

Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA

In the format provided by the
authors and unedited

Contents

List of Sections

Section S1. AI Server Power Capacity Projections based on Manufacturing Process

Section S2. AI Server Spatial Distribution based on Large-scale Data Center Locations

Section S3. Estimating On-site Energy and Water Footprints of AI Servers: Models and Materials

Section S4. Grid Interactions and Renewable Energy Penetrations

Section S5. Methodology Comparison and Validation with Existing Studies

Section S6. Nomenclature

Supplementary References

List of Figures

Figure S1 The average CoWoS capacity of each year

Figure S2 The projected annual shipment units and average rated power of top-tier AI servers

Figure S3 The projected annual shipment and accumulative capacity of top-tier AI servers.

Figure S4 AI server shipments estimation validation

Figure S5 The spatial distributions of AI servers

Figure S6 The hydropower water consumption factor of each state

Figure S7 Grid carbon factor validation

Figure S7 Grid Water factor validation

List of Tables

Table S1 DGX system information data

Table S2 PUE & WUE model inputs

Table S3 ALC adoption rate data

Table S4 Comparison of our study with existing bottom-up methodologies

S1. AI Server Power Capacity Projections based on Manufacturing Process

This study employs a bottom-up approach that centers on the Chip-on-Wafer-on-Substrate (CoWoS) technology, which is a critical bottleneck in the top-tier AI server supply chain and is nearly entirely supplied by TSMC ^{1,2}. As discussed in the **AI Server Capacity Projections** section, we define Low, Moderate, High projection for CoWoS capacity growth, AI server rated power trends, and adoption patterns of new systems to form the five defined scenarios. The assumptions underlying these influencing factors are detailed below.

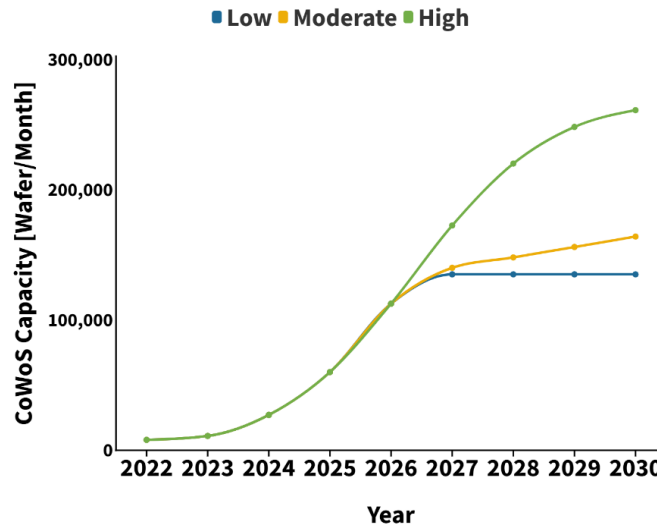


Figure S1 The average CoWoS capacity (wafer/month) of each year under the Low, Moderate, and High scenarios.

First, the CoWoS capacity growth assumptions are derived from historical data and the current planned expansion through 2026, as reported in recent analyses ³⁻⁶. According to these reports, CoWoS monthly capacity is projected to increase from 8,000 wafers to 14,000 wafers in 2023, further rising to 35,000 wafers by the end of 2024, 65,000 wafers by the end of 2025, and reaching 135,000 wafers by the end of 2026. Using this data and assuming a uniform quarterly growth rate within each year, we develop the three CoWoS projections depicted in Figure S1. The Low projection assumes that CoWoS capacity remains stagnant after 2026, representing a conservative outlook. The Moderate projection outlines a gradual increase in capacity post-2026, consistent with the rate of increase observed in 2023, reflecting constraints in manufacturing equipment availability and moderated AI server demand compared to the surge after 2023. The High projection envisions the continuation of the aggressive expansion trend beyond 2026, eventually tapering to the lower growth rates observed in earlier years by 2030. This definition reflects the

possibility of sustained demand driven by generative AI computing, effectively motivating the increase in CoWoS capacity.

Table S1 The GPU structure, rated power, GPU CoWoS size, CoWoS size, AI chip unit yield per wafer, and release year of each generation of the DGX systems. The * symbol represents the anticipated data.

GPU Structure	Rated Power	CoWoS Size	Units per Wafer	Release Year
Pascal	3.5	1.5×	37	2016
Volta	5	1.75×	32	2017
Ampere	6.5	2×	28	2020
Hooper	10.2	2×	28	2022
Blackwell	14.3	3.3×	16	2024
Rubin	N/A	4×	13*	2026*
N/A	N/A	5.5×	10*	2028*

The projected CoWoS capacity is translated into AI server shipment capacity by incorporating assumptions about rated power, chip CoWoS size, and adoption patterns of AI servers. It is important to note that only top-tier AI servers designed for large-scale AI computing tasks are produced using the CoWoS process. Low-end AI servers are estimated by using a constant value, as detailed later in this section. Given the complexity of AI server shipment mixes, this analysis focuses on the dominant elements driving capacity expansion, Nvidia’s DGX server systems. According to recent published reports, Nvidia holds over 98% of the market share for high-end AI servers, with its DGX systems (e.g., H100 and A100) representing the majority of sales ^{7,8}. As such, the evolution of DGX systems serves as a proxy for projecting the future profiles of top-tier AI servers manufactured with CoWoS technology. Table S1 summarizes the dominant DGX system chip generations, including their structure, rated power, CoWoS size, unit yield per wafer, and release years. The data are collected from official product specifications ^{9,10}, and industry reports ^{2,11}, forming the basis for projections. Specifically, information about the Rubin system (anticipated release in 2026) and its successor (anticipated release in 2028) is based on the projection plans of TSMC and current industry estimates ^{12,13}. The Rubin system is expected to feature a 4× CoWoS size, while its successor is projected to adopt a 5.5× CoWoS size.

