BERKELEY LAB
Energy Analysis & Environmental Impacts Division

2024 United States Data Center Energy Usage Report

Arman Shehabi, Sarah J. Smith, Alex Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor
Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory

December 2024

This work was supported by the DOE Office of Energy Efficiency and Renewable Energy (EERE), Office of Industrial Efficiency and Decarbonization Office (IEDO) under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

Copyright Notice

Citation

Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D. 2024. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001637

Acknowledgements

We would also like to acknowledge and thank the expert reviewers from approximately 50 organizations representing major data center equipment manufacturers and data center end users/owners that took time to answer questions from the authors and provided comments on draft versions of the report.

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

The research reported in this report was conducted by Lawrence Berkeley National Laboratory with support from the Department of Industrial Efficiency and Decarbonization 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

Executive Summary

The Energy Act of 2020 calls for the U.S. Department of Energy to make available to the public an update to Lawrence Berkeley National Laboratory's prior study entitled *United States Data Center Energy Usage Report* (2016). This report, designed to meet that Congressional request, estimates historical data center electricity consumption back to 2014, relying on previous studies and historical shipment data. This report also provides a scenario range of future demand out to 2028 based on new trends and the most recent available data. Figure ES-1 (below) provides an estimate of total U.S. data center electricity use including servers, storage, network equipment, and infrastructure from 2014 through 2028.

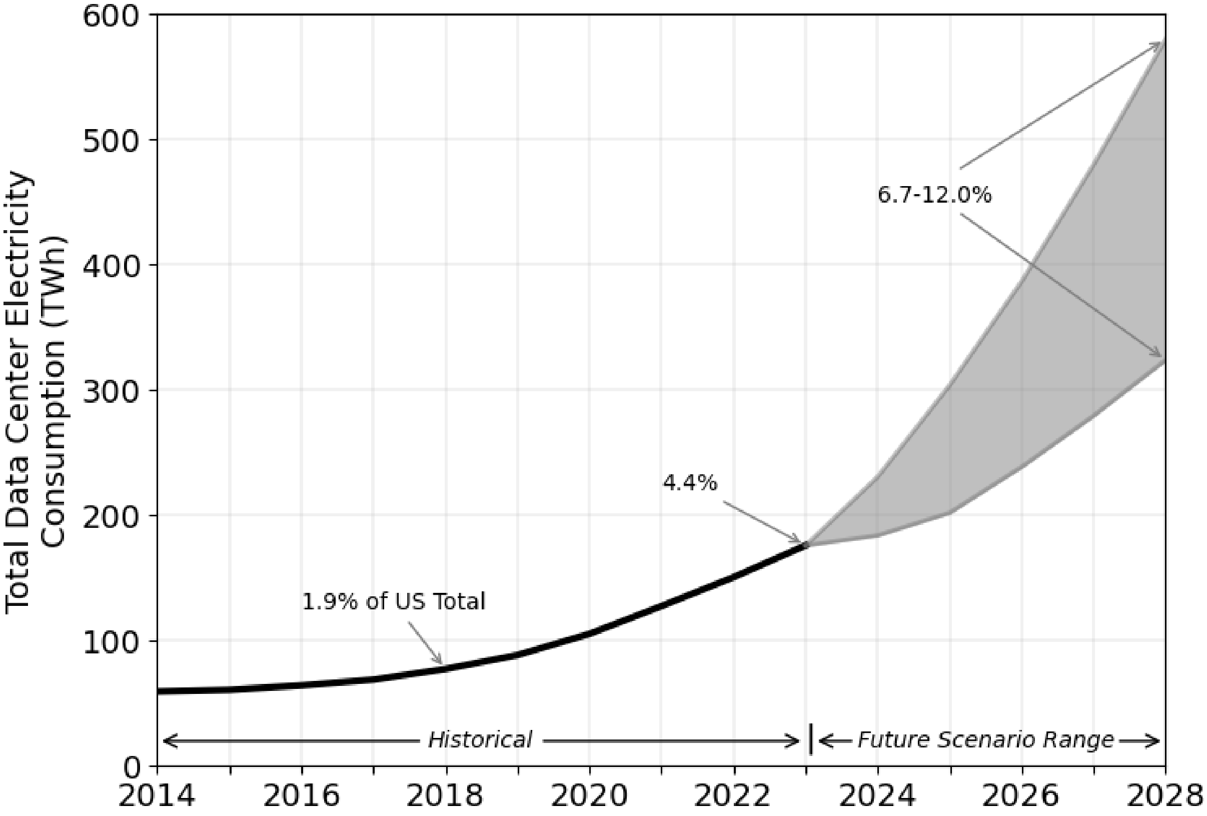


Figure ES-1. Total U.S. data center electricity use from 2014 through 2028.

As Figure ES-1 shows, U.S. data center annual energy use remained stable between 2014–2016 at about 60 TWh, continuing a minimal growth trend observed since about 2010. In 2017, the overall server installed base started growing and Graphic Processing Unit (GPU)-accelerated servers for artificial intelligence (AI) became a significant enough portion of the data center server stock that total data center electricity use began to increase again, such that by 2018 data centers consumed about 76 TWh, representing 1.9% of total annual U.S. electricity consumption. U.S. data center energy use has continued to grow at an increasing rate, reaching 176 TWh by 2023, representing 4.4% of total U.S. electricity consumption.

With significant changes observed in the data center sector in recent years, owing to the rapid emergence of AI hardware, total data center energy use after 2023 is presented as a range to reflect various scenarios. These scenarios capture ranges of future equipment shipments and operational practices, as well as variations in cooling energy use. The equipment variations are based on the assumed number of GPUs shipped each year, which depends on the future GPU demand and the ability of manufacturers to meet those demands. Average operational practices for GPU-accelerated servers represent how much computational power, and how often AI hardware in the installed base is used, to meet AI workload demand. Cooling energy use variations are based on scenarios in cooling system selection type and efficiency of those cooling systems, such as shifting to liquid base cooling or moving away from evaporative cooling. Together, the scenario variations provide a range of total data center energy estimates, with the low and high end of roughly 325 and 580 TWh in 2028, as shown in Figure ES-1. Assuming an average capacity utilization rate of 50%, this annual energy use range would translate to a total power demand for data centers between 74 and 132 GW. This annual energy use also represents 6.7% to 12.0% of total U.S. electricity consumption forecasted for 2028.

Historically, data center electricity use increased substantially from 2000–2005, roughly doubling during that period. During the early and mid-2010s, a shift from on-premise data centers to colocation or cloud facilities helped enable efficiency improvements that allowed data center electricity use to remain nearly constant at a time when the data center industry grew significantly, with a large expansion of data center services. The efficiency strategies that allowed the industry to avoid increased energy needs during this period included improved cooling and power management, increased server utilization rates, increased computational efficiencies, and reduced server idle power.

While many of these efficiency strategies continue to provide significant energy efficiency improvements in data center design and operation, the expansion of data center services into areas that require new types of hardware has ended the era of generally flat data center energy use. Most notably, the rapid growth in accelerated servers has caused current total data center energy demand to more than double between 2017 and 2023, and continued growth in the use of accelerated servers for AI services could cause further substantial increases by the end of this decade. The current and possible near-future surge in energy demand highlights the need for future research to understand the early-stage, rapidly changing AI segment of the data center industry and identify new efficiency strategies to minimize the resource impacts of this growing and increasingly significant sector in our economy.

Areas of future research identified in this report include benchmarking initiatives, collaborations with electric utilities, and technology development, all of which would be furthered by greater transparency in data center energy use, as the lack of data availability significantly limits the analysis in this report. The estimates in this report are based on a “bottom-up” energy use model that calculates total electricity use from an installed base of data center equipment. This method avoids overestimation that can be caused by tracking data center load for projects that have not yet selected a power provider, but requires many inputs and assumptions developed from limited publicly available data, proprietary market analyst data, and review by industry representatives and stakeholders. The lack of direct energy data available in a sector with rapidly evolving technologies limits the analysis in this report, especially when trying to understand and estimate future energy demand scenarios.

The results presented here indicate that the electricity consumption of U.S. data centers is currently growing at an accelerating rate. Figure ES-1 shows a compound annual growth rate of approximately 7% from 2014 to 2018, increasing to 18% between 2018 and 2023, and then ranging from 13% to 27% between 2023 and 2028. This surge in data center electricity demand, however, should be understood in the context of the much larger electricity demand that is expected to occur over the next few decades from a combination of electric vehicle adoption, onshoring of manufacturing, hydrogen utilization, and the electrification of industry and buildings. Research initiatives are needed not merely to identify strategies to meet data centers' future energy needs, but also to help stakeholders use this relatively near-term electricity demand for data centers as an opportunity to develop the leadership and strategic foundation for an economy-wide expansion of electricity infrastructure.

1. Introduction

In the United States, demand for computation has increased as the economy has digitalized. Over the last decade, the rise of expanded computational services has been accompanied by increased demand for associated cloud computing infrastructure. The data center industry's improving operational efficiency during this period initially held U.S. computing-related energy use flat while servicing greater demand (Shehabi et al. 2016; Masanet et al. 2020). The expansion of data center services into areas that require new types of hardware, such as artificial intelligence (AI) and cryptocurrency, is projected to break this trend, contributing to growth in American electricity consumption (IEA 2024). The exact magnitude of this growth remains a subject of debate due to the heterogeneity of estimation methodologies. These methodologies can be broadly categorized into three main approaches: bottom-up, top-down, and extrapolation (Mytton and Ashtine 2022). Analyzing these approaches will shed light on the strengths and limitations of each for accurate data center energy use estimation.

Bottom-up modeling delves deeply into the specificity of a data center. Using this method requires the meticulous gathering of data on individual components, such as the power draw of servers and a facility's overall efficiency. This information is then combined with estimates of equipment shipment and the number of servers already deployed (i.e., the "installed base") to calculate total energy consumption. This method provides insights into individual components of data centers and their energy use, allowing for more granular analysis. While this approach approximates a physical model of the data center, bottom-up modeling has limitations. It is heavily reliant on the accuracy of the data regarding installed base, equipment sales, and shipments, as well as accurate predictions of technological changes. Bottom-up approaches necessarily require knowledge of the underlying installed equipment base. While existing contracts for future sales mean near-term shipment forecasts are reasonably reliable, bottom-up methods cannot estimate at the same level of accuracy once they become projections, owing to variations in equipment characteristics and operation, and require periodic updates to reflect such changes in their foundational data (Malmodin et al. 2024).

Top-down modeling takes a broader view and leverages existing data from governments or industry organizations. This approach analyzes regional or national energy consumption totals, often derived from surveys and statistics. Aggregated estimates of service or product demand are then used to derive the required physical infrastructure to supply that demand. In a review of 46 studies on data center energy estimation, Mytton and Ashtine (2022) found this method to be less common, with only one example of an exclusively top-down approach found in their analysis.

The relative ease and low cost of implementation are clear advantages of a top-down model, as illustrated by an example in Malmodin et al. (2014). They utilized site-level energy measurements from various locations in Sweden, including seven large data/telecom centers, comprising measured service demand from 15 offices, 58 stores, and approximately 11,000 fixed and mobile sites. This approach was paired with bottom-up equipment statistics to extrapolate a global estimate. This approach allows for the use of defined values, including the total measured network data volumes and total energy consumption values from electricity meters, which allows for accuracy but only at a very high level of aggregation. Additionally, this approach encounters difficulties in understanding the impact of technology change compared to bottom-up approaches, due to a limited breakdown of data center components that make up the overall estimate. Functionally, top-down approaches interpolate the hardware stack based on service energy use.

The last approach, extrapolation, builds on existing estimates (either bottom-up or top-down) to forecast future energy use. Extrapolation begins with a base value of data center energy use that is then extrapolated, either using a projected annual growth rate or the basis of a service demand indicator, when normalized to a unit of service such as information technology (IT) traffic (Lei and Masanet 2021). This method projects future energy consumption based on energy intensity per unit of transmitted data by developing assumptions about overall market growth and energy efficiency improvements for future projects. While extrapolation methods are simpler and rely on less data than other methods, they rely heavily on the accuracy of the bottom-up or top-down models used, as well as assumptions about future market trends (Mytton and Ashtine 2022). These extrapolative studies are highly sensitive to errors in their input analyses, as extrapolation of those errors functionally compounds them. While they offer a useful benchmark for future demand, these methods must be used judiciously. A summary of the three main approaches to modeling data center energy use can be seen in Table 1.1.

Table 1.1. Characteristics of the Three Main Approaches to Estimating Data Center Energy Use: Bottom-Up, Top-Down, and Extrapolation

	Bottom-Up	Top-Down	Extrapolation
Key Inputs	• Equipment specification, e.g., server power draw • Data center infrastructure characteristics, e.g., power usage effectiveness (PUE) • Installed base/equipment shipment values	• Government or organization measurements, e.g., total data volume and total operator energy consumption	• Baseline year values (from bottom-up or top-down models) • Growth rate (which may factor in energy efficiency improvements and growth in data volumes or customer numbers)
Assumptions for Future Growth	• Future equipment shipment values • Future equipment trends, e.g., power density, graphics processing units (GPUs), liquid cooling • Energy efficiency improvements, e.g.,	• Relationship between demand and energy consumption, usually data volume • Future customer growth, e.g., the number of internet provider customers	• Relationship between demand and energy consumption, usually data volume • Energy efficiency improvements, e.g., PUE improvements

	PUE improvements or server power draw improvements	• Energy efficiency improvements (usually broad because of the lack of specific breakdowns such as PUE or server power)	or server power draw improvements
Main Limitations	• Availability of installed base and server equipment values • Ability to project trends in equipment types and energy efficiency improvements more than a few years into the future	• Accuracy and availability of organization data • Ability to project future trends in energy efficiency • Correlation between inputs, e.g., data consumption per customer	• Correlation between inputs, e.g., energy consumption per unit of data • Ability to project future trends in energy efficiency
Examples	• Masanet et al. (2020) • Shehabi et al. (2016) • Malmodin et al. (2014) • Koomey (2011) • Masanet et al. (2011) • Malmodin et al. (2010) • Brown, et al. (2007) • Koomey (2007)	• Malmodin et al. (2014)	• Belkhir (2018) • Andrae (2017) • Andrae and Edler (2015)

Table adapted from “Sources of data center energy estimates: A comprehensive review” (Mytton and Ashtine 2022).

Bottom-up models have historically shown more modest growth in data center energy use. In a 2007 report to Congress (Brown et al. 2007), the U.S. data center sector was estimated to have consumed about 61 billion kilowatt hours (kWh) in 2006, equivalent to 1.5% of total U.S. electricity consumption, for a total electricity cost of about $4.5 billion (in 2006 dollars). In a follow-up report (Shehabi et al. 2016), the data center sector was estimated to have consumed about 70 billion kWh in 2014, equivalent to 1.8% of total U.S. electricity consumption. This estimated level of electricity consumption is comparable to the amount of electricity consumed by approximately 6.7 million average U.S. households.

While these estimates paint a picture of a gradual rise in data center energy use, other projections suggest a much more dramatic scenario. According to Reuters, nine of the 10 largest U.S. electric utilities said data centers would be the main source of customer energy demand growth, stemming from technologies like generative AI (Kearney 2024). The electric utility Southern Company projects data centers will increase its electricity sales by 6% each year from 2025 to 2028, up from its prior predicted growth rate of 1–2%. Financial and industry analytics firms share similar findings: Boston Consulting Group estimates that data center energy use could triple from 2.5% of U.S. electricity consumption to 7.5% by 2030 (Lee 2023). Companies like Georgia Power, a subsidiary of Southern Company, have drastically increased their near-term load growth predictions based on growth from data centers (Halper 2024). However, data center operators and federal agencies are seeking greater visibility into the company’s heavily redacted resource planning filings, criticizing the utility’s approach to load forecasting. Microsoft challenged Georgia Power’s modeling methods, highlighting concerns that the model is undervaluing renewable energy while overestimating data center load for projects that have not yet selected Georgia Power as their provider (Allsup 2024).

Utilities consistently emphasize data centers, and specifically AI, as drivers of projected increases in electricity demand. Seven of eight utility forecasts highlighted AI as a key driver of demand growth (Wilson and Zimmerman 2023). Strategic consulting and financial analytics firms consistently predict expanded demand (Avelar and Donovan 2023; Anderson, Sweeney, and Canonica 2023). It is noteworthy that in parallel with predicted increases in AI-related computational demand, specialized microelectronics industry analysts have made more reserved estimates of short- and medium-term load growth due to greater emphasis on manufacturing. Industry analytics firm SemiAnalysis projected AI would account for 4.5% of 2030 electricity use in its base case estimate (Patel, Nishball, and Ontiveros 2024). Institutions specializing in projecting energy use (IEA 2017) or IT power specifically (Nelson 2022) have also updated previous analyses to incorporate AI investment (IEA 2024a; IEA 2024b; Avelar and Donovan 2023). Plots of both historical academic estimates and associated projections of data center energy use in the United States and globally can be seen in Figures 1.1 and 1.2.

We should not dismiss these projections out of hand, as the most extreme cases would constitute a significant stress on electrical infrastructure. Many underlying models on which financial analytics firms build these projections are proprietary or methodologically opaque (Goldman Sachs Research 2024; Jeff Brown 2024; Bloomberg 2024). Others are often extrapolations from utility load growth forecasts based on market research demand projections (Aljbour et al. 2024). While not meaningless, historical utility demand forecasts consistently overestimate both peak and average demand (Carvallo et al. 2018). Rigorous models of data center energy use are needed to adequately calibrate both future investment and public concern.

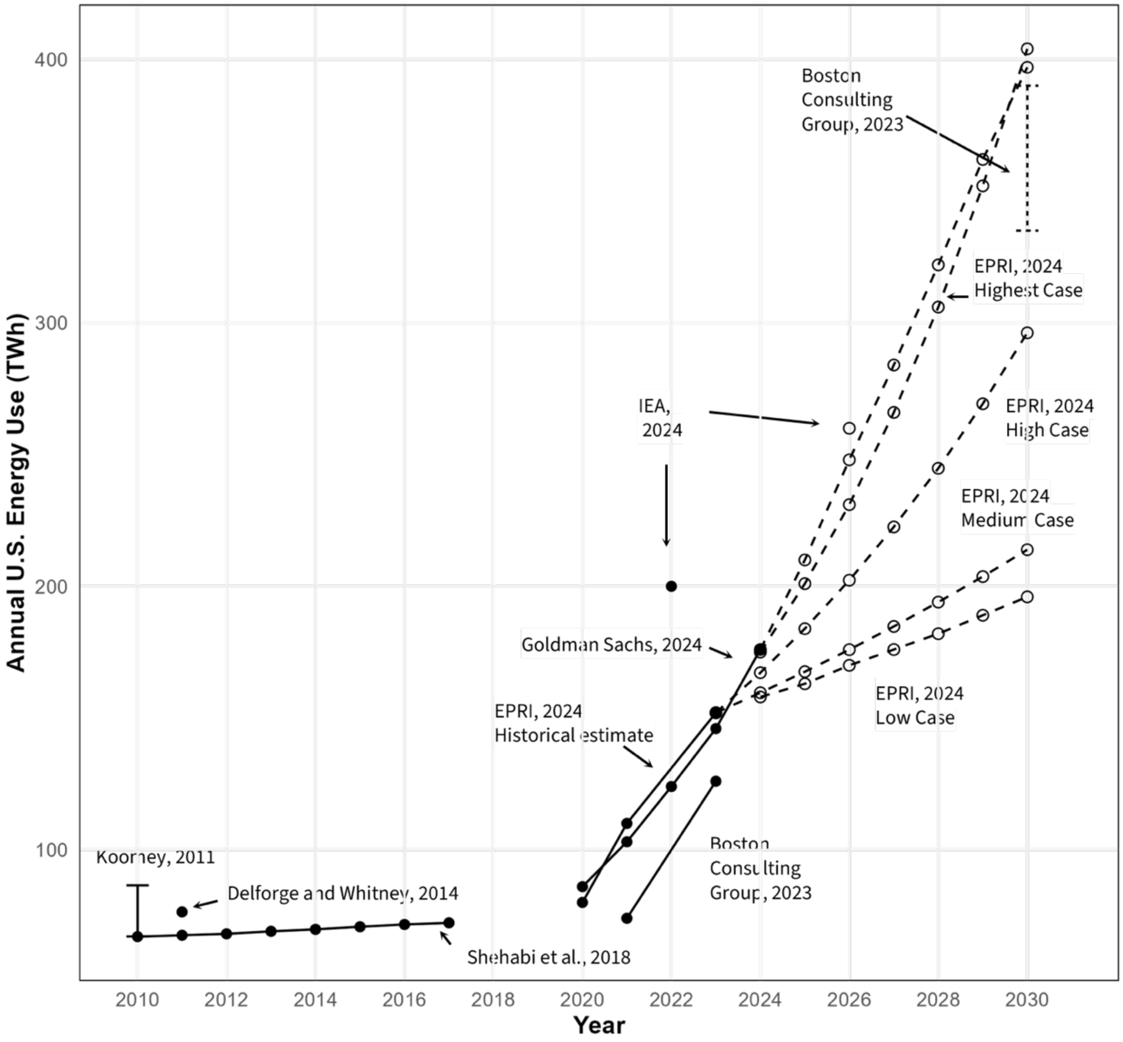


Figure 1.1. Academic and industry historical estimates of U.S. data center energy use.
Plot also includes future projections from those sources. Historical estimates are shown with solid lines, and projections are shown with dashed lines.

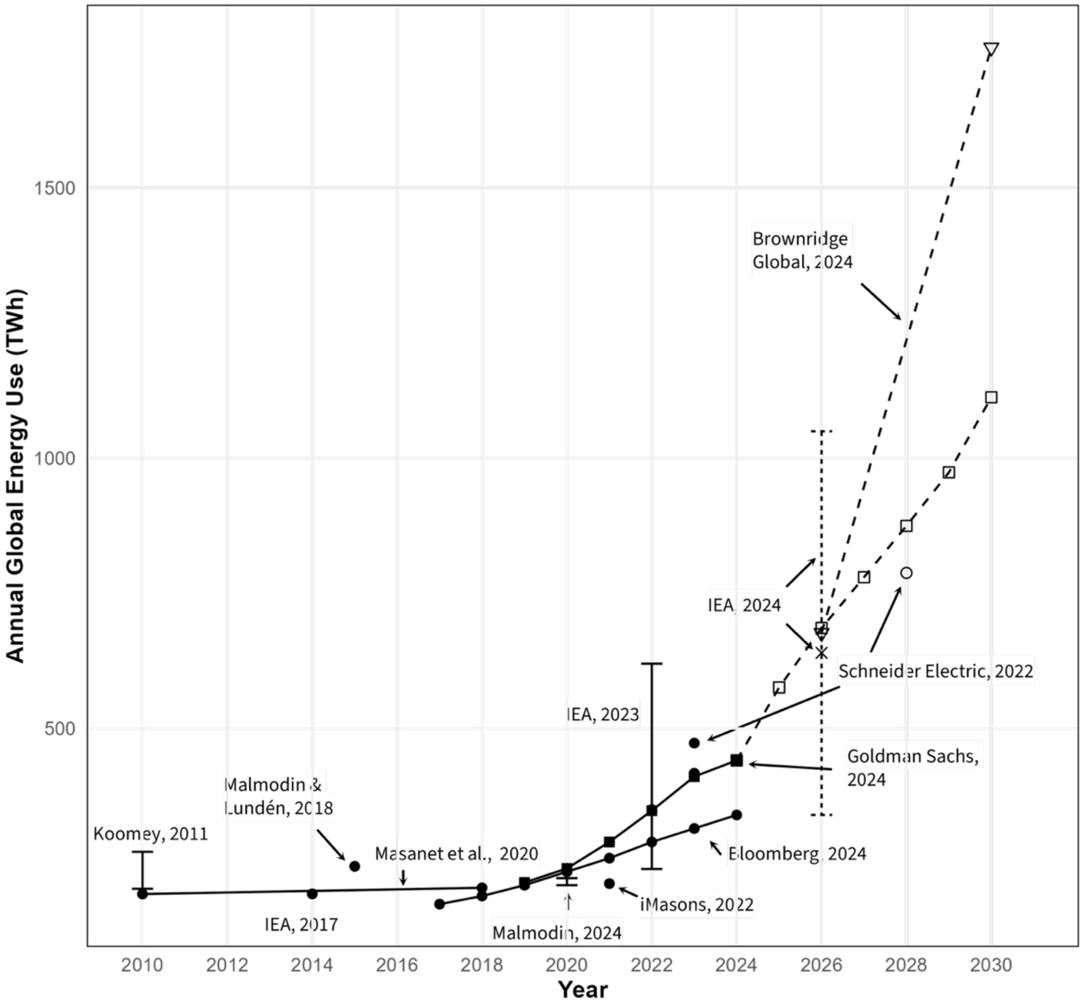


Figure 1.2. Academic and industry historical estimates of global data center energy use. *Plot also includes future projections from those sources. Historical estimates are shown with solid lines, and projections are shown with dashed lines.*

The Energy Act of 2020 (U.S. Congress 2020) calls for the Department of Energy to make available to the public an update to the *United States Data Center Energy Usage Report* from Lawrence Berkeley National Laboratory (Berkeley Lab), which estimated historical data center electricity consumption back to 2000 (Shehabi et al. 2016), hereafter referred to as “the 2016 report”. This analysis is designed to meet this Congressional request to update the 2016 report and builds on the bottom-up, technology-based modeling approach developed by Koomey, Masanet, and coauthors (Capuccio and Craver 2007; Koomey 2007; 2008; Koomey and Masanet 2021; Masanet et al. 2011; Shehabi et al. 2018). Bottom-up methods are recognized for their accuracy and robustness, stemming from their reliance on meticulous data collection regarding the IT equipment inventory within data centers (Lei, Masanet, and Koomey 2021). While limitations exist in using this approach, top-down estimates instead interpolate backwards from service demand, which has become increasingly uncoupled from electricity use in the IT sector (Malmodin and Lundén 2016). This study seeks only to estimate the direct energy use by data centers, not any underlying economic factors or transitions that may substantially change the underlying environment and technological base.

The remainder of the report is organized as follows. Section 2 describes the methodology employed in this study. Section 3 describes the categories and subcategories of IT equipment modeled in this study, including the power consumption estimates for various equipment types, how those estimates are evolving over time, and the calculation and results for the installed base of IT equipment. Section 4 then describes the different data center types considered in this study, how IT equipment is allocated across those space types, and the modeling of non-IT power and water demands (namely from the cooling systems) to estimate site power usage effectiveness (PUE) and water usage effectiveness (WUE). Section 5 then presents the resulting estimates for total electricity consumption, as well as the emissions and water footprints considering the source energy of regional electricity grids. Section 6 provides a separate analysis for cryptocurrency. Section 7 summarizes findings, highlights limitations, and provides suggestions for future research.

2. Methodology Overview

This report utilizes a “bottom-up” model, with data and assumptions starting at the equipment level being scaled up and aggregated to generate total results. Figure 2.1 shows the overall structure of the model, including input data sources and the major units of analysis. As shown in the second column, the model characterizes IT equipment as either servers, storage, or networking equipment, and it uses shipment data to generate estimates for the “installed base” of equipment in each year. The installed base represents the quantity of equipment that is being operated in a given year; these values are distinguished by subcategories of equipment, the type of data center the equipment is in, and other characteristics. Each installed base has a complementary set of assumptions regarding the power draw of the installed equipment. Like the installed base, these wattage assumptions might vary across equipment and data center types. For servers, there is an underlying model of power draw that leads to this assumption, considering rated power, maximum power, server operational time, and idle power to estimate annual average wattages.

Adding up the product of installed base and wattage across each equipment category leads to total estimates for annual IT equipment electricity consumption. This value can then be multiplied by modeled PUE and WUE values to estimate total on-site electricity and water demand across the data centers. The PUE and WUE models consider assumptions for the types of cooling systems present at different types of data centers, as well as the locations of data centers and therefore the average surrounding weather conditions. These assumptions are paired with the results from a detailed, physics-based model of cooling electricity demand, which also accounts for ancillary electrical demands such as lighting and uninterruptible power supplies (UPS) to estimate the average PUE and WUE for different types of data centers.

