GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

How Much Electricity Do Data Centers Use? LBNL vs IEA

> 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 · 2026-11-23 · how much electricity do data centers use · data center electricity consumption · data center energy use twh · lbnl data center report · iea data centre electricity demand · ai data center power consumption · gpu power to twh · data center pue · us data center electricity share

Original article: https://gpusmith.com/articles/en/how-much-electricity-do-data-centers-use



In this article

1. Executive Summary
2. Introduction and Background
3. Why the Published Numbers Disagree: Five Scope Switches
4. The Reconciliation Ledger: Every Headline Estimate by Scope
5. The Estimates in Detail: LBNL, IEA and the Nature Paper
6. From GPU Count to TWh: A Reproducible Conversion Worksheet
7. Data Analysis and Evidence: What the Numbers Imply
8. What the Estimates Do Not Tell a Local Planner
9. Reproducibility Box: Inputs, Dates and a Copy-Paste Formula
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

No single number answers "how much electricity do data centers use", because the headline figures measure different things. The two anchor estimates are **176 TWh** for United States data centers in 2023, about **4.4%** of US electricity (eta-publications.lbl.gov) (energyanalysis.lbl.gov), and **415 TWh** for data centers worldwide in 2024, about **1.5%** of global consumption (build-up.ec.europa.eu). They are consistent with each other: the International Energy Agency (IEA) puts the US at about 180 TWh in 2024, nearly 45% of the world total (build-up.ec.europa.eu). A third figure, **118 TWh**, comes from a September 2026 Communications Sustainability paper (Nature Portfolio; hereafter "the Nature paper", although it is not in the flagship journal Nature) whose body defines it as the aggregate consumption of six leading operators, not of all AI facilities worldwide (www.nature.com).

The newest US estimate moved upward. Lawrence Berkeley National Laboratory (LBNL) published a 2025 Update in June 2026 that puts 2024 US use at **192 TWh**, or 4.7% of US electricity (escholarship.org), and a 2030 Reference Case of **649 TWh**, or 11.8% (escholarship.org). The December 2024 report had projected **325 to 580 TWh** for 2028 (eta-publications.lbl.gov). Measured against 176 TWh, that is a factor of **1.85 to 3.30**, which is what "double or triple by 2028" means arithmetically (energyanalysis.lbl.gov).

Five switches explain most apparent conflicts: geography, what is counted (all data centers, AI only, or named operators), IT load versus facility load (the power usage effectiveness, PUE, multiplier), historical estimate versus scenario, and the denominator behind any percentage. The IEA labels its historical figures "modelled estimates, not measured data" (build-up.ec.europa.eu), and LBNL cautions that projections are "inherently less certain than historical estimates" (eta-publications.lbl.gov).

For a buyer sizing a GPU fleet, the Nature paper's parameter structure (accelerator count, average power, utilization, hours, PUE) reduces to one formula (www.nature.com). Using NVIDIA's rated 10.2 kW for an eight-GPU DGX H100 (docs.nvidia.com), LBNL's 1.4 national PUE for 2023 (eta-publications.lbl.gov), and a 0.60 average load fraction derived from LBNL's published power and time assumptions, a 1,000-GPU fleet draws about **9,400 MWh** a year and a 100,000-GPU fleet about **0.94 TWh** (our calculation, a training-workload

scenario at server-level rated power; an inference scenario with a 0.40 load fraction gives about 0.63 TWh for 100,000 GPUs). Moving PUE from 1.4 to 1.1 cuts that by 21%, while moving the load fraction from 0.4 to 0.8 doubles it. Capacity and energy are different units: one gigawatt running all year is 8.76 TWh, and LBNL converts its 2028 range to 74 to 132 GW only by assuming 50% utilization (eta-publications.lbl.gov).

Introduction and Background

Anyone who has searched for data center electricity use meets a scatter of figures: 58 TWh, 176 TWh, 192 TWh, 415 TWh, 945 TWh, 4.4%, 1.5%, 11.8%, and the phrase "double or triple". Each is correct for its own scope, and each is routinely quoted without it. This report treats the question as a reconciliation problem. It lays out every headline estimate with its geography, year, counting rule, load basis and method, shows the arithmetic that connects them, and then offers a reproducible way to turn a planned GPU fleet into annual megawatt-hours (MWh) and terawatt-hours (TWh).

The intended readers are operators, utility planners, journalists, policy analysts and investors who must pick one number, and private-AI buyers who need an auditable conversion from hardware counts to energy. GPU Smith is an independent private AI infrastructure engineering consultancy that advises on GPU procurement, workload sizing, power and cooling; it is not a chip maker, cloud provider or data center operator, and this report cites third-party sources. Water use is covered in a separate GPU Smith article on data center water use; cluster-level energy budgeting and interconnection are left to other articles.

A few definitions keep the rest of the report unambiguous. A **terawatt-hour (TWh)** is one trillion watt-hours, and one TWh equals 1,000 gigawatt-hours (GWh). A **gigawatt (GW)** is one billion watts, or one thousand megawatts (www.eia.gov). **Power usage effectiveness (PUE)** is total facility power divided by IT equipment power (datacenter.uptimeinstitute.com), and the IEA uses the same ratio (build-up.ec.europa.eu). ISO/IEC 30134-2, edition 2 (January 2026), calls PUE the globally recognized key performance indicator for how efficiently a data center uses energy (www.iso.org). A **capacity factor** is the ratio of electrical energy produced to what continuous full-power operation would yield over the same period (www.eia.gov).