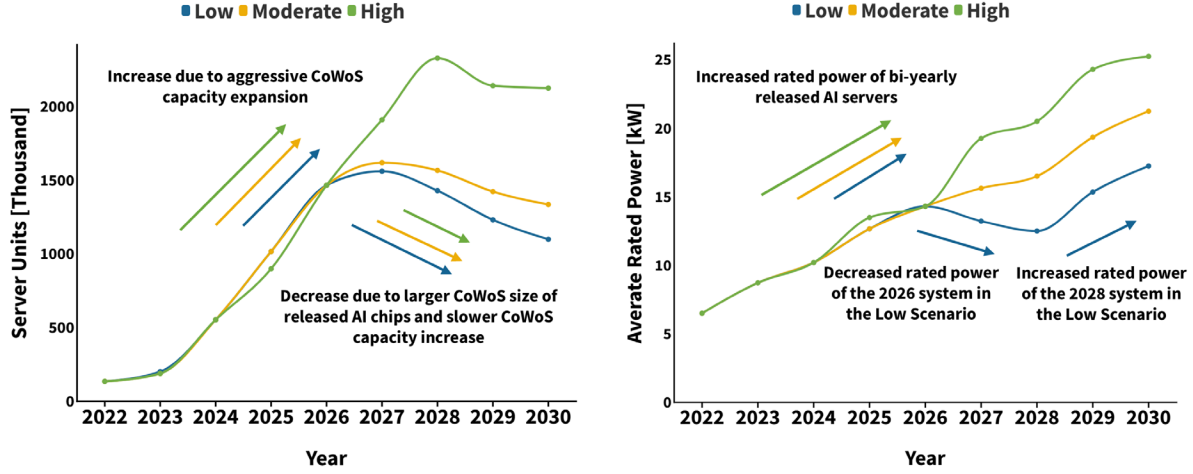


Figure S2 The projected annual shipment units (left) and average rated power (right) of top-tier AI servers. All three projections observe decreases in server units during the period due to larger AI chips and smaller CoWoS increasing rates. The Moderate and High projections forecast continuously increasing the average rated power of AI servers, while the Low Scenario predicts a decrease for the 2026 Rubin system.

The rated power of AI server systems is strongly influenced by chip CoWoS size, which is expected to increase due to larger GPUs and additional high-bandwidth memory (HBM). In this work, we assume a linear relationship between the AI server rated power and CoWoS size. This assumption is supported by the Pascal, Volta, and Ampere systems data listed in Table R2.1, strictly presenting a relationship of rated power = $6 \times \text{CoWoS size} - 5.5$. This feature ends at the Hooper system due to the applied HBM3 memory (compared to HBM2 applied in the previous three systems) and advanced tensor core design for supporting developing generative AI models¹⁴. To project future evolutions, we define a new linear relationship that will exist for systems after Hooper following its increasing pattern to the Blackwell system: rated power = $3.15 \times \text{CoWoS size} + 3.9$. Therefore, the following projections are defined:

- Moderate: The rated power for 2026 (Rubin) and 2028 systems is assumed to increase proportionally with reticle size following the increasing pattern from Hooper to Blackwell, indicating rated powers of 16.5 kW and 21.2 kW, respectively.
- Low: Rated powers for the 2026 and 2028 systems are set at 12.5 kW and 17.2 kW, reflecting lower power requirements for equivalent CoWoS sizes, as observed with Ampere and Hopper systems.

- High: Rated powers for the 2026 and 2028 systems are set at 20.5 kW and 25.2 kW, reflecting higher power requirements under equivalent CoWoS sizes based on historical patterns.

The adoption rates of newly released AI servers determine shipment mix patterns for each year. Projections are informed by the adoption trends of DGX systems from 2022 to 2024¹⁵⁻¹⁷. In 2022, Ampere systems dominated the top-tier market, with minimal shipments of the newly released Hopper system. By 2023 (the first full year after Hopper's release), it captured an estimated 60% market share, surpassing Ampere. By 2024, Hopper is expected to dominate nearly all top-tier shipments, while the Blackwell platform is not yet to be shipped after release. Using this template, the adoption scenarios for 2024, 2026, and 2028 releases are defined as follows:

- In the Moderate and Low Projections, adoption patterns follow the template: 60% of shipments in the first-year post-release, 100% in the second year, and 40% in the third year, with negligible shipments thereafter.
- In the High Projection, a lower adoption rate is assumed: 80% in the first year, 100% in the second year, and 20% in the third year, with negligible shipments thereafter.