Finally, these on-site demands are combined with models of the carbon and water intensity of the electrical grid to estimate the total greenhouse gas (GHG) footprint and water consumption related to data centers.

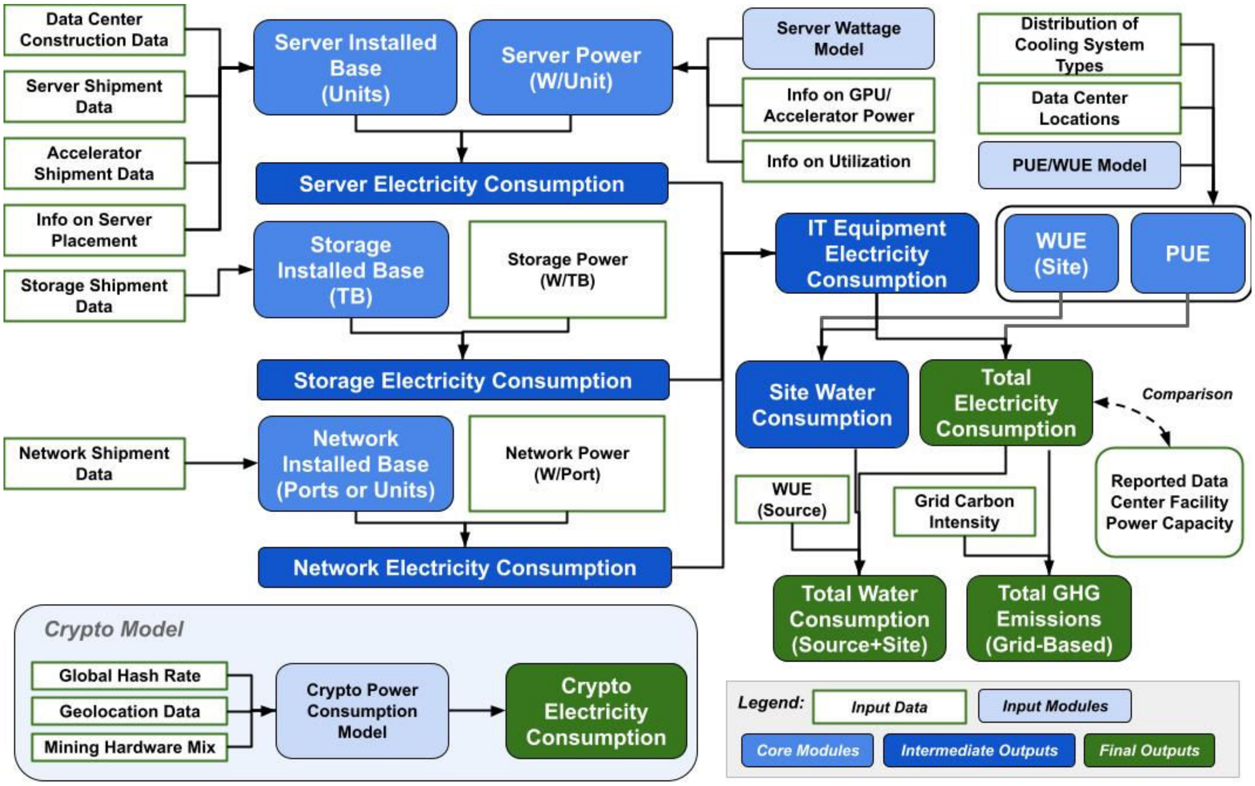

Data Center Construction Data
Server Shipment Data
Accelerator Shipment Data
Info on Server Placement
Storage Shipment Data
Server Installed Base (Units)
Server Power (W/Unit)
Server Wattage Model
Info on GPU/ Accelerator Power
Info on Utilization
Distribution of Cooling System Types
Data Center Locations
PUE/WUE Model
Server Electricity Consumption
Storage Installed Base (TB)
Storage Power (W/TB)
IT Equipment Electricity Consumption
WUE (Site)
PUE
Storage Electricity Consumption
Site Water Consumption
Total Electricity Consumption
Comparison
Reported Data Center Facility Power Capacity
Network Shipment Data
Network Installed Base (Ports or Units)
Network Power (W/Port)
WUE (Source)
Grid Carbon Intensity
Network Electricity Consumption
Total Water Consumption (Source+Site)
Total GHG Emissions (Grid-Based)
Crypto Model
Global Hash Rate
Geolocation Data
Mining Hardware Mix
Crypto Power Consumption Model
Crypto Electricity Consumption
Legend:
Input Data
Input Modules
Core Modules
Intermediate Outputs
Final Outputs

Figure 2.1. Flow chart for the data center electricity model used in this study.

The modeling framework described above is very similar to the model employed in the 2016 U.S. data center energy use study (Shehabi et al. 2016). Key updates include extensive quantification and characterization of accelerated servers used for AI applications, consideration of various cooling system types and outdoor temperatures when modeling PUE and WUE, calculation of carbon and water footprint of electricity consumed based on local grid mixes, and an estimate of electricity demand from cryptocurrency mining. Details on each of these modeling advancements will be given in the remaining sections of the report.

Historical Best Estimates and Future Scenarios

Total electricity, water, and emissions associated with U.S. data center operation are estimates for the period 2014–2028. The years prior to 2024 are historical estimates and, while there are significant variations in the equipment and operational practices of that equipment across data centers, the inputs and assumptions for the core modules during this historical period characterize best estimates of average representative values based on collected data and stakeholder feedback during the review process. Beginning in 2024, estimates are based on forecasts for equipment shipments and predictions of future operational practices, which are inherently less certain than historical estimates. This uncertainty is magnified with the recent rapid growth in accelerator shipments, which significantly affect total electricity use values given the current and anticipated power demand of GPU chips. This report addresses that uncertainty by presenting future electricity demand as a range that represents different electricity demand scenarios. The scenarios are based on possible variations in future accelerator shipments, the operational time for AI servers, the operational power draw of AI servers, and the efficiency of specific data center cooling systems. The specific variation and rationale for those variations are described in the corresponding sections in the following section.

3. IT Equipment Installed Based and Power Draw

Server Types

Servers are split into two main categories: conventional and AI specialized servers. **Conventional** servers are those with configurations of processing, memory, and networking equipment that are designed for serving legacy data center workloads. While there is significant variation in the configuration, operation, and power consumption of these servers, the primary driver of the power consumption is the number of central processing units (CPUs) present in the device. Therefore, we break out conventional servers into categories of single processor, dual processor, and multiple processor count. Traditionally, the vast majority of servers deployed at data centers were **dual processor**. In more recent years, a shift to **single processor** servers has occurred, with several “blade”-style servers being installed in a chassis (e.g., Facebook’s Yosemite design [Haken et al. 2021]). **Multiple processor count** servers are those with four or more CPUs, with some machines having 32 or even more CPUs; these servers are typically used in scientific computing, big data processing, or content delivery applications. They make up a small portion of the installed base in terms of the number of servers, but their total electricity demand can be substantial. When modeling the power draw of these machines, we include additional specificity in the CPU count (4 vs. 8 vs. 16 CPUs), but we group these servers together as multiple processor count (MPC) servers when presenting results.

AI specialized servers are further categorized as “AI accelerated” and “AI non-accelerated,” based on whether or not accelerators are present. Accelerators are additional processing units that complement the CPU function and allow the server to more quickly process large quantities of calculations in parallel. The most common accelerator type is the graphics processing unit (GPU), though other types of accelerators (e.g., ASIC, TPU) are growing in popularity. **AI non-accelerated** servers are two-processor servers that are dedicated to AI workloads but do not contain any accelerators. Compared to conventional dual processor servers, these machines typically have more robust memory and networking configurations and use higher-power CPUs. The types of data centers where they are installed and the way in which they are operated can also differ from conventional server deployments. **AI accelerated** servers contain two or more accelerators and are dedicated to AI workloads, such as training and inference for large language models. Each server includes two CPUs, modeled after the NVIDIA DGX line of servers. Within the AI accelerated category, servers are split into three groups based on the number of GPUs shipped with the servers (either 2, 4, or 8). Larger-scale AI server configurations have been announced, namely the 72-GPU rack-scale GB200 from NVIDIA (NVIDIA n.d.-a), but these are not yet being deployed and the power draw of these systems is expected to scale close to linearly from 8-GPU machines.

Server Power Draw

The actual electricity demand for a given server fluctuates throughout the day depending on the amount of work it is doing. In the past, these fluctuations were minimal, as servers consumed electricity near their maximum levels even when sitting idle, but as the power scaling ability of servers has improved, it is important to capture these fluctuations. Further, it is critical to note that servers rarely, if ever, actually draw the full wattage for which they are rated on specification sheets. These "rated power" levels are used to inform the design of power supply and electrical systems and therefore must conservatively include the absolute maximum wattage of every server component running at the exact same time. Ultimately, this creates several wattage levels to be considered when analyzing server electricity demand:

- **Rated power:** the absolute maximum possible power draw of the server, as one might find on an equipment specification sheet. This is generally determined by summing up the thermal design power (TDP) of every component within the machine.
- **Maximum power:** the observed power draw when the server is operating at its maximum computational workload. This can be found via testing benchmarks such as SPEC (SPEC n.d.) or by metering in-use servers for a sufficient amount of time such that maximum workloads are likely to have been achieved and then finding the maximum power draw.
- **Operational power:** the observed power draw when the server is operating in a "typical" workload mode. This term is most relevant for AI training servers that have a relatively predictable power draw when they are doing a training run. It is less relevant for conventional servers that are dealing with continuously variable workloads.
- **Idle power:** the power draw when a server is sitting idle with no computational demand. Lower power draws are possible with enhanced power saving settings such as "deep sleep" modes, but here we consider idle power as when the server is on and in active standby mode but not completing any computations.
- **Annual average power:** the average power draw over an entire year, considering all operating modes and power levels.

For this study, we aim to understand the total annual electricity consumption of servers. Therefore, the annual average power is our metric of interest. Average power draw is calculated in our model by approximating all of the server's operations as one of two states: operational or average. This gives us the following equation for calculating average power:

$$P_{avg} = P_{op} \cdot T_{op} + P_{idle} \cdot T_{idle}$$

Where P represents a power draw (wattage) metric and T represents the fraction of time throughout the year that the server is in a given state. For AI servers, particularly those used for training, this binary representation is reasonably representative of their behavior throughout the year: they are either conducting a training run, or they are not. The power metrics used in the equation are straightforward and as defined above, and the operational time metric is the fraction of time we expect servers to be actively conducting training runs (idle time is then 1-operational time).

For conventional servers, the operational state (and therefore power demand) is much more varied, but our choice of metrics is such that this formulation is mathematically equivalent to the method used in the prior report. We use the maximum power draw of the server to represent the operational power and use an expected average utilization level as the operational time metric. This gives us an average power draw that is equivalent to a linear interpolation between idle and max power, at a point between those two power levels that is representative of their average workload level. In other words, if we think a server is operating at 30% workload on average, we will end up with an average power that is 30% of the way between idle and maximum power. The assumptions for operational power, idle power, and operational time are described in the following subsections.

Server Operational Wattage

Conventional Servers

Figure 3.1 shows the assumed power draw of operating conventional (non-AI) servers based on the year and number of processors present. The operational power draw for single and dual processor servers shipped prior to 2014 is the maximum power for “Volume 1S” and “Volume 2S+” servers in the previous report (118 and 365 W, respectively). For 2023, the operational wattage of dual processor servers is set to 600 W based on the Green Grid Server Energy Efficiency Database (SEED) average for that year. Intermediate years between 2015 and 2023 are linearly interpolated. After 2023, the operational wattage is determined by applying the linear growth trend from the SERT database to the 2023 values, ultimately reaching 782 W by 2028.

While ultimately the average wattage from the SEED was used for the current wattage of conventional servers, substantial analysis went into exploring additional data sources and understanding the underlying trends and uncertainties in conventional server power draw. Initially, conversations with industry analysts suggested that typical dual processor servers going into data centers in 2023 had maximum power levels, on average, of 1 kW (or even more). Data from the SPEC benchmarking database (SPEC n.d.) supported this trend, with average wattages of around 750 W for 2023–2024. Generally, servers in the SPEC database are expected to be the most efficient configurations and operational settings possible, so this supported an expectation that the average server was at a wattage well above those values. However, further analysis of SPEC and SERT data showed that the SPEC servers had substantially more cores per CPU than servers in the SERT database (approximately 150 vs. 50 cores per CPU, on average, respectively). Additional data on core counts obtained from the IDC confirmed that, on average, servers being shipped to data centers have core counts more in line with the SEED data than the SPEC data. Therefore, the total wattage numbers presented in the SPEC database are not only unrepresentative due to their computationally efficient configurations, but because they are much more powerful servers than the average. While SPEC data still provides numerous valuable insights, we found the SEED data to be more representative of typical servers and therefore used the average wattages found there for our modeling.

Single processor server operational wattage is determined by following the same growth trend found in SEED, flattening after 2025. Operational wattage for 4 processor count servers is assumed to be two times that of dual processor servers, and 8 processor count server operational power draw is assumed to be two times that of 4 processor servers. Data from the SERT database supports this relative scaling between servers with various processor counts.

An additional category in the data, 16+ processor count servers, represents the remainder of the higher performance computing market. In the prior study, these servers were captured in the “high-end” category. Data on the wattage of these servers was first analyzed in 2007 by Jon Koomey (2007), who created an estimated power draw time series for 2000–2007 based on the most popular servers being shipped at that time. In the 2016 report, this trend was extrapolated to 2020 with an annual growth rate of 7%. In retrospect, this 13-year extrapolation led to an overestimation of the amount of power these servers would consume, with the high-end servers reaching 20 kW in 2020. In this study, we use a simple 10-kW power draw assumption for all years, as this is in line with the power draw of the most powerful servers on the market today. This causes some discrepancies between this study’s pre-2020 estimates and the prior study results, which will be discussed further in a later section.

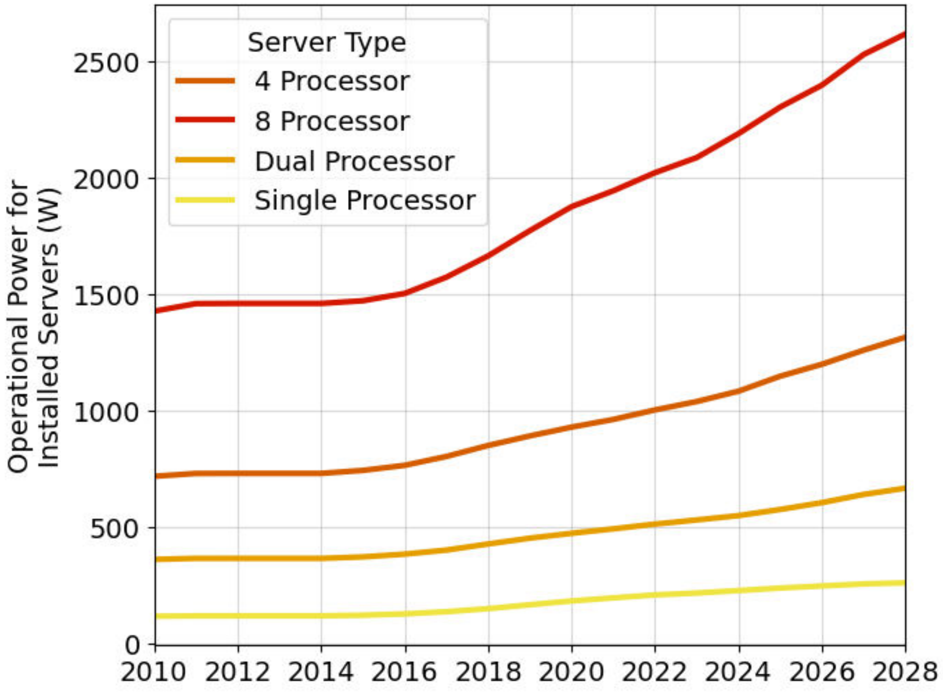


Figure 3.1. Operational power draw for conventional servers across the installed base.

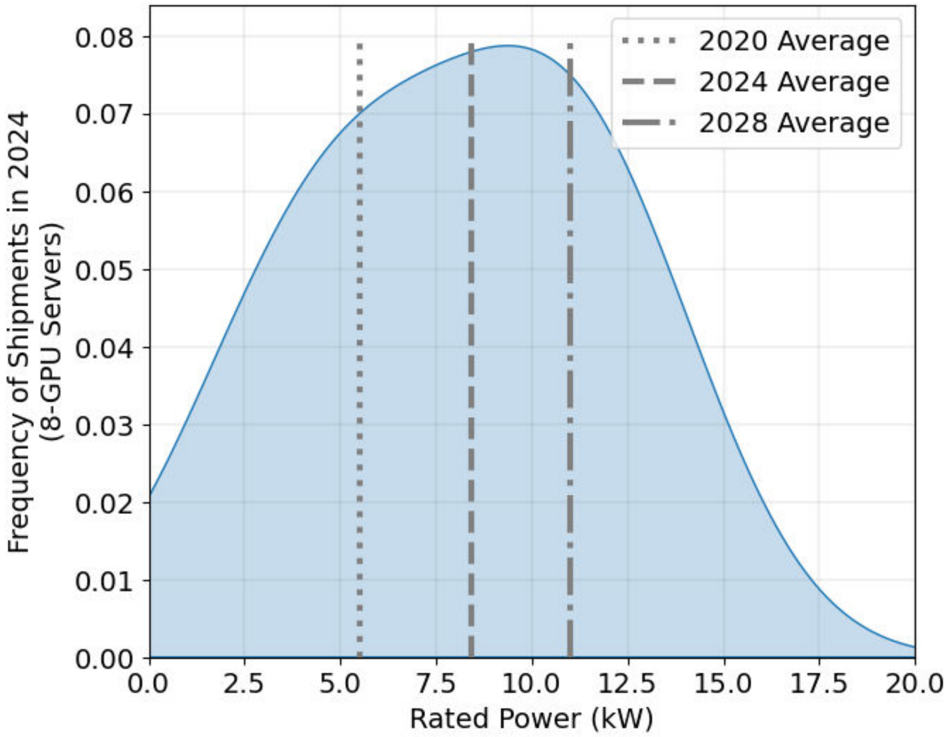


Figure 3.2. Distribution of rated power draw for shipped servers.

AI Servers

For AI servers, the operational power draw is calculated as a fraction of the rated power of the servers, which is shown by shipment year in the left panel of Figure 3.3. These values are based on the rated power for NVIDIA DGX 8-GPU systems shipped in each year, which are then averaged based on the distribution of GPU shipments in each year. For example, if in a given year, 25% of shipped GPUS are in the NVIDIA A100 class, with a rated power of 6.5 kW (NVIDIA n.d.-b), 25% are in the NVIDIA A2, A40, or L4 class, with a rated server power of 3.4 kW (Omdia, 2024), 25% are in the NVIDIA H100 class (NVIDIA n.d.-c), with a rated server power of 10.2 kW, and 25% are in the B100 class, with a rated power of 12.2 kW (Omdia, 2024), the resulting average rated power for that year would be about 8 kW (i.e., the weighted average of 6.5, 3.4, 10.2, and 12.2). Average rated power for 2020, 2024, and 2028 are shown in Figure 3.2. This results in the average rated power numbers by shipment year shown in the left panel of Figure 3.3.

For non-accelerated AI servers, we assume that the power draw in 2023 is the 90th percentile value of the dual processor server power in the SERT database, then follows same linear trend as dual processor servers. We then estimate that, during typical operation, the power draw for these servers is 70% of the rated power, leading to the values in the right panel of Figure 3.3. This is based on an analysis of measured power data from accelerated servers in workload (Newkirk 2024). This analysis found that across both single-node and multi-node AI training workloads, 8 H100 GPU nodes averaged 74% of manufacturer rated thermal design power in workload. Assuming a heterogeneous mix of workloads, some of which won't computationally saturate the hardware, we elected a value of 70%. Operational power in the years 2024 to 2028 is varied between 60% and 80% of the rated power to reflect possible differences in the future operation of these servers.

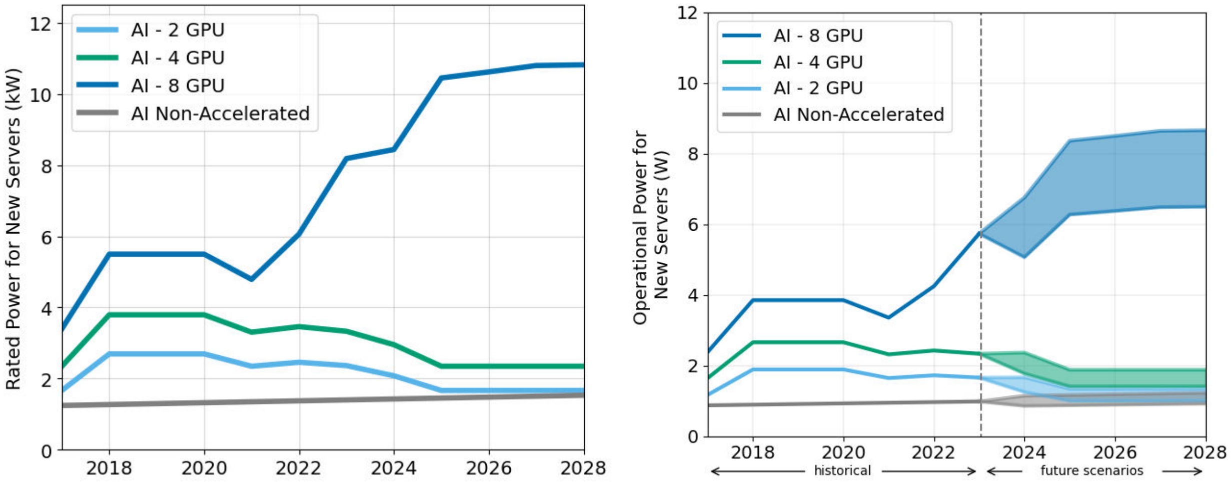


Figure 3.3. Rated power for accelerated servers (left); operational power for accelerated servers (right).

Accelerated Server Operational Parameters

The energy footprint of artificial intelligence (AI) computation differs fundamentally from traditional enterprise computing in several key aspects that directly impact modeling approaches. At its core, AI computation consists of massive quantities of parallel matrix operations that create distinctive facility-level power demands when performed at scale. These workloads are highly parallelizable and show consistent performance improvements with increased computational scale, driving the development and deployment of specialized hardware infrastructure.

This computational profile has led to a distinctive hardware stack dominated by specialized accelerators, primarily graphics processing units (GPUs) for training workloads. While these accelerators perform the bulk of computation, they operate as part of complex nodes including supervisory central processing units (CPUs), memory, and high-bandwidth interconnect. This integration means that individual components can throttle others, creating characteristic load profiles; empirical measurements consistently show that even in computationally intensive workloads, node-level power demand rarely approaches manufacturer rated maximums.
A crucial distinction exists between training and inference workloads. Training involves intensive, long-duration computation with relatively predictable power draws, typically performed in dedicated facilities with specialized hardware. This creates opportunities for efficiency optimization through facility siting and cooling system design. In contrast, inference presents more variable loads across a heterogeneous hardware mix, making it more challenging to model but increasingly important as deployment scales. The relationship between inference and training energy use continues to evolve as models grow in both scale and capability.

The consistent relationship between computational scale and AI capability improvements suggests that efficiency gains will typically be reinvested in larger models rather than reducing absolute power demand. However, the scale and capital intensity of these facilities creates strong incentives for operational efficiency, particularly in cooling and power delivery systems. These characteristics inform our modeling approach: We derive parameters from empirical measurements of operational systems, while accounting for the distinctive load profiles and efficiency characteristics of AI computation.

We collected empirical power draw data for AI training workloads on a current generation 8-H100 DGX node in collaboration with Brookhaven National Lab (Latif et al. 2024). We then integrated these results with open-source data from public benchmarking results for multinodal training, with summary statistics of these workloads shown in table 1 below. Using these measurements, we derived a statistical estimate of average node power demand in training for computationally saturated workloads of 7.9 kW or ~78% manufacturer rated power (Newkirk 2024). As commercial hardware will be running a mixture of workloads at a mixture of computational intensity, we parameterized annualized workload demand at 70% of node rated power. For a more detailed analysis of these techniques and a review of the relevant literature on AI-related computational demand, see Newkirk (2024). One finding from this work was the characteristic fluctuation in power demand for AI training as compared to more conventional HPC workloads, which can be especially taxing to grid operators. A plot of a characteristic time series for training can be seen in Figure 3.4.

While training energy use has dominated historical analysis of AI's energy footprint due to its discrete, measurable nature, inference energy consumption presents a more complex modeling challenge with greater long-term implications. The challenge stems from four interacting uncertainties: the energy efficiency of inference hardware, the mix of hardware types used (from ASICs to GPUs to CPUs), the scale and nature of user demand, and the characteristics of inference facilities.

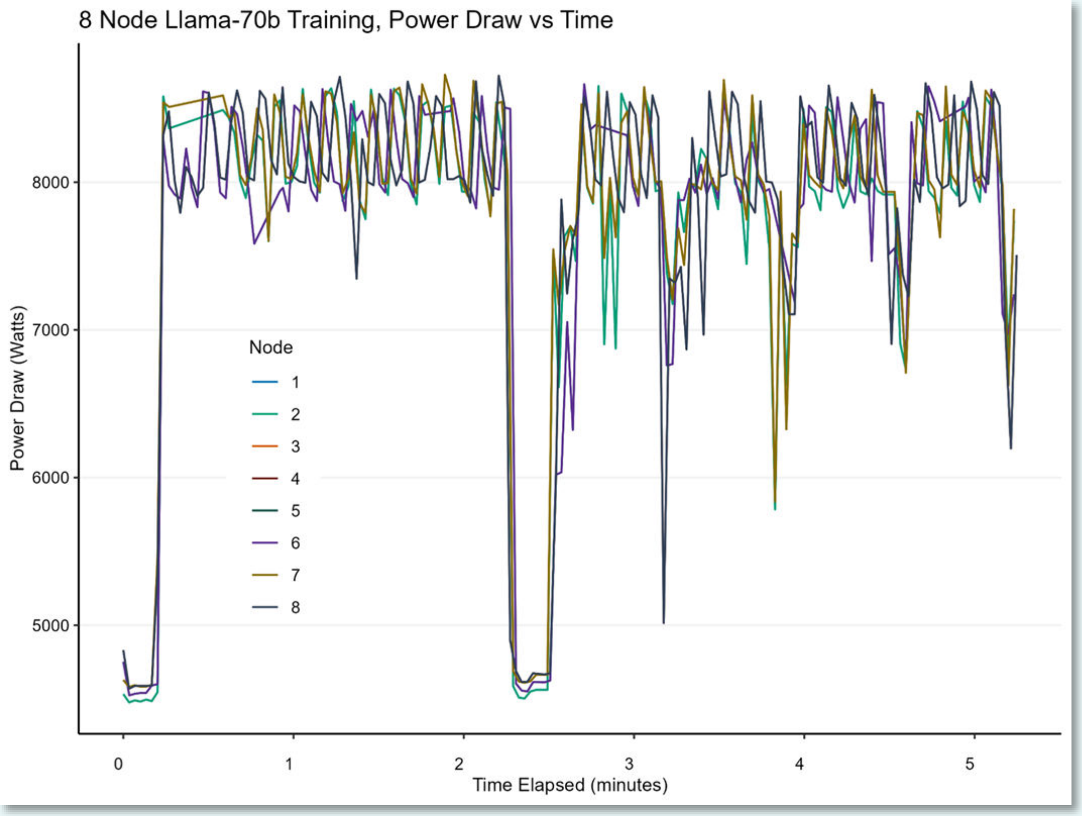


Figure 3.4. Time series of node-level power demand during Llama-70B training across 8 nodes at Sustainable Metal Cloud.

