BERKELEY LAB

Energy Analysis Division

United States Data Center Energy Usage Report: 2025 Update

Sarah J. Smith, Alex Hubbard, Alex Newkirk, Mohan Ganeshalingam, Billie Holecek, Dale Sartor, Michael Mills, and Arman Shehabi
Energy Analysis Division, Lawrence Berkeley National Laboratory

June 2026

Disclaimer

Copyright Notice

Citation

Smith, S.J., Hubbard, A., Newkirk, A., Ganeshalingam, M., Holecek, B., Sartor, D., Mills, M., Shehabi, A. 2026. United States Data Center Energy Usage Report: 2025 Update. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001758. https://doi.org/10.71468/P1RP4F

Acknowledgements

We thank the experts from organizations representing major data center equipment manufacturers and data center end users/owners who took time to answer our questions and comment on assumptions we implemented in this analysis.

Input data for this report was provided by Omdia Research, S&P Global, and the International Data Corporation.

The research in this report was conducted by Lawrence Berkeley National Laboratory with support from the Department of Energy Industrial Technologies Office. Lawrence Berkeley National Laboratory is supported by the Office of Science of the United States Department of Energy and operated under Contract Grant No. DE-AC02-05CH11231.

Table of Contents

List of Figures

List of Tables

Glossary of Terms

AI (Artificial Intelligence) – The simulation of human intelligence processes by machines, especially computer systems, to perform tasks such as visual perception, speech recognition, decision-making, and translation between languages.

ASIC (Application-Specific Integrated Circuit) – A type of semiconductor designed for a specific application rather than general-purpose use. In data centers, ASICs are often used for AI tasks due to their efficiency.

Compounded Uncertainty – A forecasting range representing the highest and lowest electricity consumption bounds, calculated by combining the most extreme values of all parameters across the sensitivity scenarios.

CPU (Central Processing Unit) – The primary component of a computer that performs most of the processing inside the computer, handling workloads and managing system tasks.

Data Center – A facility used to house computer systems and associated components, such as telecommunications and storage systems, primarily designed for data storage and processing.

Grid Interconnection Capacity – The maximum amount of electrical power a data center requests from the utility or grid operator at its point of connection to the grid, typically established through a formal interconnection request. This reflects total facility power needs (for IT, cooling, and other systems) plus design margins.

GPU (Graphics Processing Unit) – A type of processor particularly well-suited for parallel processing tasks, often used in AI applications and graphics rendering.

Idle Power – The amount of power drawn by IT equipment when it is not actively processing data but remains powered on, critical in estimating overall energy use.

Inference – The process of running a trained AI model to make predictions or generate content in response to user inputs, often latency-sensitive and dependent on real-time user demand.

Installed Base – The total number of switched-on servers or IT devices currently operational within data centers.

NERC (North American Electric Reliability Corporation) – An organization that ensures the reliability of the North American power system.

Power Capacity – The total electrical capacity that a data center can supply to servers and IT components, often expressed in gigawatts (GW).

PUE (Power Usage Effectiveness) – A measure of how efficiently a data center uses energy, specifically the ratio of total building energy usage to that used by the IT equipment alone.

Rack-Scale Architecture – A server architecture designed for high-density AI installations where power supply units and cooling resources are shared across multiple nodes within a server rack, rather than dedicated to individual servers.

Rated Power – The maximum power consumption of a server as specified by the manufacturer, typically used as a benchmark for system performance.

Reference Case – A baseline scenario in energy forecasts that relies on the most current data without modifications or adjustments.

Sensitivity Scenarios – Alternative forecasting scenarios that explore how changes in certain assumptions or parameters might impact energy use and consumption estimates.

Server – A computer or system providing data, resources, or services to other computers (clients) over a network.

TDP (Thermal Design Power) – The maximum amount of heat generated by a computer chip or component that the cooling system is designed to dissipate under any workload.

TPU (Tensor Processing Unit) – A type of Application-Specific Integrated Circuit (ASIC) developed by Google specifically for accelerating machine learning workloads.

Training – The computational process of teaching an AI model by feeding it large datasets, typically requiring high computational density and sustained server utilization.

TWh (Terawatt-hour) – A unit of energy equal to one trillion watt-hours, commonly used to measure large-scale energy consumption over time, such as in data centers.

Utilization – The average computational intensity of a server, expressed as a percentage of its maximum computing capability, indicating how effectively the server's resources are being used.

1. Executive Summary

This report updates the 2024 Data Center Energy Usage Report (2024 Report) and estimates that data centers could account for 11.8% of total U.S. electricity by 2030. The estimate also includes a range of scenarios that indicate the energy use could be between 9.5 and 15.3% of total U.S. electricity use by 2030. In comparison, the 2024 Report estimate range was 6.7% to 12.0% of total U.S. electricity by 2028.

The resulting electricity usage estimates in this Report are derived from a “bottom-up” energy use model, which determines electricity use from real-world data for planned data center IT equipment shipments (purchases), models of per-device annual electricity use and cooling system performance simulations, along with information on data center facility types and locations. The Reference Case estimate for electricity use (649 TWh in 2030) is calculated based on the current understanding of expected shipments and equipment design across the data center industry.

The estimated range for electricity use by 2030 is derived from our Sensitivity Scenarios that explore targeted adjustments from the Reference Case. These Sensitivity Scenarios are driven by alternative data sources or industry feedback suggesting parameters or input datasets that may differ from the Reference Case. The parameters and adjustments include:

- Lowering the forecasted installations of data center IT equipment (578 TWh in 2030);
- Increasing the forecasted number of specialized graphics chips shipped and installed (664 TWh in 2030);
- Reducing the assumed average operating lifetime of Artificial Intelligence (AI) chips (590 TWh in 2030); and
- Increasing the idle power and utilization rates of AI servers (782 TWh in 2030).

Additionally, the above sensitivities and additional uncertainty are combined into high and low Compounded Uncertainty scenarios, producing the ultimate estimate bounds of 521-843 TWh of U.S. data center electricity consumption in 2030.

2. Overview and Results

In December 2024, Lawrence Berkeley National Laboratory published the *2024 United States Data Center Energy Usage Report* (Shehabi et al., 2024; herein referred to as the "2024 Report"). That report estimated historical data center electricity consumption from 2014–2023 and provided a range of future demand scenarios through 2028 based on the most recent data available at the time of publication. Amid the continued growth of data centers and rapid evolution of industry technologies and facility designs, updating past findings is essential for keeping the data center industry and relevant stakeholders informed. This report updates the 2024 Report findings, leveraging the same methodology and many of the same assumptions while incorporating updated datasets for information technology (IT) equipment shipments and revised assumptions about equipment power consumption as of December 2025. Most updates made for this analysis relate to the IT equipment used for artificial intelligence (AI) applications. With this update, we also extend our future scenario timeline to 2030, providing a five-year horizon for possible outcomes based on currently available information.

Key updates are summarized as follows. In this report, we have:

- Included new server and accelerator shipment data, reflecting information available through November 2025
- Included non-commercial application-specific integrated circuit (ASIC) AI accelerators
- Updated AI server wattages reflecting newer generations of accelerators (e.g., Blackwell)
- Revised assumptions for AI inferencing power consumption parameters, including idle power fractions and utilization levels
- Developed industry-informed sensitivity cases, reflecting observed deployment strategies and vendor input to test key modeling uncertainties.

All scenarios in this study represent current trends, technologies, and practices in the data center industry. We do not model substantial departures from the current trajectory that could arise from technological breakthroughs, shifts in market conditions, or changes in external factors such as supply chains, regulatory environments, or grid power availability. While we acknowledge that such changes could influence future outcomes, quantifying them would require speculation that goes beyond available data. Therefore, this study focuses on scenarios based on current industry trends and measurable parameters.

Scenarios and Uncertainties

The 2024 Report presented a single range of future scenario values intended to capture a wide variety of uncertainties and possible changes in the trajectory of the data center industry. The upper and lower bounds of this range reflected compounding uncertainties all acting in the same direction: the high bound assumed that all parameters took on the highest values within the ranges being considered, while the low bound assumed the lowest values. While this approach illustrated a breadth of outcomes, it did not offer a central expected value or isolate the impact of individual uncertainties on the results. To provide these insights, this update presents results for five discrete scenarios, in addition to a “Compounded Uncertainty” range that is analogous to the results in the 2024 Report. Key parameter values for these scenarios are shown in Table 1 and described in detail in Section 3 of this report.

Firstly, we present a **Reference Case** based on the most informative datasets, measurements, and simulations available. Four **Sensitivity Scenarios** explore targeted adjustments from the Reference Case. These Sensitivity Scenarios are driven by alternative data sources or industry feedback suggesting parameters or input datasets may differ from the Reference Case. The Sensitivity Scenarios can be summarized as follows:

- **Consolidated Deployment**: Lowers the forecasted number of graphics processing units (GPUs) and ASICs shipped and installed in the U.S. by 15% (see Section 3.1.2)
- **High ASIC**: Increases the forecasted number of ASICs shipped and installed in the U.S. by 15% (see Section 3.1.2)
- **Low Accelerator Lifetime**: Reduces the assumed average operating lifetime of AI accelerators after shipment by one year (see Section 3.3)
- **High Inference Energy**: Increases the idle power and utilization of AI servers running inferencing workloads (see Sections 3.4.2 and 3.4.3).

Finally, the **Compounded Uncertainty** range represents the high and low electricity consumption bounds that were calculated using the highest and lowest values for each parameter across the Sensitivity Scenarios. While we do not assume these variables are inherently correlated (as it is unlikely that all factors would trend toward their extremes simultaneously), this approach provides a "stress test" of the total potential impact. For the Accelerator Lifetime and AI Server Utilization—Training parameters, additional uncertainty was included due to the limited availability of reliable data.

Table 1. Values for key parameters across scenarios.