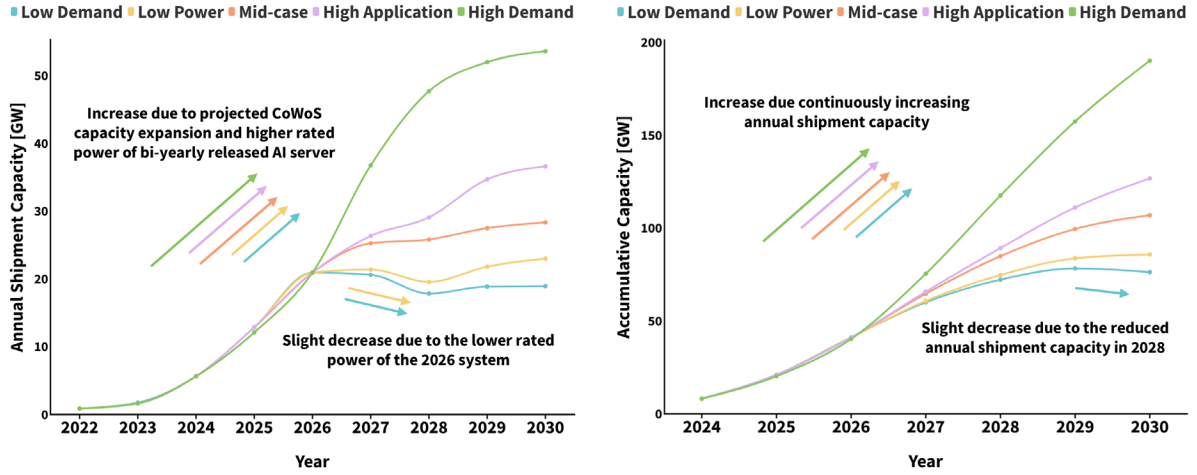


Figure S3 The projected annual shipment capacity (left) and accumulative capacity (right) of top-tier AI servers. The Low Demand and Low Power scenarios predict a decrease of annual shipment capacity starting from the 2026 system release with lower rated power. The other scenarios predict a continuous increase, while the High Demand scenario presents a more aggressive increase after the 2026 expansion.

These assumptions provide a necessary basis for evaluating future AI server capacity expansion, considering uncertainties within both technological advancements and market dynamics. Following the calculation presented in Eq (1), the projected annual shipments and

average rated power of the top-tier AI servers are presented in Figure S2. The annual top-tier AI server shipment power capacity and accumulative power capacity projections are then presented in Figure S3. A 4-year lifetime is used to calculate the accumulative power capacity, determined based on recent studies and industry estimates¹⁸⁻²¹. To further validate the effectiveness of the applied CoWoS-based approach, we have compared the estimated top-tier AI server shipment units with reported shipment data from TrendForce from 2022 to 2024^{22,23}, as presented in Figure S4. The two profiles showcase a good match.

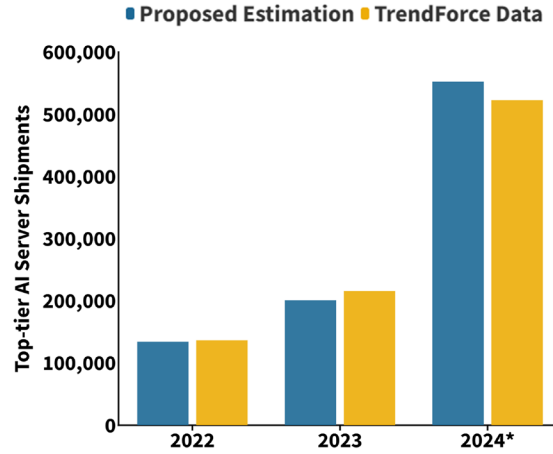


Figure S4 The comparison between the top-tier AI server shipments estimation by using the proposed method and the reported data from TrendForce^{22,23}. ‘*’ denotes that the TrendForce data is an estimated value.

To evaluate the total power capacity of AI servers, we incorporate low-end AI servers alongside top-tier systems. Due to their minor contribution to total power capacity and the more complex and uncertain market mix, the following assumptions are applied. Firstly, the annual shipment units of low-end servers are assumed to remain constant throughout the projection period, following the prediction in existing industry short-term projections with an approximate value of 540,000 units²². For power specifications, the most common enterprise cluster design is used as a reference, which features a rated power of 14 kW per rack, equipped with 22 GPUs²⁴. Low-end AI servers are assumed to be uniformly configured with an average of three GPUs per server, reflecting typical deployment patterns^{22,25}. The above definition neglects the potential increase and complex mix in low-end AI servers, introducing only a minor influence on the analysis: about 5% impact on the power capacity by 2030 with even doubling capacity of low-end AI servers considering their significantly smaller power capacity than top-tier ones. This feature is also

proved by the recent important data center report released by Lawrence Berkeley National Laboratory ²⁶.

The AI server capacity projection method is a simplified approach and relies on several key assumptions:

- The evolution of Nvidia DGX systems is representative of the top-tier AI server market.
- A constant 4-year lifetime is assumed for AI server development.
- The projection scenarios are comprehensive and rational to cover potential uncertainties in demand market development.
- The supply side is capable of producing the projected amounts of AI servers.

The Nvidia DGX system assumption is supported by its dominant position in the top-tier AI server market, which is discussed in detail in previous contents. The lifetime assumption of AI servers is deployed to address the complexity of the life cycle of AI servers and calculate their accumulative energy consumption. Its uncertainty is further illustrated in the sensitivity analysis presented in Figure 6 of the manuscript. The complex and uncertain AI server market evolution has necessitated a comprehensive modeling for the demand side. Predicting the most likely demand development scenario until 2030 is very challenging due to several factors, including the evolution of AI models, the breadth of their applications, and future improvements in server efficiency. To capture the large uncertainties, we adopt five distinct scenarios. The Highest scenario envisions a sustained and significant growth trajectory beyond the current 2026 expansion plans. Conversely, the Lowest Scenario assumes a stabilized AI server market post-2026, establishing a lower bound for projections. The final assumption largely relates to the TSMC CoWoS capacity, which represents a critical bottleneck in AI server manufacturing. This capacity serves as the foundation for our projections. However, additional factors, such as the availability of high-bandwidth memory and advancements in TSMC's 3-nanometer process, could also emerge as new constraints. While the impacts of these factors remain uncertain at present, they can be integrated into the framework as more data becomes available.