Each line represents a unique node. The characteristic square-wave pattern of transformer architecture training is evident, with regular drops in power demand across all nodes (e.g., at 2:20) corresponding to synchronization points and memory access periods. Power drops of individual nodes during compute intensive phases likely correspond to throttling as GPUs or nodes await intermediate values from other processors.

Economic incentives favor widespread model deployment to amortize high upfront training costs, suggesting inference energy use could dramatically exceed training. However, the actual energy impact will depend heavily on user adoption patterns, hardware evolution, and varying workload requirements. This makes inference energy use both potentially much larger than training and significantly harder to model with confidence. Given the modeled timeframe and these uncertainties, we parameterized inference-related power demand as half of that during training.

Table 3.1. Summary Statistics for Each Individual Workload in Our Dataset

All systems use NVIDIA H100-SXM5-80GB GPUs (8 per node). B = billion parameters, M = million. "Batch" refers to the global batch size. For each workload, average power (Pavg), maximum power (Pmax), and standard deviation (SD) are reported in kilowatts (kW). Power measurements are per-node, energy values represent total system consumption. BNL indicates Brookhaven National Laboratory data, while SMC indicates open-source benchmarking data from Sustainable Metal Cloud.

Workload	P_{avg} (kW)	P_{max} (kW)	SD (kW)	Arch.	Param.	Batch (global)	Nodes	Duration	IT Energy (kWh)
SMC - GPT3-175B (64)	7.67	8.45	0.73	DNN	175B	2048	64	0.95	479.45
SMC - Llama-70B (64)	5.92	8.76	1.5	DNN	70B	64	64	0.03	10.78
SMC - Llama-70B (8)	7.73	8.73	1.11	DNN	70B	8	8	0.09	5.43
SMC - Llama-70B (1)	6.78	7.32	1.05	DNN	70B	8	1	0.5	3.39
BNL-Llama-13B (1)	7.79	8.42	0.61	DNN	13B	8	1	8	62.36
SMC - ResNet 1 (8)	6.36	7.89	1.66	CNN	26M	3200	8	0.07	3.75
SMC - ResNet 2 (1)	6.76	6.88	0.29	CNN	26M	3200	1	0.22	1.51
BNL - Resnet 1 (1)	4.6	5.02	0.34	CNN	60M	512	1	26.8	123.41
BNL - Resnet 2 (1)	5.76	6.48	0.11	CNN	60M	4096	1	5.25	30.15

Idle Power

The method for estimating server idle power varies slightly between AI servers and conventional servers. For conventional servers, idle power is calculated as a fraction of *maximum* power that the servers draw during idle periods, while for AI servers, idle power is an assumed fraction of the server's *rated* power. As shown in Figure 3.5, we assume that conventional servers idle at 51% of their maximum power in 2014, dropping to 36% in 2023 and 27% in 2028. This assumption matches the values from the previous report through 2020, then continues the 2014-2020 linear trend through 2028. These values were compared with data found in the databases of both SPEC and SEED. The model assumption is above the general trend found in the SPEC power database, but roughly in line with SEED. As noted above, servers in the SPEC database are understood to be more efficient than the general market, so this is in line with those expectations. However, it is notable that the SPEC trend is relatively constant, if not increasing, with idle power fractions at 20-25% in recent years. This could imply a functional lower bound for idle power, relative to maximum power, that will remain constant into the future. Our current model assumption essentially states that the general server market will near this lower bound by 2028. However, this may underestimate idle power draw. For AI servers, we assume idle power equal to 20% of the server's rated power over the entire study period, based on measurement data provided by Brookhaven National Laboratory (Latif et al. 2024) of an 8-GPU NVIDIA H100 node, which showed an idle power draw of approximately 18% of the manufacturer-rated maximum.

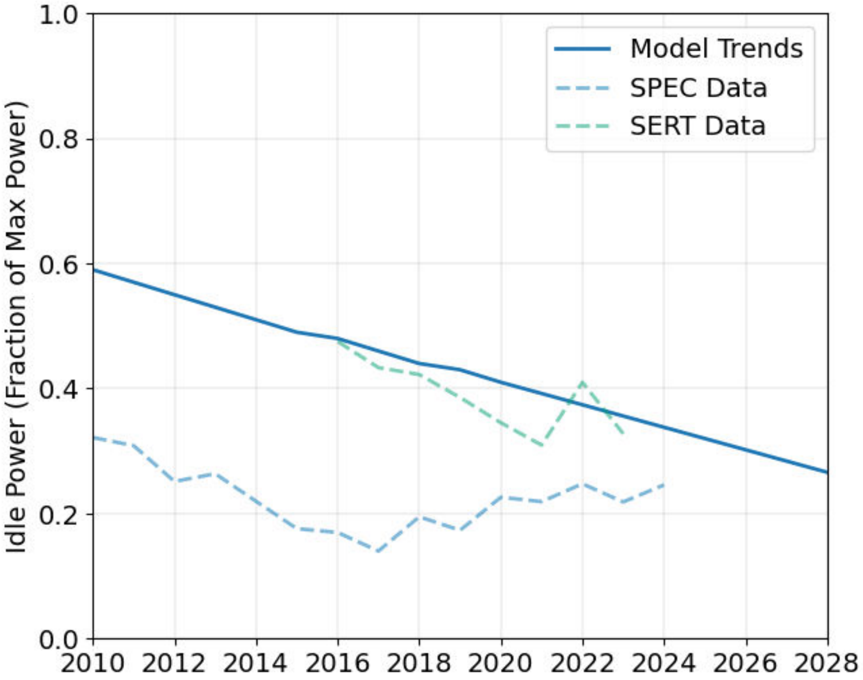


Figure 3.5. Idle power for conventional servers, as a percentage of maximum operating power.

Server Operational Time (Utilization)

Average server operational time for single, dual, and multiple processor servers varies by space type. Following trends in the 2016 report, utilization for conventional servers in all space types increases slightly through 2028. Servers in internal and small data centers average 11% utilization in 2014, rising linearly to 20% in 2027. Colocation data centers average 21% utilization in 2014, rising to 35% in 2027. Hyperscale data centers average 45% in 2014, rising to 50% in 2027. AI accelerated and non-accelerated servers doing training are assumed to have a constant 80% operational time throughout the whole period, while the same servers doing inferencing have a constant 40% operational time (see Figure 3.6). As with the operational power, we vary the operational time of training AI servers between 75% and 85%, and inference AI servers between 37.5% and 42.5%, from 2024 to 2028, to reflect possible differences in the future operation of these servers.

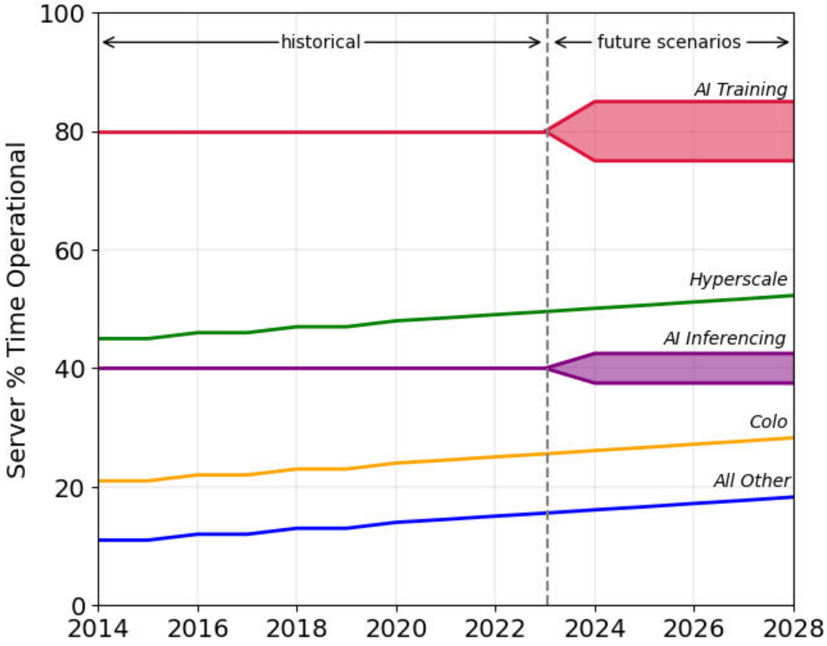


Figure 3.6. Operational time of servers given data center type.

Resulting Annual Average Power

Resulting average wattage depends on both server type and space type (for operational time). Therefore, a wide range of wattages is present in the model. As shown in Figure 3.7, servers with 8 GPUs exhibit the steepest increase, surpassing an annual average power of 5 kW by 2026, while 4-GPU and 2-GPU configurations rise more moderately, peaking above 2 kW and 1 kW, respectively. In contrast, non-accelerated AI servers and traditional processor-based servers experience relatively modest growth, with single and dual processor servers remaining the least energy-intensive, below 0.5 kW and 0.8 kW, respectively, by 2026.

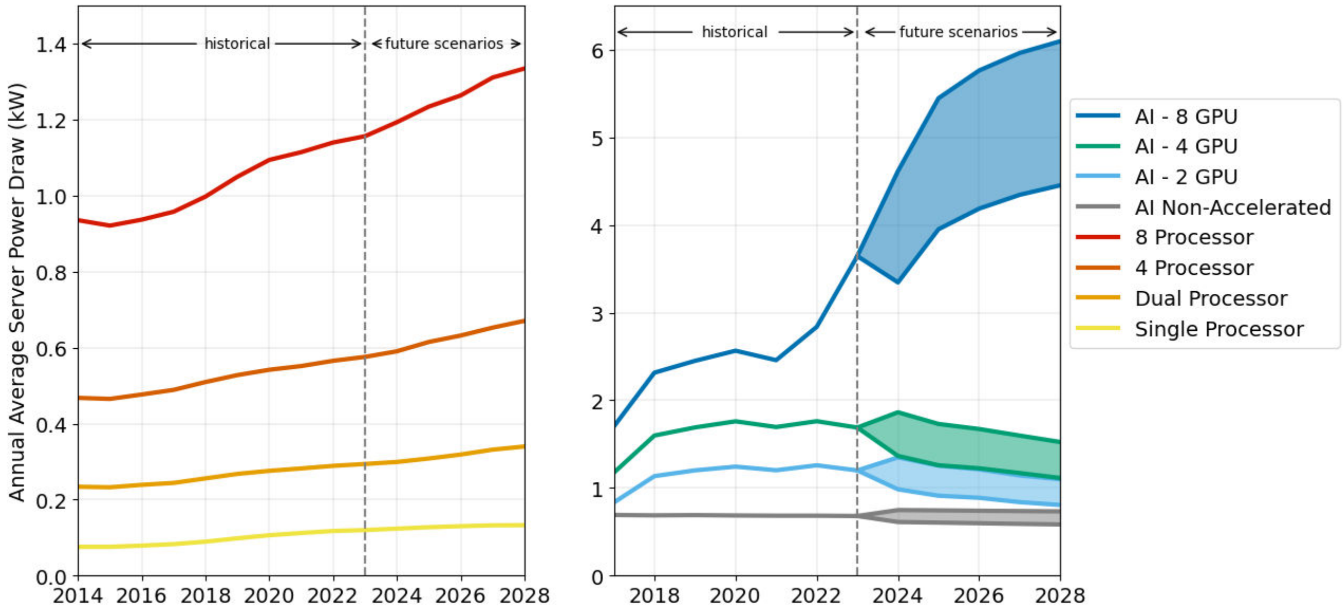


Figure 3.7. Aggregate average power draw of various server types across each analysis year.

Server Population

Total Server Shipments

Estimates for server installed base are based on IDC's "Worldwide Quarterly Server Tracker," which contains annual data for historical (2003–2023) and forecasted (2024–2028) server shipments (IDC 2024a). Shipment data are described by numerous characteristics, but the primary one used in this analysis is socket capability of shipments, representing the number of CPUs each server can be equipped with.

The number of servers shipped in the United States was relatively consistent from the mid-2000s through 2016, ranging from 2.8 to 3.4 million servers per year. After 2016, the number of shipped servers grew steadily, surpassing 6.5 million in 2022. Forecasts predict the number of shipped servers to grow to over 7.7 million in 2028, an 18% increase over six years. Much of the recent and forecasted growth is in 1-socket servers; in 2018, 1-socket servers only represented 7% of the shipments, with 89% of servers being in the "2+" socket category. In 2028, 1-socket servers are expected to be around 30% of shipments, with 2+ socket servers representing about 69%.

AI Accelerated Server Shipments

The AI accelerated server installed base is estimated using IDC's 2024 Q2 Data Center Semiconductor Consumption Report (IDC 2024b), which contains annual data for historical and forecasted CPU, GPU, and network switch shipments. GPU shipments from this 2024 report are categorized by GPU brand and model. IDC's 2023 Processors AI Tracker (IDC 2023a) was also used to estimate the accelerated server installed base. The 2023 report also provides estimated GPU shipments for the years 2017 to 2027, with each year broken out by AI activity

– either training, inferencing, or both. The ratio of inferencing and training GPUs from 2017 to 2027 is extrapolated linearly to develop a ratio for 2028, then applied to the “Data Center Semiconductor” shipment numbers. The “Training and Inferencing” category is split equally between training and inferencing. Additionally, estimates of the GPU shipment and the proportion of servers with 2, 4, or 8 GPUs that were shipped globally from 2019 to 2027 were obtained from Omdia Research (Omdia 2024).

The estimates from IDC (IDC 2023a; IDC 2024b) for the number of GPUs shipped to the United States for use in data centers is relatively flat between 2017 and 2019, with about 0.5 million units or less shipped annually. Shipments begin to rapidly grow in 2021 and continue growing through 2023. IDC’s newer (IDC 2024b) shipment forecast projects significantly more GPUs than their 2023 estimates. Omdia’s forecast for global shipment of GPUs indicate a U.S. GPU growth rate in between the two IDC estimates. Other GPU shipment forecasts reviewed (NSW 2023; TDC 2024) indicate a U.S. GPU shipment growth closer to the older IDC report (IDC 2023a), though these other forecasts were also published in a relatively older timeframe than the new IDC report (IDC 2024b). IDC’s GPU shipments are based on continued rapid growth in the sales and manufacturing of GPUs, with the newer, higher, forecasts influenced by the increased AI activity in the second half of 2024. As such, this report uses the higher and lower IDC forecasts to bound a range of possible GPU shipments that make up the accelerated server installed base of the energy use model for 2024 to 2028, with the lower bound assuming that the increased AI activity observed in the second half of 2024 does not continue and GPU shipments revert to the prior expected growth rate.

The number of accelerated servers shipped in each year are determined by using (1) the total GPU shipments range from IDC (2023a; 2024b) and (2) the distribution of 2, 4, and 8 GPU servers shipped in each year estimated by Omdia Research (2024). The distribution is applied to all GPUs except for the most advanced, such as the NVIDIA B100 or H100 classes, which are assumed to all be shipped in 8-GPU configurations. These servers are assumed to be a subset of the total 2+ socket server shipments described previously.

The allocation of shipped servers for training or inferencing is based on the number of GPUs shipped in each category provided by IDC (2023a). Servers with 8-GPUs are preferentially allocated to training, then inferencing, followed by 4 and 2 GPU servers. Once the allocation of training servers is filled, the remaining are allocated to inferencing.

AI Non-accelerated Server Shipments

We also consider servers that are configured for AI workloads but do not have GPUs or other accelerators (only two CPUs). The Omdia AI server shipment estimates (2024) also include estimates for the number of these non-accelerated servers shipped each year. From that data, we calculate the ratio of non-accelerated to accelerated servers in each year and use that to estimate the shipments in our dataset. All non-accelerated AI servers are assumed to be in the inferencing category.

AI servers are assumed to be a subset of the total servers tracked by IDC. Therefore, future scenarios contain the same total installed base of servers, but with a varying number of those servers containing additional GPUs for AI applications.

Installed Base Calculation and Results

The installed base is calculated by assuming an average lifetime for servers each year, then summing the server shipments for the corresponding number of years prior, as shown in Equation 1. For this study, the lifetime assumption for 2000–2019 is 4.4 years, increasing to 5 years in 2020, and to 5.7 years by 2023, following trends of hyperscale data centers and trends provided by Omdia (2024) and IDC (2023b), as shown in Figure 3.8. For AI accelerated and nonaccelerated servers, the same lifetime assumptions are used. The entire calendar year's worth of shipments is considered when calculating the installed base, and therefore the resulting values effectively represent the installed base that is present at the end of each calendar year. We factor this into our calculation of total electricity consumption, as discussed in the corresponding results section.

The total server installed base in 2014 was 14 million and was entirely made up of conventional servers. By the end of 2020, the total installed base reached 21 million, with AI servers accounting for 1.6 million servers, while conventional servers grew to nearly 20 million. By 2028, AI servers make up between 8 and 12 million of the total installed base, which grows to almost 37 million, depending on the number of GPU shipments from 2024 to 2028, as shown in Figure 3.9.

$$IB_y = \sum_{y-l}^{y} S_i$$

Where:

$$IB_y = Installed\ Base\ in\ year\ y$$
$$S_i = Server\ shipments\ in\ year\ i$$

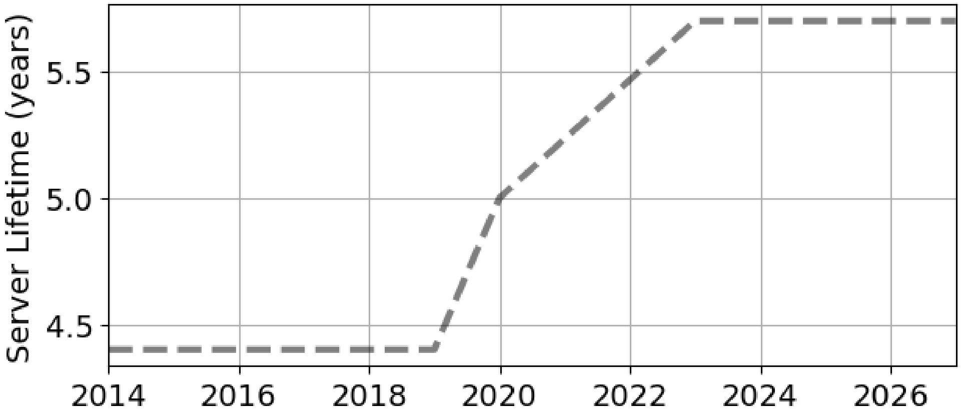


Figure 3.8. Assumed average age of servers installed in each year of our analysis.

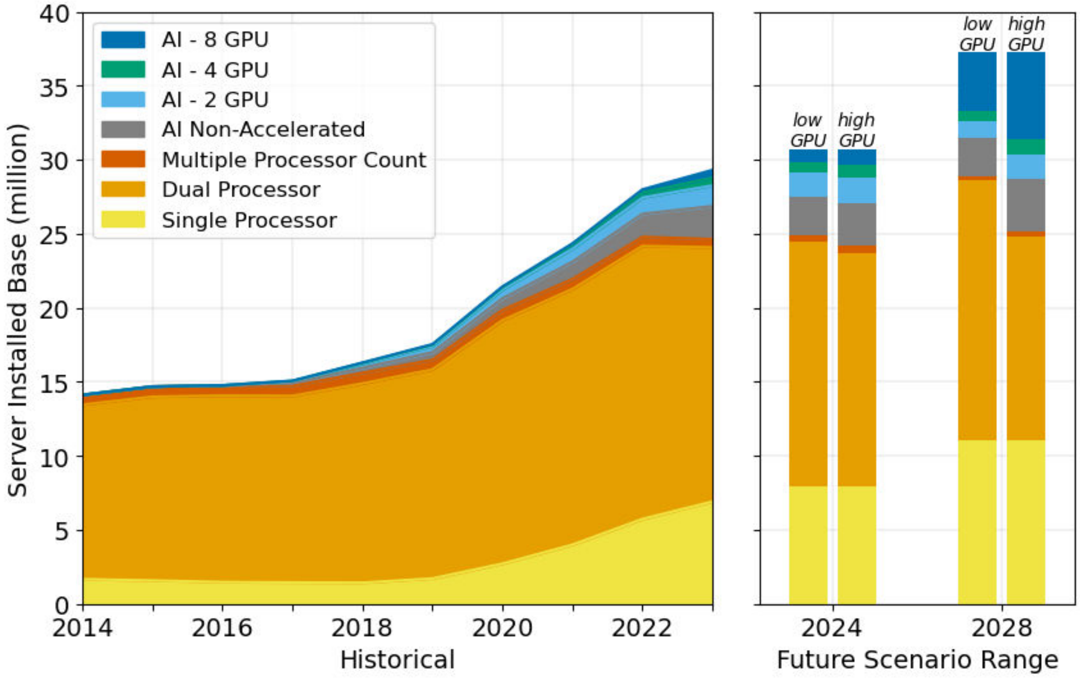


Figure 3.9. Total server installed base for 2014–2028 with higher bound shipments (left). Adjusted installed base with lower bound GPU shipments (right).

Storage and Network Types and Power Draw

Data center storage equipment is captured using shipment data from IDC's storage equipment tracker. Shipments are tracked in terms of total terabytes (TB) shipped for various storage systems and media types. The two media types that IDC categorizes shipment data by are (1) hard disk drive (HDD) and (2) flash storage, which includes solid state drives (SSDs). These TB shipments are converted to number of drives in order to be paired with power consumption estimates that are typically reported on a per-drive basis. Figure 3.10 shows the assumption used for this conversion; data for 2010–2020 is based on feedback collected from industry during the report review period; and for 2020 and beyond, we assume a 20% density increase annually.

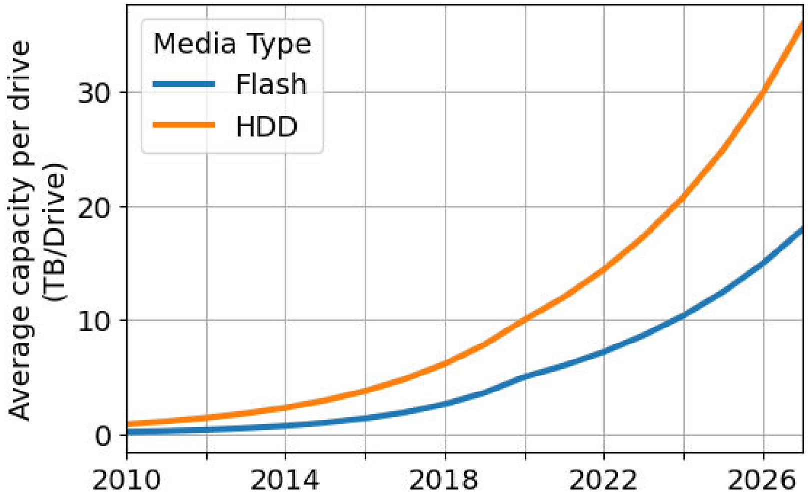


Figure 3.10. Assumed average drive capacity (TB) of new storage devices shipped in each year.

We model the energy use of these systems separately, as the way they consume electricity is functionally different. Hard disk drives consume energy anytime the disk is spinning, while flash drives primarily consume electricity when they are actively inputting or outputting data. The electricity demand of all storage is dependent on the read/write frequency and patterns, which is not feasible to define across the industry as a whole. Therefore, we rely on estimated averages and general rules of thumb for estimating the power draw of this equipment.

The wattage for HDD is estimated by using the historical trend reported in the 2016 report, which was based on datapoints for 2006 (14 W/disk) and 2015 (8.6 W/disk). For 2020 and beyond, we use a recent market research report that estimates 7 W/drive in 2020 and 6.4 W/drive in 2025 (Monroe and Johns 2024). We model the years between these datapoints as exponential curves, then extend the estimates from 2025 to 2028 using the 2020–2025 compound annual growth rate (–1.6%).

The flash storage wattage estimates are less straightforward to produce. There has been a recent transition from SSDs using traditional SATA server attachments to higher-performance NVMe configurations. While NVMe storage drives are considered to be more power efficient, they are operating at very high levels of performance, and therefore they ultimately tend to consume more electricity per drive. Similar to HDDs, our starting estimates are built on the analysis done for the prior report, which was based on a 2015 ASHRAE report and industry feedback. For 2020–2025, we leverage the findings from the same market research report (Monroe and Johns 2024), which notes the transition to NVMe-configured flash drives that are estimated to use 11 W per drive. The resulting wattage estimates for storage devices shipped in each year are shown in Figure 3.11.

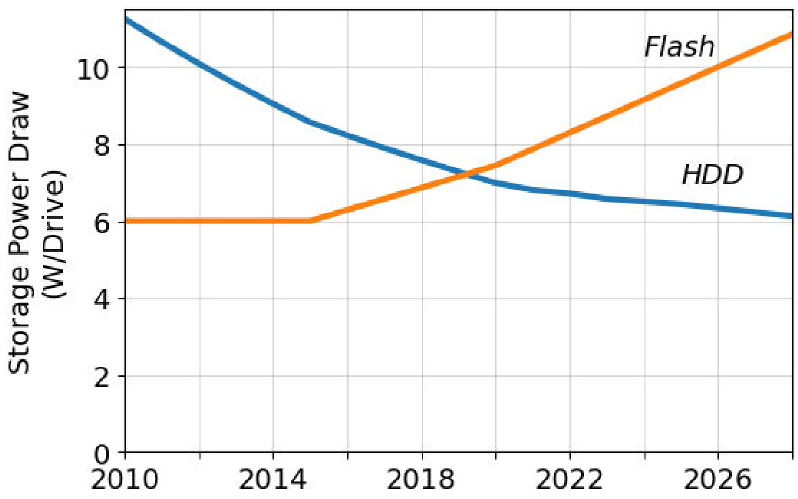


Figure 3.11. Average wattage of storage drives shipped in each year.

Network equipment is captured in two categories: ethernet switches and InfiniBand switches for AI clusters. Shipment data for ethernet switches is obtained from IDC's Network Equipment Tracker in units of number of ports. These shipments are specified for ports of various speeds. For the InfiniBand switches, we obtained data from IDC regarding AI equipment, which included the number of switch units shipped in each year. Note that other internal network hardware for AI, such as NV links, are considered part the AI server system and represented within the accelerated server category.

Average power draw per ethernet port follows the estimates used in the previous report through 2020, then is held constant through 2028, as shown in Figure 3.12. For port speeds not present in the previous report (100 Gb, 200/400 Gb, and 25/50 Gb speeds), wattage per power is assumed to scale linearly with speed. InfiniBand switch units are assumed to operate at an average of 700 W, based on the rated wattage values for 64-port units.

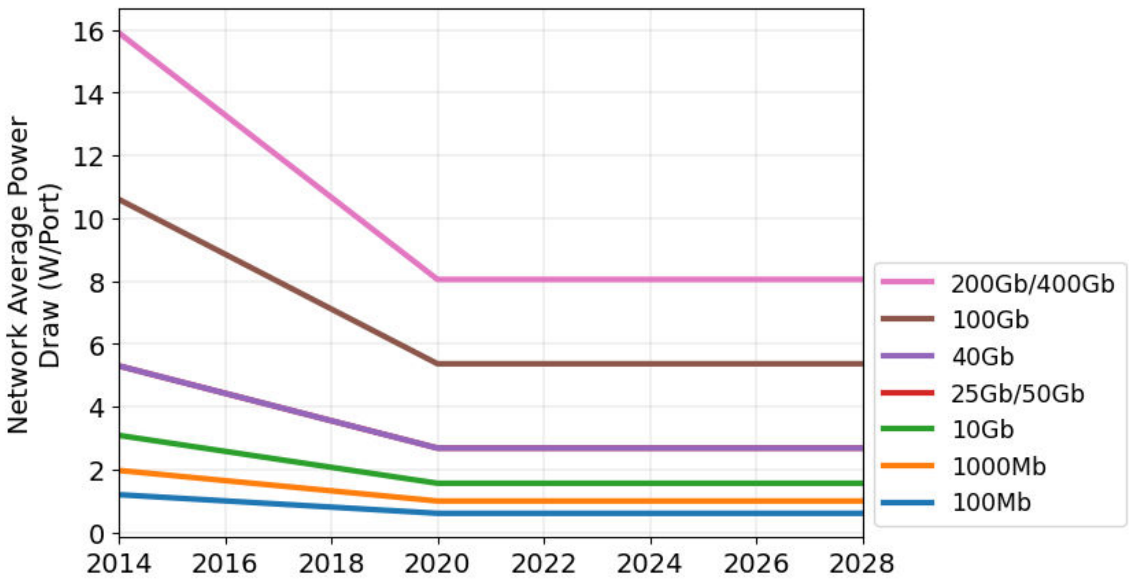


Figure 3.12. Average wattage per port for ethernet ports.