Parameters	Reference Case	Sensitivity Scenarios				Compounded Uncertainty	
		Consolidated Deployment	High ASIC	Low Accelerator Lifetime	High Inference Energy	Low	High
GPU Shipment Multiplier	1	0.85	1	1	1	0.85	1
ASIC Shipment Base Multiplier	1	0.85	1.15	1	1	0.85	1.15
Accelerator Lifetime (years)	5	5	5	4	5	4	5.5
AI Server Idle Power (% of rated power)	20	20	20	20	30	20	30
AI Server Utilization—Training (%)	80	80	80	80	80	75	85
AI Server Utilization—Inference (%)*	20	20	20	20	30	20	30

* Parameter value changes over time; 2030 value shown.
Blue shading indicates a lower value than in the Reference Case; orange shading indicates a higher value.

Total Energy Use

Figure 1 presents updated estimates of total U.S. data center electricity use from 2016–2024 and a future[1] scenario range for 2025–2030, accounting for servers, storage, network equipment, and infrastructure. Consistent with the 2024 Report, the energy use of cryptocurrency mining is not included in the scope of this report. The updated estimate of data center electricity use in 2024 is 192 terawatt-hours (TWh), accounting for 4.7% of total U.S. electricity consumption. This falls at the low end of the estimated range in the 2024 Report, with revised estimates of electricity use in prior years falling slightly below 2024 Report estimates. The change in historical values is primarily due to a reduction in reported shipments of GPUs for 2023 and 2024, along with revised assumptions for the power consumption parameters of AI inference servers (discussed in Section 3.4). Meanwhile, our current Reference Case forecast for total electricity use in 2028 is 464 TWh, which falls in the middle of the range presented in the 2024 Report.

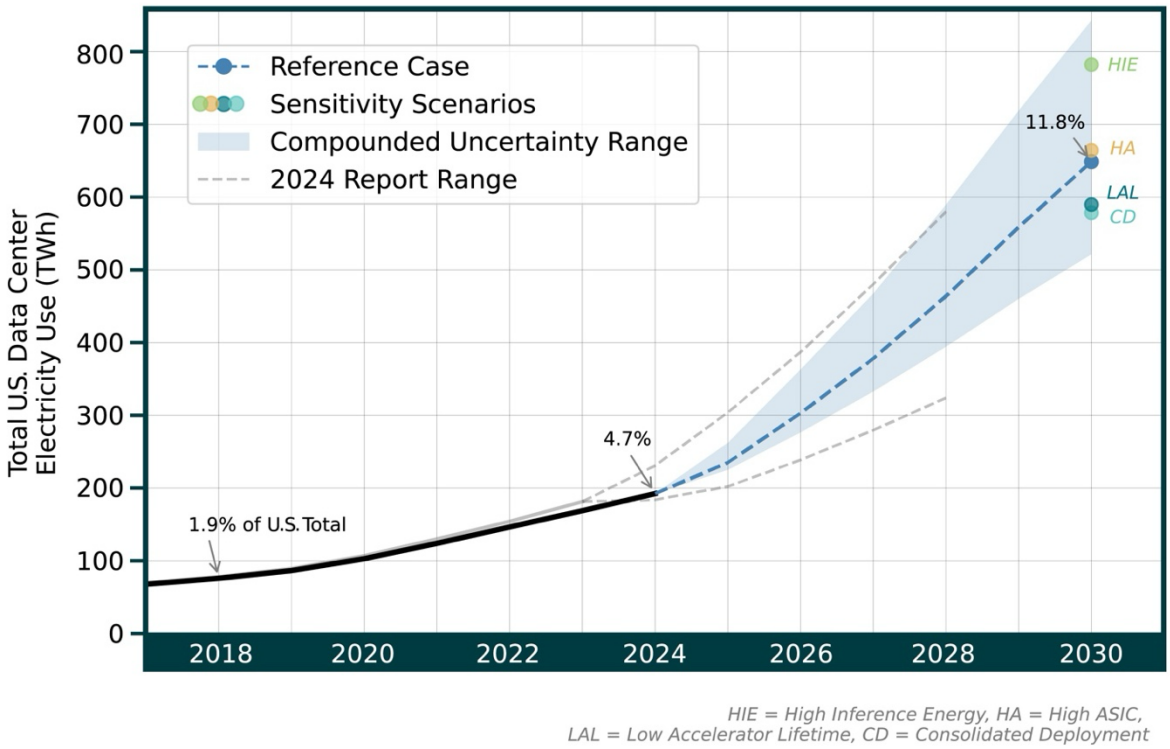

Figure 1. Total U.S. data center electricity use from 2017 through 2030

Results for 2030 across all scenarios are shown in Table 2. The Reference Case estimate for 2030 data center electricity use is 649 TWh, representing 11.8% of forecasted 2030 U.S. electricity use as presented in the 2025 North American Electric Reliability Corporation (NERC) Long-Term Reliability Assessment (NERC, 2026). Assuming the NERC values include data center energy use consistent with our analysis, their forecasts imply that non-data center load growth will increase 926 TWh (24%) from 2024–2030. This would result in data centers being responsible for 33% of total load growth over the six-year period.

[1] 2024 is considered the last historical year of this analysis because our primary IT shipment data extend only through the first three quarters of calendar year 2025. Therefore, the 2025 results include some projections.

The Consolidated Deployment and Low Accelerator Lifetime scenarios each reduce the Reference Case estimates of data center electricity consumption by approximately 10%. The High ASIC scenario yields a modest increase above the Reference Case, while the High Inference Energy scenario has the greatest impact: assumptions around increased inference idle power and utilization raise projected electricity use by more than 20% above the Reference Case. With all uncertainties combined, estimated electricity consumption ranges from -19.6% to +29.9% relative to the Reference Case. This yields a Compounded Uncertainty Range of 521–843 TWh, equal to 9.5–15.3% of total U.S. electricity use in 2030.

Table 2. 2030 data center electricity and share of U.S. electricity across all modeled scenarios.

	2030 Results		
Scenario	Data Center Electricity (TWh)	Difference from Reference Case (%)	Percent of 2030 U.S. Electricity* (%)
Reference Case	**649**	**--**	**11.8%**
Consolidated Deployment	578	- 10.8%	10.5%
High ASIC	664	+ 2.4%	12.1%
Low Accelerator Lifetime	590	- 9.1%	10.7%
High Inference Energy	782	+ 20.6 %	14.2%
Compounded Uncertainty—Low	521	- 19.6%	9.5%
Compounded Uncertainty—High	843	+ 29.9%	15.3%

* Based on the forecast in NERC's 2025 Long-Term Reliability Assessment.

Power Capacity

Data centers are often discussed in terms of power capacity. This refers to either the facility's nameplate capacity—the IT hardware load it is designed to serve, based on infrastructure like uninterruptible power supply systems—or to the grid interconnection capacity requested to supply power to the facility. It is challenging to convert between estimates of annual electricity consumption and power capacity without data on the relationships between average demand, maximum demand, facility nameplate capacity, and requested interconnection capacity (Norris, 2025). Current utilization of interconnection capacity is not well documented but is estimated to be around 50% due to high redundancy requirements and maintenance needs. However, there is a concerted effort in the data center industry to increase interconnection capacity utilization through innovations in power system design, energy balancing, and demand management (Verrus, 2026). Further, the capacity of power generation that must be added to the grid to serve new loads is not inherently equal to the interconnection capacity requested, and some data centers may forgo grid interconnection altogether and build their own generation capacity with unique redundancy requirements and utilization rates. These considerations are out of scope for our consumption-focused analysis, but they highlight the challenge of translating annual energy consumption estimates to required data center or power generation capacity.

We calculate the interconnection capacity by applying an average utilization rate to the annual

average power draw. Assuming a 50% average utilization rate, the Reference Case electricity use projections translate to 148 gigawatts (GW) of interconnection capacity needed to serve data centers in 2030. This corresponds to an average annual capacity increase of 17.4 GW/year from 2024–2030 and a compound annual growth rate of 22%. Scenarios that affect only the installed base of equipment (High ASIC, Low Accelerator Lifetime, and Consolidated Deployment) would have capacity requirements that scale proportionally with their projected electricity consumption. While electricity consumption is significantly higher under the High Inference Energy scenario, the installed equipment is the same as the Reference Case. Therefore, total capacity needs would be unchanged; the facilities in the High Inference Energy scenario would simply have higher capacity utilization rates.

Comparison with Industry Estimates

The rapid growth of U.S. data center electricity demand has prompted a wide range of forecasts from industry analysts, financial institutions, and research organizations. Historical estimates from sources including EPRI, Goldman Sachs, and Boston Consulting Group show broad agreement that U.S. data center electricity use grew from roughly 60–80 TWh around 2017 to approximately 150–200 TWh by 2023. The 2024 Report provided one of the first comprehensive academic forecasts with an explicit uncertainty range. While that estimate captured a wide envelope of possible outcomes, many subsequent industry projections have generally corroborated the expectation of significant growth, with many projecting similarly steep increases in data center electricity consumption driven by the surge in AI-related computing demand. A selection of these estimates, shown in Figure 2, illustrates how quickly the forecasting landscape has evolved over the last couple years.