Historical context matters. LBNL's bottom-up model found US data center use roughly flat at about 60 TWh across 2014 to 2016, and 76 TWh (1.9% of US electricity) in 2018 (eta-publications.lbl.gov). The report then described the end of "the era of generally flat data center energy" (eta-publications.lbl.gov). Between 2017 and 2023, LBNL reports, power demand more than doubled, largely because of growth in AI servers (energyanalysis.lbl.gov). That growth, and the uncertainty about how long it continues, is why estimates now diverge.

Information is current as of 23 November 2026. IEA figures are from the April 2025 "Energy and AI" report; later IEA updates are not reflected.

Why the Published Numbers Disagree: Five Scope Switches

Most conflicts between data center electricity figures disappear once five questions are answered. Each switch can move a number by a factor of two or more, and they compound.

Switch 1: Geography

The 176 TWh figure is US only; the 415 TWh figure is global. The IEA's regional split for 2024 is the United States 45%, China 25% and Europe 15% (build-up.ec.europa.eu). National statistics offices show how local shares differ: Ireland's Central Statistics Office reports data centers at 22% of metered electricity in 2024, up from 5% in 2015 (www.cso.ie), while Statistics Netherlands reports 5,664 GWh in 2025 (www.cbs.nl), and the European Commission cites the IEA's estimate of 70 TWh for EU data centers in 2024 (energy.ec.europa.eu). A percentage is meaningless until the denominator's geography is stated.

Switch 2: What counts as a data center

Scope varies in three ways. First, **all facilities versus AI only versus named operators**. The IEA defines data centers as facilities housing servers, storage systems and networking equipment (build-up.ec.europa.eu), and models three IT equipment types: servers, storage and in-facility network equipment (build-up.ec.europa.eu). The wide-area network that connects data centers to users falls outside that model (build-up.ec.europa.eu). Second, **size and ownership**. LBNL's taxonomy runs from telco edge closets and micro data centers up to hyperscale campuses (eta-publications.lbl.gov), so small rooms are in scope. Third, **cryptocurrency**. LBNL analyses it separately with a top-down method (eta-publications.lbl.gov), and the 2025 Update keeps mining out of the data center total (escholarship.org). The IEA also treats crypto as separate, putting it at around 125 TWh in 2023 (build-up.ec.europa.eu); Cambridge researchers put Bitcoin mining at 138 TWh a year, about 0.5% of global consumption, extrapolated from reported data (www.jbs.cam.ac.uk).

Switch 3: IT load versus facility load

IT load is the electricity reaching servers, storage and network gear. Facility load adds cooling, power conversion and lighting. LBNL multiplies installed-base IT energy by modeled PUE to reach facility totals (eta-publications.lbl.gov), and infrastructure was 40% of total data center electricity in 2014 and 30% in 2023 (eta-publications.lbl.gov). The Nature paper's formulation also applies PUE to convert IT energy to data center energy (www.nature.com). A GPU-only calculation understates facility load by the PUE factor and by the non-GPU server components.

Switch 4: Historical estimate versus scenario

LBNL's 176 TWh is a modeled historical estimate; 325 to 580 TWh is a scenario range. IEA's 415 TWh is a modeled 2024 estimate and 945 TWh a Base Case. The Nature paper's 239 to 295 TWh is a 2030 range built from three growth scenarios of 15%, 25% and 35% a year for new AI load (www.nature.com). Mixing a historical number with a projection in one sentence is the most common source of apparent contradiction.

Switch 5: The denominator

Percentages depend on what is in the bottom of the fraction. LBNL's 4.4% for 2023 uses US electricity consumption (www.energy.gov). The Electric Power Research Institute (EPRI) uses US electricity **generation** and projects 4.6% to 9.1% by 2030 in its 2024 paper (restservice.epri.com). The IEA's 1.5% uses global electricity consumption. The same 2024 US data center load looks different against retail sales of 3,975 TWh (www.eia.gov) than against net generation of 4,309 TWh (www.eia.gov): 192 TWh is 4.8% of the first and 4.5% of the second (our calculation).

These five switches are the organizing logic for the ledger below. When a figure is quoted, a careful reader asks which cell of the ledger it occupies.

The Reconciliation Ledger: Every Headline Estimate by Scope

Table 1 below lists the headline estimates a reader is most likely to meet, with the scope switches resolved for each. Figures in the "Counts" column are what the publisher states it includes; where a source does not say, the cell says so rather than guessing.