Storage and Network Population

From 2014 to 2020, HDD accounts for most storage equipment, with Flash storage contributing a relatively small but steadily increasing proportion. By 2023, flash storage grows to account for 25% of the total, then continues to grow to 41% by 2028. The total installed base across both technologies reaches nearly 340 million drives by the end of 2028, as shown in Figure 3.13.

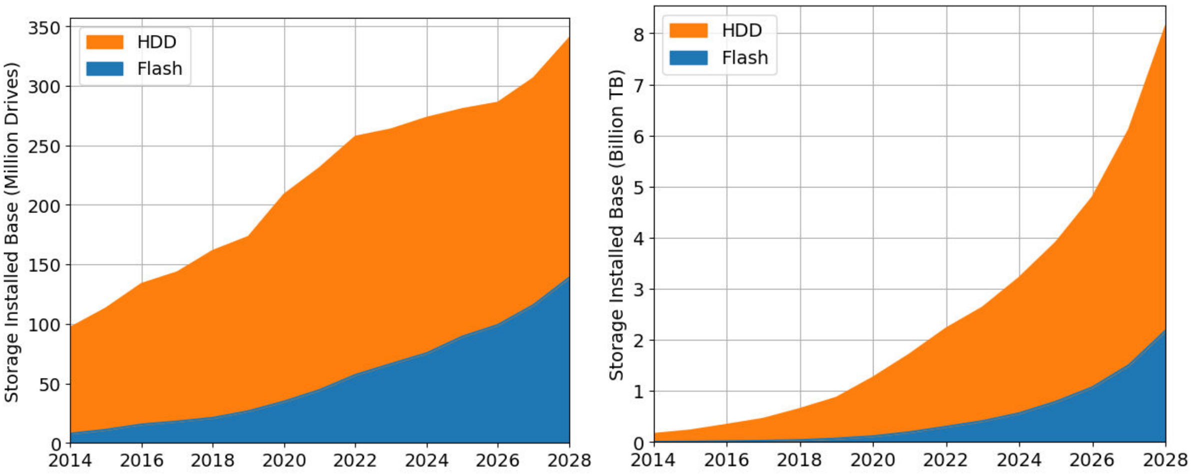


Figure 3.13. Installed base of storage devices in drive units (left) and TB capacity (right).

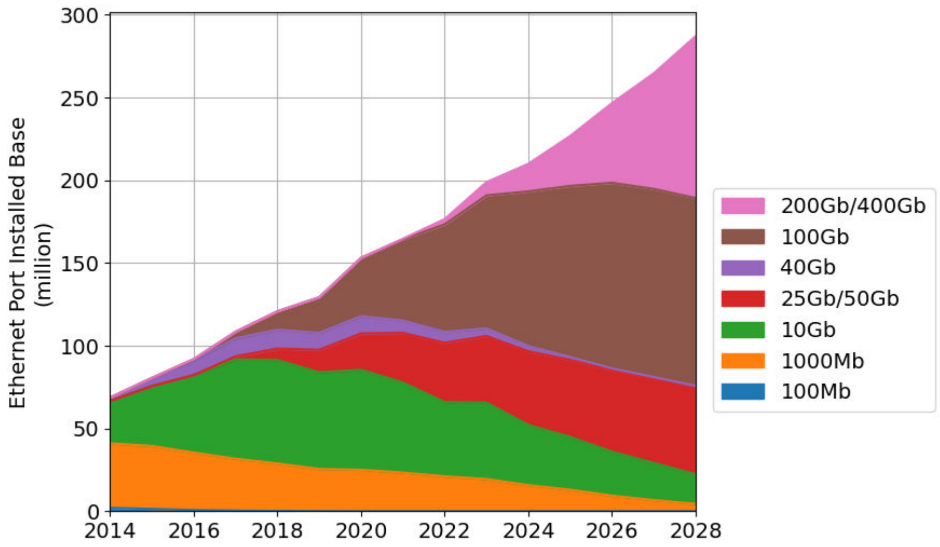


Figure 3.14. Installed base for data center ethernet switch ports by port speed.

Figure 3.14 shows the installed base of ethernet ports broken out by speed in the end of each year. Higher speed ports (> 50GB) begin to dominate the installed base around 2023, accounting for 64% of the total. Higher speeds continue to grow through 2028, with 200/400Gb reaching 34% and 100GB reaching 40% of the total. InfiniBand switch units grow rapidly, beginning in 2020 with under 2000 units and reaching 1.8 million units in 2028.

4. Data Center Infrastructure

Data Center Classifications

Although this study utilizes an IT equipment-based bottom-up model, it is still important to understand the types of data center facilities that IT equipment is being installed in, as different space types can have different operational characteristics as well as different cooling systems, which impact total electricity demand. In the 2016 report, our space types were defined by size and ownership/business (internal vs. service provider) categories. In this study, the space types are defined to align with new data available from IDC regarding the number, square footage, rack count, and other metrics of built data centers over time (IDC 2023b). These space types are described in Table 4.1. While these space types are not explicitly defined by size, for the most part they can be considered as either small, with an average square footage less than 150 (Telco Edge, Commercial Edge, SMB, and Enterprise Branch), midsize, with an average square footage of 2700 (Internal) and 6900 (Comms SPs), or large-scale, with average square footages of 11,000 for the colocation space types and 30,000 for hyperscale facilities (average square footage per module, not per entire facility/campus).

Table 4.1. Data Center Space Types Considered in This Study

Space Type	Description
Telco Edge	Deployment of small closets/rooms to micro data centers and network infrastructure by communications companies as points of presence throughout their network
Commercial Edge	Network closets, server rooms, and micro-data centers deployed to support modern digital, infrastructure, and software delivery services to edge locations for commercial (focused on customer and business operations) and industrial (focused on supply chain and channel operations)
Small and Medium Businesses (SMB)	SMB deployments in their own internal facilities
Enterprise Branch	Classic remote and branch office (ROBO) deployments for large enterprises in their own internal facilities (network closets, server rooms)
Internal	Data centers run by enterprises, internally, for their own use
Communications Service Providers (Comms SPs)	Data centers run by telecommunications/cable companies to support internal services required to enable provision of communications technology services to their customers
Colocation – Sm/Med Scale	Data centers built by local colocation companies typically providing retail leasing at smaller scale
Colocation – Large Scale	Data centers built by major colocation companies providing wholesale and retail colocation leasing, typically deploying large and mega datacenters
Hyperscale	Data centers built by companies that deploy internet services and platforms at massive scale

Data centers are not uniformly spread across the United States. These data centers house computing equipment that operates continuously to support on-demand network requests, ranging from small setups in closets to massive warehouses with thousands of servers. Internal data centers are typically integrated into larger buildings and managed by businesses with their own IT systems. In contrast, multitenant data centers, also known as colocation centers, provide space for companies to host their hardware offsite, offering essential services such as power, cooling, security, and networking. Concurrently, hyperscale data centers, which are large-scale facilities operated by major technology companies, are experiencing rapid growth, with Amazon, Microsoft, and Google collectively owning more than half of them (Synergy 2024).

While internal data centers are scattered across the United States, distinct clusters emerge for colocation and hyperscale data center locations. These data centers strategically position themselves close to their clients and cloud services users to ensure high availability and low response times. Factors such as proximity to population centers, electricity cost, network infrastructures, and local utility prices influence their location choices (Goiri et al. 2011). Information availability on data center location and size varies depending on the type and owner. We identified likely locations of in-house small and midsize data centers following the methodology provided in Ganeshalingam et al. (2017) reports, and we updated the method with latest available data. Detailed information on colocation and hyperscale data centers is derived from commercial compilations that get support and input from data center service providers. A comprehensive geospatial dataset of over 3,000 data centers was developed through direct engagement with data center operators, analysis of open-source data center inventories, and extensive manual web searches.

The 451 Research Datacenter KnowledgeBase, managed by S&P Global Market Intelligence, (S&P Global 2024) offers comprehensive insights into data center locations, services, and utilization, covering both colocation and wholesale facilities. This dataset was integrated into our inventory to enhance the depth and accuracy of our spatial distribution analysis, enabling a comprehensive assessment of energy demand distribution by data center types across the U.S. According to the latest estimates, Virginia hosts the highest electricity demand associated with data centers in the United States, serving as the primary hub for both colocation and hyperscale data centers, followed by California and Texas.

Distribution of Servers Across Data Center Types

Total number of servers in each data center type is calculated based on the distribution shown in Figure 4.1 which is from an IDC "build" dataset that includes number of racks in each space type, plus an estimate of how many servers are typically on each rack in various data center types. In the early 2010s, internal data centers dominated the industry, with hyperscale and colocation providers housing less than 15% of the servers. In the 2016 report, we estimated that in 2014, 24% of servers were present in hyperscale data centers, and that by 2020, this number would rise to 40%. The current data estimates a slower shift to hyperscale facilities (8% in 2014, 25% in 2020), but a more aggressive overall shift when considering colocation and hyperscale facilities together (20% in 2014, 57% in 2020). With the substantial build out of large-scale facilities for AI and other purposes, we now estimate that 74% of servers were present in these space types in 2023, and this proportion will rise to 85% by 2028.

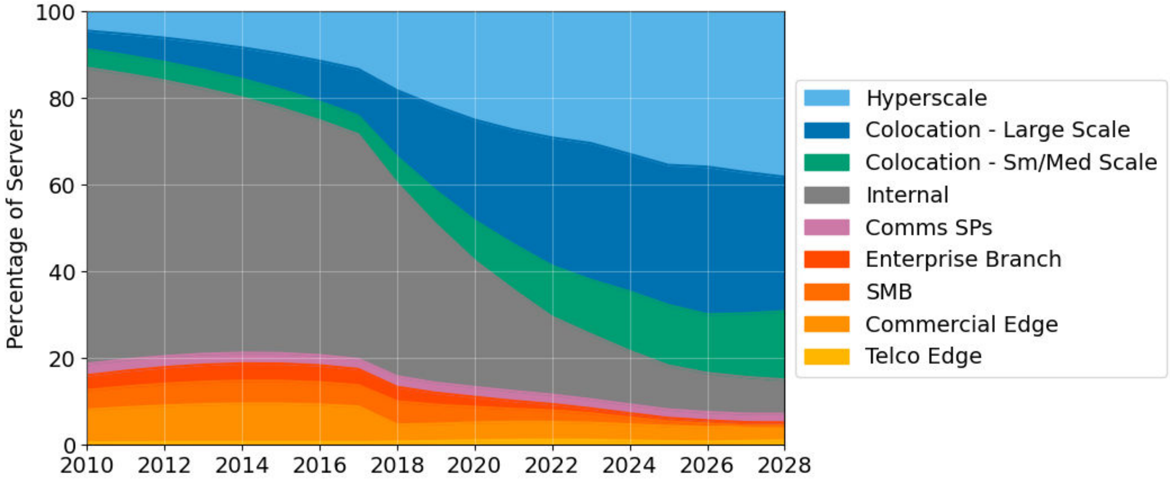


Figure 4.1. Distribution of servers by data center type.

We use the distribution in Figure 4.1 to assign IT equipment in our model into each space type. First, we calculate the total number of servers in each space type by multiplying the total installed base by the distribution. All AI accelerated servers are allocated evenly to the Hyperscale and Colocation-Large Scale categories. These servers are then broken out separately into an “AI” category to capture the unique cooling system configurations used for AI servers (a mix of liquid and air cooling). The total load calculated for AI servers is then distributed evenly between Hyperscale and Colocation-Large Scale. AI nonaccelerated servers go 80% into Hyperscale and 20% into Colocation-Large Scale. Then, multiple processor count servers go into Hyperscale, Colocation, and Internal space types proportional to how many servers are left in each category. Finally, single and dual processor servers are assigned to space types proportionate to how many spaces remain.

Distribution of Cooling Systems Across Data Center Space Types

The distribution of cooling systems across the United States is determined by combining market data with insights from industry experts. Specifically, we acquired market data from the Dell’Oro Group, which details annual manufacturing revenue segmented by primary heat rejection methods (such as chilled water, direct evaporative, indirect evaporative, and direct expansion) in North America. These market insights were then integrated with system costs (in dollars per megawatt) to estimate the distribution of major cooling system categories for each year.

The market data offers an overview of the distribution of major cooling system categories, while insights from experts allow for a more detailed breakdown into specific systems, as outlined in this report. To achieve this, we collected data and insights from industry experts and formulated an optimization problem to estimate the nationwide distribution of cooling systems defined in this study. This optimization problem aims to align the estimated distribution of cooling systems with the major cooling system categories identified in market data, while adhering to constraints established by industry expertise. The result is a comprehensive breakdown of cooling systems, categorized by data center space type across the United States for each year (see Figure 4.2 for cooling systems distribution in 2023).

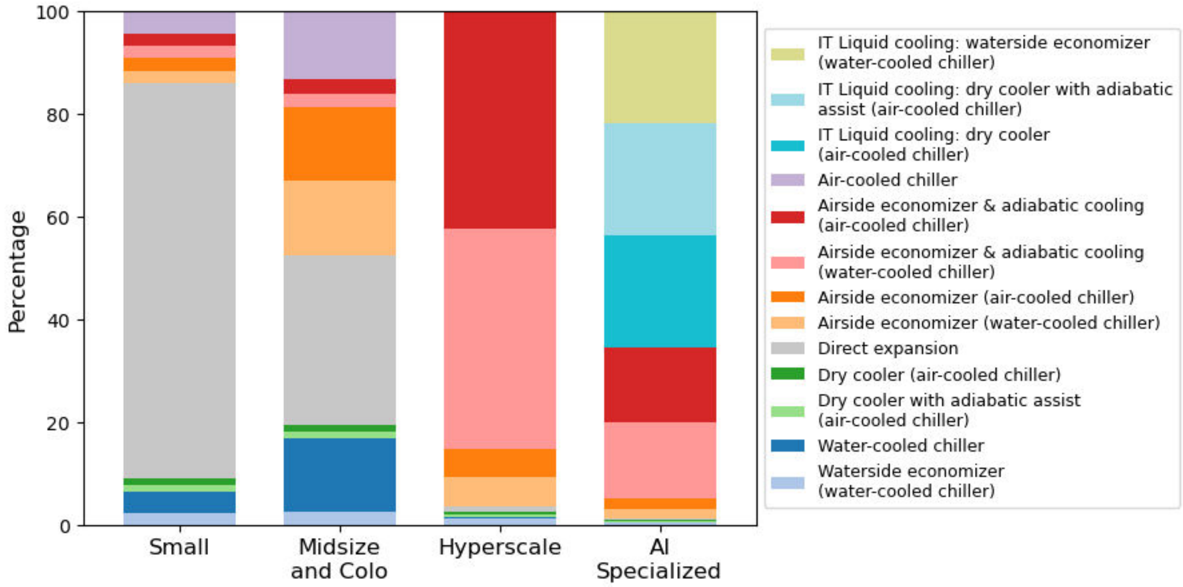


Figure 4.2. Data center cooling systems distribution by space category in 2023.

Data Center Infrastructure Electricity and Water Modeling

Method Overview

Two key metrics are used when describing the resource intensity of data center facility infrastructure. Power Usage Effectiveness (PUE) is defined as the total electricity demand of the data center divided by the electricity demand of the IT equipment. Water Usage Effectiveness is similarly defined as the total water consumption of the data center divided by the electricity demand of the IT equipment. In this section, we discuss the on-site (or “direct”) water consumption, sometimes referred to as “WUE (site)” primarily associated with cooling infrastructure, as opposed to the water consumption from the electricity generation (“WUE (source)”). PUE is technically dimensionless (kWh/kWh), while WUE is reported in terms of liters per kWh.

The PUE and WUE metrics of data centers are influenced by various factors, including cooling systems, operational practices, and climatic conditions. For this report, these metrics are simulated using established thermodynamics-based models (Lei et al. 2023; Lei and Masanet 2022). Simulation results are then paired with assumptions for the locations of data centers across the United States, and the cooling systems deployed across different space types, in order to develop estimates for overall PUE and WUE averages; this modeling framework is summarized in Figure 4.3. The inputs and outputs of the thermodynamic simulation model are described in the blue box on the left of Figure 4.3. First, operational characteristics are determined for a total of 18 simulation cases, which represent a combination of data center space type and cooling system type (described further below). This model considers the cooling system at the facility as well as other infrastructure equipment including UPS, power transformation and distribution, fans, and pumps. For each case, 50 simulation scenarios are developed drawing from underlying distributions for each parameter in the model. Each scenario is then run through the simulation model using data from each weather station in the United States. This results in a dataset of PUE and WUE (site) results for over 1 million simulations.

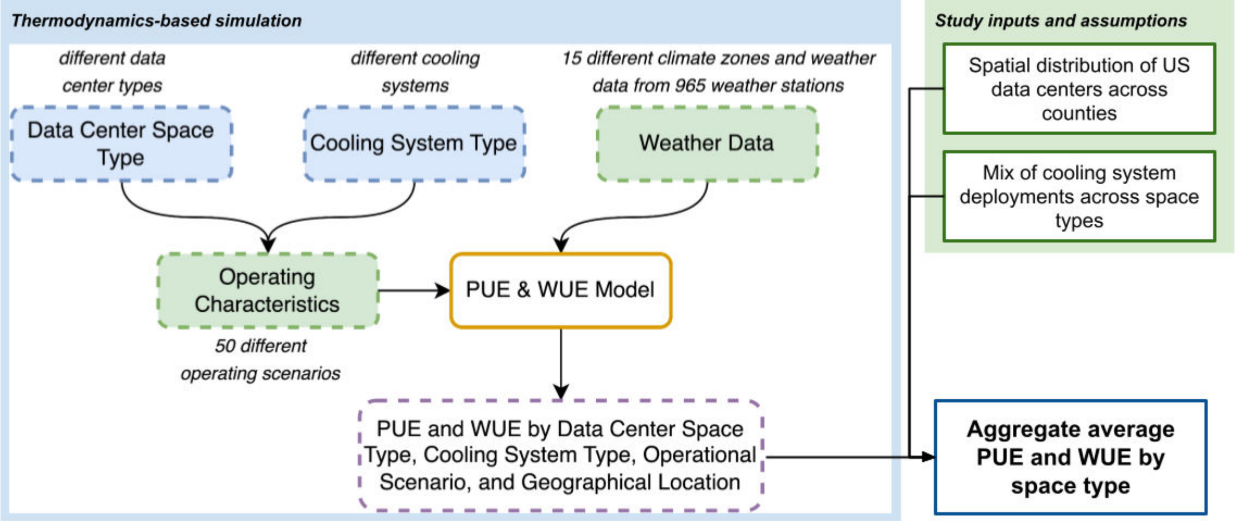

Thermodynamics-based simulation
different data center types
different cooling systems
15 different climate zones and weather data from 965 weather stations
Data Center Space Type
Cooling System Type
Weather Data
Operating Characteristics
50 different operating scenarios
PUE & WUE Model
PUE and WUE by Data Center Space Type, Cooling System Type, Operational Scenario, and Geographical Location
Study inputs and assumptions
Spatial distribution of US data centers across counties
Mix of cooling system deployments across space types
Aggregate average PUE and WUE by space type

Figure 4.3. Flow chart of methodology for modeling PUE and site WUE across data center types.

Simulation Modeling and Cooling Systems Modeled

Nine cooling systems commonly implemented in the United States were considered, comprising various combinations of refrigeration units (including direct expansion, air-cooled chiller, and water-cooled chiller), economizer use (including airside and waterside economizers), dry/adiabatic cooling, and liquid cooling options, as summarized in Table 4.2 below.

Table 4.2. Major Cooling Systems Considered in this Study

Cooling System Type	Notes
Direct expansion system	Direct expansion systems, typically called computer room air conditioners (CRACs), are commonly employed in small to midsize data centers because of their simplicity and cost-effectiveness. In these systems, refrigerant circulates directly through indoor coils to cool the data center air.
Air-cooled chiller	Air-cooled chillers are widely used across data centers of varying sizes. These systems employ air-cooled condensers to remove heat from the system and are commonly paired with computer room air handler units to maintain ideal IT operating conditions. They are the preferred option where the first cost is a major factor, in situations where minimal on-site water consumption is crucial, or where space limitations prevent the installation of cooling towers.
Water-cooled chiller	Water-cooled chillers are widely used in data centers, owing to their high efficiency and capacity to manage substantial cooling requirements. These systems use water-cooled condensers to extract heat from the system, subsequently

	releasing it into the environment via cooling towers. They are typically integrated with computer room air handler units to ensure optimal IT operating conditions. However, the water evaporation process in cooling towers has raised concerns regarding data center water consumption and availability at the local level.
Airside economizer (air- or water-cooled chiller)	Airside economizer systems leverage outdoor air to cool the internal space of a data center during favorable weather conditions. This approach effectively reduces the data center's cooling energy consumption and operational costs. Additionally, an air-cooled or water-cooled chiller system can serve as the supplementary cooling when the airside economizer alone cannot meet the cooling demands.
Waterside economizer (water-cooled chiller)	Waterside economizer systems minimize energy consumption and operational costs by using cool water drawn from natural resources (such as lakes, rivers, or the sea) or produced by cooling towers when the outdoor air is cool and dry, thereby reducing or eliminating reliance on mechanical refrigeration. In instances where the waterside economizer is insufficient, a water-cooled chiller provides supplementary cooling. Like with water-cooled chillers, local water consumption remains an issue.
Dry cooler with or without adiabatic assist (air- or water-cooled chiller)	Dry coolers reject heat to the ambient air and enable “free” cooling during favorable weather conditions, with or without adiabatic assistance. When the outdoor dry bulb temperature is low, the dry coolers efficiently reject heat using ambient air alone. With adiabatic assistance, spraying or circulating water through pads upstream of the heat exchanger coils enhances cooling, especially when the dry bulb temperature is high and the wet bulb temperature is low (typically in dry seasons). This system conserves water compared to evaporative cooling towers. An air- or water-cooled chiller system provides additional cooling when needed.
Airside economizer & adiabatic cooling (air- or water-cooled chiller)	Like the airside economizer (air- or water-cooled chiller), this configuration seeks to minimize or eliminate mechanical refrigeration by leveraging the adiabatic process, wherein water evaporates directly into the supply air to provide cooling. This approach is highly energy-efficient and cost-effective, and in certain climate zones, it can entirely replace the need for supplementary cooling throughout the year. Additionally, potential use of an air- or water-cooled chiller system can provide supplementary cooling when necessary (e.g., during heat waves or periods of increased cooling demands when the airside economizer and adiabatic cooling together cannot meet the cooling requirements).
IT liquid cooling: dry cooler with or without adiabatic assist (air-cooled chiller)	Dry coolers with or without adiabatic assist, as described above, can be used to minimize water consumption and mechanical refrigeration while maintaining high energy efficiency (“free,” compressor-less cooling). The higher

	operating temperature of liquid cooling systems facilitates better water and energy efficiency/performance (less chiller run hours and less adiabatic assist). Dry coolers can effectively meet data center cooling needs for much of the year, particularly in regions with favorable climates. This provides an energy-efficient and water-friendly cooling solution for data center infrastructure. In instances where dry coolers cannot fully meet the cooling requirements, an air- or water-cooled chiller system is deployed for supplementary cooling. For purposes of our simulation, an air-cooled chiller was assumed.
IT liquid cooling: waterside economizer (water-cooled chiller)	Liquid IT cooling is an emerging technology for cooling dense IT equipment (e.g., AI). While IT liquid cooling can use conventional chiller plants, an advantage of liquid cooling is the ability to elevate the cooling temperature and take greater advantage of “free” cooling. Therefore, the base case is a waterside economizer. A waterside economizer system decreases or eliminates dependence on mechanical refrigeration and is well suited for liquid cooling (e.g., rear door heat exchanger, direct-to-chip, and immersion). Liquid-cooled IT systems can often be operated at higher water/refrigerant temperatures than air-cooled IT systems, and waterside economizers can effectively fulfill data center cooling requirements for much of the year, especially in regions with favorable climates. When necessary, an air- or water-cooled chiller system is deployed for supplementary cooling. For purposes of our simulation, a water-cooled chiller was assumed. While highly energy-efficient, this system consumes substantial amounts of water, like all evaporative cooling systems.

Note: (1) parentheses denote a supplementary cooling system that may not actually be deployed in favorable climate zones. For example, in the scenario of an airside economizer (air- or water-cooled chiller), the air- or water-cooled chiller system would only be used when the airside economizer alone cannot meet the data center's cooling needs; (2) systems whose names do not begin with 'liquid cooling' represent air-cooled IT systems.

Various data center operating characteristics that influence PUE and WUE values were considered in the simulations conducted for this report. These characteristics encompass equipment efficiencies (such as UPS, power transformation and distribution, fans, and pumps) as well as indoor environment set points within the data center (such as dry bulb temperature, relative humidity, and facility water temperature), which vary across different data center categories defined in this report. Generally, smaller categories of data centers, such as commercial edge, enterprise branch, small- to medium-size business (SMB), and telco edge, operate with lower UPS and airflow efficiencies and have narrower allowable ranges for temperature and relative humidity. In contrast, hyperscale and AI data centers tend to operate with higher UPS and airflow efficiencies, as well as wider allowable temperature and relative humidity ranges, leading to a significant portion of free cooling throughout the year. Data centers falling between these two categories, such as comms service providers (SPs), internal, and colocation, typically have UPS and airflow efficiencies, as well as allowable temperature and humidity ranges, that lie somewhere in between. In this report, uncertainty ranges of the best and worst operating characteristics, drawn from several relevant studies (Lei et al. 2023; Lei and Masanet 2022, 2020), were employed for PUE and WUE simulations of each data center type (see Table 4.3 for assumptions regarding the major determinants of PUE and WUE). These simulation results were subsequently reviewed by industry experts and meticulously adjusted based on industry-reported PUE and WUE values until they aligned with expected ranges.

Table 4.3. Model Assumptions Regarding the Major Determinates of PUE and WUE by Data Center Space Type/AI

Space Type	UPS Efficiencies	ASHRAE Thermal Envelope (ASHRAE 2021)	ASHRAE Liquid Cooling Class (ASHRAE 2014)
• **Commercial Edge** • **Enterprise Branch** • **SMB** • **Telco Edge**	77–85%	Recommended	N/A
• **Comms SPs** • **Internal** • **Colocation - Sm/Med Scale** • **Colocation - Large Scale**	80–94%	A1	N/A
• **Hyperscale** • **AI (IT Air Cooling)**	90–99%	A2	N/A
• **AI (IT Liquid Cooling)**	90–99%	N/A	W45

To accurately capture variations in PUE and WUE values across different climate zones and account for micro-level geographical differences in climate, typical meteorological year (TMY) climate data were collected from different weather stations situated across the United States, comprising hourly measurements of dry bulb temperature, relative humidity, and atmospheric pressures throughout the year. In total, climate data from 965 weather stations were collected, representing evenly distributed geolocations across the U.S.

Using these location specific TMY climate data, we conducted a large-scale simulation to evaluate annual average PUE and WUE across a broad spectrum of data center space types and cooling systems (refer to Figure 4.4 for a simplified flowchart illustrating the simulation process). Specifically, for each data center space type and cooling system type considered, we randomly sampled 50 operational scenarios from the uncertainty ranges of best and worst operating characteristics, where each scenario represents real-world data center operational practices. These simulations provide valuable insights into annual average PUE and WUE values for diverse data center settings across the U.S., accounting for various space types, cooling system types, operational scenarios, and climate conditions.