S2. AI Server Spatial Distribution Based on Large-scale Data Center Locations

The first step in evaluating the on-site environmental costs of AI servers involves estimating the allocation ratio of AI servers in each region. Initially, 53% of the total AI server capacity is allocated to the U.S. following the current power capacity distribution of large-scale data centers

²⁷. The sensitivity analysis of this factor is presented in Figure S6 of the manuscript. To further capture the current and future distributions of AI server capacity among U.S. states, we further compiled data on large-scale data centers belonging to major purchasers of top-tier AI servers, including Google, Meta, Microsoft, AWS, XAI, and Tesla ^{16,17}. The collected dataset includes 386 existing large-scale data centers (or clusters) and 146 planned data centers scheduled for construction between 2025 and 2030. These data were sourced from official resources and industry reports ²⁸⁻³². The dataset is included in our GitHub repository <https://github.com/PEESEgroup/US-AI-Server-Analysis>. To estimate the installation potential of each data center, we also collected building area data for these facilities. For data centers without available building area information, we used the average building area calculated based on our dataset as a proxy, following a common statistical method applied in many previous works ^{33,34}. The building area is considered a good index for evaluating the distribution pattern because it represents the potential of installing new AI servers and also are widely available through official resources. Additionally, for data centers under construction, the expected completion year was recorded for yearly resolution in our projections.

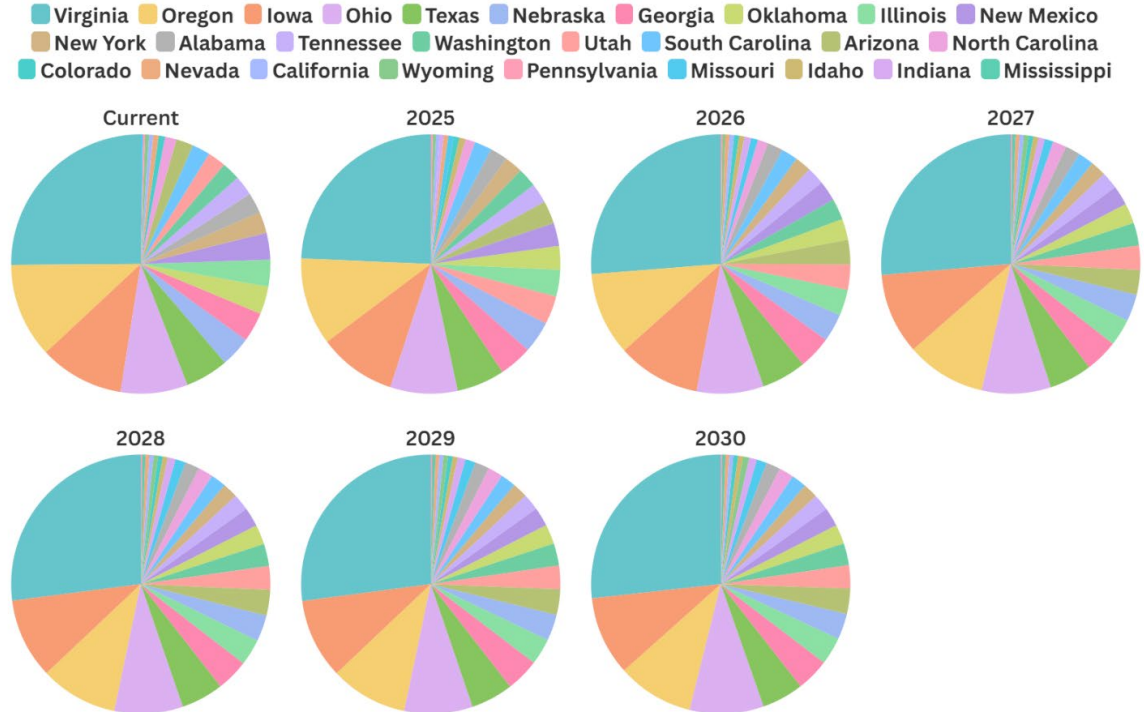


Figure S5 The spatial distributions of AI servers each year based on the building area distributions of large-scale data centers belonging to major AI server purchasers.

The spatial distribution of AI servers for each year was calculated based on the total building area of all data centers within each state. These distributions are presented in Figure S5. Projections for future distributions were derived by summing the building areas of existing data centers with those of new facilities completed by or within the projection year. While quite a few large-scale data centers are expected to be operational by 2030, the distributions show no significant shifts between 2024 and 2030. Given these findings and the inherent uncertainties in directly incorporating in-construction data centers into the distribution models, we project that the current spatial distribution will remain constant from 2024 to 2030.