Estimate	Publisher, date	Geography and year	Counts	Load basis and method
176 TWh (4.4% of US electricity); **325 to 580 TWh** in 2028 (6.7 to 12.0%)	LBNL, December 2024 (eta-publications.lbl.gov)	US, 2023 and 2028	All space types from edge closets to hyperscale; crypto analysed separately (eta-publications.lbl.gov)	Facility load: bottom-up from server shipments times power times modeled PUE (eta-publications.lbl.gov)
192 TWh (4.7%) in 2024; **464 TWh** in 2028; **649 TWh** (521 to 843) in 2030	LBNL 2025 Update, June 2026 (escholarship.org)	US, 2024 to 2030	Same scope; crypto still excluded (escholarship.org); adds non-commercial ASIC accelerators	Same bottom-up method, updated shipments and AI server power (escholarship.org)
415 TWh (1.5% of global); **945 TWh** in 2030 Base Case; 700 to 1,700 TWh in 2035	IEA "Energy and AI", April 2025 (build-up.ec.europa.eu)	Global, 2024 to 2035; US about 180 TWh	Servers, storage, in-facility network, plus cooling and infrastructure; wide-area network outside scope	Facility load; modelled estimates, not measurements
118 TWh in 2024; **239 to 295 TWh** in 2030	Chen et al., Communications Sustainability (Nature Portfolio), 21 September 2026 (www.nature.com)	Six named operators, global	Aggregate electricity of six leading operators, with a rising AI share (www.nature.com)	Facility load: GPU-hours style formulation times PUE (www.nature.com); parameters extracted by a language model from corporate documents
4% to 5% of US electricity today; **9% to 17%** by 2030	EPRI, 2026 update (esca.epri.com)	US, 2030	Not stated in the source summary; lists crypto among drivers it covers	State-level capacity, construction and announced plans (esca.epri.com)
22% of metered electricity in 2024 (6,969 GWh)	Central Statistics Office, Ireland, June 2025 (www.cso.ie)	Ireland, 2024	Metered data center consumption	Measured meter data, not a model
5,664 GWh in 2025, provisional	Statistics Netherlands, 2026 (www.cbs.nl)	Netherlands, 2025	Excludes in-house data centers of organisations such as universities and hospitals (www.cbs.nl)	Official statistics

The ledger shows three patterns. First, the US estimates agree across publishers: LBNL's 176 TWh for 2023 and 192 TWh for 2024 sit beside the IEA's roughly 180 TWh for 2024, and the gap between 180 and 192 is 6.7%, well inside the uncertainty both sources acknowledge (our calculation). Second, the global figure of 415 TWh and the six-operator figure of 118 TWh cannot be cleanly reconciled: 118 is 28.4% of 415 (our calculation), close to the 28% implied by the Nature paper's own cross-check assumptions, since it takes hyperscale as 70% of global data center electricity (www.nature.com) and the top six operators as 40% of hyperscale (www.nature.com), and 0.70 times 0.40 is 0.28. That benchmark is the authors' assumption-driven check, not independent confirmation. Third, only the national statistics offices report metered data; everything else in the table is a model.

The definition of the 118 TWh figure matters, so this section states which one the article uses. The paper's abstract describes data centers built for artificial intelligence and gives 118 TWh for 2024 rising to 239 to 295 TWh by 2030, about 1% of global electricity demand (www.nature.com), while the body defines the figure as total electricity of the six operators' data centers with the AI share rising over time (www.nature.com). This article follows the body's definition: a six-operator figure, not an all-AI-worldwide figure and not a GPU-only figure. The authors state that excluding smaller cloud providers, colocation, telecom and other operators may understate total AI-related demand (www.nature.com).

The Estimates in Detail: LBNL, IEA and the Nature Paper

LBNL: the US bottom-up model

LBNL starts from installed servers, built from shipment data, and multiplies by per-device power and then by PUE. For AI servers it uses the rated power of NVIDIA DGX-class eight-GPU systems, for example 10.2 kW for the H100 class (eta-publications.lbl.gov). Operating power is set at 70% of rated power, based on measured H100 nodes averaging 74% of rated thermal design power (eta-publications.lbl.gov), and varied between 60% and 80% for 2024 to 2028. Idle power is 20% of rated, based on Brookhaven National Laboratory measurements of an eight-GPU H100 node (eta-publications.lbl.gov), and inference servers are assumed to operate 40% of the time (eta-publications.lbl.gov).

GPU-accelerated server electricity rose from under 2 TWh in 2017 to over 40 TWh in 2023 (eta-publications.lbl.gov), against total server electricity of nearly 100 TWh in 2023 (eta-publications.lbl.gov). LBNL's national average PUE falls from 1.6 in 2014 to 1.4 in 2023 (eta-publications.lbl.gov) and to between 1.15 and 1.35 by 2028, though the authors caution that real-world PUE will likely be worse than simulated (eta-publications.lbl.gov).

The 2025 Update revised history as well as the future. It attributes slightly lower prior-year estimates primarily to a reduction in reported GPU shipments for 2023 and 2024, along with revised power-consumption assumptions for AI inference servers (escholarship.org) and states 14% growth in 2024 (escholarship.org). Dividing 192 TWh by 1.14 implies about 168 TWh for 2023, against the original 176 TWh (our calculation; the text does not print a revised 2023 value). The Reference Case reaches 464 TWh in 2028, mid-range within the earlier 325 to 580 TWh, and in 2030 AI servers are 84% of server energy and 55% of total data center energy (escholarship.org).