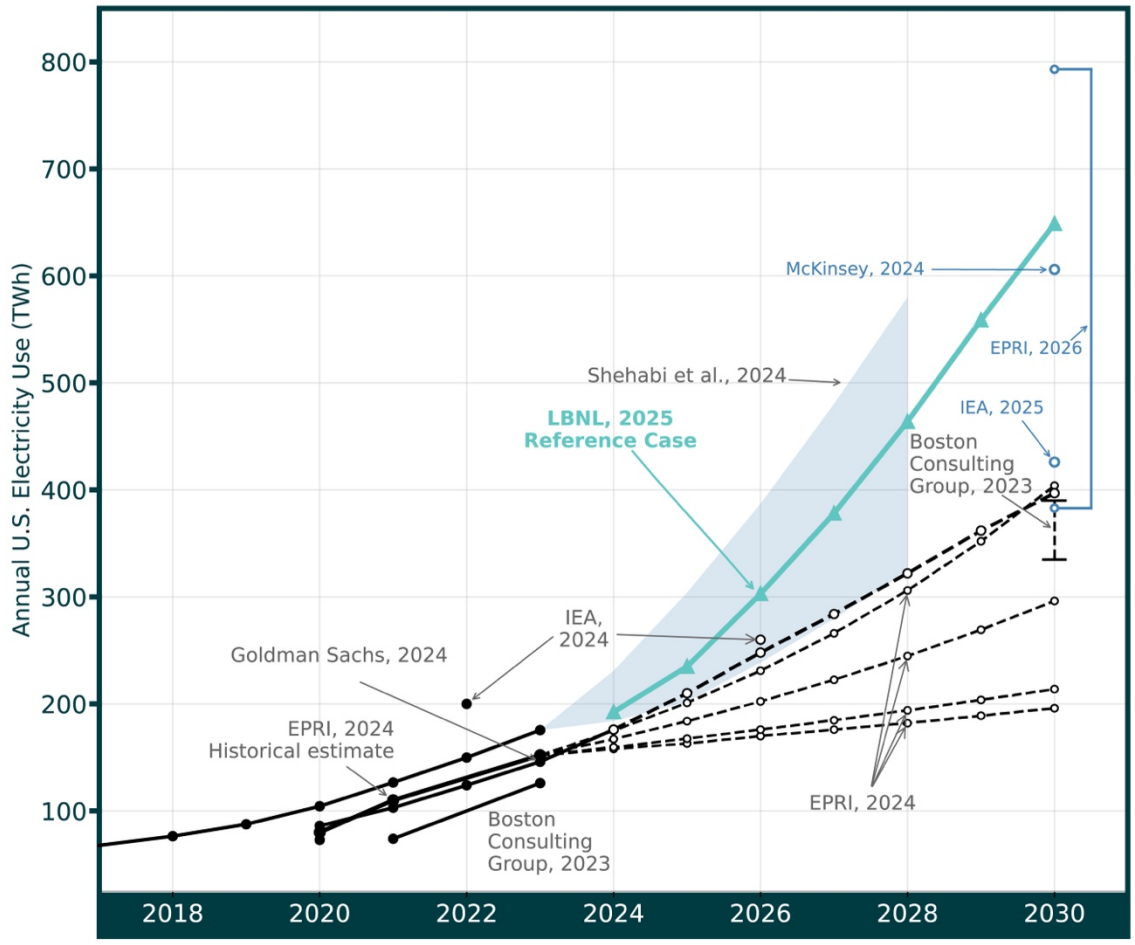

Figure 2. Updated academic and industry estimates of U.S. data center electricity use
This plot includes historical estimates (solid lines) and future projections (dashed lines) from multiple sources including LBNL's 2024 Report (Shehabi et al., 2024). The Reference Case from this 2025 Update is shown as a solid green line. Projections for calendar year 2030 from other sources developed after the publication of the 2024 Report are shown in blue.

The 2025 Reference Case projects U.S. data center electricity use rising steeply through 2030. This trajectory places the estimate above the central path of the earlier 2024 Report range and towards its upper bound. Among industry estimates for 2030, the Reference Case falls in the upper-middle portion of the distribution—closely aligned with McKinsey (2024) and the updated EPRI (2026) estimate, which spans a wide range from roughly 390 to 800 TWh, and above IEA (2025) at approximately 430 TWh. IEA's projection for 2030, published in early 2025 and developed using a methodology broadly consistent with this report, is substantially lower than LBNL's updated Reference Case estimate. This divergence highlights the sensitivity of long-term demand forecasts to assumptions about server and accelerator shipment trajectories (IEA, 2025). The LBNL 2025 Reference Case, based on the latest available data, seeks to reduce the uncertainty associated with the range of assumptions embedded in these diverse industry outlooks.

3. Analysis Updates

The estimates in this report are derived from a "bottom-up" energy use model that calculates total electricity use from an installed base of data center equipment. Figure 3 shows the model structure, including input data sources and the major units of analysis. As shown in the second column, the model characterizes IT equipment as either **servers** (compute equipment that runs workloads), **storage** (equipment that stores data), or **networking equipment** (hardware that moves data between systems), and uses shipment data to generate estimates for the installed base of equipment in each year. The model also applies assumptions regarding the power draw of installed equipment, and these wattage assumptions may vary across equipment and data center types. For servers, there is an underlying model of power draw that considers rated power, maximum power, server utilization, and idle power to estimate annual average wattages. Total annual IT equipment electricity consumption is estimated by summing the product of installed base and unit wattage across each equipment category. These electricity consumption estimates are multiplied by modeled power usage effectiveness (PUE) values to estimate total electricity demand across the data centers.

A full description of the model is presented in the 2024 Report (Shehabi et al., 2024). This section describes the updates that have been made to the 2024 Report methodology, assumptions, and input datasets. Methodological elements not discussed here are unchanged from the 2024 Report. Key calculations from the methodology are provided for quick reference in the Appendix.

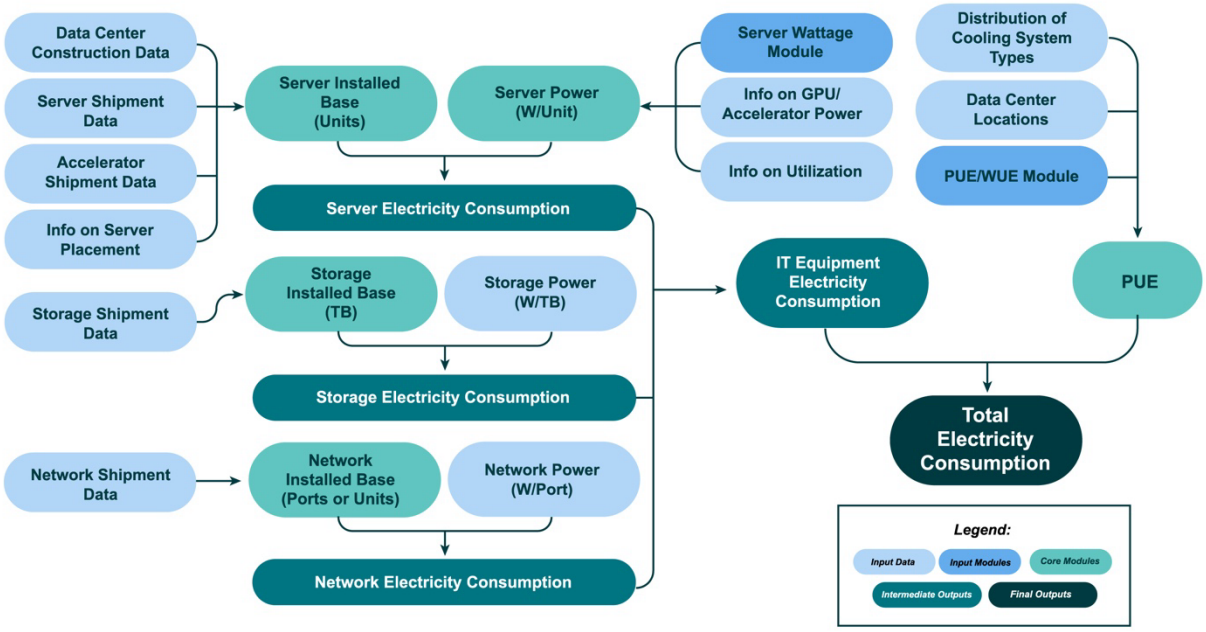

Figure 3. Flow chart for the data center electricity model used in this study

3.1 Server and Accelerator Shipments

3.1.1 Total Server Shipments

Inputs for total server shipments come from the International Data Corporation (IDC)'s Worldwide Quarterly Server Tracker, which contains annual data for historical (2003–2024) and forecasted (2025–2029) server shipments (IDC, 2025b). The 2030 values in this study are based on a linear extrapolation of the annual growth rates for individual subcategories of servers from 2024–2029. Shipment data are described by numerous characteristics; this analysis primarily uses socket capability, representing the number of central processing units (CPUs) each server can be equipped with, and product detail, which defines the form factor and typical deployment of each server.

The number of shipped servers is projected to grow from 5.9 million in 2024 to 9.1 million in 2028 and 10.6 million in 2030, representing an 80% increase from 2024 to 2030. For comparison, the updated 2028 forecast value is 20% higher than the projection used in the 2024 Report. Much of the recent and forecasted growth is in 1-socket servers: in 2018, 1-socket servers only represented 7% of annual total shipments, with 89% of servers in the "2+" socket category (and the remainder having 4 sockets or more). In 2028, 1-socket servers are expected to be approximately 30% of shipments, with 2+ socket servers representing about 69%. This growth in 1-socket server shipments reflects the increasing capabilities of individual CPUs. Core counts are rising among all processors, with some individual processors reaching 192 cores in 2025 (SPEC, 2026), allowing single-processor systems to accommodate the same workloads previously allocated to dual-processor systems.

Despite this growth in 1-socket servers, dual processor servers still dominate server shipments, growing to nearly seven million shipments in 2030. Rack-scale servers, such as Nvidia's NV72 configuration (NVIDIA, 2025), grew substantially: zero servers existed in this category in 2018, but by 2025 they represented over 4% of shipped servers. By 2030, these servers make up nearly 17% of all shipped servers. Rack-scale servers are designed for high-density AI-optimized installations and are typically deployed at large scales, with shared power and cooling supplies across nodes. Server shipments in the multi-processor categories (four or more CPUs, without accelerators) fall substantially through 2030 and make up less than 0.5% of shipped servers in 2030. This reflects the shift to accelerated servers for advanced workloads, moving away from CPU-driven compute.

3.1.2 AI Accelerator Shipments

Data for shipments of AI accelerators comes from IDC's 2025 Q3 Data Center Semiconductor Consumption Report (IDC, 2025a), which contains annual data for historical and forecasted CPUs, AI accelerators (such as GPUs), and network switch shipments through 2030. The analogous dataset used in the 2024 Report focused only on GPU accelerators; this update expands the accounting to all AI accelerators, namely ASICs. Additionally, estimates of global GPU and ASIC shipments from 2019 to 2026 were obtained from Omdia Research (Omdia, 2024, 2025). While this report focuses on U.S. data centers, Omdia's global estimates offer a useful comparison to IDC's and help us understand various perspectives and forecasts.