S3. Estimating On-site Energy and Water Footprints of AI Servers: Models and Materials

The actual electricity usage of installed AI servers is calculated following the method introduced in Section **AI Server Electricity Usage Calculation**. The subsequent step involves calculating the on-site energy consumption and water footprint of AI servers based on the PUE and WUE estimation of AI data centers. As mentioned in the Section **Assessment of State-level Energy, Water, and Carbon Footprints of AI Servers**, the average climate data of each state is calculated by using existing resource³⁵. A statistical and thermodynamic model developed in the previous work^{36,37} is then utilized to compute PUE and WUE values based on state-level average climate data. The model inputs descriptions and ranges are listed in Table S2, which are collected from literature³⁶⁻⁴⁰. Specifically, two types of economizers, including airside economizers (AE) and waterside economizers (WE), can be adopted in the AI data centers. The best and worst practices for PUE and WUE values are calculated based on the parameter range listed in Table S2 and applied in the Section **Achieving Higher Energy and Water Usage Efficiency in AI Data Centers**. For instance, the best PUE values are calculated based on the following optimization problem:

$$\begin{aligned} \min_x \quad & PUE \\ \text{s.t.} \quad & [PUE, WUE] = F(x) \\ & x \in \mathbf{X} \end{aligned} \tag{S1}$$

The input x includes system inputs listed in Table S2 and the selection of economizer. Due to the lack of a specific cooling design for AI data centers, the PUE and WUE values of the base case are calculated by averaging the best and worst values.

Table S2. Inputs for the PUE and WUE estimation model. AE, CRAH, DLC, PTD, and WE represent airside economizers, computer room air handlers, direct liquid cooling, power transformation and distribution, and water economizers, respectively.

Inputs	Units	AE	WE	AE+DLC	WE+DLC	Resources
Outside Air Temperature	°C	Climate Data	Climate Data	Climate Data	Climate Data	35
Outside Air Relative Humidity	%	Climate Data	Climate Data	Climate Data	Climate Data	35
Outside Air Pressure	Pa	Climate Data	Climate Data	Climate Data	Climate Data	35
UPS efficiency	%	90 – 99	90 – 99	90 – 99	90 – 99	36,37,39
Percentage of power loss in PTD system	%	0 - 2	0 - 2	0 - 2	0 - 2	36,37,39
Lighting power to IT power ratio	%	0 - 0.2	0 - 0.2	0 - 0.2	0 - 0.2	36,37,39
Temperature difference (supply/return CRAH air)	°C	13.9 - 19.4	13.9 - 19.4	13.9 - 19.4	13.9 - 19.4	36,37,39
Fan pressure (CRAH)	Pa	300 - 1000	300 - 700	300 - 1000	300 - 700	36,37,39
Fan efficiency (CRAH)	%	65 - 90	65 - 90	65 - 90	65 - 90	36,37,39
Pump pressure (humidification pump)	kPa	6300 - 7700	6300 - 7700	6300 - 7700	6300 - 7700	36,37,39
Pump efficiency (humidification pump)	%	0.6 - 0.9	0.6 - 0.9	0.6 - 0.9	0.6 - 0.9	36,37,39
Approach temperature (cooling tower)	°C	2.8 - 6.7	2.8 - 6.7	2.8 - 6.7	2.8 - 6.7	36,37,39
Chiller partial load factor	-	0.2 - 0.8	0.2 - 0.8	0.2 - 0.8	0.2 - 0.8	36,37,39
Temperature difference (supply/return facility system water)	°C	5 - 10	5 - 10	5 - 10	5 - 10	36,37,39
Pump pressure (chiller pump)	kPa	114.9 - 172.4	114.9 - 172.4	114.9 - 172.4	114.9 - 172.4	36,37,39
Pump efficiency (chiller pump)	%	60 – 80	60 – 80	60 – 80	60 – 80	36,37,39
Temperature difference (supply/return cooling tower water)	°C	4 - 6	4 - 6	4 - 6	4 - 6	36,37,39
Pump pressure (cooling tower)	kPa	166.9 - 250.4	166.9 - 250.4	166.9 - 250.4	166.9 - 250.4	36,37,39
Pump efficiency (cooling tower)	%	60 – 80	60 – 80	60 – 80	60 – 80	36,37,39
Windage loss of water as a percentage of cooling tower mass flow rate	%	0.005 - 0.5	0.005 - 0.5	0.005 - 0.5	0.005 - 0.5	36,37,39
Cycles of concentration	-	3 – 15	3 – 15	3 – 15	3 – 15	36,37,39
Fan pressure (cooling tower)	Pa	100 - 400	100 - 400	100 - 400	100 - 400	36,37,39
Fan efficiency (cooling tower)	%	65 - 90	65 - 90	65 - 90	65 - 90	36,37,39
Sensible heat ratio	%	95 - 99	95 - 99	95 - 99	95 - 99	36,37,39
Liquid-gas ratio (cooling tower)	-	0.2 - 4	0.2 - 4	0.2 - 4	0.2 - 4	36,37,39
Supply air dry bulb setpoint (higher bound)	°C	27 – 35	27 – 35	27 – 35	27 – 35	36,37,39

Supply air dry bulb setpoint (lower bound)	°C	10 - 18	10 - 18	10 - 18	10 - 18	36,37,39
Supply air dew point setpoint (higher bound)	°C	15 - 27	15 - 27	15 - 27	15 - 27	36,37,39
Supply air dew point setpoint (lower bound)	°C	-12 - -9	-12 - -9	-12 - -9	-12 - -9	36,37,39
Supply air relative humidity setpoint (higher bound)	%	65 - 95	65 - 90	65 - 95	65 - 90	36,37,39
Supply air relative humidity setpoint (lower bound)	%	8 - 20	8 - 20	8 - 20	8 - 20	36,37,39
COP relative error to the regressed value	%	-11 - 11	-11 - 11	-11 - 11	-11 - 11	36,37,39
Heat exchanger effectiveness (CRAH cooling coils)	-	N/A	0.7 - 0.9	N/A	0.7 - 0.9	36,37,39
Approach temperature (economizer heat exchanger)	°C	N/A	1.7 - 2.8	N/A	1.7 - 2.8	36,37,39
Pump pressure (waterside economizer pump)	kPa	N/A	114.9 - 172.4	N/A	114.9 - 172.4	36,37,39
Pump efficiency (waterside economizer pump)	%	N/A	60 - 80	N/A	60 - 80	36,37,39
Coolant density	kg/m ³	N/A	N/A	1400 - 1855	1400 - 1855	38,40
Coolant flow rate	L/min	N/A	N/A	0.5 – 2.5	0.5 – 2.5	38,40
Pump pressure (coolant)	kPa	N/A	N/A	114.9 - 172.4	114.9 - 172.4	38,40
Pump efficiency (coolant)	%	N/A	N/A	60 - 80	60 - 80	38,40