IEA: the global model

The IEA reports that data center electricity use has grown about 12% a year since 2017, yet its share of global demand rose only from 1% in 2005 to 1.5% in 2024 (build-up.ec.europa.eu). The 2030 Base Case is about 945 TWh (build-up.ec.europa.eu); the Lift-Off Case exceeds 1,260 TWh (build-up.ec.europa.eu), High Efficiency is about 800 TWh and Headwinds about 670 TWh (build-up.ec.europa.eu). For 2035 the range is 700 to 1,700 TWh, with a Base Case near 1,200 TWh (build-up.ec.europa.eu).

On AI, the IEA uses accelerated servers as a proxy: 24% of server electricity and 15% of total data center demand in 2024 (build-up.ec.europa.eu), projected to grow 30% a year in the Base Case and to account for almost half of the net increase to 2030 (build-up.ec.europa.eu). It also says there is no comprehensive data on the AI share (build-up.ec.europa.eu). The US adds about 240 TWh (up 130%) and China about 175 TWh (up 170%) by 2030 (build-up.ec.europa.eu), together nearly 80% of global growth (build-up.ec.europa.eu), yet data centers are less than 10% of global demand growth in the Base Case (build-up.ec.europa.eu). Our sum of 180 and 240 gives roughly 420 TWh for the US in 2030, well below LBNL's 649 TWh Reference Case, which its authors place above the central path of their earlier range (escholarship.org).

The Nature paper: six operators and a GPU-hours formulation

Chen et al. estimate that aggregate electricity consumption by six leading operators rises from roughly 118 TWh in 2024 to 239 to 295 TWh in 2030 (www.nature.com), a compound annual growth rate of roughly 13 to 17% (www.nature.com), with existing-stock consumption growing 10% a year in all scenarios (www.nature.com). Its Table 1 defines the baseline parameters of the GPU-hours-based formulation (www.nature.com) but lists symbols and definitions rather than per-chip wattages; the only numeric utilization hint is that inference utilization is typically 0.2 to 0.4 (www.nature.com). The paper benchmarks itself against the IEA's roughly 415 TWh in 2024 and 945 TWh in 2030 (www.nature.com). Its limitations include the exclusion of smaller cloud providers and colocation (www.nature.com) and the extraction of parameters by a language model from corporate documents (www.nature.com), so it is best read as a scenario study built on disclosed operator data, not a metered total.

EPRI and national statistics as cross-checks

EPRI's 2024 white paper projected 4.6% to 9.1% of US generation by 2030 (restservice.epri.com); its 2026 update projects 9% to 17% of US electricity, up from 4% to 5% today (esca.epri.com). EIA's reference-case analysis estimates computing at 8% of commercial-sector electricity in 2024 (www.eia.gov).

From GPU Count to TWh: A Reproducible Conversion Worksheet

The conversion follows the parameter structure of the Nature paper: accelerator counts, average power draw, utilization, operating hours and PUE (www.nature.com). The formula in kilowatt units is:

TWh per year = N x P x L x H x PUE / 1,000,000,000

where **N** is the number of GPUs, **P** is rated power per GPU in kW, **L** is the average load fraction, **H** is hours in the period (8,760 for a non-leap year), and **PUE** is the facility multiplier. The divisor converts kilowatt-hours to terawatt-hours. For MWh, divide by 1,000 instead.

Inputs, with sources

Table 2 lists each input with the value used, the published range, and the source. Rated power values are vendor maxima, not measurements.

Input	Value used	Published range or alternatives	Source and note
P, GPU-only rated power	0.70 kW (H100 SXM)	H100 NVL 350 to 400 W; H200 NVL up to 600 W; A100 400 W standard; MI300X 750 W peak	NVIDIA lists up to 700 W configurable for H100 SXM (www.nvidia.com), 350 to 400 W for H100 NVL (www.nvidia.com), up to 600 W for H200 NVL (www.nvidia.com), 400 W for A100 standard (www.nvidia.com); AMD lists 750 W peak for the MI300X (www.amd.com)
P, server-level rated power per GPU	1.275 kW (10.2 kW divided by 8)	DGX B200 14.3 kW for 8 GPUs gives 1.79 kW per GPU	The DGX H100 guide gives 10.2 kW max (docs.nvidia.com); the DGX B200 guide gives 14.3 kW max (docs.nvidia.com) for 8 B200 GPUs (docs.nvidia.com); our division
Rack-scale power	Not used in examples	About 1.67 kW per GPU slot in a 72-GPU GB200 rack	NVIDIA's DGX GB200 guide gives approximately 120 kW per rack (docs.nvidia.com); our division
L, average load fraction	0.60 training; 0.40 inference	0.20 to 0.80	Derived below from LBNL's published assumptions; Nature's Table 1 gives 0.2 to 0.4 as typical inference utilization, which is compute utilization rather than average power fraction, so it is only a qualitative cross-check on L (www.nature.com)
H, hours	8,760	8,784 in a leap year	EIA's convention of 8,760 hours per year (www.eia.gov)
PUE	1.4 (base) and 1.1 (efficient)	1.09 to 1.56	LBNL national average 1.4 in 2023 (eta-publications.lbl.gov); Google fleet 1.09 (www.datacenters.google); Uptime survey average 1.52 in 2026 (datacenter.uptimeinstitute.com) and 1.54 in 2025 (datacenter.uptimeinstitute.com); LBNL 2025 Update AI facilities 1.145 in 2024 (escholarship.org)

The key distinction is between GPU-only and server-level power. A DGX H100 draws its 10.2 kW maximum across eight GPUs plus CPUs, memory, networking and fans, so 1.275 kW per GPU is 82% higher than the 0.70 kW GPU rating (our calculation). Counting only the GPU board understates IT load; applying the full server rating at all times overstates it.

