capacity separately.** Compare peak coincident IT and facility demand against design and contract limits without substituting annual PUE for a peak engineering model.
9. **Add cost if sourced.** Calculate time-of-use energy blocks and monthly demand charges from the buyer's tariff.
10. **Add emissions if sourced.** Multiply location-specific MWh by the applicable dated factor and report methodology.

The data model should preserve both raw and transformed fields. Dell telemetry can separate CPU, memory, fan, PCIe and storage power, with subsystem metrics sampled every five seconds (dell.com). Redfish standardizes properties for power-consumption metrics (redfish.dmtf.org). These streams can inform group averages, but retaining source resolution prevents false precision in the final annual total.

Cost module

The minimum cost formula is:

Energy cost = sum(facility kWh in tariff block x $/kWh)

Demand cost = sum(monthly billed demand kW x $/kW)

Total electricity cost = energy cost + demand cost + fixed charges + taxes and adjustments

Energy charges are based on kWh consumed (energy.gov). Delivered voltage often determines the applicable rate first (energy.gov). A ratchet may allow one peak to influence demand charges for the rest of the year (energy.gov).

Do not use a national average as a site budget. EIA reported a **2025 U.S. average retail price of 13.63 cents/kWh**, but state averages ranged from **8.20 cents in North Dakota to 35.72 cents in Hawaii** (eia.gov) (eia.gov). EIA directs users to utility tariff schedules and demand charges (eia.gov).

Emissions module

The simplest location-based equation is:

Emissions tCO2e = facility MWh x kg CO2e/MWh / 1,000

EPA recommends total-output emission rates in lb/MWh when estimating emissions attributable to electricity use (epa.gov). Its guidance also converts kWh to MWh by dividing by 1,000 (epa.gov). The Greenhouse Gas Protocol says a location-based method reflects average emissions (ghgprotocol.org). Where qualifying contractual instruments exist, it requires scope 2 reporting in two ways (ghgprotocol.org). Renewable claims therefore belong in emissions accounting, not as a PUE adjustment.

Portable CSV schema

A downloadable implementation can use six related CSV files:

- **inventory.csv:** `asset_group,model,quantity,in_service,out_service,max_input_kw,source_url,source_date,label,owner`
- **telemetry.csv:** `timestamp,asset_group,average_kw,interval_minutes,meter_id,quality_flag`
- **phases.csv:** `phase,start,end,hours,asset_group,power_basis,load_factor,average_kw,it_mwh`
- **pue.csv:** `period_start,period_end,total_energy_mwh,it_energy_mwh,pue,boundary_id,measurement_category`
- **tariff.csv:** `effective_date,block,start_time,end_time,energy_rate_per_kwh,demand_rate_per_kw,fixed_charge`
- **scenario.csv:** `scenario,input_name,value,unit,label,confidence,owner,review_date`

Validation rules should reject overlapping phases for the same asset group, unexplained gaps, percentages outside 0 to 100, negative energy, PUE below 1 without an explicit review, and period mismatches. They should warn, not silently repair, when a source date predates the modeled hardware configuration.

Data Analysis and Evidence

This section demonstrates the arithmetic with a **fictional hypothetical example**. It is not a benchmark, forecast, vendor claim or recommended default. The example assumes 100 servers at a vendor maximum of 10.2 kW each, plus separate network and management loads. Its four phases total exactly **8,760 hours**.

Table 2 calculates annual IT energy by operating phase.

Phase (Hypothetical Example)	Hours	Server load basis	Other IT kW	Total IT kW	IT MWh
Training	4,000	82% of 1,020 kW	30	866.4	3,465.60
Inference	3,000	55% of 1,020 kW	30	591.0	1,773.00
Idle reservation	1,500	18% of 1,020 kW	20	203.6	305.40
Maintenance	260	10% of 1,020 kW	10	112.0	29.12
Total	**8,760**	**Varies by phase**	**Varies**	**Not additive**	**5,573.12**

The result is **5,573.12 IT MWh**. A nameplate-only upper planning case, with all 100 servers and 30 kW of other IT held at maximum for every hour, would be **9,198 IT MWh**. The large difference is why the model should show nameplate and workload-scheduled scenarios separately.

Assume, solely for this hypothetical example, a matched annual PUE of **1.25**. Facility energy is **6,966.40 MWh**, and overhead is **1,393.28 MWh**. That ratio must not be carried into another facility without its measurement boundary. DOE guidance specifically cautions that climate and reliability levels matter when comparing PUE (eere.energy.gov).