PUE and WUE Results

Figure 4.4 and Figure 4.5 depict the uncertainty ranges of annual average PUE and WUE values for the data center space types and cooling systems examined in this study. The lower limits represent data centers with best practice efficiency values in favorable climates, while the upper limits correspond to poor efficiency values in hot, humid climates.

Figure 4.4 reveals a striking variation in PUE values across different types of data center spaces, reflecting the diverse efficiency practices implemented within each category. PUE values vary according to cooling system types, albeit not as significantly as the variance among different types of spaces. Utilization of economizers, adiabatic cooling, or dry coolers can contribute to lower PUE values, highlighting the impact of different cooling methodologies on overall energy efficiency.

Note that simulated PUEs assume systems are commissioned and operate as designed. This is rarely the case, and generally system performance does not meet design expectations. For example, it is not uncommon to find multiple computer room air conditioning units in a data center “fighting” with each other—some humidifying while others are dehumidifying. Therefore, the actual PUEs (and WUEs) will likely not be as good as estimated, however, since the uncertainty ranges are primarily drawn from studies regarding data centers in the earlier years of our historical range of 2014-2023 (Lei et al. 2023; Lei and Masanet 2022, 2020), and generally data center PUE has improved during that period (Uptime 2024), the median of the uncertainty values was assumed to be representative for data centers through 2023. For future years, 2024–2028, the PUE and WUE values for the data center space types were varied by a range of +/- 10% to reflect possible differences in the future type and operation and cooling systems.

As illustrated in Figure 4.4, WUE values display variability across different data center space types, reflecting distinct efficiency practices. However, a significant contrast emerges among various cooling system types. Data centers employing water-cooled chiller systems without economizers exhibit the highest WUE, largely attributed to substantial cooling tower water usage.

This trend extends to data centers using waterside economizers (water-cooled chillers), for both IT air-cooled facilities and IT liquid-cooled facilities. Waterside economizers play a pivotal role in reducing WUE values by eliminating or reducing heat from compressors. Moreover, in IT liquid-cooled data centers, WUE values can be further reduced through improved heat transfer and elevated coolant/facility water temperature setpoints, enhancing the utilization of waterside economizers throughout the year.

Despite these optimizations, the indispensable role of evaporative cooling in dissipating internal heat from IT devices remains evident, therefore contributing to significant water consumption. To mitigate water consumption, data centers can employ airside economizers, as demonstrated by facilities utilizing either an airside economizer (water-cooled chiller) or airside economizer with adiabatic cooling (water-cooled chiller). In these scenarios, the implementation of airside economizers allows for the shutdown of chilled water systems during favorable weather conditions, resulting in substantial water conservation. Notably, the application of airside economizers with adiabatic cooling (water-cooled chiller) is prevalent among hyperscale data centers. Their optimized operational practices enable extensive use of airside economizers to support data center cooling, with sporadic adiabatic cooling resulting in minimal water consumption.

Furthermore, data centers using dry coolers with adiabatic assist exhibit the highest water consumption among the remaining scenarios, where water consumption primarily occurs during the adiabatic pre-cooling process when wet mode operations are activated. Lastly, the remaining cases display minimal WUE values, primarily influenced by occasional humidification requirements.

The above discussion does not imply low site WUEs are necessarily good. In many cases, there are tradeoffs between low PUEs and low site WUEs. For example, water-cooled chillers and other evaporation-based cooling systems are generally more energy efficient than an air-cooled chiller or other waterless systems. While air-cooled chillers use no water, they use more energy. There are two WUE metrics: site and source. Site WUE only measures water at the facility level, while source WUE accounts for the water required to generate the electricity that is used by the facility. Including both site and source WUE reflects the true water cost of data centers, but its calculation is complex and highly dependent on the source of electricity.

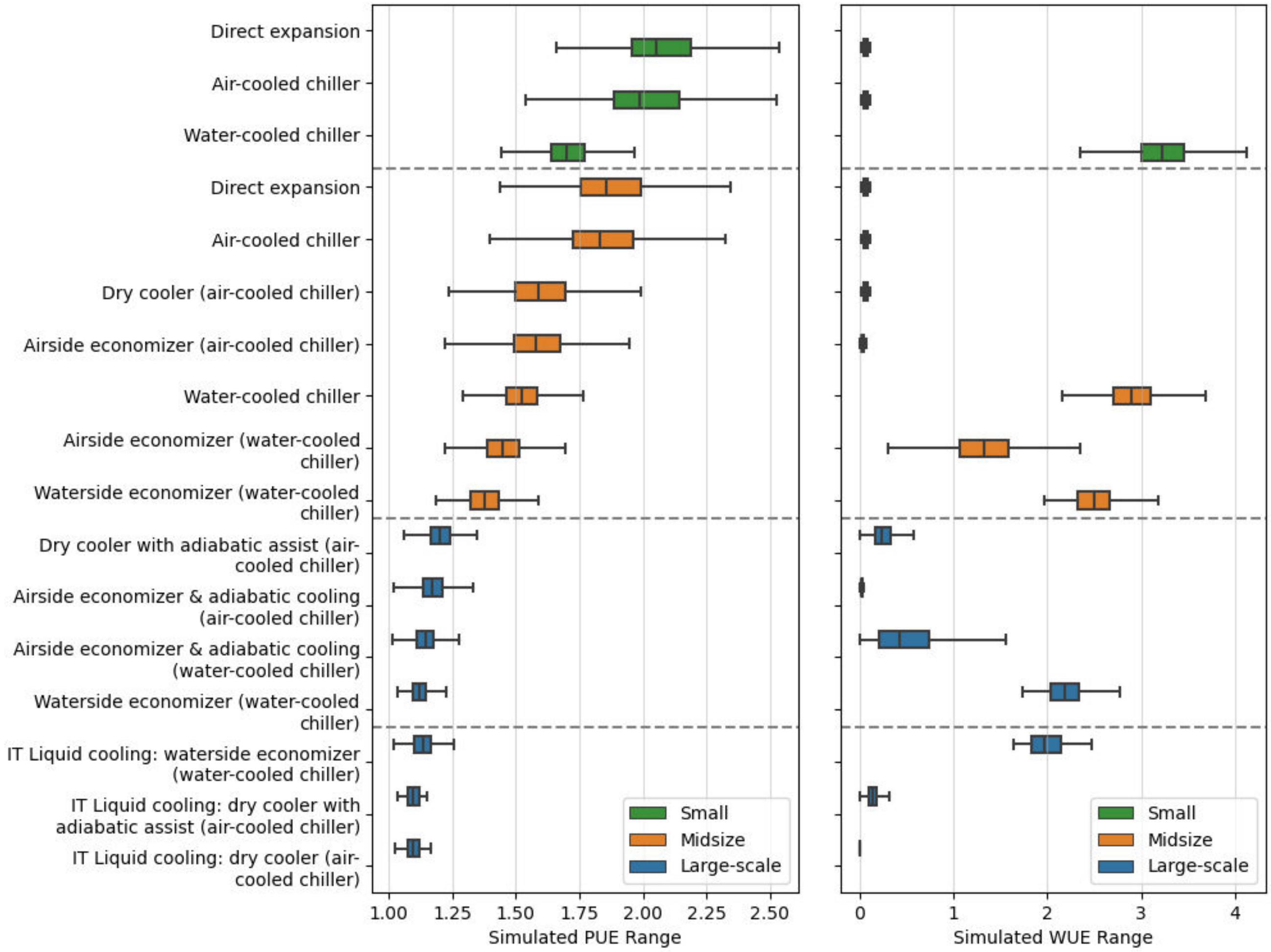


Figure 4.4. Simulated PUE and WUE (site) ranges by data center cooling system and space type.

Aggregate PUE and WUE for each data center space type is calculated by combining (1) the simulation results presented in Figure 4.4 (2) the distribution of cooling systems, discussed above and shown for a sample year in Figure 4.2 and (3) the assumed locational distribution of different data center space types, discussed above. This leads to the aggregate space type PUE and WUE values shown in Figure 4.5. The ranges shown in the figure represent the possible range of the aggregate metric, considering the 10th and 90th percentiles of the simulation results described above; the ranges do *not* represent the full spectrum of individual facility PUEs within the space type category. For example, while colocation facilities as a whole are not assumed to typically deploy the lowest-PUE cooling systems (i.e., those with adiabatic cooling), such systems on certainty in some colocation facilities. The PUE of a single colocation facility might therefore fall well below the range of the aggregate metric shown here.

As discussed above, significant uncertainty exists regarding the “true” PUE of existing data centers. The simulation modeling here assumes facilities are operated as designed, and makes assumptions about temperature setpoints, outdoor weather conditions, and other characteristics that might differ from reality. Notably, more data is needed to appropriately model the “airside economizer & adiabatic cooling (air-cooled chiller)” system, and the water usage associated with it. The current simulations are set up such that the adiabatic cooling system is not used most of the year, with most of the cooling being supplied by the airside economizer. This optimal operation might not be reflective of the true practices with these systems, and therefore the WUE of this system (shown in Figure 4.4) may be low. Some hyperscale facilities report WUE values for similar systems of 0.1-0.3 L/kWh, which is much higher than the values simulated here, though the specifics of those systems are not known. For context, if this cooling system has a median WUE of 0.2 L/kWh instead of the values modeled here, the aggregate median WUE for the hyperscale category would increase from 0.32 to 0.40, which is still within the range of uncertainty shown in Figure 4.5.

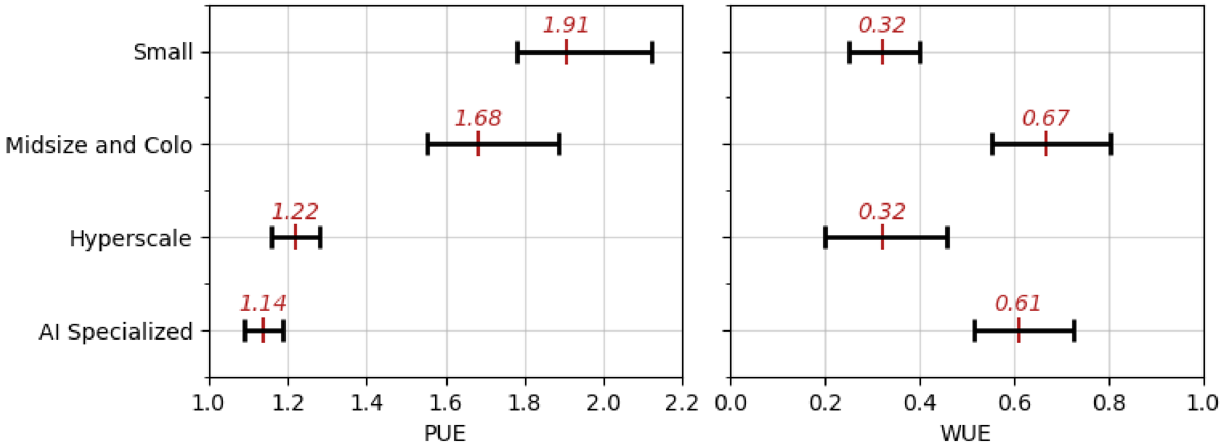


Figure 4.5. Aggregate PUE and WUE across space type categories considering the facility locations and mix of cooling systems present in 2023.

The resulting annual average PUE falls from 1.6 in 2014 to 1.4 in 2023, as shown in Figure 4.6. This decline is primarily due to the shift towards larger data centers (hyperscale, colocation) that have a lower PUE. By 2023, the average PUE across hyperscale and colocation data centers is under 1.4, with around 75% of servers installed in those data centers. This trend continues, and by 2028, the average PUE falls to between 1.15 and 1.35, driven again by the shift into more energy efficient hyperscale and colocation facilities, combined with the increase in liquid-cooled AI servers.

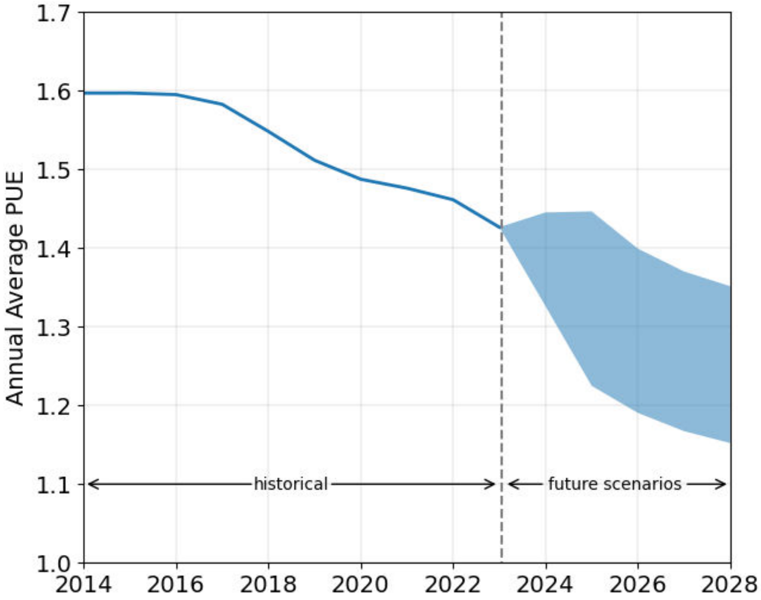


Figure 4.6. Annual average PUE across all U.S. data centers.

The shift toward hyperscale and colocation data centers results in an increase in the overall average WUE, which stays just over 0.36 L/kWh through 2023, as shown in Figure 4.7. After 2023, the average WUE rises slightly, reaching between 0.45 and 0.48 L/kWh, reflecting increasing WUEs in the hyperscale and colocation data centers, along with the increased water consumption of liquid-cooled systems.

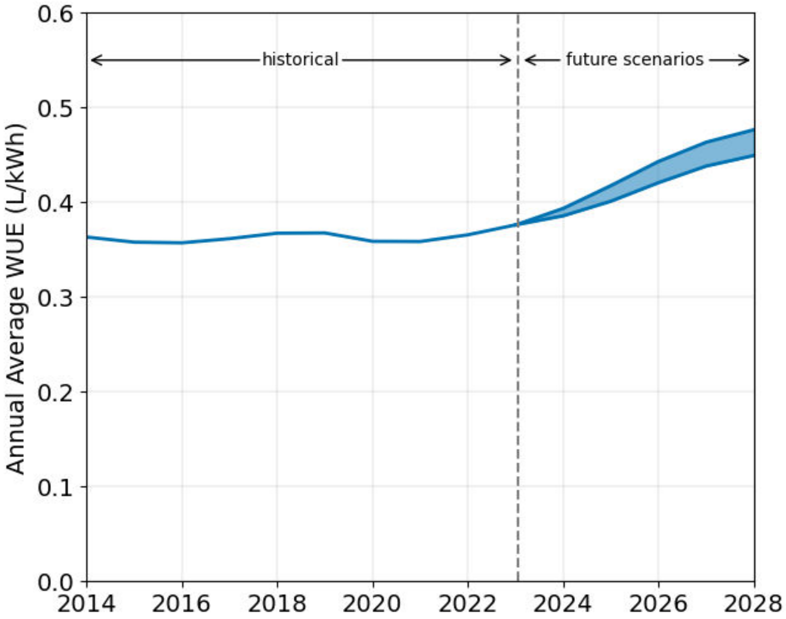


Figure 4.7. Annual average site WUE across all U.S. data centers.

5. Total Data Center Electricity/Water/Carbon Estimates

Server Electricity Use

The total annual server energy use from 2014 to 2023 is presented in Figure 5.1, along with a future scenario range of server energy use through 2028. Server energy usage grew from about 30 terawatt-hours (TWh) in 2014 to nearly 100 TWh in 2023, more than tripling during that period. A large portion of this increase came from GPU-accelerated AI servers, which grew in energy usage from less than 2 TWh in 2017 to more than 40 TWh in 2023. Conventional servers, primarily dual processor servers, increased significantly during the same period as well, doubling from about 30 TWh to nearly 60 TWh.

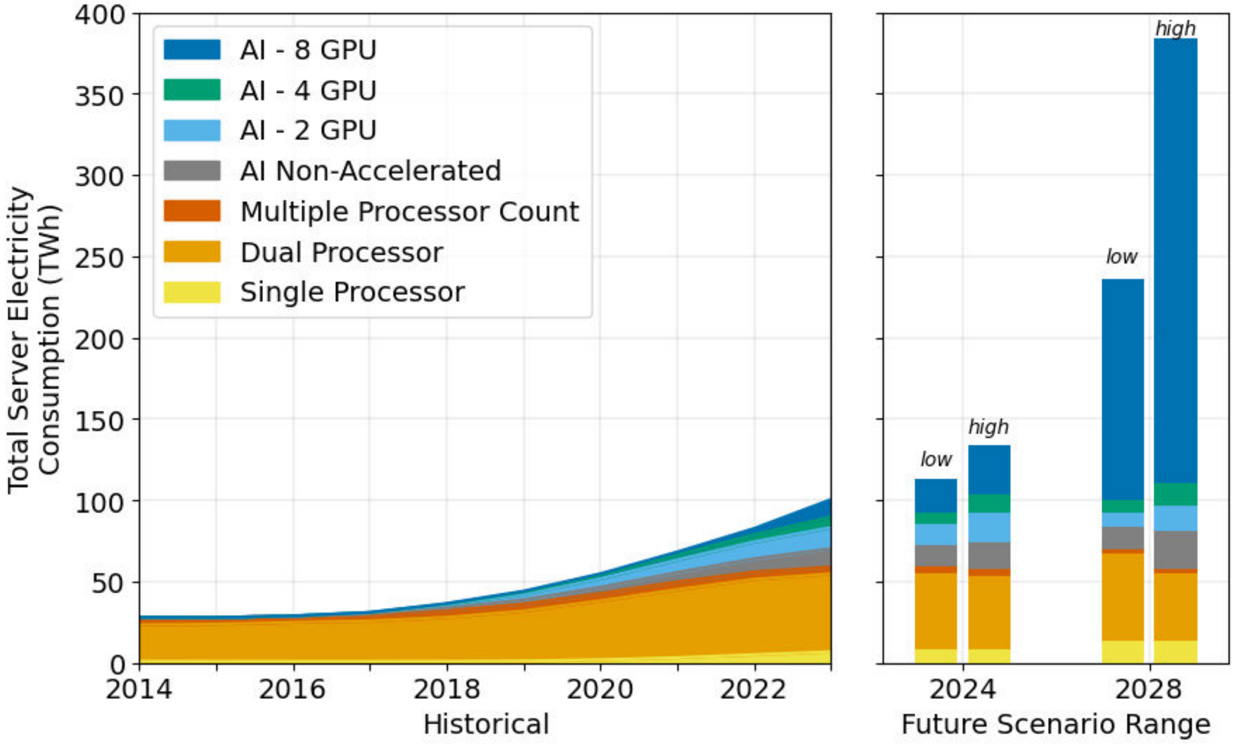


Figure 5.1. Server annual electricity usage by type.

After 2023, server energy use is presented as a range to reflect various scenarios of future equipment shipments and operational practices. Specifically, the count of future GPU shipments varies within the higher and lower bounds previously described in the AI Accelerated Server Shipments section. Second, the average operational power of AI servers varies between 60% and 80% of the rated power as noted in the Server Operational Wattage section. Finally, the average operational time of AI servers varies between 75% to 85% of the year as noted in the Server Operational Time section. Together the variations of these inputs provide a range of operation energy, with the low and high end representing about 240 and 380 TWh in 2028, respectively, as shown in Figure 5.1.

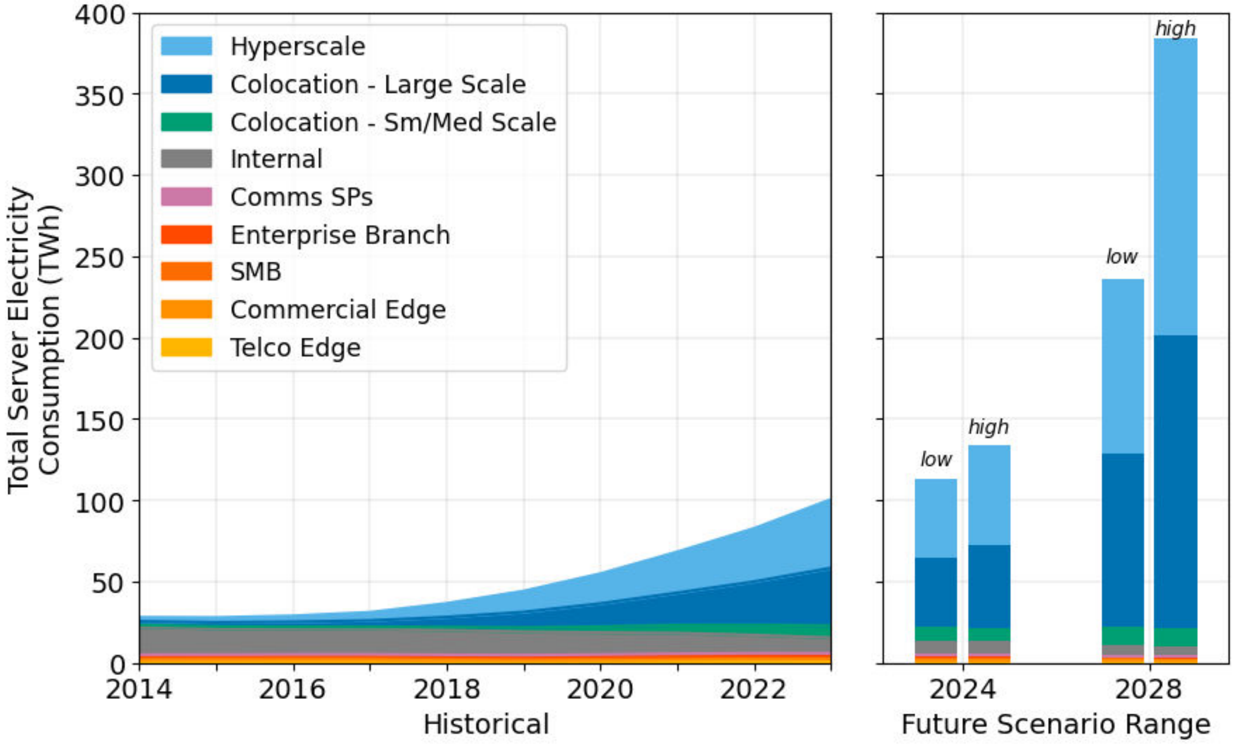


Figure 5.2. Server annual electricity use by space type.

Figure 5.2 presents total annual server energy use allocated by data center space type. In 2014, over 60% of server energy consumption was in internal data centers. By 2023, this fell to nearly 10%, with hyperscale and colocation data centers accounting for almost 80%.

After 2023, internal data centers' share of server energy continued to fall, reaching below 2% by 2028. Hyperscale and colocation data centers continue to grow in proportion, and by 2028 it is expected that hyperscale and colocation will account for over 90% of server energy consumption, primarily driven by AI workloads.

Most AI servers are allocated to inferencing throughout the entire period, and these servers make up nearly 60% of the AI server energy usage in 2023. By 2028, training server energy consumption surpasses inferencing, as more of the higher TDP GPUs are allocated towards training, consuming between 50% and 53% of the total AI server energy.

Storage and Network Electricity Use

Storage equipment electricity use grows steadily after 2014, nearly quadrupling by 2028, as shown in Figure 5.3. HDD units accounted for 94% of the total electricity in 2014, with flash making up just 6% of the 7 TWh total. By 2023, flash units had grown to 25% of the 16 TWh total. This growth continues and, by 2028, flash units make up 50% of the 22 TWh total. HDD electricity usage peaked in 2023 at 12.3 TWh, then falls slightly, reaching 11 TWh in 2028.

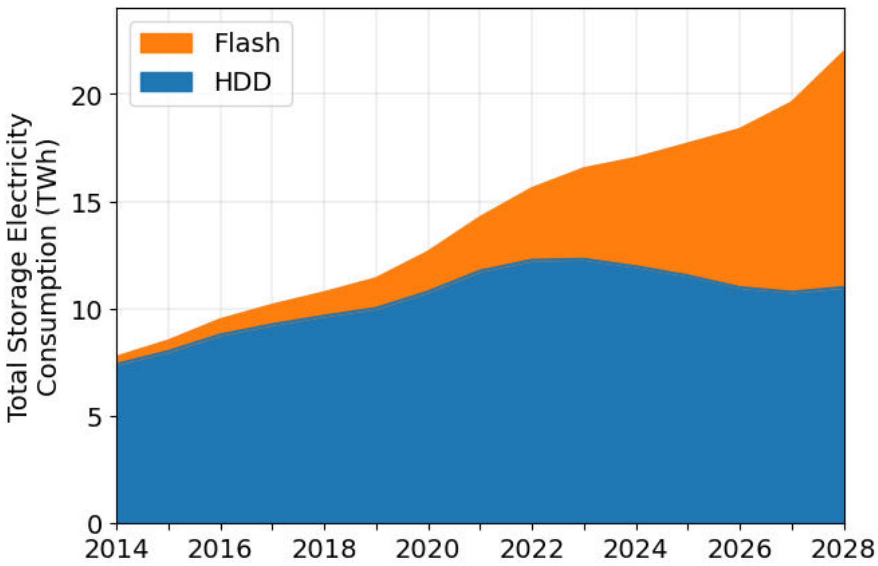


Figure 5.3. Storage equipment annual electricity use by type.

Network electricity also sees rapid growth through 2028, ultimately reaching 23 TWh, as shown in Figure 5.4. This growth is largely driven by the introduction of InfiniBand switch units, which account for 45% of the total in 2028, growing from just under 10% in 2023. Note that other internal network hardware for AI, such as NV links, are considered part the AI server system and represented within the accelerated server category. 200/400Gb ports also grow rapidly, from 6% in 2023 to 26% by 2028. Port speeds less than 50Gb see a decline in total energy usage. In 2023, these speeds accounted for 15% of total energy usage, but in 2028 they fall to under 2%.

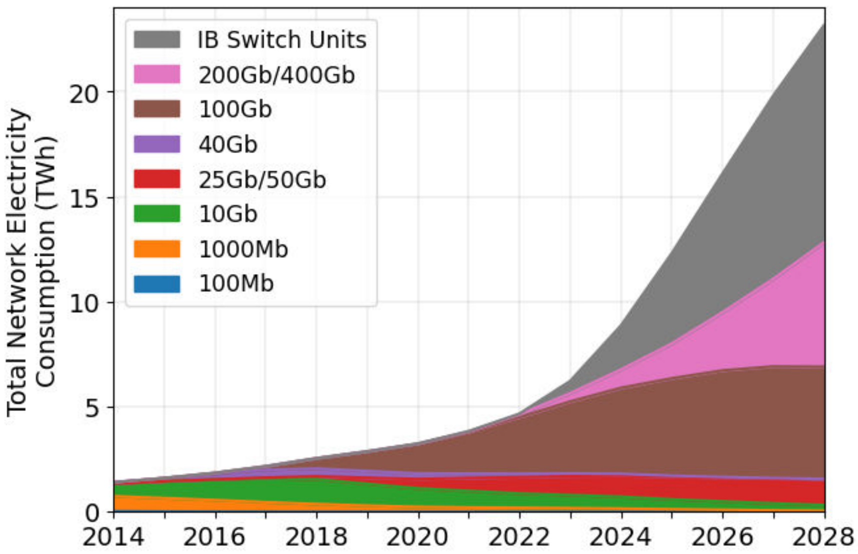


Figure 5.4. Network electricity use by port speed, with InfiniBand switches.

Total Electricity Use

Figure 5.5 presents total annual data center energy use from 2014 to 2023, along with a future scenario range of total data center energy use in 2024 and 2028. Data center energy use remained fairly stable between 2014–2016 at about 60 TWh. Energy use began to increase as the amount of accelerated AI servers in the server stock began to become significant in 2017, and by 2018 data centers consumed about 76 TWh, representing 1.9% of total U.S. electricity consumption. U.S. data center energy use continued to grow at an increasing rate, reaching 176 TWh by 2023, representing 4.4% of total U.S. electricity consumption.