Accelerator shipments totaled just over seven million units in 2024, consisting of 73% GPUs and 27% ASICs. The GPU mix in 2024 was dominated by H100/H200 chips, with smaller shares of Blackwell chips and AMD models (IDC, 2025a). ASIC shipments in 2024 were primarily composed of nearly equal shares of Google Tensor Processing Units (TPUs) and Meta Training and Inference Accelerator (MTIA) units, with smaller numbers of Amazon Web Services (AWS) Trainium and Inferentia units (Omdia, 2025). By 2030, total accelerator shipments are projected to reach 19 million units. ASIC shipments represent an increasing share through 2026, then return to 2024 levels by 2030. In 2030, over 80% of GPUs shipped are advanced GPUs, such as Blackwell, Vera Rubin, and AMD MI400 (IDC, 2025a).

Accelerator Shipment Uncertainty and Scenarios

Estimating shipments of accelerators on a unit basis is challenging, and both historical and forecasted estimates are uncertain. Analysts tracking this market use multiple data sources to develop their forecasts, including direct reporting from server manufacturers on current and planned production, foundry supply and capital expenditure data, high-bandwidth memory production data from component suppliers, and consultation with experts throughout the ecosystem to verify assumptions (IDC, 2026). These datasets also rely on estimating and forecasting total revenue from accelerator sales and then converting to units by modeling the average selling price of various products and the estimated mix of product categories shipped in each quarter. While these datasets reflect the best information available at the time they are produced, there is inherent uncertainty in the forecasts, especially beyond the horizon for product orders of six months to two years. Estimating the ASIC market is particularly challenging, as these chips are custom-designed for specific customers rather than standardized commercial products. While the manufacturers of ASICs are public companies, they typically report custom chip revenue in aggregate rather than customer-specific volumes. The companies designing and deploying these ASICs rarely disclose detailed shipment or deployment figures for their infrastructure.

IDC's unit-level semiconductor dataset is updated approximately every six months, and Berkeley Lab has obtained the last three releases (November 2024, May 2025, and December 2025) in addition to a prior generation of the analysis from fall 2023. Similarly, Berkeley Lab obtained Omdia research from June 2024 and November 2025. With this historical context, variation across the datasets provided insight on the inherent uncertainty in the data. We make the following observations across these datasets:

- More recent shipment forecasts project significantly fewer GPU units than the prior estimates used in the 2024 Report, with recent estimates for shipments of GPUs and ASICs in 2030 reduced by as much as half (IDC, 2025a). Some of these revisions can be explained by the evolving forecast functional unit (see next section), but frequent changes and uncertainties remain.
- Revenue projections, however, have increased substantially between the mid-2025 and December 2025 datasets. These contrasting changes reflect a substantial shift in the accelerator market to fewer, more expensive (and more powerful) units being shipped.
- IDC estimates ASIC shipment data at the global level and uses these same global totals as their U.S. forecast. While these chips are being designed and deployed by American companies, not all ASICs are installed in the U.S.; Google, for example, sells computing

time on TPUs across Hong Kong, the Netherlands, and Taiwan (Google, 2026). However, no data is available to indicate if non-U.S. ASIC installations are negligible or if shipments should be adjusted downwards for our U.S.-focused analysis.
- Omdia currently forecasts many more global ASIC shipments than IDC. However, the two datasets may differ in scope, with Omdia including non-AI ASIC applications.

Based on these observations, we defined AI accelerator shipment scenarios to illustrate how evolving forecasts may influence future energy use. The Reference Case uses the most recent IDC data as-is, including the assumption that all global ASIC shipments go to the U.S. The Consolidated Deployment scenario reduces forecasted GPU and ASIC shipments by 15% to account for i) the possible continuation of downward revisions from past forecasts, reflecting continued improvements in the computational density of accelerator units; and ii) the possible impact of non-U.S. installations of accelerators not accounted for in the datasets. Finally, the High ASIC scenario increases forecasted ASIC shipments by 15% to represent values closer to Omdia's estimates.

It should be noted that AI servers are assumed to be a subset of the total servers tracked by IDC. In the 2024 Report, the overall number of servers was held constant regardless of the number of GPU shipments, such that a reduction in GPU shipments would cause an increase in conventional non-accelerated servers. In this update, adjustments to accelerator shipments from the Reference Case cause a direct change in the number of accelerated servers being shipped without influencing the assumed values for non-accelerated server shipments.

Functional Unit for GPU Forecasts

As previously described, analysts tracking the GPU market make several assumptions when developing forecasts, particularly when estimating unit quantities. GPU product development and release cycles are accelerating, and while some information is available on upcoming GPU generations, uncertainties remain regarding pricing, specifications, and timing. As a result, IDC analysts advised us that they essentially forecast future shipments of "high-end" GPUs using a functional unit of Nvidia B300-equivalents in terms of die count and die size. Because the market is currently constrained by die supply, aggregate die-level production is expected to remain consistent with the current estimates even as future GPU units incorporate more and/or larger die. And because power consumption scales with die size, modeling the forecasted high-end accelerators as B300-equivalents provides a reasonable basis for estimating future power consumption (IDC, 2026). The corresponding unit power assumptions are discussed further in Section 3.4.1.

3.2 AI Server Configurations

Our model considers the quantity and power consumption of servers configured for AI applications with 0, 2, 4, or 8 GPUs. ASIC-accelerated servers are modeled as a single average configuration based on the annual mix of individual models shipped, provided by Omdia. Prior to 2026, we assume an average of six ASICs per server. This decreases to an average of five ASICs per server from 2026–2030, as 4-ASIC configurations become more prevalent. It should be noted that, in this analysis, we consider each node in a rack-scale system to count as one server. Therefore, a product like the NVL72 rack, which contains 72 GPUs across 18 nodes, is

considered to have 18 servers in a 4-GPU configuration. For a given year, we take the total number of GPUs shipped and allocate them across these configurations to quantify the number of accelerated servers shipped in that year. These allocations are informed primarily by Omdia analysis, which estimates shipments of specific server models for 2024–2026 (Omdia, 2025). For 2019–2024, we use a prior version of Omdia's analysis that reports shipments by number of GPUs (Omdia, 2024).

GPUs are allocated into these server configurations based on the GPU type and shipment year. For advanced GPUs (i.e., Blackwell series), we assume that 55% are deployed in 4-GPU servers in 2024, increasing to 75% in 2026 and 100% in 2030. The remainder of advanced GPUs are deployed in 8-GPU systems, such as Nvidia's HGX. While some HGX systems are deployed with Blackwell chips, they are otherwise primarily deployed with H-series GPUs, which are not included in IDC's advanced GPU category. Non-"advanced" GPUs are allocated across 2-, 4-, and 8-GPU server configurations. For this GPU category, the share of 2-GPU servers declines from 27% in 2024 to 9% in 2030, while 8-GPU servers grow from 37% to 55% over the same period. The share of 4-GPU servers remains relatively stable at approximately 36%. Taken together, these allocations imply shipments of over 1.5 million accelerated servers to the U.S. in 2024, rising to 4.4 million in 2030. As advanced GPUs come to dominate GPU shipments by 2030, the overall system mix shifts heavily towards 4-GPU server configurations.

We also account for AI-optimized servers without GPUs or other accelerators, which are shipped with only two CPUs. These systems are treated separately from conventional 2-CPU servers because they are configured with much higher memory capacity and networking capabilities, resulting in higher energy use. Omdia's AI server shipment estimates include annual projections for these non-accelerated servers (Omdia, 2024, 2025). Using that data, we calculate the ratio of non-accelerated to accelerated servers and apply that ratio to the calculated accelerated server shipments in our model. The resulting mix of server configurations installed in each year of the analysis is shown in Figure 4 (see Section 3.3).

3.3 AI Server Lifetime and Installed Base

The installed base of AI servers is estimated by applying an average server lifetime to annual shipments. For non-accelerated servers, we retain the lifetime assumptions used in the 2024 Report: 4.4 years through 2019, increasing to 5.7 years in 2023, and then remaining constant. In the 2024 Report, the same assumption was applied to accelerated servers, reflecting the lack of data at the time to differentiate expected lifetimes by server type. In this update, we revise that assumption: for accelerated servers, we assume a constant lifetime of 5 years in the Reference Case, consistent with typical lifetimes of present-day server stock reported by IDC and Omdia.

The Low Accelerator Lifetime sensitivity scenario explores the implications of a shorter average lifetime of 4 years for accelerated servers. This scenario is motivated by recent statements from industry experts citing high early-life failure rates and 1–3 year lifetimes for some deployments (Shilov, 2024; Stanley, 2025). At the same time, there is clear evidence of accelerated servers operating for longer than 5 years (Leswing, 2025; Trueman, 2025). Importantly, the calculations in this study are based on *average* lifetimes across the installed base. If reports of failure rates and short lifetimes for some use cases are accurate, it could imply a lower average lifetime even

if some systems remain in operation for many years. Our shorter lifetime assumption reflects uncertainty in the expected average lifetime, rather than a scenario in which industry practices shift to shorter lifetimes than those currently observed. Consistent with this interpretation, annual accelerator shipments are held constant across scenarios, reflecting that current purchasing decisions already incorporate expectations about useful lifetimes. A shorter assumed lifetime reduces the number of servers installed and operating in any given year, which in turn results in lower annual energy use. In the Compounded Uncertainty high bound, we raise the accelerator lifetime slightly to 5.5 years, given the limited information on actual industry retirement practices and the possibility that lifetimes are being extended to similar levels as conventional servers.

Figure 4 shows the resulting installed base of servers over the study period. In 2018, the total installed base was 16.3 million servers, consisting almost entirely of conventional servers. By 2024, the installed base grew to 30.9 million, including 4.2 million accelerated servers. Estimates for 2030 vary by scenario, with differences driven by GPU and ASIC shipment adjustment factors and the assumed lifetime of accelerators. In the Reference Case, AI servers reach nearly 19 million by 2030, representing nearly 40% of the total installed base. The Low Lifetime scenario projects the smallest AI server base, with 16 million AI servers in 2030, reflecting the shorter assumed accelerator lifetime. In contrast, the High ASIC scenario increases ASIC shipments by 15%, resulting in 19.5 million accelerated servers in 2030. The conventional server installed base is held constant across scenarios, with a 2030 value of 29.6 million. This represents 60–68% of the total server installed base, depending on the number of accelerated servers shipped each year. Non-accelerated AI servers are calculated based on a fixed ratio relative to accelerated servers, and the total number of non-accelerated servers (AI and conventional combined) is held constant at the Reference Case level across scenarios.