As an important part of the data center infrastructure, the adoption of ALC during AI development is also considered in our work. We primarily focus on the adoption of immersion cooling, which is expected to see widespread applications in future AI servers^{41,42}. The following assumptions are applied to ALC considerations:

- Future ALC adoption within AI data centers will consist entirely of immersion cooling.
- The current adoption rate of ALC is estimated at approximately 5%, with compound annual growth rates (CAGR) for the worst, base, and best-case scenarios set at 0%, 20%, and 50%, respectively.

Immersion cooling is projected to be the primary driver of ALC adoption in AI data centers due to its superior ability to handle large rack densities in hyperscale data centers compared to the cold plate liquid cooling method⁴². The PUE and WUE values for immersion cooling were calculated using the statistical model and coolant parameters detailed in Table S2. The current adoption rate and future CAGR projections are crucial in shaping our projection scenarios. However, given the

lack of systematic industry data, we derive initial estimates from open-source market reports as summarized in Table S3. The CAGR settings are applied to cover the lower and upper bounds of potential developments for addressing the uncertainties.

The SUO adoption within AI data centers can improve GPU utilization in active GPUs and reduce idle server units by scheduling computing jobs and shutting down idle servers. Considering the concerns of exceeding thermal design power when increasing utilization in active servers ⁴³, we only consider the r_{active} improvement method to project the best, base, and worst case scenarios:

- Best scenario: r_{active} will be 90% for inference and 95% for training by 2030 as the best-reported results in previous studies ^{44,45}.
- Base scenario: r_{active} will be over 70% for inference and 92.5% for training by 2030, assuming half of the data centers achieve the best-reported results in previous studies ^{44,45}.
- Worst scenario: r_{active} will stay the same value by 2030.

Table S3 Data resources of ALC adoption rate

Resource	Base Year ALC market Share	CAGR	Period	Reference
Dell'Oro Group	5%	45.6%	2022 - 2026	46
Mordor Intelligence	4.77%	26.15%	2024 - 2029	47,48
Markets and Markets	3.15%	12.8%	2023 - 2031	49,50
Global Market Insights	6.92%	17%	2024 - 2032	51,52
Maximize Market Research	5%	26.6%	2024 - 2030	53,54

S4. Grid Interactions and Renewable Energy Penetrations

The grid data, including water and carbon factors, are generated from the Regional Energy Deployment System (ReEDs) model ⁵⁵. By applying this model and the related baseline projection data ⁵⁶, we have obtained the grid carbon and water data based on balancing authority (BA)-level grid projections. Specifically, the projected AI data center load is also included within the ReEDs model. The input settings, defined load data files, and outputs for the ReEDs model are provided in our GitHub repository. The calculation process below explains how the water and carbon

footprints of each state are obtained in this work and especially supports the results presented in Sections **Location Matters: Influences of AI Server Spatial Distribution** and **Taking Advantage of Renewable Energy Penetration within Grid**. The grid carbon factor and grid water factor of each state are calculated as follows:

$$CF_i = \sum_k \sum_m \frac{CF_{i,m} \cdot NG_{i,k,m}}{NG_{i,k}}, k \in \mathbf{K}_i, m \in \mathbf{M} \quad (\text{S2})$$

$$WF_i = \sum_k \sum_m \frac{WF_{i,m,thermo} \cdot NG_{i,k,m}}{NG_{i,k}} + WF_{i,hydro,evap} \cdot \frac{NG_{hydro,i}}{NG_i}, k \in \mathbf{K}_i, m \in \mathbf{M} \quad (\text{S3})$$

CF , NG and WF represent the carbon factor, net generation, and water factor, respectively. Subscripts i , k , and m represent the state index, balancing area index, and energy resource index. The water factor calculation includes water consumption from thermoelectric generation and water evaporation from hydropower. CF and WF_{thermo} are determined based on the ReEDs model's default national-level resolution settings⁵⁵, to enable the problem feasibility and reduce computational intensity. The hydropower evaporation rates, presented in Figure S6, are sourced from a U.S. hydroelectric water consumption estimation study⁵⁷. Furthermore, the total energy carbon emission and water footprint of each state are further calculated as follows:

$$CE_i = E_i \cdot PUE_i \cdot CF_i, i \in \mathbf{I} \quad (\text{S4})$$

$$W_i = E_i \cdot WUE_i + E_i \cdot WF_i, i \in \mathbf{I} \quad (\text{S5})$$

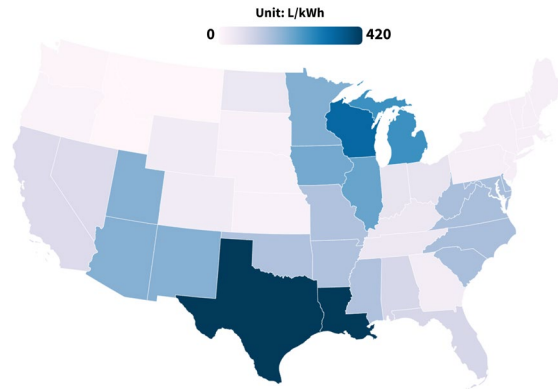


Figure S6. The hydropower water consumption factor of each state. The data is obtained from previous works⁵⁷.