Deriving the load fraction without inventing numbers

LBNL gives the building blocks: operating power of 70% of rated power (eta-publications.lbl.gov), idle power of 20% of rated (eta-publications.lbl.gov), and operating time of 80% for training and 40% for inference (eta-publications.lbl.gov). A time-weighted average is L = 0.70 x t + 0.20 x (1 minus t). For training with t = 0.80 this gives 0.60; for inference with t = 0.40 it gives 0.40 (our simplification of LBNL's method, not a quoted formula). Independent measurements show real draw varies by phase: the ML.ENERGY benchmark paper finds LLMs

and VLMs consume significantly less power than the GPU's thermal design power (arxiv.org), and a 2026 arXiv study cites prefill at 80 to 90% of thermal design power and decode at 40 to 60% (arxiv.org). These support a load fraction well below 1.0 but do not replace metering your own workload.

Worked examples (Hypothetical Example)

The following fleets are illustrations, not client data. Both use server-level power of 1.275 kW per GPU, L = 0.60, H = 8,760 and PUE = 1.4. LBNL's operating, idle and time-share parameters are fleet-level modelling assumptions, so a metered site may differ widely from these results.

- **1,000-GPU fleet:** 1,000 x 1.275 x 0.60 x 8,760 x 1.4 = about 9,382 MWh per year, or 0.0094 TWh. The average facility draw is about 1.07 MW (9,382 MWh divided by 8,760 hours).
- **100,000-GPU fleet:** the same arithmetic gives about 938,000 MWh, or 0.94 TWh per year, and an average facility draw of about 107 MW.
- **One H100 server-level GPU:** about 9.4 MWh a year. A single 700 W GPU running flat out all year with no overhead would use 6.1 MWh.
- **Variants of the 100,000-GPU fleet:** with GPU-only P = 0.70 kW the result is about 0.52 TWh; with inference L = 0.40 it is about 0.63 TWh.

For scale, one million server-level H100 equivalents under the same assumptions would use about 9.4 TWh a year, so GPU counts alone cannot reach the 415 TWh global figure; the IEA attributes 15% of 2024 data center electricity to accelerated servers (build-up.ec.europa.eu).

Sensitivity: PUE and load fraction

Table 3 varies the two parameters buyers most often debate for the 100,000-GPU fleet, holding server-level power at 1.275 kW per GPU.

Annual TWh, 100,000 GPUs	PUE 1.1	PUE 1.4	Ratio, 1.4 to 1.1
L = 0.40 (inference-heavy)	0.491	0.625	1.27
L = 0.60 (training-heavy)	0.737	0.938	1.27
L = 0.80 (sustained)	0.983	1.251	1.27
Ratio, L 0.80 to 0.40	2.00	2.00	

The table shows that the load fraction matters about twice as much as PUE in relative terms: doubling L doubles energy, while moving PUE from 1.1 to 1.4 adds 27%. The full range, from 0.49 to 1.25 TWh, spans a factor of 2.55 for identical hardware (our calculation). The PUE bounds are defensible: 1.1 is near Google's fleet average of 1.09 and LBNL's 1.145 for AI-serving facilities, while 1.4 matches LBNL's 2023 national average and sits below the Uptime survey average of 1.52.

Unit hygiene: GW, TWh and capacity factor

Gigawatts measure power and terawatt-hours measure energy. One GW running continuously for a year is 8.76 TWh (1 GW x 8,760 hours, our arithmetic from EIA's hours convention (www.eia.gov) and its gigawatt definition (www.eia.gov)). LBNL shows the bridge explicitly: its 325 to 580 TWh range for 2028 translates to 74 to 132 GW if the average capacity utilization is 50% (eta-publications.lbl.gov). Check: 325 TWh divided by 8.76 is 37.1 GW average, and dividing by 0.5 gives 74 GW; 580 TWh gives 66.2 GW and 132 GW. The 2025 Update's 649 TWh in 2030 converts to 148 GW of interconnection capacity (escholarship.org), which matches 649 / 8.76 / 0.5 (our check).

Nameplate and average draw are different again. A 2026 arXiv case study of a 240-server A100 facility says sizing at thermal design power would give 1.19 MW against a generated peak of 0.75 MW (arxiv.org), and concludes that nameplate provisioning overstates required interconnection capacity (arxiv.org). Utility-reported load factors give a field reference: Dominion reported an 82% load factor for large data centers in 2024 (gridstrategiesllc.com). A planner who converts a GW announcement to TWh at 100% load factor will overstate energy; one who converts a TWh estimate to GW without dividing by a load factor will understate capacity.