After 2023, total data center energy use is presented as a range to reflect various scenarios of future equipment shipments and operational practices, as well as variations in cooling energy use. The equipment variations are based on the range of GPU shipments, average operational power and operational time of GPU-accelerated servers, as previously outlined for the future range of server electricity use. The cooling energy use variations are based on scenarios in cooling system selection type and efficiency of those cooling systems, as previously outlined in the infrastructure section of this report. Together the scenario variations provide a range of total data center energy, with the low and high end representing about 325 and 580 TWh in 2028, as shown in Figure 5.5, representing 6.7% to 12.0% of total U.S. electricity consumption.

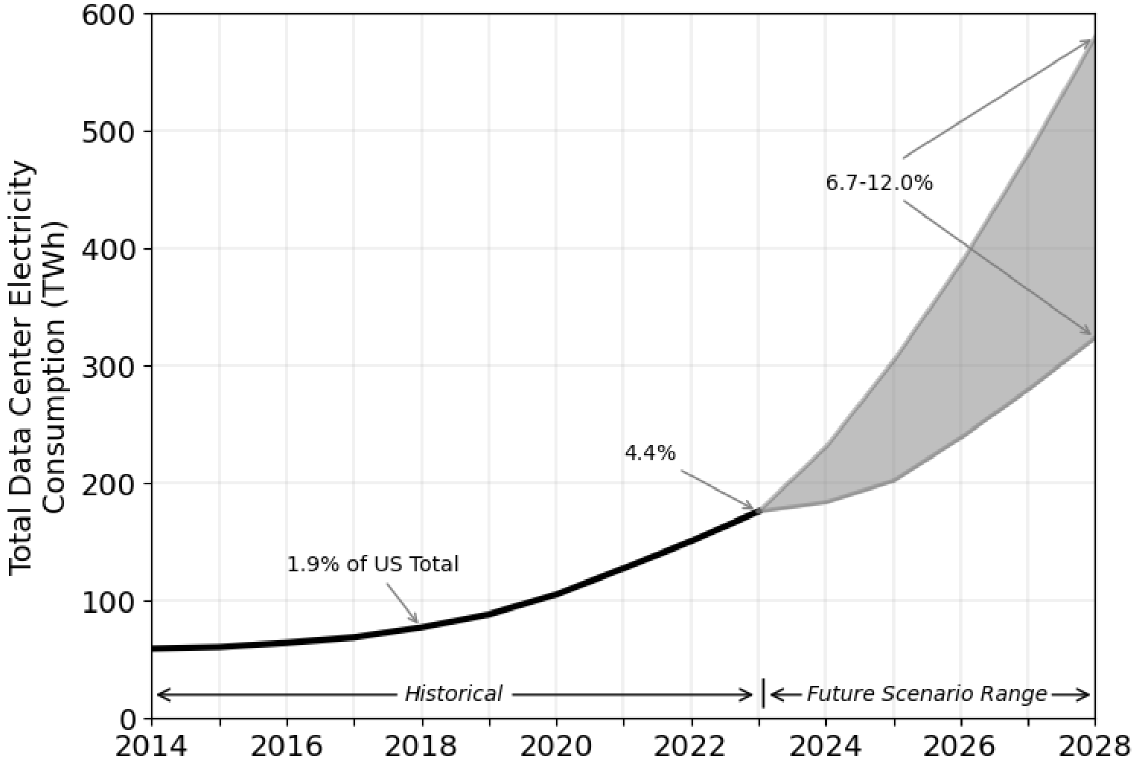


Figure 5.5. Total data center electricity use from 2014 through 2028.

Figure 5.6 presents the total historical and future ranges of U.S. data center energy use allocated by data center equipment, showing that total energy use growth between 2014 and

2023 is driven by both a rapid proliferation of AI servers and well as continued growth in conventional server energy demand. Storage electricity continues to increase, but at a rate slower than server growth, causing the storage proportion of total data center energy to begin to slightly decrease as AI servers begin to impact the total server stock. Networking electricity use remains flat around 3% from 2014 to 2022, then begins to grow in 2023 with the introduction of InfiniBand switch units. Infrastructure energy accounts for 40% of total electricity in 2014, then falls to 30% in 2023 as the national average PUE improves, as previously shown in Figure 4.7. The falling PUE is driven by two main factors, PUE improvements across all data center types and the increase in proportion of servers installed in facilities with lower PUE (hyperscale, colocation), as shown in Figure 5.7. Additionally, Figure 5.6 shows that the total range of energy use growth for 2024 and 2028 is highly dependent on the quantity and operation of AI servers.

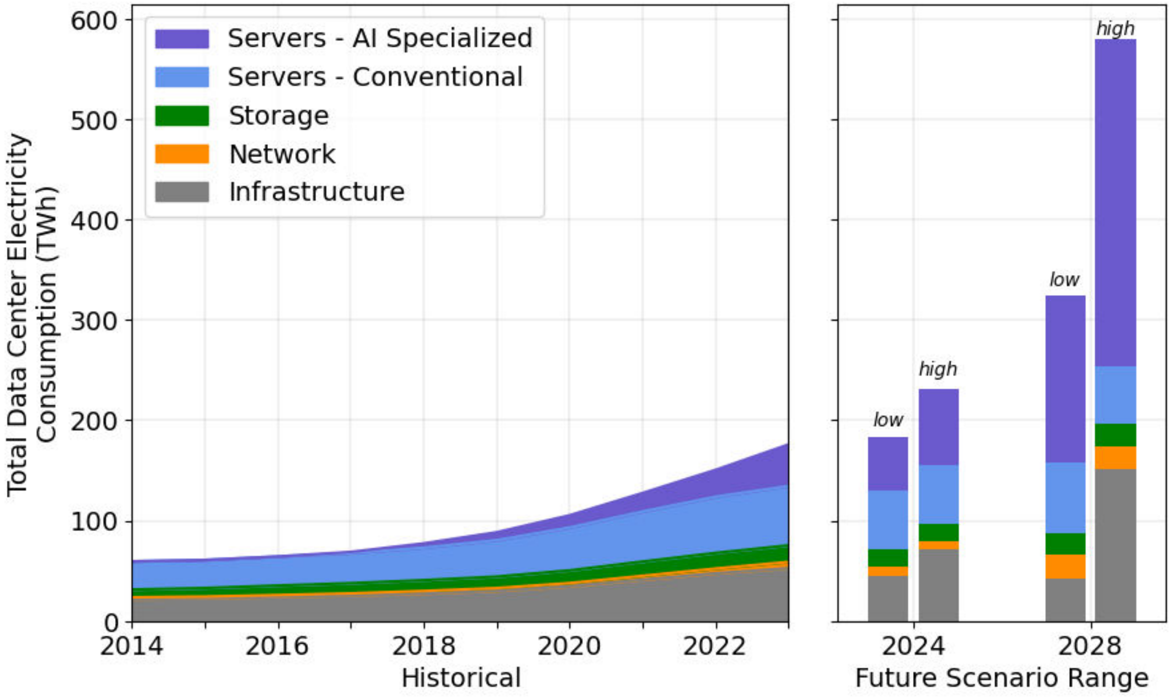


Figure 5.6. Total data center electricity use from 2014 through 2028 by equipment type.

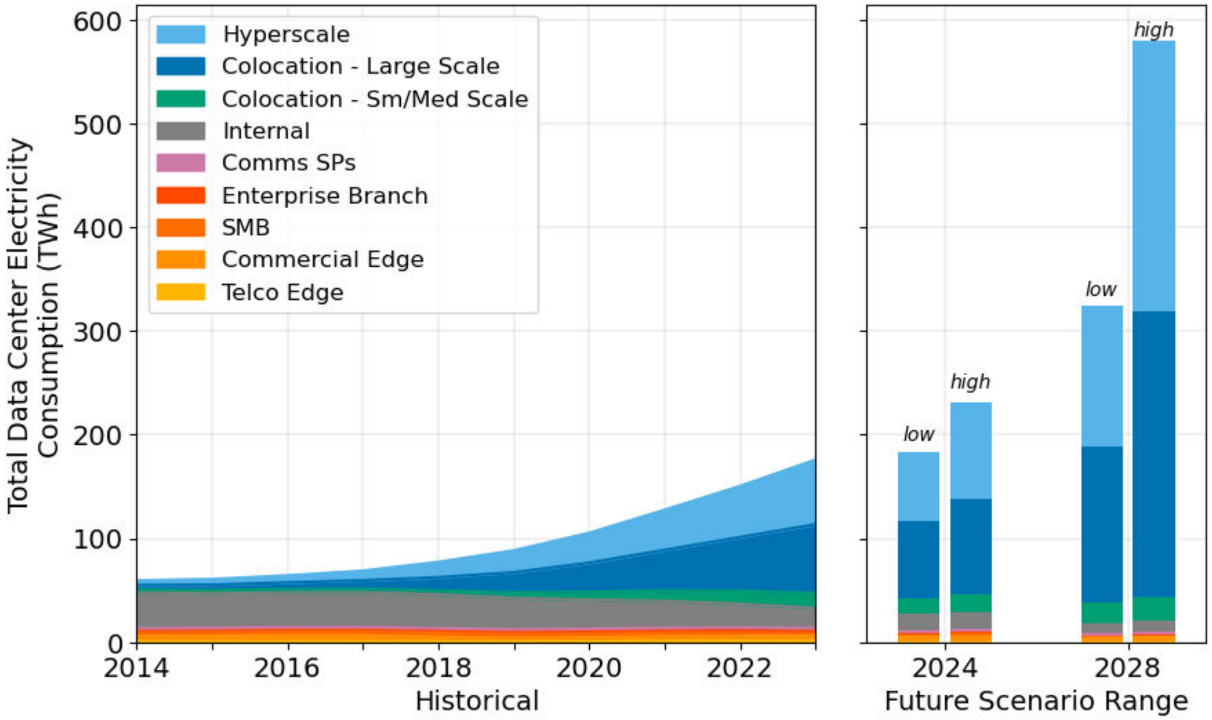


Figure 5.7. Total data center electricity use from 2014 through 2028 by space type.

Comparison to Previous Report

The prior study conducted by Shehabi et al. (2016) covered the 2000–2020 time period. For 2015–2020, values were projected scenarios based on expected equipment shipment trends and several operational and efficiency scenarios. Substantial differences exist between the values presented in that report and what we find today. Even looking at the historical base year of that study, 2014, we have slightly different estimates for total electricity consumption. This difference is driven by a few key factors. First, updated shipment data showed a slight decrease (a few percent) in units sent, as compared to data we received at the time. Additionally, our definition of space types changed slightly, and we obtained much more granular analyst estimates for the floor space and number of racks in various types of data centers. Ultimately, this led us to estimate that more servers were still present in internal and small data centers, which run servers at lower utilization levels, than we previously thought; this lowered the total estimated power consumption of servers. Finally, we altered the categorization of servers to be based on processor count rather than price point (i.e., volume, midrange, and high-end categories in the prior report). In doing so, we revisited our power assumption for servers that previously appeared in our high-end category and now largely exist in our multiple processor count category (in the subcategories of 8 or 16+ processors). For these servers, we previously extrapolated the wattage trend over a very long time period using an annual percentage growth rate; this led to very high wattage estimates that are not in line with any servers we have seen produced and shipped since then. While these servers are a very small portion of the total server installed base, assuming these high wattages increased our results meaningfully over the new estimates.

About one-third of the difference between the 2014 values in the two studies is due to the differences in server electricity consumption described above. This difference is compounded by PUE, since any change in IT electricity also changes the electricity consumed. Further, the current study has a lower aggregate PUE across the industry as compared to the 2016 study; this is both because we now assume fewer servers had shifted to hyperscale data centers, as mentioned above, and because our current PUE modeling assumes systems are operated as designed, leading to more efficient values.

The "Current Trends" scenario in the 2016 report displayed a future where data center electricity use remained constant through 2020. Additional efficiency scenarios showed the potential for electricity use of the industry to decrease throughout this time period. The scenarios in the 2016 report did not capture the rise of AI, which has brought a fundamental change in the industry and the demand for computing services. Therefore, the current study estimates historical electricity use for 2018–2020 that is higher than any of the 2016 report scenario results. A comparison of the 2016 report and this current analysis is presented in Figure 5.8.

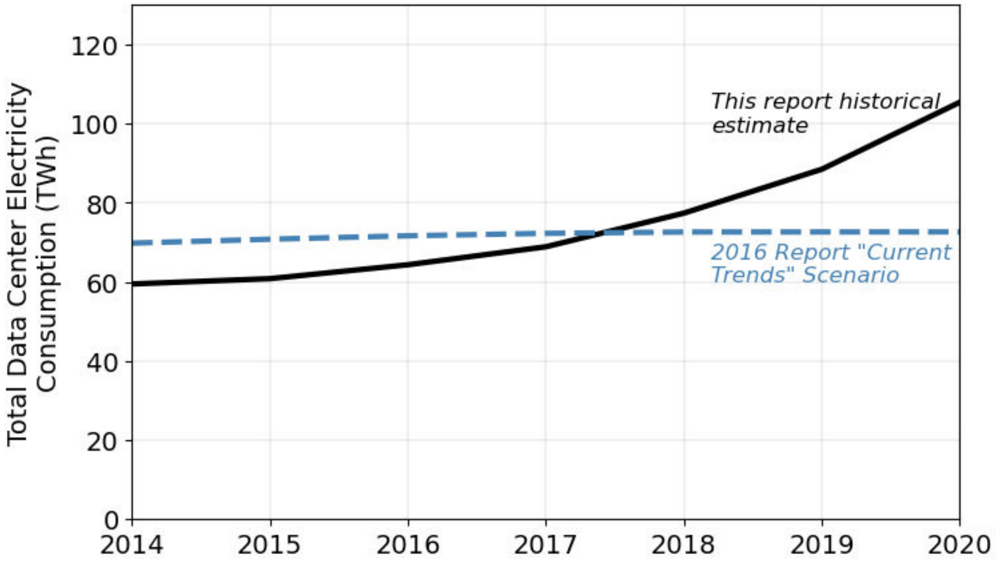


Figure 5.8. Total electricity use for 2014–2020 as estimated in this report, compared to the "Current Trends" scenario from the 2016 report.

Water and Emission Impacts

Direct Water Consumption

Figure 5.9 shows the growth in direct water usage in data centers. In 2014, data centers consumed 21.2 billion liters of water, with 64% in internal data centers. By 2023, hyperscale and colocation account for 84% of the 66-billion-liter total, while internal data centers fell to just 12%, driven by water efficiency improvements. These trends are expected to continue through 2028, with internal data centers falling to just 2% of the total. Hyperscale data centers in 2028 are expected to consume between 60 and 124 billion liters.

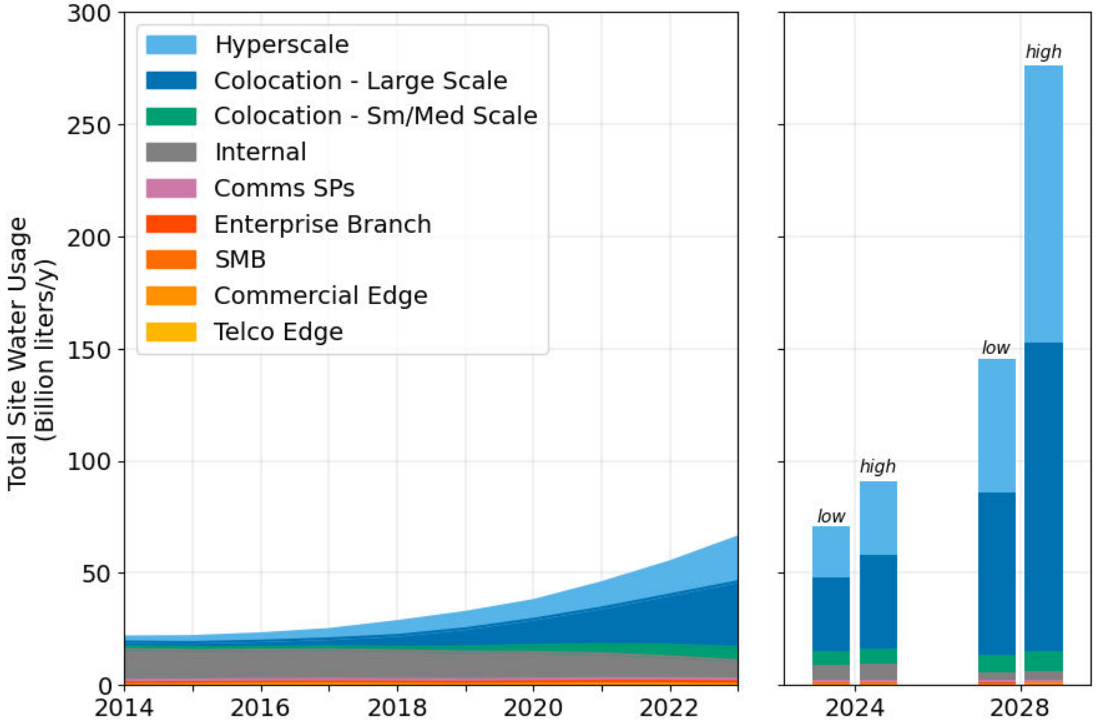


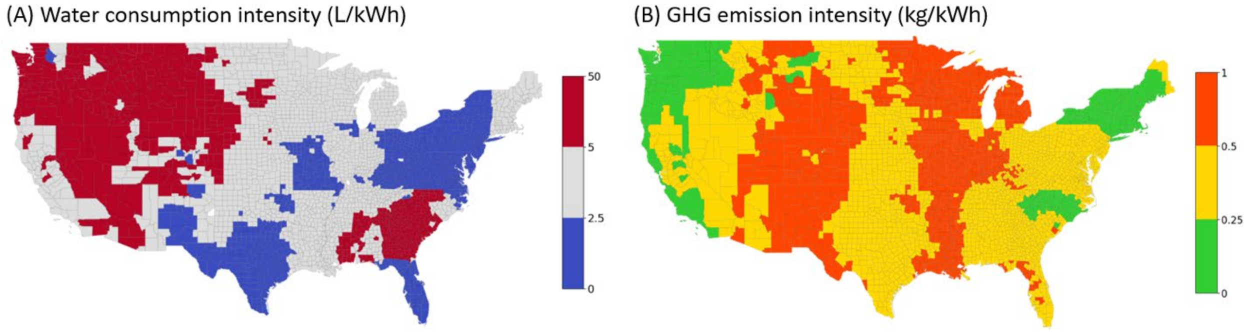
Figure 5.9. Direct water consumption by data center type.

Indirect Water Consumption and Emissions from Electricity Use

Indirect water use and GHG emissions associated with electricity use represent impacts occurring at the power generation source. The U.S. has over 12,000 utility-scale power plants using fossil fuels, nuclear, and renewables, each with unique water and emissions footprints (EIA 2022). Water usage during power generation varies based on the type of energy and the plant's efficiency. Water consumption refers to the amount of withdrawn water that is permanently removed from the immediate water cycle due to evaporation or other irreversible processes. Thermoelectric plants require significant cooling, while hydroelectric reservoirs lose water through open surface evaporation, leading to water consumption that impacts local resources. GHG emissions are primarily attributed to combustion from fossil-fuel-based (coal, natural gas, and petroleum) electricity generators, which account for more than 99% of emissions associated with electricity generation (EIA 2022). The water consumption and GHG emissions embedded in electricity use depend on the fuel type, technology, and the location of generation.

Balancing authorities manage the electrical grid, ensuring that electricity demand and supply balance within a specific portion of it. They oversee generation and coordinate electricity transfers with neighboring areas. Balancing authorities represent the most detailed level of the electricity grid, managing real-time electricity supply and demand through internal generation and external transfers. The water and emissions associated with electricity used by data centers and other end users depends on the power plants supplying their electricity, influenced by the fuel mix and efficiency managed by the balancing authorities. The U.S. Energy Information Administration (EIA) gathers and disseminates detailed data on U.S. power plants, thereby enhancing our understanding of the U.S. electricity generation portfolio and its environmental impact. We used detailed electricity interchange data between balancing authorities from the EIA hourly electric grid monitor (EIA 930 2024) and combined it with power plant-level electricity generation (EIA 923 2024), water consumption (EIA 2024), and GHG data (EPA 2024) to estimate the environmental impact associated with a unit of supplied electricity for each balancing authority. The detailed methodology, intensity factors, and related scripts can be found in Siddik et al. (2024).

The water and emissions associated with electricity use can be quantified using water consumption intensity, which measures the volume of water consumed per unit of electricity used, and GHG emission intensity, which measures the mass of GHG per unit of electricity used. Each county is assigned a balancing authority based on its geographic location, linking the environmental impact of electricity use to the end users. Figure 5.10 illustrates the spatial variation of the annual average water consumption intensity and GHG emission intensity factors across U.S. counties, derived from the balancing authorities governing those areas. This visualization reveals significant, yet uneven, variability across different regions of the country.


(A) Water consumption intensity (L/kWh)

(B) GHG emission intensity (kg/kWh)

Figure 5.10. Water consumption and GHG emission intensity factors of electricity use by county.

As noted earlier, U.S. data centers consumed approximately 176 TWh in 2023. The total indirect water footprint of U.S. data centers is nearly 800 billion liters, attributed to water consumed indirectly through electricity use, based on the regional electricity grid mix for U.S. data center locations. Concurrently, total GHG emissions for that same electricity grid mix would be 61 billion kilograms of CO_2 equivalent. In 2023, energy consumption by a data center translates to national average of 4.52 L/kWh of indirect water consumption, along with 0.34 kg/kWh emissions. In comparison, the average water intensity factor for electricity use in the U.S. overall is 4.35 L/kWh, while the emission factor is 0.35 kg/kWh of CO_2 equivalent.

It is important to note that the methodology used here to calculate indirect water and emission impacts does not incorporate any power purchase agreements between individual data center facilities and their electricity providers or on-site “behind the meter” generation, which could significantly affect water consumption and emissions estimates, depending on the electricity source. Nevertheless, due to the unavailability of facility-level data, we are constrained to assume the same electricity grid mix as that provided by the local balancing authority for all data centers within its jurisdiction.

With the projected growth of data centers' energy use in the coming years, indirect water consumption and emissions are also expected to increase; however, future estimates of water consumption and emissions from electricity generation should consider potential future changes in the electricity mix. Decarbonization of the power sector is a critical requirement to achieving the U.S. 2050 net-zero GHG emissions goal (Grubert 2020), which may accelerate the retiring of fossil fuel-fired power plants by 2035 and transitioning to newer generation technologies like renewable energy and new firm clean power, such as scalable nuclear generation. Such decarbonization efforts will inevitably impact the future indirect water consumption and GHG emissions associated with data center electricity consumption.

6. Total Cryptocurrency Electricity Use

Introduction

Cryptocurrency mining represents a specialized form of computational infrastructure that differs substantially from traditional data centers. At its core, mining involves repeatedly generating cryptographic hashes—unique fixed-length codes derived from transaction data—until one meets specific criteria set by the network. The rate at which these hashes can be generated, measured in hashes per second, determines a miner's probability of earning rewards. While typical data centers process diverse workloads with varying computational demands, cryptocurrency mining facilities are single-purpose installations dedicated solely to this hash generation process through specialized hardware. These application-specific integrated circuits (ASICs) operate continuously at maximum capacity, with their combined hashrate determining their share of network rewards. Miners use these chips to validate network transactions, receiving newly minted cryptocurrency as compensation when they successfully generate a qualifying hash.

This analysis adopts a top-down methodology rather than the bottom-up approach employed elsewhere in this report, necessitated by two fundamental constraints. First, cryptocurrency mining operates on entirely different IT hardware than conventional data centers. As a result, our primary IT equipment shipment data—which forms the foundation of our bottom-up analyses for traditional data centers—does not capture the specialized ASICs used in mining operations. Second, the cryptocurrency mining industry's limited transparency makes comprehensive facility-level data collection impractical.

To address these constraints, we estimate U.S. energy consumption by analyzing the global Bitcoin network's total hashrate and hardware efficiency characteristics, then determine the U.S. share based on geographic distribution data. This methodology enables us to develop robust estimates despite data limitations, providing a continuous historical view of energy consumption patterns that can be validated against available reference points.

Methodology

This analysis estimates the historical energy consumption of Bitcoin mining activity in the U.S. by combining publicly available data on global network hashrate, hardware efficiency, and the estimated U.S. share of the global hashrate. Hardware efficiency estimates are obtained from the Cambridge Bitcoin Electricity Consumption Index (CBECI) web portal,[1] which provides lower bound, best-guess, and upper bound estimates of mining hardware efficiency in joules per terahash (J/TH). All CBECI parameters applied in this analysis are sourced from their $0.05/kWh cost scenario.

The CBECI hardware efficiency estimates are based on a comprehensive analysis of over 100 distinct Bitcoin ASIC models launched since 2013. To better reflect real-world market conditions, the CBECI methodology limits hardware manufacturers to focus on the three major producers (Bitmain, MicroBT, and Canaan), which are estimated to have a combined market share exceeding 85%. This filter excludes “exotic” devices with minimal sales impact and is not applied to hardware released before July 2014. Additionally, a five-year maximum economic lifetime is introduced for hardware to avoid overstating power demand by including nominally profitable but likely obsolete older hardware. The CBECI analysis also incorporates expert input to adjust manufacturer efficiency specifications to better

[1] https://ccaf.io/cbnsi/cbeci

reflect real-world operating conditions and variation.

The location of Bitcoin mining is another area of uncertainty. Historically, miners have sought inexpensive electricity and a permissive regulatory environment. In 2021, the Chinese government imposed a ban on crypto-mining. In order to estimate the energy footprint of U.S. crypto mining, it is necessary to estimate the share of global mining occurring in the United States. Geographic hashrate distribution data is sourced from the CBECI mining map, which provides monthly estimates of each country's share of the global hashrate. The CBECI mining map tracks Bitcoin's geographic hashrate distribution by collecting and aggregating IP address data from participating mining pools, which connect mining facility operators to pool servers, providing a sample that has historically represented 32–38% of total Bitcoin hashrate since the map's launch in September 2019. In this analysis, the U.S. share of global hashrate is assumed to be the mean of the monthly shares prior to May 2021 (5.98%) for months where the CBECI estimate is unavailable before that date. For missing months after May 2021, the mean of post-China-ban shares is used (34.82%). These pre- and post-ban U.S. hashrates share parameters that can be adjusted as needed. Lack of transparency in mining location remains a significant source of uncertainty in estimating cryptocurrency carbon emissions (de Vries et al. 2022).

Based on this geographic share and global hashrate, we estimate the monthly hashrate in the United States. The monthly U.S. hashrate is then converted to an estimated monthly information technology (IT) energy consumption in terawatt-hours (TWh) using the CBECI hardware efficiency values. To account for the total network energy consumption, upper and lower bound estimates are calculated based on assumed power usage effectiveness (PUE) ratios of mining facilities. These estimates are obtained by multiplying the IT hardware energy consumption by PUE ratios of 1.1 and 1.2. These PUE assumptions align with analysis from Siddik, Amaya, and Marston (2023), who found typical PUE values for mining operations range from 1.05 to 1.20. For lower bounding scenarios, hardware is assumed to be only the most efficient models, operating with the lowest PUE. For upper bounding scenarios, the CBECI profitability threshold and upper bound PUE are used.

The CBECI best-guess estimate assumes that miners use a weighted basket of profitable hardware, with the prevalence of each model depending on its deployment date. The profitability threshold is determined by the assumed electricity price. Hardware is assigned a weight between 0 and 1 based on the time elapsed since its estimated deployment, with newer devices receiving higher weights. A two-month lag is assumed between hardware release and deployment. This approach aligns with Lei, Masanet, and Koomey (2021), who emphasize the importance of using time-period appropriate technology data given how rapidly mining hardware evolves. The weighted average efficiency of the hardware basket is then used to calculate the network power demand and electricity consumption, assuming a PUE of 1.10 for all facilities. These estimates were updated (Messina 2023) in response to a novel method of hardware composition estimation developed by Helmy et al. (2023).