To further validate the effectiveness of the grid model applied in this study, we have collected high-resolution data from the U.S. Energy Information Administration (EIA) for comparison. Figures S7 and S8 present the comparison of the grid carbon factor and grid thermoelectric water factor from 2020 to 2023, respectively. The reference carbon factors are sourced from EIA state-

level grid data ⁵⁸, which are extracted from their real-time hourly 930 monitoring system ⁵⁹. The reference grid thermoelectric water factor is derived from methods established in previous literature for estimating historical water factor ⁶⁰, which utilized the EIA power plant-level water and generation data ^{61,62}. The carbon factors showcase good fits in 2021, 2022, and 2023, and observe some deviations in 2020. This mismatch can be derived from multiple factors, including evolving EIA state-level estimation approach ⁶³, uses of low-resolution data in the ReEDS model, and instability of the solving process.

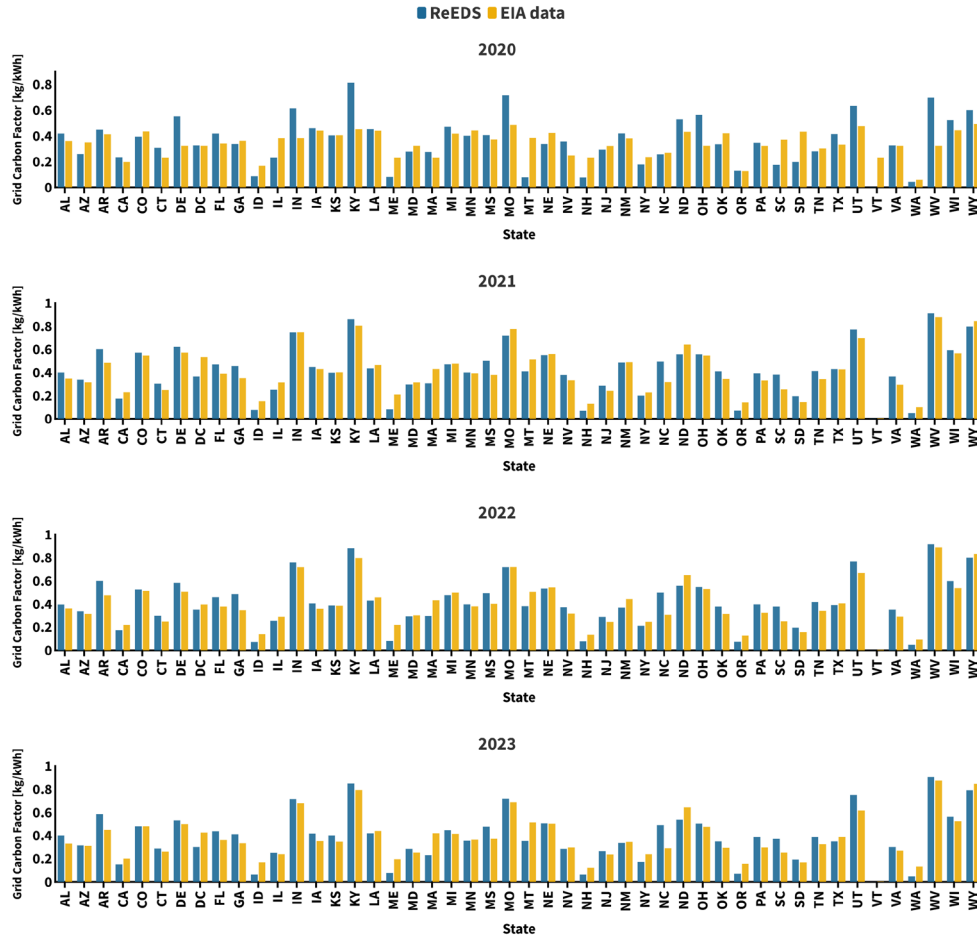


Figure S7. The comparison of the grid carbon factors. Blue columns are the data generated from the applied model. Yellow columns are the EIA state-level data ⁵⁸.

On the other hand, the grid thermoelectric factor comparison is generally consistent, but mismatches appear in several states due to several factors:

- The ReEDS model used in this study applies a national average water consumption rate for each fuel type. While this approach improves computational feasibility, it reduces accuracy

particularly for grid water factors, since water consumption varies significantly with local climate and technology adoption.

- The thermoelectric data reported by the EIA is less mature than its real-time carbon factor monitoring system. During the comparison period, several power plants did not report their water consumption values, which undermines the credibility of the dataset.

It is possible to improve the performance of the ReEDS model by incorporating BA-level data for once-through, recirculating, dry, and cooling pond water consumption rates for each fuel type. However, such data is currently difficult to organize. Moreover, relying on insufficient BA-level data could introduce additional uncertainty and further involve difficulty in solving the optimization problem for grid projection. Based on these comparison results, we believe the ReEDS model remains a sufficient tool for generating future grid development scenarios. Further validations and definitions of the ReEDS model can be found in their official documents⁶⁴.