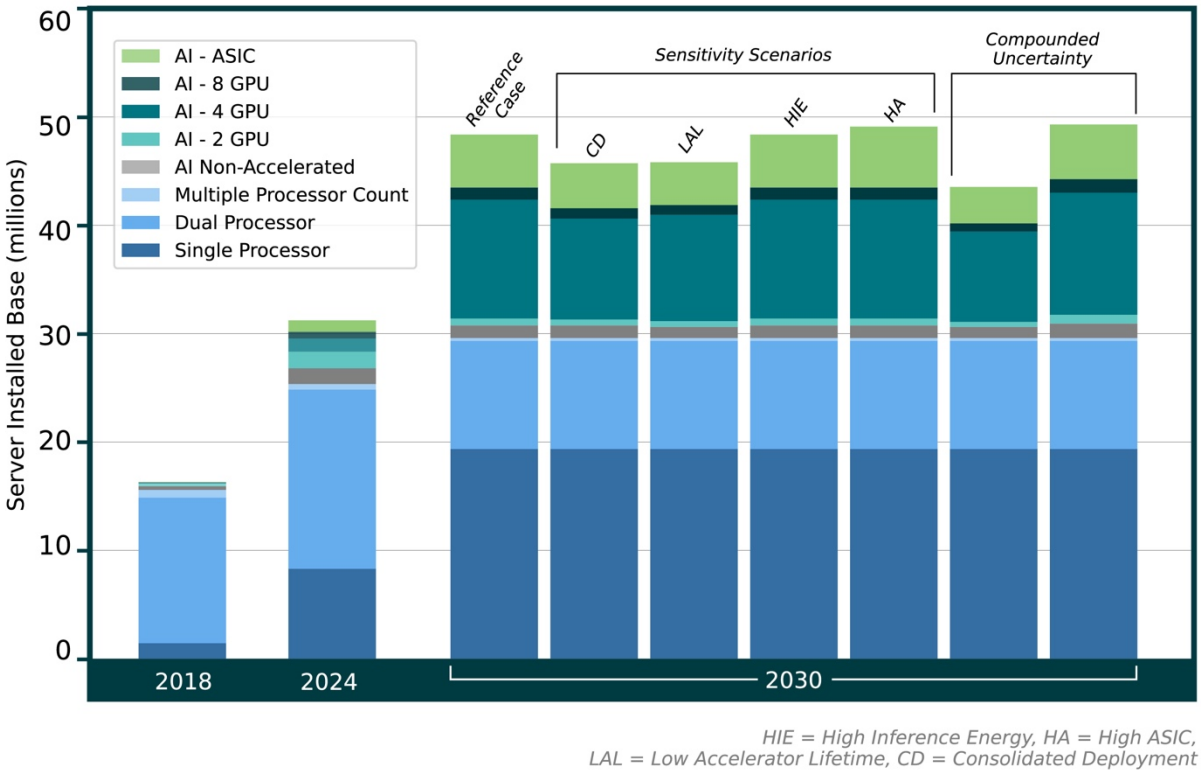

Figure 4. Total installed base of servers across years and scenarios

3.4 AI Server Power Draw

3.4.1 Rated Power

Calculations for server electricity use start by modeling the average rated power for all servers in a given category shipped each year. Here, "rated power" refers to the maximum possible power draw of the server, as reported by manufacturer specification sheets. This is generally determined by summing the thermal design power (TDP) of every component within the system. The 2024 Report considered AI servers equipped with GPUs through the H100 generation. In this update, we revise GPU rated power estimates to reflect more recent accelerators including Nvidia H200, B100, B200, and B300, and AMD MI355 and MI400. We also model the rated power of ASIC-accelerated servers, considering systems with AWS Inferentia v1 and Trainium v1–v3, and Google TPU v6–v7. While specifications are available for upcoming Nvidia Rubin chips, all future shipments of advanced GPUs are treated as B300-equivalents, as described in Section 3.1.2. After 2026, when all advanced GPU shipments are modeled as B300-equivalents, GPU server rated power is still assumed to increase slightly to account for growing memory requirements.

Rated power assumptions for Nvidia-accelerated servers are drawn from publicly available product specification sheets. For servers with accelerators designed by AMD, AWS, and Google, we rely on power estimates developed by Omdia. These power estimates for specific server designs are combined with assumptions regarding the mix of server generations and

configurations being shipped in each year. The result is an annual average rated power for each accelerated server configuration (ASIC, 2-GPU, 4-GPU, and 8-GPU servers), calculated as a shipment-weighted average across GPU models deployed in a given year. The resulting annual rated power values are shown in Figure 5.

The rated power of 8-GPU servers grows rapidly through 2025, driven by the introduction of high-TDP Blackwell GPUs in the DGX architecture. Beginning in 2026, however, the primary server architecture shifts to the NV72 configuration, which is composed of 4-GPU nodes. As a result, the share of high-TDP GPUs in 8-GPU servers declines. Any continued shipments of 8-GPU configurations after 2026 are assumed to be a mixture of mid-power H-series GPUs with a declining share of Blackwell GPUs, leading to a decline in rated power for 8-GPU servers from 2026–2030. The rated power of 4-GPU servers also increases through 2026, reflecting the increase in high-TDP GPUs. After 2026, rated power is held constant through 2030, consistent with the B300-equivalent functional unit of advanced GPUs shipped in these years (see Section 3.1.2). As with GPUs, ASIC accelerator TDPs increase rapidly through 2026; ASIC server rated power grows accordingly and is held constant thereafter.

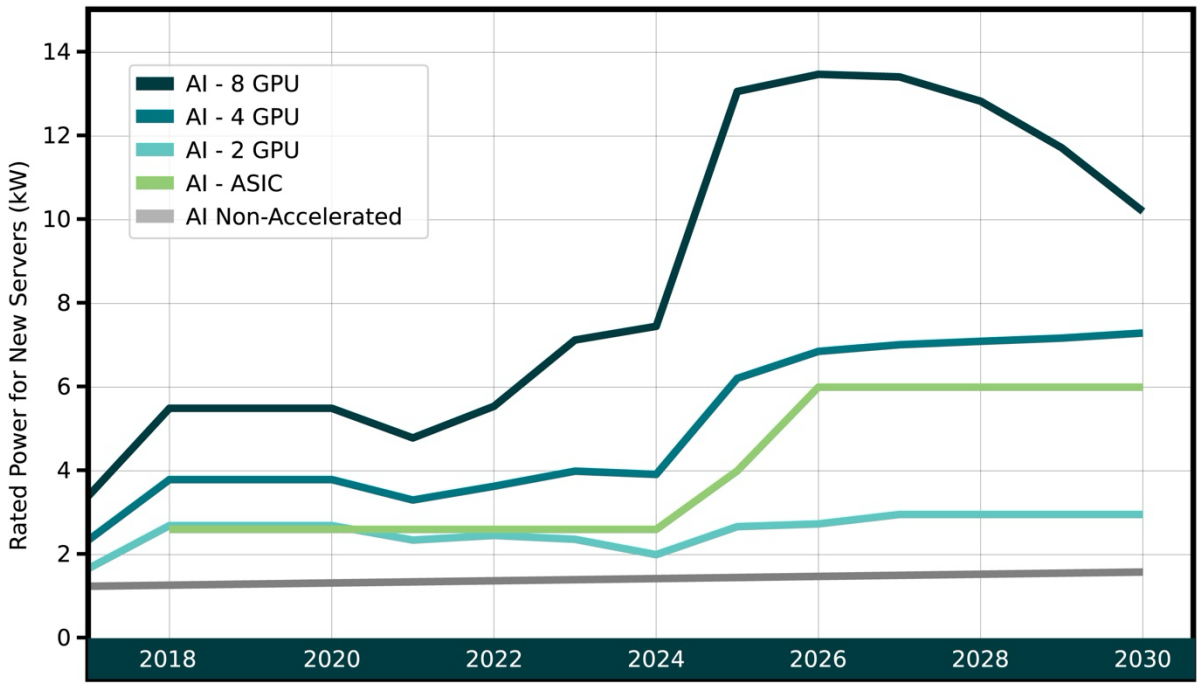

Figure 5. Average rated power draw of AI servers shipped each year from 2017 to 2030

3.4.2 Idle Power

Assumptions for the idle power of conventional servers are consistent with those used in the 2024 Report, which reflect a slight decrease over time due to improvements in power management technologies and operational practices. In this update, we make a minor refinement by setting a floor of 30% on the ratio of idle to maximum power (see Figure 6, left panel). As a result, the idle power fraction of conventional servers remains constant after 2026. We assume that idling AI servers draw power at 20% of their rated capacity, based on measurement data

reported by Brookhaven National Laboratory (Newkirk et al., 2025). This idle-to-rated power ratio is held constant across all years of the analysis, consistent with assumptions in the 2024 Report. Idle power assumptions have a substantial effect on overall energy use estimates, particularly given the low utilization rates assumed for AI inference workloads (see Section 3.4.3).

However, recent feedback from industry experts indicates that, in commercial AI operations, idle power may constitute a much higher fraction of rated power than that indicated by available measurement data. Idle power is influenced by server power management settings, which may be adjusted from factory settings to keep servers in a more active state to reduce latency. This increases energy consumption during idle periods relative to systems operating under more aggressive power management settings. The industry experts hypothesized that idle power could reach 50% of rated power in AI data centers, particularly those serving latency-sensitive inference workloads. In contrast, a follow-up conversation with a hardware architect at a major semiconductor manufacturer confirmed our original 20% idle-to-rated power ratio, though it was not confirmed if this ratio may change in practice once the hardware is installed and managed by end users. Given this uncertainty, we explore the impact of higher idle power levels in a sensitivity analysis. Specifically, the High Inference Energy scenario assumes an idle power level of 30% of rated power for AI servers, aligning their idle power ratio with that of conventional servers. These assumptions are shown in the right panel of Figure 6.