The U.S. Energy Information Administration (EIA) recently conducted a comprehensive assessment of cryptocurrency mining's electricity consumption, employing both top-down and bottom-up approaches (Morey, McGrath, and Minato 2024). Their top-down analysis utilized CBECI data to estimate U.S. mining's share of global activity, while their bottom-up approach identified 137 mining facilities across 21 states, with detailed power capacity data available for 101 of these sites. By assuming an 80% utilization rate for their documented 10,275 MW of mining capacity, the EIA estimated annual electricity consumption of approximately 70 TWh for 2023. The bottom-up estimate from the EIA is shown alongside our estimates of U.S. annual mining energy use in the following plot, providing an independent point of calibration for our methodology.

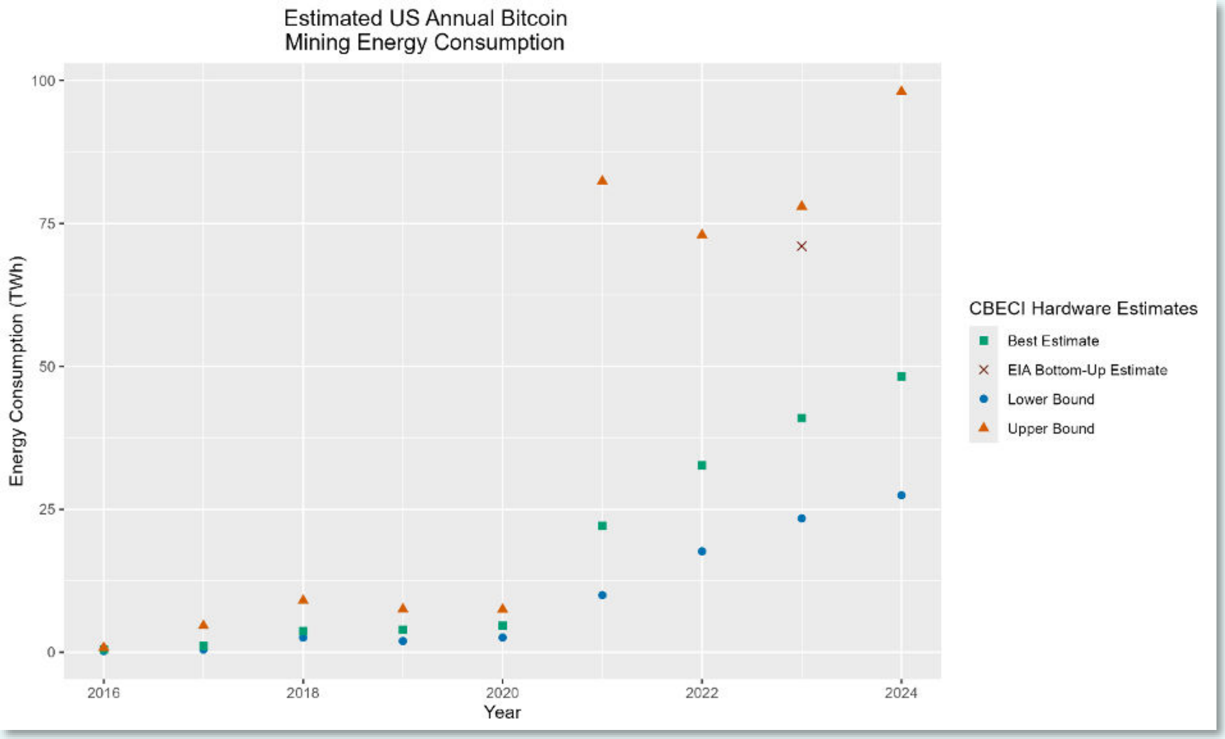


Figure 6.1. Annual energy consumption estimates for U.S. Bitcoin mining from 2016 to 2024. The graph presents three CBECI-based estimates (best estimate, lower bound, and upper bound) showing the growth in energy usage over time. The EIA's 2023 bottom-up estimate of 70 TWh (marked with an X) falls between CBECI's best estimate and upper bound projections, providing independent validation of the estimation methodology.

While our approach aims to more realistically represent the hardware mix and its evolution over time compared to a simple equally weighted average, it has some limitations. The weighting scheme is based on assumed average hardware lifetimes and depreciation schedules rather than direct data on the actual hardware in use. The two-month deployment lag and 1.10 PUE are also generalizations that may not capture the full diversity of miner behavior and facility efficiency. Finally, the CBECI methodology relies solely on public data and does not incorporate direct input from miners on their installed hardware base, though it does use expert judgment to adjust device efficiency assumptions.

Our analysis solely considers Bitcoin mining related energy use for several reasons. First, Bitcoin accounts for the vast majority of proof-of-work-based cryptocurrency mining activity. Second, cryptocurrencies based on alternative consensus mechanisms like proof-of-stake use dramatically less energy (OSTP 2022). Based on current market dynamics and the technological underpinnings of these digital assets, our assessment is that Bitcoin accounts for the overwhelming majority of cryptocurrency-related energy use in the United States (McDonald 2022).

Focusing on Bitcoin allows for a more accurate and targeted analysis of the largest contributor to mining energy consumption. However, this methodology could be extended to other proof-of-work cryptocurrencies if they gain significant adoption and there is sufficient data on their network hashrates, hardware efficiency, and geographic distribution of mining activity becomes available.

While this methodology provides a data-driven approach to estimate bitcoin mining energy consumption under a range of plausible assumptions, it could be further improved by incorporating direct data from mining industry participants. Granular information from Bitcoin miners on their specific hardware types, efficiencies, facility PUE ratios, and capacity utilization would enable more precise estimates.

Such industry data could include:

1. Detailed inventories of ASIC models used, their rated efficiencies, and deployment volumes over time
2. Measured PUE ratios and environmental operating conditions (e.g., ambient temperature, cooling setup) of mining facilities
3. Capacity utilization rates and hashrate contribution of different facilities and hardware types
4. Retirement and replacement rates of older hardware models
5. Energy procurement and consumption data from utility bills or smart meter readings.

Incorporating this type of direct, verifiable data from mining operations would greatly improve the accuracy and granularity of energy consumption estimates. It would enable more realistic modeling of the evolution of the mining hardware mix, efficiency gains from new ASICs and facility upgrades, and geographic shifts in mining activity. Engaging mining industry participants to voluntarily share anonymized, aggregate data would be a valuable extension of this analysis.

To project future energy consumption, we developed linear regression models linking Bitcoin price to U.S. mining energy consumption across the CBECI hardware efficiency scenarios. The models demonstrate strong statistical relationships between price and energy consumption, with R-squared values of 0.8251 for the lower bound estimate and 0.9011 for the upper bound estimate. These high correlations, each with a high level of statistical significance (p>.0001), are consistent with theoretical predictions (Helmy et al. 2023; OSTP 2022; Messina 2023), as mining profitability—and thus energy consumption—should closely track Bitcoin price. Time series for Bitcoin price, and our lower bound, best guess, and upper bound estimates are shown below in Figure 6.2.

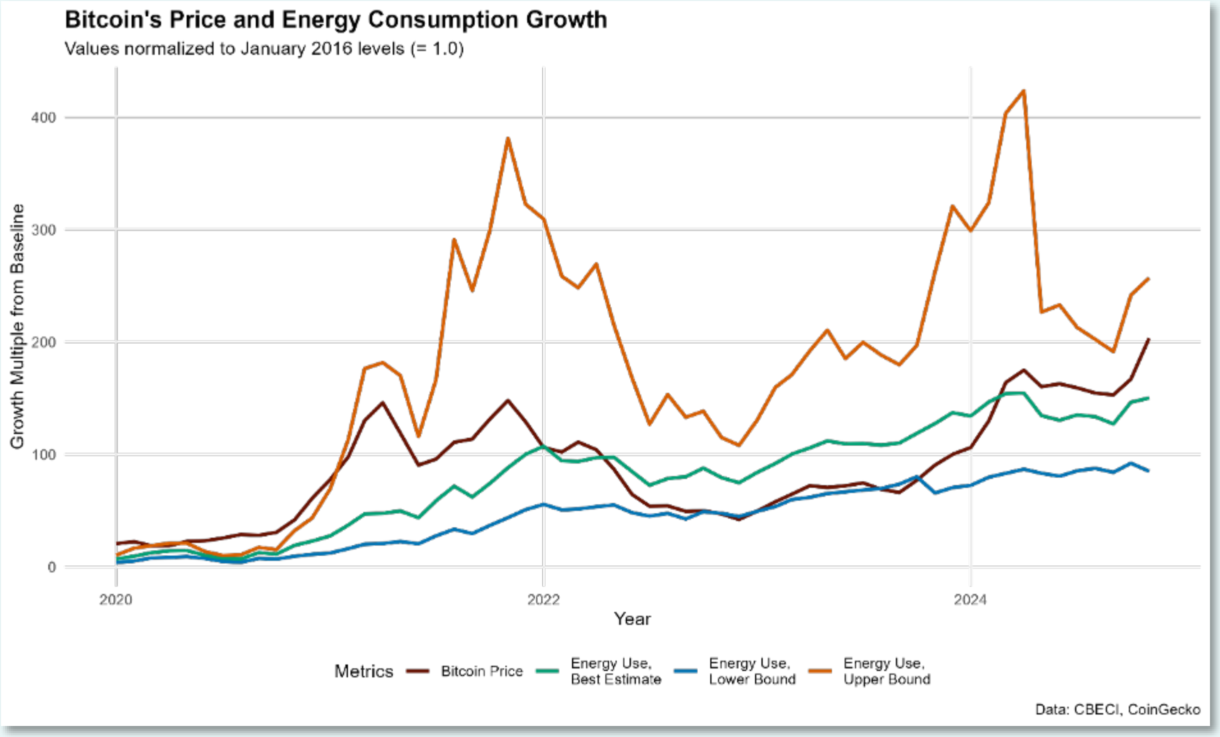


Figure 6.2. Relationship between Bitcoin's price and estimated energy consumption from 2020 to 2024, normalized to January 2016 price and best estimate energy use.
The graph shows three energy consumption scenarios (lower bound, best estimate, and upper bound) plotted against Bitcoin price movements. The upper bound estimate demonstrates the highest volatility, closely tracking price fluctuations, while the lower bound estimate shows more gradual growth. All metrics indicate substantial growth from baseline, with energy consumption estimates ranging from roughly 90x to 250x of their 2016 levels by 2024. Data sourced from the Cambridge Bitcoin Electricity Consumption Index (CBECI) and CoinGecko.

Using our best-guess hardware efficiency scenario, the model shows that price variations explain approximately 86% of the observed changes in energy consumption ($R^2 = 0.86$). The model coefficient indicates that for every $1,000 increase in Bitcoin price, U.S. mining energy consumption increases by approximately 0.058 TWh per month. While direct measurement of mining energy consumption would be preferable, the strong theoretical foundation and high statistical significance of the price-consumption relationship make this approach the most reliable option given currently available data.

To construct future scenarios, we analyzed historical Bitcoin price movements over rolling four-year periods to understand typical growth patterns. Based on this analysis, we modeled two scenarios: a conservative case where Bitcoin price doubles by the end of 2028, and an aggressive case where price increases five-fold over the same period. While actual Bitcoin prices exhibit significant short-term volatility, we model price growth as a smooth exponential increase to focus on long-term trends rather than short-term fluctuations. These price trajectories were then input into our regression model to estimate corresponding energy consumption under the CBECI's lower bound, best-guess, and upper bound hardware efficiency assumptions.

This approach was informed by Papp et al. (2023), who established a clear relationship between Bitcoin price movements and carbon emissions, finding a long-run price elasticity of

0.33–0.40 and a larger short-run elasticity of 0.69—-0.71. Rather than assuming a carbon intensity, we instead derived a relationship between modeled energy use and price, implicitly imputing hardware efficiency improvements, which allowed for more robust estimates (Koomey and Masanet 2021). These methods allow us to bound future energy consumption estimates based on historically observed price relationships while acknowledging the inherent uncertainty in both price movements and the price-energy use relationship. Plots of these estimates for both the moderate and high growth scenarios are shown below.

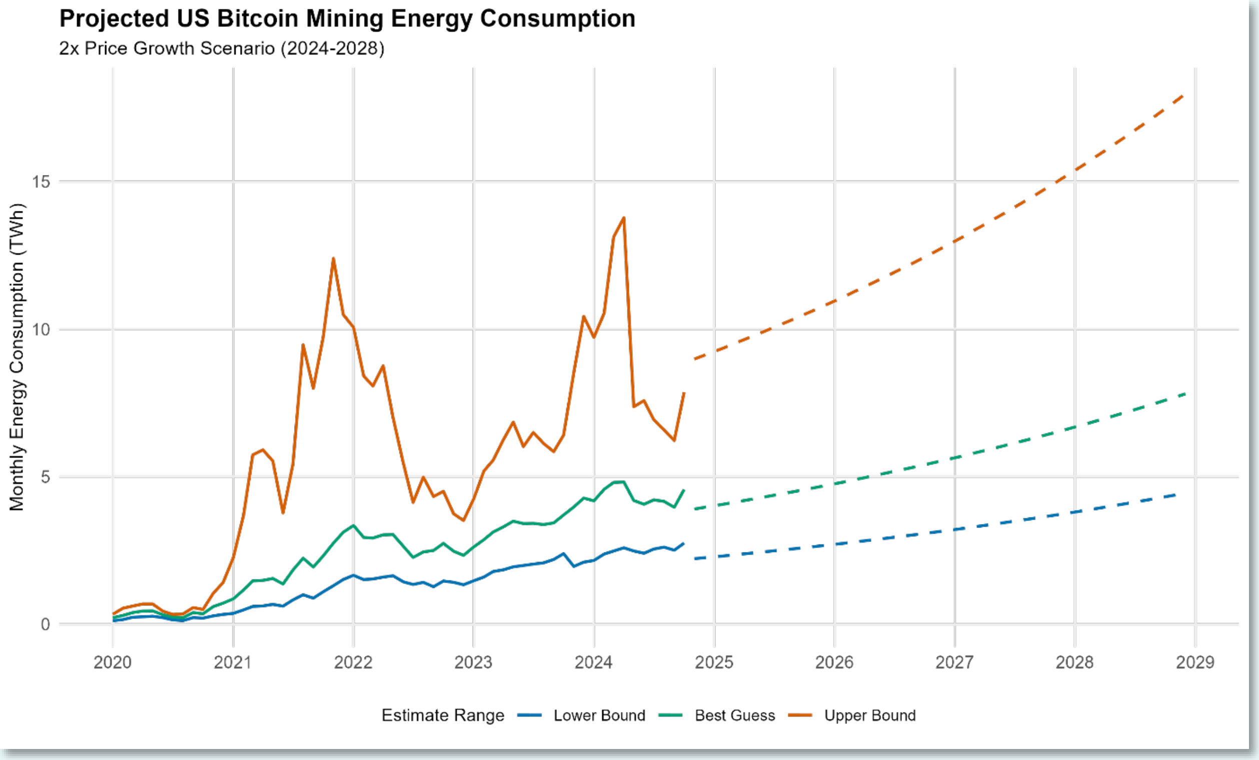


Figure 6.3. Projected U.S. Bitcoin mining energy consumption under a moderate price growth scenario (2024–2028).
Historical data (solid lines) shows actual energy consumption estimates from 2017–2024, while projected values (dotted lines) assume Bitcoin price increases five-fold by 2028. The three trajectories represent different hardware efficiency scenarios: lower bound (most efficient hardware, optimal PUE), median (mixed hardware efficiency, average PUE), and upper bound (profitability threshold hardware, higher PUE).

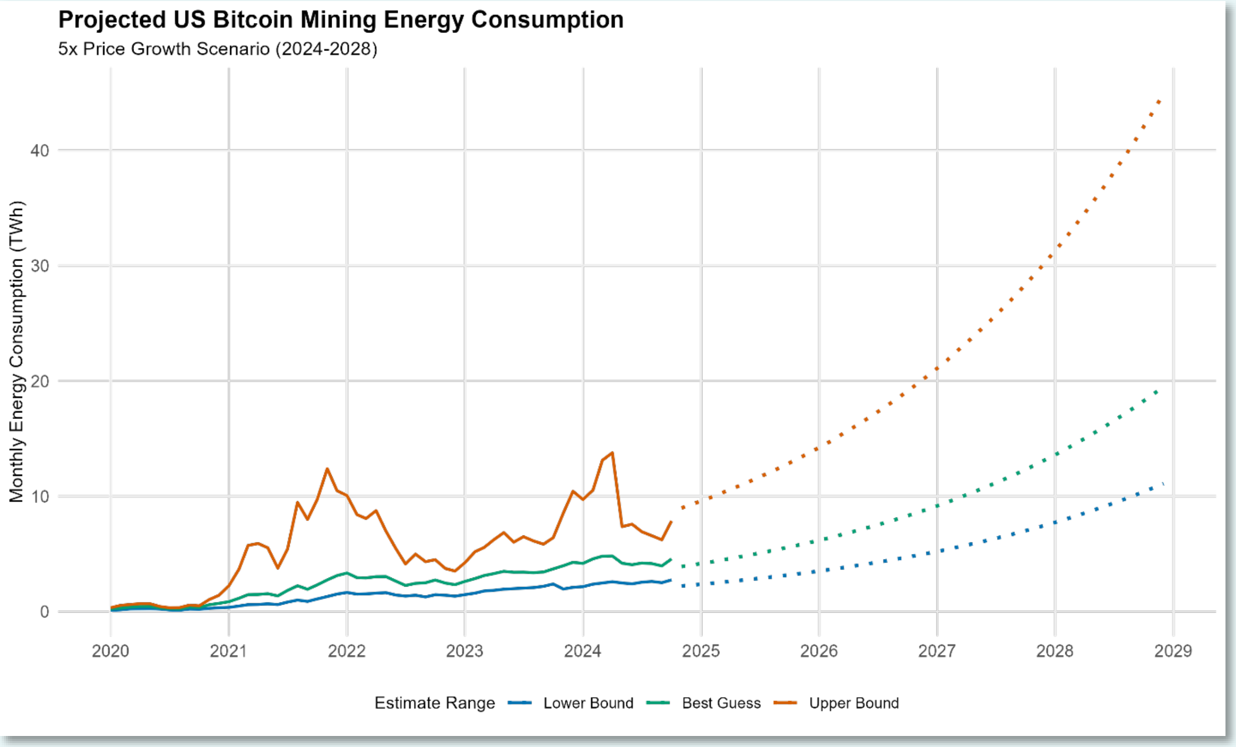


Figure 6.4. Projected U.S. Bitcoin mining energy consumption under an aggressive price growth scenario (2024–2028).
Historical data (solid lines) shows actual energy consumption estimates from 2017–2024, while projected values (dotted lines) assume Bitcoin price increases five-fold by 2028. The three trajectories represent different hardware efficiency scenarios: lower bound (most efficient hardware, optimal PUE), median (mixed hardware efficiency, average PUE), and upper bound (profitability threshold hardware, higher PUE).

As an input into our overall model of data center energy use, we selected the CBECI best-guess based estimate. Without more detailed data on the underlying hardware composition, we view this approach as the most robust estimate of mining hardware efficiency. That the International Energy Agency's bottom-up estimate for 2023 was much closer to the upper bound suggests that this may, in fact, be a conservative estimate hardware composition (International Energy Agency 2024). Future data collection efforts to better characterize the actual mining hardware in operation would help reduce this uncertainty.

The insights from this work highlight the urgent need for enhanced monitoring and reporting frameworks in the cryptocurrency mining sector. While our methodology provides a reliable estimation approach, the industry's opacity continues to hamper precise forecasting and planning. As digital assets become increasingly mainstream, policymakers and utilities must develop more sophisticated tools for tracking and managing mining-related energy demand to ensure grid stability and meet decarbonization goals.

References

Helmy, K., Nuzzi, L., Mead, A., and Waters, K.. 2023. The Signal & the Nonce. Coin Market.

International Energy Agency. 2024. Electricity 2024. Paris: International Energy Agency. https://www.iea.org/reports/electricity-2024.

Koomey, J.G., and Masanet, E. 2021. Does Not Compute: Avoiding Pitfalls Assessing the Internet's Energy and Carbon Impacts." Joule 5 (7): 1625–28. https://doi.org/10.1016/j.joule.2021.05.007.

Lei, N., Masanet, E., and Koomey, J.G. 2021. Best Practices for Analyzing the Direct Energy Use of Blockchain Technology Systems: Review and Policy Recommendations. Energy Policy 156 (September):112422. https://doi.org/10.1016/j.enpol.2021.112422.

McDonald, K. 2022. Ethereum Emissions: A Bottom-up Estimate. arXiv. https://doi.org/10.48550/arXiv.2112.01238.

Messina, I. 2023. Bitcoin Electricity Consumption: An Improved Assessment - News & Insight. Cambridge Judge Business School. August 31, 2023. https://www.jbs.cam.ac.uk/2023/bitcoin-electricity-consumption/.

Morey, M., McGrath, G., and Minato, H. 2024. Tracking Electricity Consumption from U.S. Cryptocurrency Mining Operations. https://www.eia.gov/todayinenergy/detail.php?id=61364.

OSTP. 2022. Climate and Energy Implications of Crypto-Assets in the United States. Washington, D.C.: White House Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2022/09/09-2022-Crypto-Assets-and-Climate-Report.pdf#page=2.48.

Papp, A., Almond, D., and Zhang, S. 2023. Bitcoin and Carbon Dioxide Emissions: Evidence from Daily Production Decisions. Journal of Public Economics 227 (November):105003. https://doi.org/10.1016/j.jpubeco.2023.105003.

Siddik, M.A.B., Amaya, M., and Marston, L.T. 2023. The Water and Carbon Footprint of Cryptocurrencies and Conventional Currencies. Journal of Cleaner Production 411 (July):137268. https://doi.org/10.1016/j.jclepro.2023.137268.

de Vries, A., Gallersdörfer, U., Klaaßen, L., and Stoll, C. 2022. Revisiting Bitcoin's Carbon Footprint. Joule 6 (3): 498–502. https://doi.org/10.1016/j.joule.2022.02.005.

7. Conclusions, Limitations, and Future Work

The 2024 updates to the Berkeley Lab data center energy use model described in this report support previous analyses showing near-constant U.S. data center energy use in the early to mid-2010s, despite substantial growth in data center services. The efficiency strategies that allowed the industry to avoid increased energy needs during this period included improved cooling and power delivery efficiency, increased server utilization rates, increased computational efficiencies, and reduced server idle power, much of which were enabled by a shift toward larger, cloud-based, hyperscale data centers. This report shows that the emergence of accelerated servers became a significant enough portion of the data center server stock to begin increasing total data center energy use again by 2017, driving a new trend that led to an approximate tripling of energy use from 2014 to 2023. Future energy demand scenarios indicate further growth based on a range of possible equipment shipments and operational practices of accelerated computing to support Artificial Intelligence (AI) services.

These insights are based on a “bottom-up” energy use model that requires inputs and assumptions developed from limited publicly available data, proprietary market analyst data, and review by industry representatives and stakeholders. The lack of direct energy data available in a sector with rapidly evolving technologies limits the analysis in this report, especially when trying to understand and estimate future energy demand scenarios. Some key limitations and opportunities for future research to address them are outlined below.

Benchmarking Initiatives

The Berkeley Lab data center energy use model is based on representative equipment, equipment configurations, and operational practices that may change over time and can increase the underlying uncertainty in the modeling results. Providing an empirically sound characterization of U.S. data center energy and water use and identifying opportunities for data center optimization require ongoing data collection and reporting on facility size, energy and water consumption, power source, cooling system types, backup power system information, and installed IT equipment characteristics, along with other available characteristics. This information can be used to provide broader insights into industry trends, estimates of energy/water/carbon impact, improved or new efficiency metrics, and updates on efficiency strategies. This information can also facilitate greater data and knowledge sharing, which can improve data center modeling by the broader energy analysis community. Approaches include the following:

<u>More frequent data center energy use reports.</u> This 2024 Data Center Energy Use Report is the third such report released over nearly 20 years. Providing reports on an annual or biannual basis would allow energy modelers to more frequently update inputs and assumptions, continuously improve its modeling structure and capabilities, and work with stakeholders to validate those inputs, ultimately providing increased accurately and flexibility to match changes in the data center sector. In short, the sector and its energy demands are quickly evolving, and a tighter feedback loop is needed to support timely market insights and forecasts so

policymakers and market actors can anticipate impacts and make balanced, informed decisions.

Energy/resource data repository. The lack of primary performance and utilization data indicates that much greater transparency is needed around data centers. Very few companies report actual data center electricity use and virtually none report it in context of IT characteristics such as compute capacities, average system configurations, and workload types. These details are often considered proprietary, but novel data sharing arrangements could address these concerns by developing a repository for companies to provide energy use data that would be anonymized and aggregated for public release through coordination with entities that collect and anonymize data for other industries.

Metric development. Comparing and benchmarking the efficiency of data centers is hampered by the lack of sufficient metrics. PUE only measures the efficiency of the infrastructure supporting a data center and indicates nothing about the efficiency of the IT equipment itself. Other metrics, such as ITUE, are meant to capture IT equipment efficiency but have seen little uptake and may need expansion to accommodate accelerated hardware. Computation efficiency alone provides little insight into energy efficiency opportunities without understanding how those computations are applied to different workloads. While the need for new metrics is a known challenge, data center operators, standards organizations, and researchers should continue to work together to develop a set of energy and water performance metrics specific to different workloads or service segments to enable more accurate comparisons, benchmark setting, and targets.

Data center equipment testbed. Direct IT equipment measurements by working with universities, national labs, as well as industry partners would allow for the collection of operational characteristics based on physical attributes (e.g., server cores, DRAM, network ports, storage, etc.) rather than relying on overall power averages from the data vendors. Measured data could then be linked to feed simulation models of energy that could be applied to a wider combination of physical characteristics and provide a stronger relationship between rated power and operational power, idle power, and power variation at different utilization levels.

Utility/ISO Collaborations

In this report, the data center carbon and water impacts associated with different electricity sources are based on regional electricity sources of the utility grid. While U.S. data centers are increasingly engaging in onsite electricity generation, power purchase agreements, and trade in various carbon credits, the lack of transparency around the details of these activities for all U.S. data center operators prevents including in the overall analysis. Along with limiting the scope of this report, this lack of transparency highlights that data center growth is occurring with little consideration for how best to integrate these emergent loads with the expansion of electricity generation/transmission or for broader community development. Research and technical support at the intersection between utilities/ISOs and data center owners/operators could guide where and how data centers are sited, how renewable contracts are developed, and coordinate efforts at improving efficiency and load flexibility. Approaches include the following:

Balancing risk and cost allocation needs through smart contracts. Acceleration of data center deployments is likely to require substantial investment in new or expanded power infrastructure. The magnitude of that investment and how much is on the customer vs. grid side of the meter is highly uncertain. If investments are made on the grid side but the expected load fails to show up, ratepayers could be unduly burdened by cost recovery. Some utilities are requesting regulators to approve rate structures that transfer all this risk to data centers, with the potential consequence of slowing down deployment. Research from energy economists, data center modelers, grid modelers, as well as others is needed to identify key risks for existing customers, data centers, and utilities, explore existing contractual arrangements, and propose novel methods for risk-sharing and cost recovery

Demand bidding as a coordination mechanism. Wholesale markets have developed methods to incentivize long term deployment of supply side resources for adequacy. Research could explore “demand bidding” mechanisms where large loads would bid their future demand needs, becoming part of a demand-side interconnection queue.

Support co-investment in new firm clean power. The data center industry has shown interest and leadership in implementing real-time renewable energy and zero carbon power, including battery storage resources. Future research efforts should include working with utilities and data center companies to develop strategies that can quantify the potential costs and benefits of investing in large-scale, customer-driven renewable energy projects and new firm clean power, including scalable nuclear generation.