Table 3 shows arithmetic sensitivity around the fictional base case. These are user-selected deltas, not observed benchmarks.

Scenario (Hypothetical Example)	Changed input	IT MWh	PUE	Facility MWh	Change from base
Low-PUE sensitivity	PUE reduced by 0.10	5,573.12	1.15	6,409.09	-557.31 MWh
Base	Scheduled phases as shown	**5,573.12**	**1.25**	**6,966.40**	**0**
PUE sensitivity	PUE increased by 0.05	5,573.12	1.30	7,245.06	+278.66 MWh
Server-load sensitivity	Every server phase increased by 10 percentage points	6,466.64	1.25	8,083.30	+1,116.90 MWh
Nameplate-only ceiling	1,050 kW for 8,760 hours	9,198.00	1.25	11,497.50	+4,531.10 MWh

The sensitivity makes two levers comparable in the same unit. A **0.05 PUE increase** adds **278.66 facility MWh** to the base. A ten-percentage-point increase in server power across all phases adds **1,116.90 facility MWh**. This does not make workload efficiency a PUE component; it shows why both inputs must remain visible.

For cost illustration only, a flat hypothetical rate of **$0.10/kWh** would make base-case energy charges **$696,640**, before demand, fixed, tax or adjustment items. At a hypothetical location-based factor of **350 kg CO2e/MWh**, base emissions would be **2,438.24 tCO2e**. Both figures must remain blank in a real model until the buyer supplies a tariff and a dated applicable emissions factor.

Implications and Future Directions

The calculator's main procurement value is the list of what is still unknown. A decision-ready workbook should expose:

- **Capacity gap:** Contracted kW versus coincident engineering peak, including commissioning and redundancy tests.
- **Energy gap:** Scheduled IT MWh versus facility MWh with a matched, documented PUE.
- **Evidence gap:** Vendor specifications or assumptions that should be replaced by interval measurements.
- **Commercial gap:** Missing tariff blocks, billed-demand method, ratchets, losses, fixed charges and taxes.
- **Carbon gap:** Missing grid region, factor year, reporting method and contractual-instrument treatment.
- **Schedule gap:** Hardware deployment dates, workload ramps and maintenance windows without owners.

Over time, scenario inputs should be replaced with operating evidence. LBNL guidance calls for regular collection of accurate, repeatable data (navigator.lbl.gov). Baseline data should be measured before a change is made (navigator.lbl.gov). After commissioning, compare actual results with expected results rather than overwriting the original assumptions (navigator.lbl.gov).

Future versions can improve temporal detail without changing the accounting structure. Hourly workload forecasts can map to time-of-use prices. Rack telemetry can be reconciled to PDU and UPS meters. Weather-normalized facility models can replace a single annual PUE when credible interval facility data exist. The fundamental discipline remains the same: preserve boundaries, provenance, time alignment and uncertainty.

Frequently Asked Questions (FAQs)

How do you calculate GPU cluster power usage and annual MWh?

For each phase, multiply average IT kW by phase hours and divide by 1,000. Sum the phase MWh. If the schedule represents a full non-leap year, confirm that it totals **8,760 hours**. Keep a nameplate-only scenario separate from a measured or workload-scheduled scenario.

How do you convert MW to annual MWh for a data center?

Multiply average MW by operating hours. A constant 1 MW across a non-leap year equals **8,760 MWh**. For a changing GPU load, calculate each phase separately and sum the results. Do not use installed MW capacity as average demand unless the scenario is explicitly a nameplate ceiling.

How should PUE be used in a GPU cluster calculator?

Multiply IT energy by a PUE measured or contractually specified for the same facility boundary and period. Do not use PUE to scale GPU TDP directly, compare unlike measurement categories, or claim workload efficiency. PUE is the ratio of facility energy to IT energy, not a server benchmark.

Can server load factor replace measured power?

It can support an explicitly hypothetical planning scenario. It should not be presented as measured demand. Use a phase-specific factor applied to whole-system maximum input, include non-server IT separately, and replace the estimate when interval data become available.

How does a GPU cluster electricity cost calculator work?

Apply the exact tariff to facility energy by time block, then add billed-demand, fixed and applicable adjustment items. Do not multiply annual MWh by a national average and call the result a budget. The OpenEI Utility Rate Database is a free source of U.S. rate-structure information (openei.org), but the executed utility or colocation contract controls.