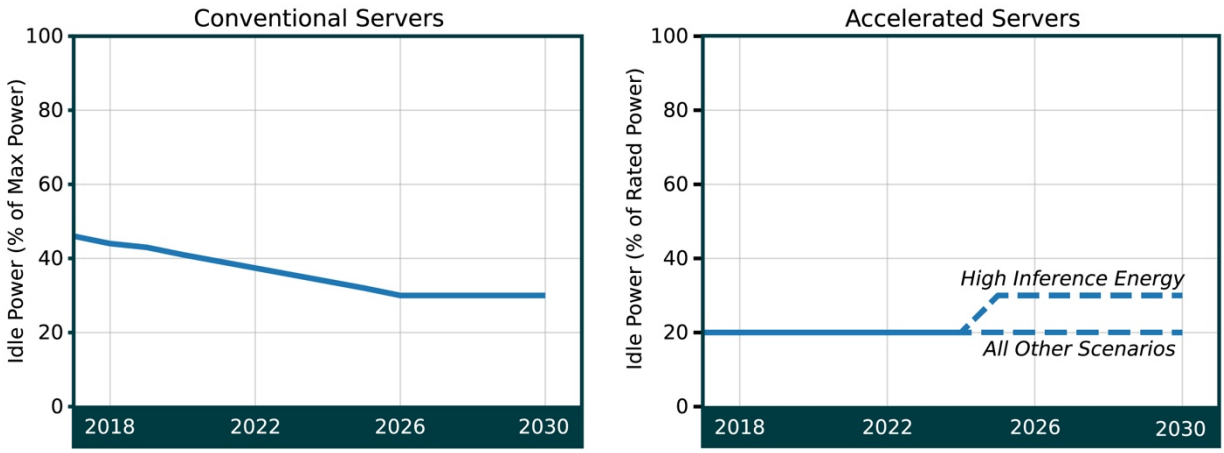

Figure 6. Idle power assumptions for conventional and accelerated servers

3.4.3 Server Utilization

In this report, “utilization” refers to the average computational intensity of a server throughout the course of a year, expressed as a fraction of its maximum possible computational level. For an AI training cluster, utilization reflects a time-weighted average of idling (0% utilization), active training (~100% utilization), and intervals devoted to other tasks, such as communication, that occur at partial utilization levels. In the 2024 Report, AI training servers were assumed to operate at an average utilization of 80% (referred to in that report as “operational time”), while the same servers performing inference workloads were assumed to operate at an average utilization of 40%. These assumptions were informed by conversations with data center industry experts, but there was little-to-no measured data available to verify them.

Recent feedback from industry experts indicates that the utilization of AI inference servers is likely much lower than assumed in the 2024 Report. The reason is that inference workloads are largely user-facing and highly dependent on behavioral demand patterns, requiring systems to be provisioned for peak usage when large numbers of users simultaneously query models. Thus, outside of those peak periods, servers may operate at relatively low utilization rates. Industry feedback indicates that average server utilization rates may be similar to those observed in enterprise data centers running conventional workloads, which the 2024 Report assumed would increase gradually from 10% in 2014 to 20% in 2030. While this feedback was corroborated by multiple experts, there continues to be a lack of verifiable data to confirm typical inference utilization levels. Given this uncertainty, we explore a range of inference utilization rates in this study:

- **Conventional server** utilization assumptions are consistent with those in the 2024 Report and vary by data center space type.
- **AI training** server utilization is assumed to be 80% in the Reference Case and all sensitivity scenarios. We vary this assumption in the Compounded Uncertainty range to reflect uncertainty of the available data, using a low value of 75% and a high value of 85%.
- **AI inference** utilization is assumed to follow the levels of conventional servers in enterprise data centers from the 2024 Report, reaching 20% by 2030. In the High Inference Energy sensitivity scenario, utilization instead follows the 2024 Report's assumption for conventional servers in colocation data centers, which is 10% higher across all years.

These utilization scenarios and their implications are illustrated in Figure 7.

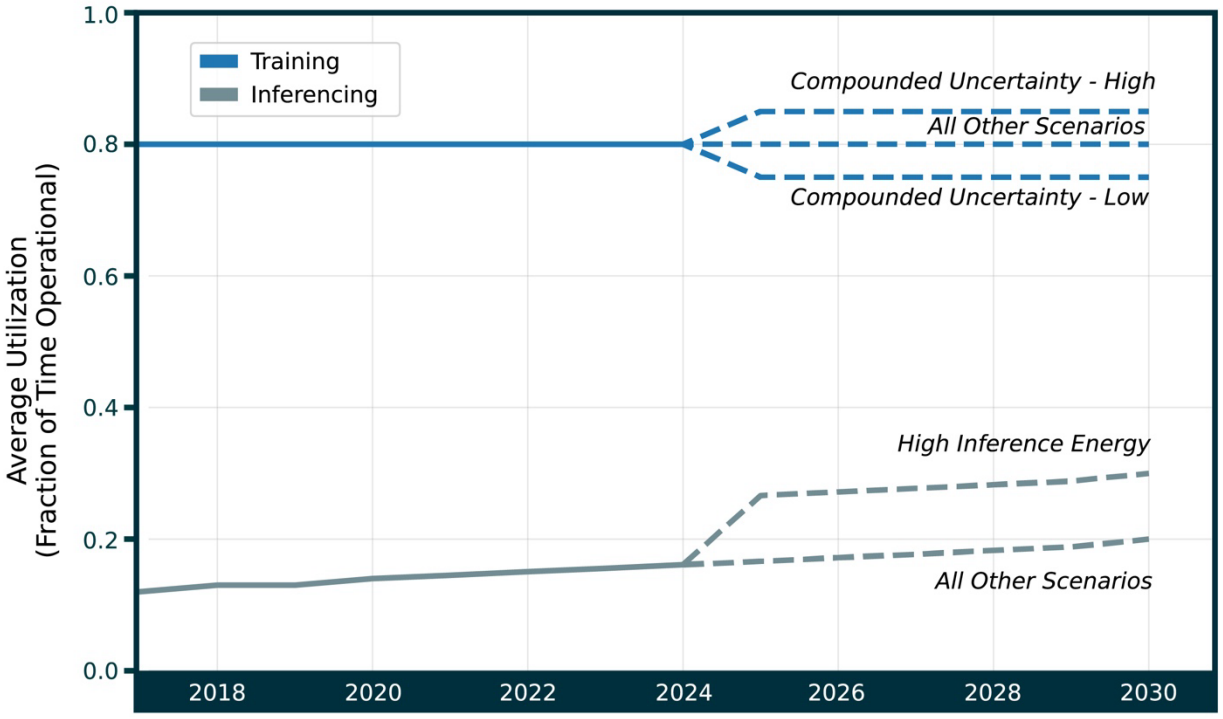

Figure 7. Utilization of AI servers by workload

To estimate average utilization for all AI servers, we weight the values described above by an assumed distribution of workloads between training and inferencing. This workload distribution assumption only affects utilization calculations and does not affect other aspects of the model. In this update, we revise the approach used in the 2024 Report by allocating workloads based on the share of total AI server *power* devoted to training versus inference, rather than on the number of server *units*. Based on information provided by Omdia, we assume that approximately 35% of the AI server power draw in 2024 was for training, declining to roughly 20% in 2030. This trend is in line with industry feedback indicating that inference is accounting for an increasing share of overall AI activity and spending. Non-accelerated AI servers are allocated entirely to inference workloads, following the methodology of the 2024 Report. This reflects fundamental differences in computational requirements: training workloads require the parallel processing capabilities of GPUs and ASICs, making CPU-only servers unsuitable for most model training tasks. In contrast, inference workloads, particularly for less computationally intensive models or lower-volume inference applications, can be effectively performed on CPU-only servers without dedicated accelerators.

3.4.4 Resulting Annual Average Power

Rated power, idle power fraction, operational power fraction, and utilization are combined to calculate the average power consumption of servers. Effectively, average power is the average of operational power and idle power, weighted by the utilization. The calculation is as follows:

*Operational Power = Rated Power * Operational Power Fraction*

*Idle Power = Rated Power * Idle Power Fraction*
*Average Power = (Operational Power * Utilization) + (Idle Power * (1-Utilization))*

These averages, across all installed servers, are shown in Figure 8. Overall, the resulting average power trajectories mirror the trends observed for rated power (shown in Figure 5 for servers shipped in a given year), with a time lag reflecting the lifetime and turnover of the server stock. After 2024, results are shown for the Reference Case and the High Inference Energy scenario, which together provide bounds on the future outcomes modeled in this study. Average power values differ slightly across other scenarios due to alternative assumptions for shipments and server lifetimes, but do not vary meaningfully from the Reference Case.