Data Analysis and Evidence: What the Numbers Imply

This section tests the headline figures against each other using only arithmetic on the sourced values. All ratios and growth rates are our calculations unless a source states them.

What "double or triple by 2028" means

The Berkeley Lab announcement says data center load growth "is projected to double or triple by 2028" (energyanalysis.lbl.gov). The underlying numbers are 176 TWh in 2023 and a 2028 range of 325 to 580 TWh (eta-publications.lbl.gov). The ratios are 1.85 and 3.30, so "double" is a slightly generous description of the low end (1.85) and "triple" of the high end (3.30). The implied compound annual growth over five years is 13.1% at the low end and 26.9% at the high end. LBNL's 76 TWh for 2018 (eta-publications.lbl.gov) to 176 TWh in 2023 is a factor of 2.3, or 18% a year, so the low scenario assumes a slower pace than the recent past and the high scenario a faster one.

In the 2025 Update, 192 TWh in 2024 to 649 TWh in 2030 is a factor of 3.38 and 22.5% a year; to 464 TWh in 2028 it is a factor of 2.42. The 2028 value falls inside the earlier 325 to 580 TWh range, so the newer work narrows rather than overturns the earlier scenario.

Global growth in the IEA and Nature scenarios

The IEA Base Case rises from 415 to 945 TWh between 2024 and 2030, a factor of 2.28 and a compound rate of 14.7% a year, a little above the 12% a year the agency reports since 2017 (build-up.ec.europa.eu). The Nature paper's six-operator range rises from 118 to 239 or 295 TWh, factors of 2.03 and 2.50 and 12.5% and 16.5% a year. If the six operators are 28% of global demand throughout, the IEA Base Case implies about 265 TWh for them in 2030, inside the paper's 239 to 295 TWh range; the authors make the same cross-check with a 70% hyperscale share (www.nature.com) and a 40% top-six share (www.nature.com).

Percentages under different denominators

The bullets below compare shares directly, because the point is the arithmetic rather than a new data set.

- **US 2023, LBNL:** 176 TWh is 4.4% of US consumption (www.energy.gov), implying a denominator near 4,000 TWh.
- **US 2024, LBNL 2025 Update:** 192 TWh is 4.7% (escholarship.org), implying a denominator near 4,085 TWh; against EIA 2024 retail sales of 3,975 TWh (www.eia.gov) it would be 4.8%, and against net generation of 4,309 TWh (www.eia.gov) it would be 4.5%.
- **Global 2024, IEA:** 415 TWh is 1.5% (build-up.ec.europa.eu), implying a denominator near 27,700 TWh. Ember's global electricity demand for 2024 is 30,856 TWh (ember-energy.org), against which 415 TWh is 1.35%. The difference is a definitional choice about what counts as demand, not an error.
- **Global 2030, IEA:** 945 TWh (build-up.ec.europa.eu) is, in the Nature paper's restatement, about 3% of global power consumption by 2030 against about 1.5% in 2024 (www.nature.com).
- **Ireland 2024:** 22% of metered electricity (www.cso.ie), about fifteen times the global share.

Quote a share together with its numerator, denominator and year, and never compare a generation-based percentage with a consumption-based one.

Where AI sits within the total

Sources give three different "AI" measures that should not be added. LBNL's more than 40 TWh for US GPU-accelerated servers in 2023 (eta-publications.lbl.gov) is server-level energy. The IEA's 15% share for accelerated servers in 2024 (build-up.ec.europa.eu) applied to 415 TWh gives about 62 TWh worldwide (our calculation), also a server-level proxy. The Nature paper's 118 TWh is total six-operator electricity with a rising AI share, not an AI-only total (www.nature.com). In LBNL's 2030 Reference Case, AI servers are 55% of total data center energy, so the AI share is expected to move from a minority to a majority of US data center load within the decade.

Strength of the evidence

Metered national statistics (Ireland, the Netherlands) are the most reliable but national and, in the Dutch case, exclude some in-house facilities (www.cbs.nl). Bottom-up models depend on shipment data, and the IEA notes that very few governments mandate comprehensive reporting (build-up.ec.europa.eu). Scenario studies such as the Nature paper rest on growth assumptions (www.nature.com). Weight a metered figure above a modeled historical figure, and that above any projection.

What the Estimates Do Not Tell a Local Planner

National and global totals hide where load lands. EPRI found that fifteen states account for 80% of US data center load (restservice.epri.com), and the Nature paper finds that fewer than ten regions hold nearly two-thirds of projected 2030 demand among the six operators it studies (www.nature.com). EIA's 2024 analysis found commercial electricity demand growth concentrated in a handful of states, with Virginia adding 14 billion kWh and Texas 13 billion kWh from 2019 to 2023 (www.eia.gov), and notes that some larger facilities can consume as much electricity as a medium-sized power plant (www.eia.gov).