What is the difference between an AI data center power consumption calculator and this energy calculator?

A power calculator focuses on coincident kW for electrical and cooling capacity. An energy calculator integrates changing power over time to produce kWh or MWh. A complete planning workbook contains both modules and never uses annual energy to prove peak capacity.

Should renewable-energy purchases reduce PUE?

No. PUE accounts for energy crossing the defined boundary. Renewable procurement belongs in emissions and contractual reporting. Removing renewable electricity from the numerator would change the efficiency ratio without changing the facility's physical energy use.

Conclusion

An auditable GPU cluster energy model has a simple mathematical core and a demanding evidence structure. First freeze the dated inventory. Then describe operating phases whose hours sum to the analysis period. Calculate IT energy from kW and hours, add network, storage and management loads, and only then apply a period-matched, boundary-defined PUE to obtain facility MWh.

Keep five outputs distinct: **peak IT kW**, **peak facility kW**, **IT MWh**, **facility MWh**, and **overhead MWh**. Add cost only from the applicable tariff and emissions only from a dated, location-appropriate factor. Show nameplate, measured-low, measured-base and measured-high scenarios without disguising assumptions as benchmarks.

The most useful result is not a single total. It is a reproducible workbook that states which inputs were measured, vendor-specified, buyer-supplied or hypothetical, who owns each unresolved field, and how actual readings will be reconciled against the plan. That structure gives facilities, procurement, finance and investment teams a common basis for capacity decisions, annual budgets and due diligence.

Source links

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

1. https://www.eia.gov/energyexplained/electricity/measuring-electricity.php
2. https://aa.usno.navy.mil/faq/asa_glossary
3. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html
4. https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html
5. https://www.thegreengrid.org/resources/glossary?combine=pue
6. https://www.iso.org/obp/ui?_escaped_fragment_=iso%3Astd%3Aiso-iec%3A30134%3A-2%3Aed-2%3Av1%3Aen
7. https://www1.eere.energy.gov/manufacturing/datacenters/pdfs/task_force_metrics_recommendations.pdf
8. https://www.energy.gov/cmei/femp/evaluating-your-utility-rate-options
9. https://www.epa.gov/system/files/documents/2025-01/using-egrid-to-determine-emissions.pdf
10. https://gpusmith.com/articles/en/data-center-electrical-design-gpu-clusters
11. https://gpusmith.com/articles/en/gpu-data-center-cooling-design
12. https://eta.lbl.gov/news/berkeley-lab-report-evaluates-increase-electricity-demand-data-centers
13. https://gpusmith.com/
14. https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/planning.html
15. https://www.nvidia.com/en-eu/data-center/h200/
16. https://www.arista.com/assets/data/pdf/Datasheets/7060X6-Datasheet.pdf

17. https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/hw/n9364e-sg2-q/cisco-nexus-9364e-sg2-q-nx-os-mode-switch-hardware-installation-guide/m_n93xxx_system_specs.html
18. https://www.spec.org/osg/power/docs/SPECpower_ssj2008-Result_File_Fields.html
19. https://docs.nvidia.com/deploy/nvidia-smi/
20. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A02024R1364-20240517
21. https://datacenters.lbl.gov/sites/default/files/WP49-PUE%20A%20Comprehensive%20Examination%20of%20the%20Metric_v6.pdf
22. https://betterbuildingssolutioncenter.energy.gov/sites/default/files/attachments/Data_Center_Metering_and_Resource_Guide.pdf
23. https://www.energystar.gov/sites/default/files/2024-09/LicensedProfessionalsGuide_July2024.pdf
24. https://portfoliomanager.energystar.gov/pm/help
25. https://www.dell.com/support/manuals/en-gm/poweredge-xe9680/idrac_telemetry_reference_guide_pub/x86-subsystem-power-metrics?guid=guid-b40c8e07-9f85-4cc3-abd5-9a2c218b7633&lang=en-us
26. https://redfish.dmtf.org/redfish/schema_index
27. https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php
28. https://www.eia.gov/tools/faqs/faq.php?id=20&t=3
29. https://ghgprotocol.org/sites/default/files/standards/Scope%202%20Guidance_Final_Sept26.pdf
30. https://navigator.lbl.gov/guidance/task/8
31. https://navigator.lbl.gov/guidance/task/21/download
32. https://openei.org/wiki/Utility_Rate_Database

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