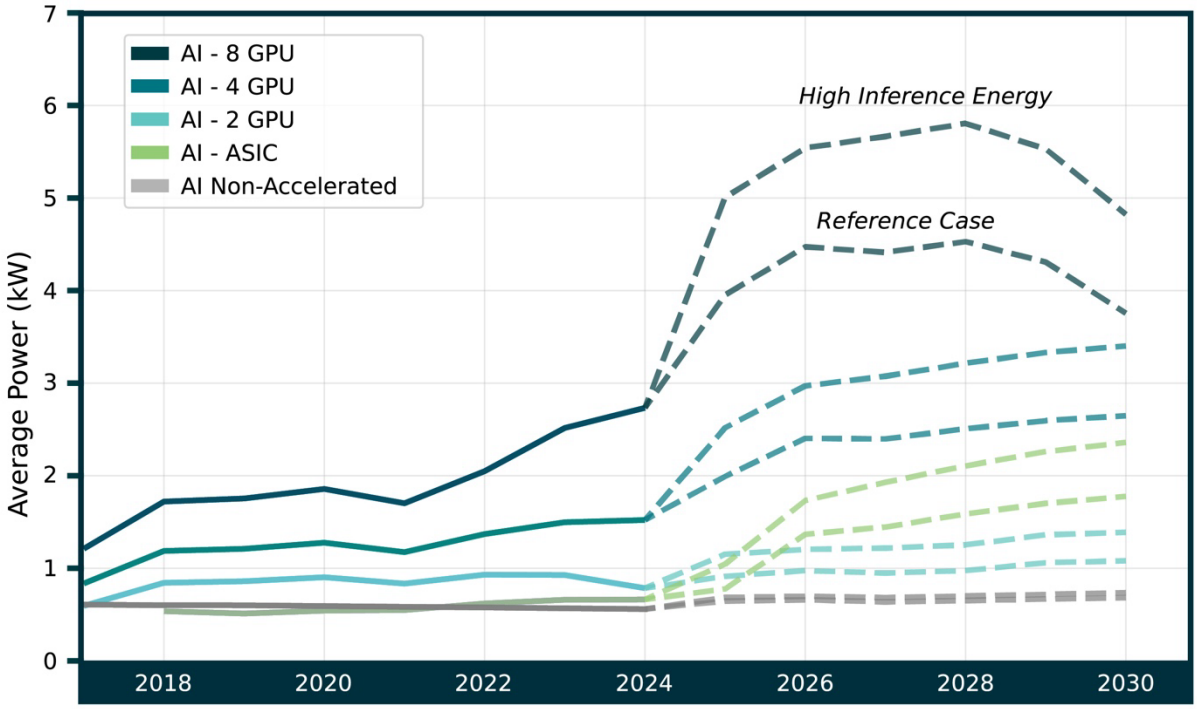

Figure 8. Aggregate annual average power draw of AI server types

In the Reference Case, the average annual power of 8-GPU servers rises steeply through 2026, surpassing 4 kilowatts (kW). In comparison, 4-GPU and 2-GPU server configurations rise less steeply, peaking just over 2 kW and 1 kW, respectively. After 2026, the average power of 8-GPU servers largely levels off through 2029, reflecting a shift in the mix of deployed configurations, with higher power accelerators increasingly deployed in 4-GPU systems (see Section 3.2 for discussion of configuration trends). In contrast, non-accelerated AI servers experience relatively modest growth in power draw. ASIC-accelerated servers follow a similar trajectory to 4-GPU servers, rising from 950 watts (W) in 2024 to over 2 kW in 2030. In 2030, average annual power draw is roughly 28% higher in the High Inference Intensity scenario than the Reference Case.

3.5 Storage, Networking, and Facility Infrastructure

In addition to servers, our model accounts for standalone storage and networking equipment (i.e., devices external to servers), as well as facility infrastructure energy requirements, including cooling systems and power distribution losses. For these three categories, this update leverages the data, assumptions, and methodologies used in the 2024 Report. Where necessary, values are extrapolated from 2028 to 2030. The extrapolation methods (linear or exponential change, or constant value) were chosen on a case-by-case basis, based on the given dataset being extrapolated.

Data center storage and networking systems are evolving, with networking equipment in particular becoming increasingly power intensive. To assess whether the assumptions retained from the 2024 Report remain appropriate, we compared our modeled fraction of total data center energy use attributable to storage and networking devices with independent estimates from Omdia, which estimates global power capacity requirements for these same equipment categories (Omdia, 2025). The relative energy demand across servers, storage, and networking equipment was very similar in the two models. Based on this consistency, we retained our 2024 Report assumptions in this study.

Cooling systems, which dominate facility infrastructure energy use, are also evolving rapidly to meet the increasing power densities of new AI systems. Facility infrastructure energy performance is commonly characterized using power usage effectiveness (PUE), defined as the ratio of total facility energy use (including cooling and power distribution) to IT equipment energy use (excluding facility infrastructure). Using the methodology from the 2024 Report, we estimate an average national PUE of 1.145 in 2024 for facilities serving AI equipment, declining modestly to 1.136 in 2030. Despite the evolving cooling system designs, we do not expect meaningful, industry-wide changes in the PUEs for data centers serving AI workloads compared to our previous modeling results. Therefore, evolving cooling system designs do not materially affect the total energy use estimates presented in this study.

3.6 Resulting Electricity Use by Equipment Type

Figure 9 shows the resulting U.S. data center electricity use by data center equipment category. Historical estimates for 2018 and 2024 are shown, along with 2030 values for each scenario modeled. Between 2018 and 2024, growth in total data center energy use is driven by both rapid proliferation of AI servers and continued growth in conventional server energy demand. Storage electricity continues to increase, but at a much slower rate than server energy, leading to a declining share of total energy use. Networking electricity consumption grows slightly, from 3.4% of total data center electricity in 2018 to 4.5% in 2024, due in part to the introduction of InfiniBand switch units. Infrastructure consumes 36% of total data center electricity in 2018, declining to 31% in 2024 as the national average PUE improves from 1.55 to 1.45. This improvement primarily reflects the continued shift in server energy use into the types of data centers that have the lowest PUEs, namely those deploying liquid cooling systems for AI servers.

In 2030, AI servers dominate energy consumption across all scenarios. In the Reference Case, AI servers account for 84% of total server energy use and 55% of total data center energy use. Storage energy use continues to decline as a share of the total. Networking maintains a roughly constant share of energy use, reflecting the substantial networking requirements of AI facilities, which drive networking energy growth roughly in line with server demand. Aggregate PUE continues to decline, reaching 1.36 in 2030 in the Reference Case. This is reflected in the declining share of total electricity use going to the Infrastructure category. The analysis does not adjust the energy use of conventional servers, storage or networking equipment, or PUE across the scenarios. Therefore, differences in the mix of energy use across equipment types are driven entirely by changes in AI server energy use. For example, in the Consolidated Deployment scenario, AI servers represent 53% of total data center energy use, while in the High Inference Energy scenario they account for 59%.

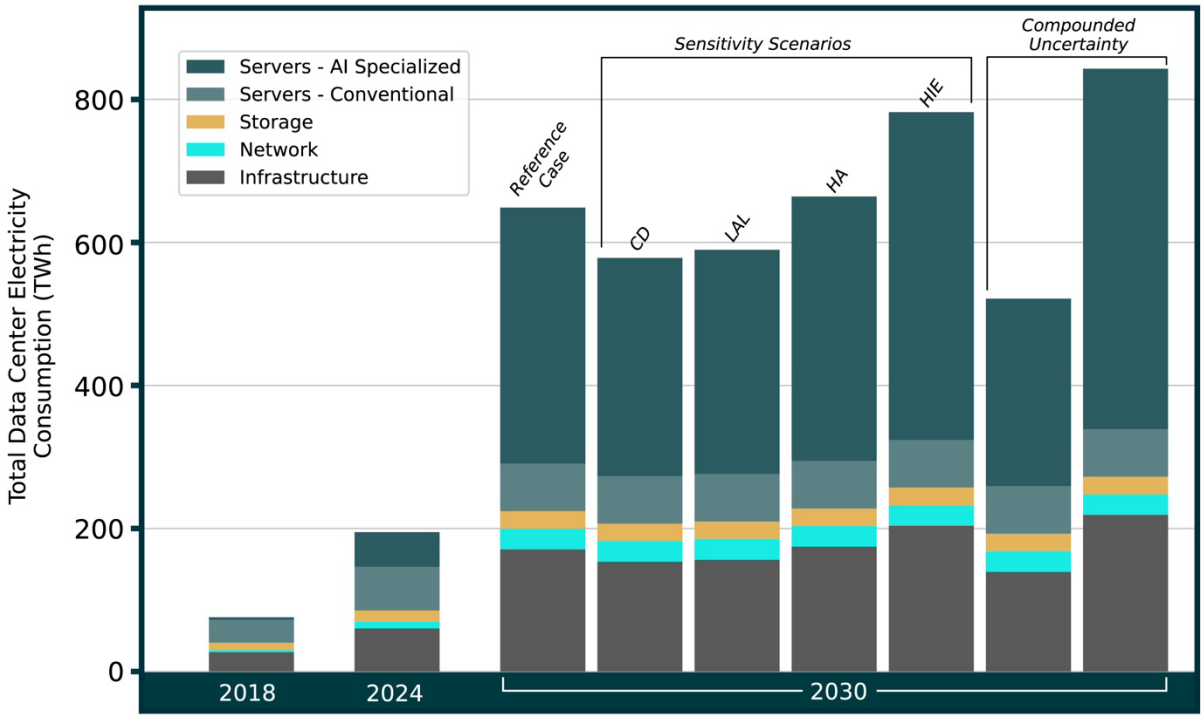

Figure 9. Total data center electricity use from 2018 to 2030 by equipment type

4. Conclusion

This 2025 update to Berkeley Lab's U.S. data center energy use analysis indicates a continued growth trend consistent with the prior 2024 Report. The continued deployment of GPU- and ASIC-accelerated servers, with ever-increasing computing capabilities and power consumption, led to a 14% increase in U.S. data center electricity use between 2023 and 2024. In the near term, growth accelerates, with Reference Case energy consumption increasing 22% from 2024 to 2025 and 29% from 2025 to 2026. Looking further out, Reference Case growth continues at a compound annual growth rate of 21%. Section 2 describes these results expressed in TWh, GW, and percent of U.S. electricity consumption. Increases in energy use are primarily driven by growth in both the quantity and rated power of accelerated servers shipped each year. Although successive generations of computing hardware improve in energy efficiency (in terms of number of computations performed per unit of energy), the scale and growth of computational demand more than offset these efficiency gains, leading to continued increases in absolute electricity consumption.

Significant uncertainties and data gaps remain in these estimates. To address these, we present four curated sensitivity scenarios that vary parameters when industry feedback or other information warranted modifications to Reference Case data and assumptions. These scenarios reflect cases where available data may differ from real-world practices or could plausibly change in the future, or where conflicting data sources exist. Across these scenarios, estimated energy use varies from -11% to +21% relative to the Reference Case. These uncertainties could plausibly compound, leading to larger variations in outcomes (roughly -20% to +30%). While we do not assume these variables are inherently correlated, this approach provides a "stress test" of their total potential impact on U.S. data center energy consumption. Further, several categories of uncertainty are not addressed in this update report. In the 2024 Report, we included uncertainty in AI server operational power levels (i.e., the power consumption at maximum utilization) and in PUE values. While these uncertainties may still exist to some extent, they are not included in this update because we have reliable measured or simulated data informing these parameters, and no information or feedback indicating they should be modified in either direction. Finally, as with any forecasting analysis, broader uncertainties remain regarding industry trends and practices, technology innovation affecting data center design and operation, and exogenous factors that could either boost or limit industry growth.