Utility forecasts add a further layer of uncertainty that national energy estimates do not resolve. Grid Strategies reports that roughly 90 GW of 166 GW of forecast five-year peak load growth is linked to data centers (gridstrategiesllc.com), and judges that the data center portion of utility load forecasts is likely overstated by roughly 25 GW (gridstrategiesllc.com). A queue request in gigawatts is therefore neither an energy figure nor a prediction of what will be built.

Planners should use the national figures for context and growth rates, and site-specific interconnection and demand-response analysis for siting, which is outside this report's scope.

Reproducibility Box: Inputs, Dates and a Copy-Paste Formula

Every input used in this report is listed below with its source and retrieval date (all retrieved 23 November 2026).

- **Formula (kWh basis):** annual_TWh = N * P_kW * L * 8760 * PUE / 1e9
- **Formula (MWh basis):** annual_MWh = N * P_kW * L * 8760 * PUE / 1000
- **Spreadsheet form:** =N*P*L*8760*PUE/1E9
- **P_kW:** 0.70 for an H100 SXM board (www.nvidia.com); 1.275 for server-level DGX H100 divided by eight GPUs (docs.nvidia.com); 1.79 for DGX B200 (docs.nvidia.com).
- **L:** 0.70 x t + 0.20 x (1 minus t), with operating power 70% of rated, idle power 20%, and t = 0.80 for training or 0.40 for inference (LBNL assumptions cited in the worksheet).
- **PUE:** 1.4 base (LBNL 2023 national average); 1.1 efficient, near Google's reported fleet average (www.datacenters.google).
- **Hours:** 8,760 (www.eia.gov).
- **GW to TWh:** TWh = GW x 8.76 x load factor, with LBNL assuming 50% utilization in its GW conversion.
- **Output check:** 1,000 GPUs at the base inputs give 9,382 MWh; 100,000 GPUs give 0.938 TWh.
- **Retrieval caveats:** LBNL figures are from its December 2024 report and 2025 Update; IEA figures are from the April 2025 Energy and AI report, as reproduced on the European Commission BUILD UP portal; vendor maxima are not measurements.

To audit a vendor quote, replace P and L with measured values and use the metered facility PUE.

Implications and Future Directions

Three implications follow for readers deciding which number to use.

First, choose the estimate by scope, not by size. For US questions about total load, use LBNL's 2025 Update, with 192 TWh for 2024 and a 2030 Reference Case of 649 TWh within a 521 to 843 TWh range (escholarship.org); the 2025 Update's own Reference Case sits above the central path of its earlier range (escholarship.org). For global questions, use the IEA's 415 TWh in 2024 and 945 TWh Base Case in 2030, with the Lift-Off and Headwinds cases as bounds (build-up.ec.europa.eu). For hyperscaler load, the Nature paper is the only source here built for six named operators.

Second, expect revisions. The 2025 Update revised prior history downward because reported GPU shipments fell for 2023 and 2024 and AI inference server power assumptions were revised (escholarship.org) while raising the 2030 projection above the central path of its earlier range, and the IEA publishes four explicit cases and says the spread of uncertainty widens further by 2035 (build-up.ec.europa.eu). LBNL itself recommends more frequent reporting, noting that its 2024 report was only the third in nearly 20 years (eta-publications.lbl.gov). Hardware also changes the conversion: a DGX B200 is rated 14.3 kW for eight GPUs against 10.2 kW for the H100 generation (docs.nvidia.com), while PUE at the most efficient operators is already near 1.1 (www.datacenters.google).

Third, for private-AI buyers, the key inputs are the load fraction (a factor of two across plausible workload mixes), the PUE of the hosting facility (a factor of about 1.3), and the rated power of the chosen server generation. GPU Smith, as an independent adviser rather than a hardware or facility supplier, treats these as the inputs worth measuring before procurement, and suggests any fleet energy figure be accompanied by its formula and inputs, as in the reproducibility box. Readers needing engineering sign-off for a specific site should consult a qualified electrical engineer and the local utility.

Frequently Asked Questions (FAQs)

How much electricity do data centers use worldwide?

The IEA estimates about 415 TWh in 2024, or about 1.5% of global electricity consumption (build-up.ec.europa.eu), reaching about 945 TWh by 2030 in its Base Case. Against Ember's 30,856 TWh of 2024 global demand the share would be 1.35% (ember-energy.org).

What is the data center electricity consumption in TWh in the United States?

LBNL estimates 176 TWh in 2023 (eta-publications.lbl.gov) and, in its 2025 Update, 192 TWh in 2024 (escholarship.org). The IEA's independent estimate is about 180 TWh for 2024 (build-up.ec.europa.eu).

What percent of US electricity do data centers use?

LBNL gives 4.4% for 2023 and 4.7% for 2024, and projects 6.7% to 12.0% by 2028 (eta-publications.lbl.gov) and 11.8% by 2030 (escholarship.org). EPRI's 2026 update states 4% to 5% today and 9% to 17% by 2030 (esca.epri.com).

What does the LBNL data center energy report say?

The December 2024 report says US data center use climbed from 58 TWh in 2014 to 176 TWh in 2023 (energyanalysis.lbl.gov) and could reach 325 to 580 TWh in 2028. The June 2026 update adds 192 TWh for 2024 and 649 TWh for 2030.

What does the IEA say about data centre electricity demand in 2030?