Quantifying the elasticity of cloud and AI services. Anecdotal evidence suggests recent increases in the cost of cloud services, including fees for services such as virtualization, are causing some customers to rethink their cloud migration plans and even consider repatriating workloads to legacy data centers. The power demands for AI and cloud services are tied to the cost and perceived value of those services, but this relationship is not well understood. Research is needed to understand the price sensitivity of these services as a barometer for the magnitude and timing of related power demands.

Technology Development

In previous reports, the Berkeley Lab data center energy use model has been used to generate various efficiency scenarios for traditional IT equipment and data center infrastructure. In the current report, the use of accelerated IT equipment for AI applications is at a nascent stage where industry practices are still developing, and a broader array of efficiency strategies need to be identified. At the same time, advanced cooling technologies and strategies are being developed and deployed to provide much higher cooling capacities compared to the past. Since the emergence of accelerated IT equipment for AI applications is driving future energy growth, the future energy use in this report is presented as a range based on possible changes in operational practices and the rate of growth in the installed base for accelerated servers, as well as the cooling technologies that may be deployed for heat removal. This lack of AI-specific efficiency scenarios in the current report highlights that AI is changing data center requirements in ways that call for new research and development initiatives to unlock the next generation of efficiency measures. These strategies could include efficient software algorithms, application specific chips, new cooling system designs, and network configurations. Approaches include the following:

Demand flexibility capability studies. Research is needed to identify opportunities for data centers to exercise load flexibility to limit their impacts on the grid and lower their carbon footprint. Various data center computing workloads could be identified and parameterized by their service requirements, and IT equipment utilization levels could be better understood. Data center energy models must be expanded to include the spatial and temporal resolution needed to assess load flexibility. Additionally, the costs and benefits of load flexibility to data center operators should be understood to identify incentive structures for realizing flexibility potential.

Monitoring protocol and validation of the data center tools. Collection of detailed performance and operational data from well-instrumented data centers could include advanced metering infrastructure at the hardware and rack levels and support the expansion and validation of building energy and system modeling software to include the physics and dynamics of emerging technologies, such as advanced liquid cooling designs and new computing hardware.

Efficient software and algorithms. Future efficiency scenarios should include broad classes of energy efficient algorithms (Patterson et al. 2022; Leiserson et al. 2020), including communication avoiding algorithms, reduced precision and mixed-precision algorithms, and randomized algorithms such as Hardamard matrix methods.

Emerging technologies. Innovation occurs constantly in the IT sector but can be hard to capture in future scenarios without credible data on the technical, energy, and economic performance of emerging technologies. Moving forward, advances in processor designs, new materials, system integration approaches, and other data center aspects could provide substantial efficiency benefits. Data standards and collection initiatives that can help provide public datasets of key characteristics necessary for confident modeling (e.g., adoption readiness levels, technology characteristics, energy requirements, and costs) would help energy modelers consider emerging technologies in future scenarios.

Beyond 2028

While significant energy efficiency improvements in data center design and operation have occurred over the past decade, the expansion of data center services into areas that require new types of hardware, such as AI and cryptocurrency, has ended the era of generally flat data center energy use. Most notably the recent rapid growth in accelerated servers has caused current total data center energy demand to more than double between 2017 and 2023, and continued growth in the use of accelerated servers for AI services could cause further substantial increases by the end of this decade. This surge in energy demand highlights the need for future research to understand this rapidly changing industry and identify new efficiency strategies to minimize the environmental impacts of this growing and increasingly significant portion of our overall economy.

Furthermore, when looking beyond 2028, the current surge in data center electricity demand should be put in the context of the much larger electricity demand expected over the next few decades from a combination of electric vehicle adoption, onshoring of manufacturing, hydrogen utilization, and the electrification of industry and buildings. Research initiatives are needed not just to identify strategies to meet data centers’ future energy needs but also to help stakeholders use this relatively near-term electricity demand for data centers as an opportunity to develop the leadership and a foundation for an economy-wide electricity infrastructure expansion.

Citations and Bibliography

Alben, J. 2024. Computing in the Era of Generative AI. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 67, 26–28. https://doi.org/10.1109/ISSCC49657.2024.10454562.

Aljbour, J., Wilson, T., and Patel, P. 2024. Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption. Electric Power Research Institute, May 2024. https://www.epri.com/research/products/000000003002028905.

Allsup, Maeve. 2024. Microsoft Says Georgia May Be Overestimating Data Center Load Growth. Latitude Media, April 12, 2024. https://www.latitudemedia.com/news/microsoft-says-georgia-may-be-overestimating-data-center-load-growth.

Anderson, J. and Sweeney, D. 2023. Power of AI: Wild predictions of power demand from AI put industry on edge. S&P Global Commodity Insights, October 16, 2023. https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/electric-power/101623-power-of-ai-wild-predictions-of-power-demand-from-ai-put-industry-on-edge.

Andrae A. and Edler T. 2015. On global electricity usage of communication technology: trends to 2030. Challenges, 6(1):117–157. https://doi.org/10.3390/challe6010117.

Andrae A. 2017. Total consumer power consumption forecast. Presented at the Nordic Digital Business Summit.

Aschenbrenner, L. 2024. Situational Awareness: The Decade Ahead. https://situational-awareness.ai/.

Avelar, V., Donovan, P., Torell, W., and Arango, M. 2023. The AI Disruption: Challenges and Guidance for Data Center Design. Schneider Electric.

Belkhir, L. and Elmeligi. A. 2018. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. Journal of Cleaner Production. 177:448-463. https://doi.org/10.1016/j.jclepro.2017.12.239.

Ben-Nun, T. and Hoefler, T. 2019. Demystifying Parallel and Distributed Deep Learning: An In-depth Concurrency Analysis. ACM Computing Surveys, 52(4), 65:1-65:43. https://doi.org/10.1145/3320060.

Berthelot, A., Jay, M., Lefevre, L., and Caron, E. 2023. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. https://inria.hal.science/hal-04346102.

Bloomberg. 2024. AI Is Wreaking Havoc on Global Power Systems. Bloomberg.com, June 21, 2024. https://www.bloomberg.com/graphics/2024-ai-data-centers-power-grids/.

Bouza, L., Bugeau, A., and Lannelongue, L. 2023. How to estimate carbon footprint when training deep learning models? A guide and review. Environmental Research Communications, 5(11), 115014. https://doi.org/10.1088/2515-7620/acf81b.

Brown, Jeff. 2024. Data Center Electricity Consumption Is about to Quadruple. Jeff Brown. June 4, 2024. https://www.brownridge.com/data-center-electricity-consumption-is-about-to-quadruple/.

Brown, R.E., Masanet, E., Nordman, B., Tschudi, B., Shehabi, A., Stanley, J., Koomey, J., Sartor, D., and Chan, P. 2007. Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431. LBNL-363E, 929723. Lawrence Berkeley National Lab. http://www.osti.gov/servlets/purl/929723-4d6s1A/.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. 2020. Language Models are Few-Shot Learners (arXiv:2005.14165). arXiv.

https://doi.org/10.48550/arXiv.2005.14165.

Buchanan, B. 2020. The AI Triad and What It Means for National Security Strategy. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/.

Capuccio, D, and L Craver. 2007. “The Data Center Power and Cooling Challenge.” The Gartner Group.

Carvallo, J. P., Larsen, P. H., Sanstad, A. H., and Goldman, C. A. 2018. Long term load forecasting accuracy in electric utility integrated resource planning. Energy Policy. https://doi.org/10.1016/j.enpol.2018.04.060.

Caspart, R., Ziegler, S., Weyrauch, A., Obermaier, H., Raffeiner, S., Schuhmacher, L. P., Scholtyssek, J., Trofimova, D., Nolden, M., Reinartz, I., Isensee, F., Götz, M., and Debus, C. 2022. Precise Energy Consumption Measurements of Heterogeneous Artificial Intelligence Workloads. In H. Anzt, A. Bienz, P. Luszczek, & M. Baboulin (Eds.), High Performance Computing. ISC High Performance 2022 International Workshops (pp. 108–121). Springer International Publishing. https://doi.org/10.1007/978-3-031-23220-6_8.

Chien, A. A., Lin, L., Nguyen, H., Rao, V., Sharma, T., and Wijayawardana, R. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). Proceedings of the 2nd Workshop on Sustainable Computer Systems, 1–7. https://doi.org/10.1145/3604930.3605705.

Clark, D. 2024. NVIDIA, Powered by A.I. Boom, Reports Soaring Revenue and Profits. The New York Times, May 22, 2024. https://www.nytimes.com/2024/05/22/technology/nvidia-quarterly-earnings-results.html.

de Vries, A. 2023. The growing energy footprint of artificial intelligence. Joule. https://doi.org/10.1016/j.joule.2023.09.004.

del Rey, S., Martínez-Fernández, S., Cruz, L., and Franch, X. 2023. Do DL models and training environments have an impact on energy consumption? 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 150–158. https://doi.org/10.1109/SEAA60479.2023.00031.

Deng, L. 2018. Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook. IEEE Signal Processing Magazine, 35(1), 180–177. https://doi.org/10.1109/MSP.2017.2762725.

Desislavov, R., Martínez-Plumed, F., and Hernández-Orallo, J. 2023. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems, 38, 100857. https://doi.org/10.1016/j.suscom.2023.100857.

Epstein, D. 2019. Range: Why Generalists Triumph in a Specialized World (Illustrated edition). Riverhead Books.

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., and Jiang, L. 2023. LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models (arXiv:2309.14393). arXiv. https://doi.org/10.48550/arXiv.2309.14393.

Farrell, S., Emani, M., Balma, J., Drescher, L., Drozd, A., Fink, A., Fox, G., Kanter, D., Kurth, T., Mattson, P., Mu, D., Ruhela, A., Sato, K., Shirahata, K., Tabaru, T., Tsaris, A., Balewski, J., Cumming, B., Danjo, T., … Yin, J. 2021. MLPerfTM HPC: A Holistic Benchmark Suite for Scientific Machine Learning on HPC Systems. 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC), 33–45. https://doi.org/10.1109/MLHPC54614.2021.00009.

Goldman Sachs Research. 2024. AI Is Poised to Drive 160% Increase in Data Center Power Demand. May 14, 2024. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand.

Google. 2018. Moving toward 24x7 Carbon-Free Energy at Google Data Centers: Progress and Insights. October.

https://storage.googleapis.com/gweb-sustainability.appspot.com/pdf/24x7-carbon-free-energy-data-centers.pdf.

Govind, A., Bhalachandra, S., Zhao, Z., Rrapaj, E., Austin, B., and Nam, H. A. 2023. Comparing Power Signatures of HPC Workloads: Machine Learning vs Simulation. Proceedings of the SC ’23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis, 1890–1893. https://doi.org/10.1145/3624062.3624274.

Haenlein, M., and Kaplan, A. 2019. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. California Management Review, 61(4), 5–14. https://doi.org/10.1177/0008125619864925.

Haken, M., Clow, J., Li, Y., Wei, B., Chong, D., Thoon, KY, Vemuri, K., Westhauser, T., Shetty, P., Chan, A., Wu, H., Mehta, D.S., and Liao, S. 2021. Open Compute Project: Yosemite V3: Facebook Multi-Node Server Platform Design Specification, 1v16. https://www.opencompute.org/documents/ocp-yosemite-v3-platform-design-specification-1v16-pdf.

Halper, E. 2024. Amid explosive demand, America is running out of power. Washington Post, March 7, 2024. https://www.washingtonpost.com/business/2024/03/07/ai-data-centers-power/.

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., Fawzy, M., Jia, B., Jia, Y., Kalro, A., Law, J., Lee, K., Lu, J., Noordhuis, P., Smelyanskiy, M., Xiong, L., and Wang, X. 2018. Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), 620–629. https://doi.org/10.1109/HPCA.2018.00059.

Hoefler, T., Hendel, A., and Roweth, D. 2022. The Convergence of Hyperscale Data Center and High-Performance Computing Networks. Computer, 55(7), 29–37. https://doi.org/10.1109/MC.2022.3158437.

Hu, K. 2023. ChatGPT sets record for fastest-growing user base—Analyst note. Reuters, February 2, 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Huerta, E. A., Khan, A., Davis, E., Bushell, C., Gropp, W. D., Katz, D. S., Kindratenko, V., Koric, S., Kramer, W. T. C., McGinty, B., McHenry, K., and Saxton, A. 2020. Convergence of artificial intelligence and high performance computing on NSF-supported cyberinfrastructure. Journal of Big Data, 7(1). Scopus. https://doi.org/10.1186/s40537-020-00361-2.

International Data Corporation (IDC). 2024a. IDC Worldwide Quarterly Server Tracker Q22024, Framingham, MA, September.

International Data Corporation (IDC). 2024b. IDC Data Center Semiconductor Consumption CPU-GPU-DPU-Networking 11122024, Framingham, MA, November.

International Data Corporation (IDC). 2023a. IDC Processors and AI Unit Spreadsheet 08282023, Framingham, MA, August.

International Data Corporation (IDC). 2023b. Datacenter Deployment and Spend Forecast- 1H 2023, Framingham, MA, August.

International Energy Agency (IEA). 2017. “Digitalisation and Energy.” https://www.iea.org/reports/digitalisation-and-energy.

International Energy Agency (IEA). 2024a. *Electricity 2024*. Paris: International Energy Agency. https://www.iea.org/reports/electricity-2024.

International Energy Agency (IEA). 2024b. “World Energy Outlook 2024.” World Energy Outlook. Paris: International Energy Agency. https://www.iea.org/reports/world-energy-outlook-2024.

Jones, N. 2018. How to Stop Data Centres From Gobbling Up the World’s Electricity. Nature, 561(7722), 163–166. https://doi.org/10.1038/d41586-018-06610-y.

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., … Yoon, D. H. 2017. In-Datacenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture, 1–12. https://doi.org/10.1145/3079856.3080246.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. 2020. Scaling Laws for Neural Language Models (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361.

Kearney, L. 2024. Data centers, EVs to significantly boost US power load by 2030, consultancy says. Reuters, June 25, 2024. https://www.reuters.com/business/energy/data-centers-evs-significantly-boost-us-power-load-by-2030-consultancy-says-2024-06-25/.

Kelechi, A. H., Alsharif, M. H., Bameyi, O. J., Ezra, P. J., Joseph, I. K., Atayero, A.-A., Geem, Z. W., and Hong, J. 2020. Artificial Intelligence: An Energy Efficiency Tool for Enhanced High performance computing. Symmetry, 12(6), Article 6. https://doi.org/10.3390/sym12061029.

Khan, S. M. and Mann, A. 2020. AI Chips: What They Are and Why They Matter (pp. 22–23). Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/.

Koomey, Jonathan. 2007. Estimating Total Power Consumption by Servers in the US and the World. https://www.researchgate.net/publication/228365136_Estimating_Total_Power_Consumption_by_Servers_in_the_US_and_the_World.

Koomey, Jonathan G. 2008. “Worldwide Electricity Used in Data Centers.” Environmental Research Letters 3 (3): 034008. https://doi.org/10.1088/1748-9326/3/3/034008.

Koomey J. 2011. Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times, 9:161.

Koomey, Jonathan. 2015. A primer on the energy efficiency of computing. In Physics of Sustainable Energy III: Using Energy Efficiently and Producing it Renewably (Proceedings from a Conference Held March 8-9, 2014 in Berkeley, CA). Edited by R. H. Knapp Jr., B. G. Levi and D. M. Kammen. Melville, NY: American Institute of Physics (AIP Proceedings). pp. 82-89. https://www.mediafire.com/file/t63app03o9uca14/koomeyprimeroncomputingefficiencyforAPS2015-final.pdf.

Koomey, Jonathan G., and Eric Masanet. 2021. Does Not Compute: Avoiding Pitfalls Assessing the Internet’s Energy and Carbon Impacts. Joule 5 (7): 1625–28. https://doi.org/10.1016/j.joule.2021.05.007.

Koomey, Jonathan, and Karl Freund. 2024. Digital twins for digital infrastructure: The key to optimizing data center operations. Koomey Analytics and Cambrian AI Research. September. https://www.mediafire.com/file_premium/gs0unfjvcbb1ii3/Digital_twins_for_digital_infrastructure-FINAL.pdf.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. 2019. Quantifying the Carbon Emissions of Machine Learning (arXiv:1910.09700). arXiv. https://doi.org/10.48550/arXiv.1910.09700.

Latif, I., Newkirk, A.C., Carbone, M.R., Munir, A., Lin, Y., Koomey, J., Yu, X., and Dong, Z. 2024. Empirical Measurements of AI Training Power Demand on a GPU-Accelerated Node. arXiv. https://doi.org/10.48550/arXiv.2412.08602.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. Nature, 521(7553), Article 7553. https://doi.org/10.1038/nature14539.

Lee, V. 2023. The Impact of GenAI on Electricity: How GenAI is Fueling the Data Center Boom in the U.S. September 13, 2023. https://www.linkedin.com/pulse/impact-genai-electricity-how-fueling-data-center-boom-vivian-lee/.

Lei, N. and Masanet, E.R. 2021. Global data center energy demand and strategies to conserve energy. Data Center Handbook, H. Geng (Ed.). https://doi.org/10.1002/9781119597537.ch2.

Lei, N., Masanet, E., and Koomey, J.G.. 2021. Best Practices for Analyzing the Direct Energy Use of Blockchain Technology Systems: Review and Policy Recommendations. Energy Policy 156 (September):112422. https://doi.org/10.1016/j.enpol.2021.112422.

Leiserson, C.E., Thompson, N.C., Emer, J.S., Kuszmaul, B.C., Lampson, B.W., Sanchez, D., and Schardl, T.B. 2020. There’s plenty of room at the top: What will drive computer performance after Moore’s law? Science. vol. 368, no. 6495. https://doi.org/10.1126/science.aam9744.

Lin, W., Adetomi, A., and Arslan, T. 2021. Low-Power Ultra-Small Edge AI Accelerators for Image Recognition with Convolution Neural Networks: Analysis and Future Directions. Electronics, 10(17), Article 17. https://doi.org/10.3390/electronics10172048.

Luccioni, A. S., Jernite, Y., and Strubell, E. 2024. Power Hungry Processing: Watts Driving the Cost of AI Deployment? Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 85–99. https://doi.org/10.1145/3630106.3658542.

Luccioni, A. S., Viguier, S., and Ligozat, A.-L. 2023. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. Journal of Machine Learning Research, 24(253), 1–15. https://doi.org/10.48550/arXiv.2211.02001

Ma, R., Georganas, E., Heinecke, A., Gribok, S., Boutros, A., and Nurvitadhi, E. 2022. FPGA-Based AI Smart NICs for Scalable Distributed AI Training Systems. IEEE Computer Architecture Letters, 21(2), 49–52. https://doi.org/10.1109/LCA.2022.3189207.

Masanet, E.R., Brown, R.E., Shehabi, A., Koomey, J.G., and Nordman, B. 2011. Estimating the energy use and efficiency potential of US data centers. Proc IEEE 2011;99(8):1440–1453.

Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. 2020. Recalibrating global data center energy-use estimates. Science, 367(6481), 984–986. https://doi.org/10.1126/science.aba3758.

Malmodin, J., Moberg, A., Lunden, D., Finnveden G., and Lovehagen, N. 2010. Greenhouse Gas Emissions and Operational Electricity Use in the ICT and Entertainment & Media Sectors. Journal of Industrial Ecology, 14:770-790. https://doi.org/10.1111/j.1530-9290.2010.00278.x.

Malmodin, J., Lundén, D., Moberg, Å., Andersson, G., and Nilsson, M. 2014. Life Cycle Assessment of ICT. Journal of Industrial Ecology, 18: 829-845. https://doi.org/10.1111/jiec.12145.

Malmodin, Jens, and Dag Lundén. 2016. The Energy and Carbon Footprint of the ICT and E&M Sector in Sweden 1990-2015 and Beyond. Proceedings of ICT for Sustainability 2016, 209–218. Atlantis Press. https://doi.org/10.2991/ict4s-16.2016.25.

Malmodin, J., Lövehagen, N., Bergmark, P., and Lundén, D. 2024. ICT sector electricity consumption and greenhouse gas emissions – 2020 outcome. Telecommunications Policy, 48(3), 102701. https://doi.org/10.1016/j.telpol.2023.102701.

McDonald, J., Li, B., Frey, N., Tiwari, D., Gadepally, V., and Samsi, S. 2022. Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models. Findings of the Association for Computational Linguistics: NAACL 2022, 1962–1970. https://doi.org/10.18653/v1/2022.findings-naacl.151.

Miller, C. 2022. Chip War: The Fight for the World's Most Critical Technology. Scribner.

Monroe, J. and Johns, B. 2024. The Sustainable Preservation of Enterprise Data. Further Market Research. https://www.bradjohnsconsulting.com/_files/ugd/8b8555_78e13c6f9b454179af76877d67fbb9db.pdf.

Montgomerie-Corcoran, A., Venieris, S. I., and Bouganis, C.-S. 2019. Power-Aware FPGA Mapping of Convolutional Neural Networks. 2019 International Conference on Field Programmable Technology (ICFPT), 327–330. https://doi.org/10.1109/ICFPT47387.2019.00059.

Moorhead, P. 2023. Google Provides More Details On Its Cloud Generative AI Play At Next Event. Forbes, September 12, 2023. https://www.forbes.com/sites/patrickmoorhead/2023/09/12/google-provides-more-details-on-its-cloud-generative-ai-play-at-next-event/.

Moss, S. 2023. Meta details AI data center redesign that led to facilities being scrapped. May 18, 2023. https://www.datacenterdynamics.com/en/analysis/meta-details-ai-data-center-redesign-that-led-to-facilities-being-scrapped/.

Mytton, D. and Ashtine, M. 2022. Sources of data center energy estimates: A comprehensive review. Joule. 6(9): 2032-2056. https://doi.org/10.1016/j.joule.2022.07.011.

Nelson, D. 2022. Defining The Digital Infrastructure Industry | InterGlobix Magazine. February 10, 2022. https://www.interglobixmagazine.com/defining-the-digital-infrastructure-industry/.

Newkirk, A.C. 2024. Evaluating the Node Level Power Draw of Production AI Workloads on Commercial Hardware. Presented at the 2024 Open Compute Project Global Summit, San Jose California, October 16. https://www.opencompute.org/events/past-events/2024-ocp-global-summit.

NSW. 2023. AI datacenter industry model v2.0 – less, but more expensive chips, with more memory content. NewStreet Research, March 27.

NVIDIA. n.d.-a. NVIDIA. GB200 NVL72: Powering the new era of computing. NVIDIA. https://www.nvidia.com/en-us/data-center/gb200-nvl72/. Accessed December 17, 2024.

NVIDIA. n.d.-b. Introduction to DGX A100. NVIDIA Documentation. https://docs.nvidia.com/dgx/dgxa100-user-guide/introduction-to-dgxa100.html Accessed: December 18 2024.

NVIDIA. n.d.-c. Introduction to DGX H100. NVIDIA Documentation. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html Accessed: December 18, 2024.

Omdia, 2024. Omdia Socket and Co-processor estimates for 2022-2027. Informa TechTarget, Inc.

Patel, D. and Nishball, D. 2024. NVIDIA Blackwell Perf TCO Analysis—B100 vs B200 vs GB200NVL72. April 10, 2024. https://www.semianalysis.com/p/nvidia-blackwell-perf-tco-analysis.

Patel, D., Nishball, D., and Ontiveros, J. E. 2024. AI Datacenter Energy Dilemma—Race for AI Datacenter Space. March 13, 2024. https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race.

Patel, P., Choukse, E., Zhang, C., Goiri, Í., Warrier, B., Mahalingam, N., and Bianchini, R. 2024. Characterizing Power Management Opportunities for LLMs in the Cloud. Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, 3, 207–222. https://doi.org/10.1145/3620666.3651329.

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink (arXiv:2204.05149). arXiv. https://doi.org/10.48550/arXiv.2204.05149.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. 2021. Carbon Emissions and Large Neural Network Training (arXiv:2104.10350). arXiv. https://doi.org/10.48550/arXiv.2104.10350.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. 2017. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. International Journal of Automation and Computing, 14(5), 503–519. https://doi.org/10.1007/s11633-017-1054-2.

Qasaimeh, M., Denolf, K., Lo, J., Vissers, K., Zambreno, J., and Jones, P. H. 2019. Comparing Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernels. 2019 IEEE International Conference on Embedded Software and Systems (ICESS), 1–8. https://doi.org/10.1109/ICESS.2019.8782524.

Ren, J. and Xu, L. 2015. On Vectorization of Deep Convolutional Neural Networks for Vision Tasks. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1), Article 1. https://doi.org/10.1609/aaai.v29i1.9488.

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. 2023 IEEE High Performance Extreme Computing Conference (HPEC), 1–9. https://doi.org/10.1109/HPEC58863.2023.10363447.

Santos, D. C. dos. 2022. Understanding the Energy Consumption of HPC Scale Artificial Intelligence (arXiv:2212.00582). arXiv. https://doi.org/10.48550/arXiv.2212.00582.

Shankar, S. and Reuther, A. 2022. Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications. 2022 IEEE High Performance Extreme Computing Conference (HPEC), 1–8. https://doi.org/10.1109/HPEC55821.2022.9926296.

Shehabi, A., Smith, S.J., Horner, N., Azevedo, I., Brown, R., Koomey, J., Masanet, E., Sartor, D., Herrlin, M., and Lintner, W. 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775.

Shehabi, A., Smith, S. J., Masanet, E., and Koomey, J. 2018. Data center growth in the United States: Decoupling the demand for services from electricity use. Environmental Research Letters, 13(12), 124030. https://doi.org/10.1088/1748-9326/aaec9c.

Siddik, M. A. B., Shehabi, A., and Marston, L. 2021. The environmental footprint of data centers in the United States. Environmental Research Letters, 16(6), 064017. https://doi.org/10.1088/1748-9326/abfba1.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), Article 7587. https://doi.org/10.1038/nature16961.

SPEC. n.d. SPEC's Benchmarks and Tools. Standard Performance Evaluation Corporation. https://open.spec.org/benchmarks.html. Accessed December 17, 2024.

Stanley, J.R., Brill, K.G., and Koomey, J. 2007. Four Metrics Define Data Center "Greenness" Enabling Users to Quantify "Energy Efficiency for Profit" Initiatives. Santa Fe, NM: The Uptime Institute. http://www.uptimeinstitute.org.

Strubell, E., Ganesh, A., and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP (arXiv:1906.02243). arXiv. https://doi.org/10.48550/arXiv.1906.02243.

Sutton, R. 2019. The bitter lesson. Incomplete Ideas.

S & P Global. 2024. 451 Research Datacenter Knowledgebase. S&P Global Market Intelligence. https://www.spglobal.com/market-intelligence/en/solutions/datacenter-knowledgebase.

TDC, 2024. The next datacenter decade, driven by acceleration & GenAI - ahead of the curve. TD Cownen Securities. May 24.

U.S. Congress. 2020. *Energy Act of 2020*. Public Law No: 116-260. Available at: https://www.congress.gov/bill/116th-congress/house-bill/133.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30.

Verdecchia, R., Sallou, J., and Cruz, L. 2023. A systematic review of Green AI. WIREs Data Mining and Knowledge Discovery, 13(4), e1507. https://doi.org/10.1002/widm.1507.

Weng, L. and Brockman, G. 2022. Techniques for training large neural networks. June. 9, 2022. https://openai.com/research/techniques-for-training-large-neural-networks.

Wilson, J. D. and Zimmerman, Z. 2023. The Era of Flat Power Demand is Over (p. 29). Grid Strategies.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., and Bai, C. 2022. Sustainable ai: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems, 4, 795–813.

Zhang, Y., Zhao, Y., Dai, S., Nie, B., Ma, H., Li, J., Miao, Q., Jin, Y., Tan, L., and Ding, Y. 2022. Cooling technologies for data centres and telecommunication base stations – A comprehensive review. Journal of Cleaner Production, 334, 130280. https://doi.org/10.1016/j.jclepro.2021.130280.