Future research should continue to reduce these uncertainties through data collection in partnership with the data center industry, direct measurements and testing of hardware and facility infrastructure systems, and expanded use of advanced simulation capabilities. Energy efficiency strategies should continue to be explored across all scales—from chip design, IT hardware system architecture, and computing resource management and scheduling to facility power distribution, backup power systems, and cooling technologies. Beyond annual energy estimates, future research should leverage expertise across DOE offices to characterize the temporal variability of data center electricity demand to better understand how these loads can be served by the power grid and/or on-site electricity generation resources. Relatedly, potential mechanisms for adjusting the timing of these loads (i.e., demand flexibility) should be explored and evaluated to increase the amount of computational demand that can be served with a given set of energy resources. As the data center industry continues to grow rapidly, it is increasingly

important to also understand how this growth interacts with, and may be constrained by, supply chain capacity, workforce availability, and water resources. With robust, ongoing analysis of U.S. data center energy use and expanded research into the broader data center ecosystem, including interactions with the power grid, supply chains, and workforce, the U.S. can make strategic, informed decisions to enable the continued deployment of AI technologies and meet the computing demands of the future.

Citations

Electric Power Research Institute (EPRI). (2026, February). Powering Intelligence 2026: Updated Scenarios of U.S. Data Center Electricity Use and Power Strategies (White Paper No. 3002034696). https://www.epri.com/research/products/000000003002034696.

Google. (2026). Cloud TPU Pricing. Google Cloud. https://cloud.google.com/tpu/pricing.

IDC. (2025a). Datacenter Semiconductor Consumption. International Data Corporation.

IDC. (2025b). Worldwide Quarterly Server Tracker Q3 2024. International Data Corporation.

IDC. (2026). Direct communication with Brandon Hoff, Executive Analyst, Enabling Technologies & Semiconductors: Accelerated Compute and Networking [Personal communication].

International Energy Agency. (2025). Energy and AI. https://www.iea.org/reports/energy-and-ai

Latif, I., Newkirk, A. C., Carbone, M. R., Munir, A., Lin, Y., Koomey, J., Yu, X., & Dong, Z. (2024). Empirical Measurements of AI Training Power Demand on a GPU-Accelerated Node (No. arXiv:2412.08602). arXiv. https://doi.org/10.48550/arXiv.2412.08602.

Leswing, K. (2025, November 14). The question everyone in AI is asking: How long before a GPU depreciates? CNBC. https://www.cnbc.com/2025/11/14/ai-gpu-depreciation-coreweave-nvidia-michael-burry.html.

McKinsey & Company. (2024, November 6). AI's power binge. https://www.mckinsey.com/featured-insights/week-in-charts/ais-power-binge.

NERC. (2026). 2025 Electricity Supply & Demand [Dataset]. https://www.nerc.com/programs/reliability-assessment--performance-analysis/electricity-supply--demand.

Newkirk, A. C., Fernandez, J., Koomey, J., Latif, I., Strubell, E., Shehabi, A., & Samaras, C. (2025). Empirically-calibrated H100 node power models for accurate AI training energy estimation. Environmental Research: Energy, 2(4), 045016. https://doi.org/10.1088/2753-3751/ae2486.

Norris, T. (2025, July 24). The Puzzle of Low Data Center Utilization Rates. https://www.powerpolicy.net/p/the-puzzle-of-low-data-center-utilization.

NVIDIA. (2025). NVIDIA GB200 NVL72 Product Page. https://www.nvidia.com/en-us/data-center/gb200-nvl72/ .

Omdia. (2025). Data Center Compute Research—Server Workload Tracker. Informa TechTarget, Inc.

Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A. C., Siddik, M. A. B., Holocek, B., Koomey, J. G., Masanet, E. R., & Sartor, D. (2024). United States Data Center Energy Usage Report. Lawrence Berkeley National Lab.

Shilov, A. (2024, October 24). Datacenter GPU service life can be surprisingly short—Only one to three years is expected according to unnamed Google architect. Tom’s Hardware. https://www.tomshardware.com/pc-components/gpus/datacenter-gpu-service-life-can-be-surprisingly-short-only-one-to-three-years-is-expected-according-to-unnamed-google-architect.

Stanley, B. (2025, November 21). Why GPU Useful Life Is the Most Misunderstood Variable in AI Economics. The Stanley Laman Group. https://www.stanleylaman.com/signals-and-noise/gpus-how-long-do-they-really-last.

Trueman, C. (2025, October 30). Google says TPU demand is outstripping supply, claims 8yr old hardware iterations have “100% utilization.” Data Center Dynamics. https://www.datacenterdynamics.com/en/news/google-says-tpu-demand-is-outstripping-supply-claims-8yr-old-hardware-iterations-have-100-utilization/.

Verrus. (2026). Beyond PUE: Rethinking Efficient Data Center Design in an Era of Power Scarcity. https://verrusdata.com/news/beyond-pue.

Appendix 1: Data Sources

Citation	Source	Source Type	Notes
IDC. (2025a). *Datacenter Semiconductor Consumption*.	International Data Corporation	Market Research	Tracking of AI accelerator shipments. Purchased by LBL.
IDC. (2025b). *Worldwide Quarterly Server Tracker Q3 2024*.	International Data Corporation	Market Research	Tracking of server shipments and sales. Purchased by LBL.
Omdia. (2024). *Data Center Compute Research—Server Silicon Tracker*.	Informa TechTarget	Market Research	Proprietary silicon market research. Paid product; complimentary access given to LBL.
Omdia. (2025). *Data Center Compute Research—Server Workload Tracker*.	Informa TechTarget	Market Research	Analysis of specific workload distributions. Paid product; complimentary access given to LBL.
Latif, I., et al. (2024). *Empirical Measurements of AI Training Power Demand...*	arXiv / Research Team	Academic Preprint	Detailed measurements on GPU-accelerated node power. Link
Leswing, K. (2025, Nov 14). *The question everyone in AI is asking: How long before a GPU depreciates?*	CNBC	News Article	Covers industry debate on hardware lifespan. Link

Citation	Source	Source Type	Notes
Google. (2026). *Cloud TPU Pricing*.	Google Cloud	Technical/ Corporate	Current market pricing for TPU usage. Link
NERC. (2026). *2025 Electricity Supply & Demand*.	North American Electric Reliability Corp.	Dataset	Forecasts for grid capacity and demand. Link
Norris, T. (2025, July 24). *The Puzzle of Low Data Center Utilization Rates*.	Power Policy	Industry Analysis	Analysis of efficiency vs. capacity. Link
Shehabi, A., et al. (2024). *United States Data Center Energy Usage Report*.	LBNL	Research Report	Primary energy usage study for U.S. facilities. Link
Shilov, A. (2024, Oct 24). *Datacenter GPU service life can be surprisingly short...*	Tom's Hardware	News Article	Reports on 1–3 year service life expectations. Link
Stanley, B. (2025, Nov 21). *Why GPU Useful Life Is the Most Misunderstood Variable...*	Stanley Laman Group	Financial Analysis	Financial implications of hardware depreciation. Link
OCP Workshop, October 2025	Open Compute Project	Industry Feedback	LBNL hosted a workshop at the OCP Global Summit to solicit industry feedback on model assumptions. Participants provided detailed information on accelerated server power consumption, ASIC market penetration, future GPU models,

Citation	Source	Source Type	Notes
			and deployment patterns for AI systems.
Data Center Infrastructure and Operations Expert	Digital Infrastructure Services	Industry Feedback	Provided comprehensive feedback across several topics, with a focus on differentiating the lifetime and idle power for various server types, utilization rates, and cooling technologies
Engineer	Digital Infrastructure Retailer	Industry Feedback	Noted that lifetimes for AI servers may vary depending on the workload (training vs. inference) and that training nodes may cascade into secondary clusters, resulting in a longer useful life. Also provided information on the use of liquid cooling in AI systems.
Senior Principal Engineer	Semiconductor Manufacturer	Industry Feedback	Provided feedback on conventional server idle power projections. Noted workload considerations for utilization metrics.
Vice President	Digital Infrastructure Services	Industry Feedback	Provided feedback on cooling systems for AI and environmental factors affecting energy and water consumption of those systems.
John Shalf	LBNL, NERSC	National Laboratory Staff	Provided insight into server operational practices that may impact idle power and utilization, referencing insights from contacts at cloud computing companies. Noted industry practices for demand forecasting and utilization for compute nodes. Highlighted the shift towards inference workloads rather than training.

Appendix 2: Relevant Methodology from 2024 Report

This appendix summarizes the core modeling framework used in this update. The methods are unchanged from those presented in the 2024 United States Data Center Energy Usage Report.

The equations below are provided for reference.

2.1 Installed Base Calculation

Installed base in year *t* is calculated as:

$$IB_t = IB_{t-1} + Shipments_t - Retirements_t$$

Retirements are modeled using a fixed service lifetime (L):

$$Retirements_t = Shipments_{t-L}$$

Where:

- IB_t= installed base in year t
- $Shipments_t$= annual equipment shipments
- L= assumed service lifetime

This framework is applied separately to:

- Conventional servers
- AI-accelerated servers
- Accelerator devices (if modeled separately)

2.2 Server Power Draw

2.2.1 Conventional Servers

Annual average power:

$$P_{avg} = P_{idle} + U \cdot (P_{max} - P_{idle})$$

Where:

- P_{idle}= idle power, calculated as a fraction of maximum power
- P_{max}= maximum power
- U= average utilization

2,2,2 Accelerated Servers

Operational power is modeled as:

$$P_{op} = \alpha \cdot P_{rated}$$

Annual average power:

$$P_{avg} = T_{op} \cdot P_{op} + (1 - T_{op}) \cdot P_{idle}$$

Where:

- T_{op}= fraction of time in operational state
- α= operational scaling factor

2.3 Annual IT Electricity Consumption

Total IT electricity consumption:

$$E_{IT,t} = \sum_{i} (IB_{i,t} \cdot P_{avg,i,t} \cdot 8760)$$

Where:

- 8760 = hours per year
- i= equipment category

2.4 Total Data Center Electricity

Total site electricity:

$$E_{total,t} = E_{IT,t} \cdot PUE_t$$

Where:

- PUE_t= power usage effectiveness for the applicable data center type