The IEA Base Case is about 945 TWh in 2030 (build-up.ec.europa.eu); Lift-Off exceeds 1,260 TWh and Headwinds is about 670 TWh (build-up.ec.europa.eu).

How much electricity does AI use in data centers?

No source gives one agreed AI-only global total. The IEA says there is no comprehensive data on the AI share and uses accelerated servers at 15% of 2024 demand as a proxy (build-up.ec.europa.eu); LBNL reports more than 40 TWh for US GPU-accelerated servers in 2023; the Nature paper reports six operators at about 118 TWh in 2024, with a rising AI share.

How do I convert GPU power consumption to TWh?

Use TWh = N x P x L x 8,760 x PUE / 1e9, as in the worksheet above. For 100,000 H100-class GPUs at server-level power, a 0.60 load fraction and a PUE of 1.4, the result is about 0.94 TWh a year.

How many TWh does an H100 GPU use per year?

A 700 W H100 (www.nvidia.com) running continuously uses 6.1 MWh a year before overhead. At server-level power, a 0.60 load fraction and PUE 1.4, about 9.4 MWh, or 0.0000094 TWh (our calculation).

Do these estimates include crypto mining?

No. LBNL analyses it separately (eta-publications.lbl.gov) and the 2025 Update keeps it out of the total; the IEA puts crypto at around 125 TWh in 2023 outside data centers (build-up.ec.europa.eu).

Conclusion

The headline figures for data center electricity are not in conflict once their scope is read. LBNL's 176 TWh and 192 TWh describe US data centers in 2023 and 2024, about 4.4% and 4.7% of US electricity. The IEA's 415 TWh describes the world in 2024, about 1.5% of global consumption, with the US at about 180 TWh of it. The Nature paper's 118 TWh describes six large operators, not all AI facilities, and not GPUs alone. "Double or triple by 2028" is 1.85 to 3.30 times the 2023 level, and the newest LBNL Reference Case lies inside that earlier range for 2028.

A reader who must pick one number should answer four questions first: which geography, which facilities, IT or facility load, and which year and denominator. A buyer who must turn a GPU fleet into energy can use one formula with five inputs, report the inputs, and expect the answer to move by a factor of about 2.5 across plausible load fractions and PUE values. Because the national and global totals reviewed are modeled rather than metered, cite ranges, state the scope, and revisit the figures as LBNL and the IEA publish updates.

Source links

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

1. https://eta-publications.lbl.gov/sites/default/files/2024-12/us_data_center_energy_usage_report_lbnl-2001637.pdf
2. https://energyanalysis.lbl.gov/news/berkeley-lab-report-evaluates-increase-electricity-demand-data-centers
3. https://build-up.ec.europa.eu/system/files/2025-04/PChtNmAZVk_11_04_2025_145800.pdf
4. https://www.nature.com/articles/s44458-026-00152-5
5. https://escholarship.org/content/qt33m6w3x0/qt33m6w3x0.pdf

6. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html
7. https://www.eia.gov/tools/glossary/index.php?id=gigawatt
8. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2025.Annual.Survey.Report.pdf?version=0
9. https://www.iso.org/standard/85172.html
10. https://www.eia.gov/tools/glossary/index.php?id=Capacity%20factor
11. https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2024/keyfindings
12. https://www.cbs.nl/en-gb/news/2026/47/data-centres-consume-5-percent-of-the-netherlands-electricity
13. https://energy.ec.europa.eu/news/focus-data-centres-energy-hungry-challenge-2025-11-17_en
14. https://www.jbs.cam.ac.uk/2025/cambridge-study-sustainable-energy-rising-in-bitcoin-mining
15. https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-centers
16. https://restservice.epri.com/publicdownload/000000003002028905/0/Product
17. https://www.eia.gov/electricity/state/unitedstates
18. https://esca.epri.com/products/3002034696
19. https://www.nature.com/articles/s44458-026-00152-5/tables/1
20. https://www.eia.gov/todayinenergy/detail.php?id=65564
21. https://www.nvidia.com/en-us/data-center/h100
22. https://www.nvidia.com/en-us/data-center/h200
23. https://www.nvidia.com/en-us/data-center/a100
24. https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html
25. https://docs.nvidia.com/dgx/dgxb200-user-guide/introduction-to-dgxb200.html
26. https://docs.nvidia.com/dgx/dgxgb200-user-guide/hardware.html
27. https://www.eia.gov/outlooks/aeo/electricity_generation/pdf/LCOE_methodology.pdf
28. https://www.datacenters.google/efficiency
29. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/UptimeInstitute.GlobalDataCenterSurvey.2026.pdf
30. https://arxiv.org/html/2505.06371
31. https://arxiv.org/html/2603.18383
32. https://arxiv.org/pdf/2603.18383
33. https://gridstrategiesllc.com/wp-content/uploads/Grid-Strategies-National-Load-Growth-Report-2025.pdf
34. https://ember-energy.org/latest-insights/global-electricity-review-2025/global-electricity-trends
35. https://www.eia.gov/todayinenergy/detail.php?id=62409
36. https://www.eia.gov/todayinenergy/detail.php?id=63344

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.