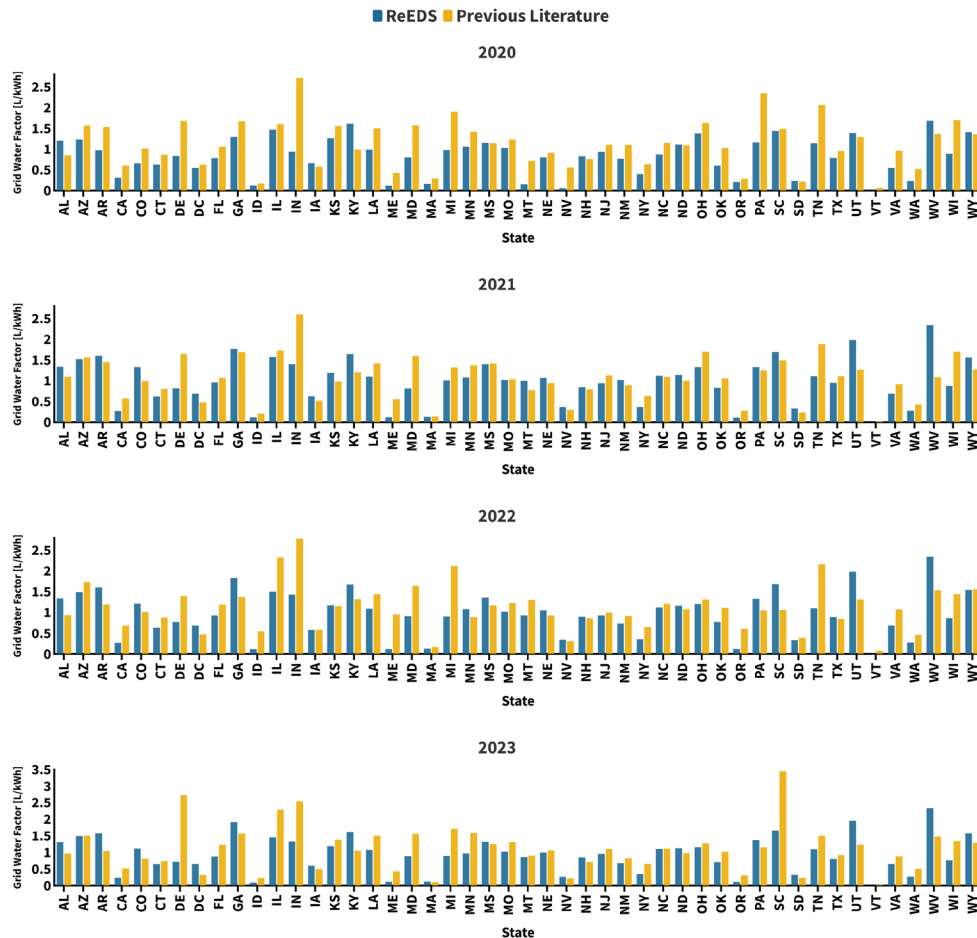


Figure S8. The comparison of the grid thermoelectric water factors. Blue columns are the data generated from the applied model. Yellow columns are the data generated from a high-resolution approach in previous literature ⁶⁰ based on EIA data ^{61,62}.

S5. Methodology Comparison and Validation with Existing Studies

In this study, we adopt a bottom-up approach, building upon prior modeling efforts to capture the unique characteristics of AI server deployment. The approach mainly involves four steps to outline the environmental footprints of servers:

- The estimation of the total shipment capacity of servers
- The calculation of the real electricity usage of servers based on the shipment capacity
- The data center's on-site energy consumption/water footprint is calculated by multiplying PUE/WUE by the estimated server electricity usage.
- The Scope 2 water footprint/carbon emission is calculated by multiplying the grid factors by the data center's on-site energy consumption.

Based on the four key steps, a detailed comparison between our methodology and existing studies is provided in Table 3.1. In particular, several contributions from Lawrence Berkeley National Laboratory (LBNL) are highlighted due to their recognized expertise in the field. LBNL has played a leading role in developing bottom-up estimation methodologies and has produced official reports for U.S. energy agencies, making their work a foundational reference in data center energy analysis. We emphasize that the end-of-2024 release by LBNL ²⁶ and the 2025 report by the International Energy Agency (IEA) ⁶⁵, both developed in parallel with our work and released during the submission process, currently represent the most reliable references for assessing the energy implications of the ongoing AI surge. These two works are built upon LBNL's modeling foundation developed over the past decade, including their seminal 2016 report ⁶⁶ and subsequent 2018 updates ^{67,68}. Our methodology is largely consistent with these studies. Specifically, our analysis adopts the same utilization-based electricity usage model ⁶⁹, as well as the statistical and thermodynamic approaches ³⁶ for estimating PUE and WUE, consistent with both the LBNL and IEA methodologies. The primary methodological distinction in our study lies in the development of an open-source AI server shipment estimation approach based on a bottleneck analysis of CoWoS (Chip-on-Wafer-on-Substrate) manufacturing, and the usage of the Regional Energy Deployment System (ReEDS) model ⁵⁵ for estimating future grid dynamics. We have compared

the ReEDS model with high resolution EIA in **Section S4**. Here, we further verify the deployment of CoWoS-based projections in the following contents through a comprehensive comparison:

1). Empirical validation: We have compared the estimated top-tier AI server shipment units with reported shipment data from the TrendForce tracker from 2022 to 2024, as presented in Figure S4. The comparison demonstrates strong agreement between our CoWoS-based estimates and the TrendForce data. To the best of our knowledge, the TrendForce dataset represents the most reliable open-source reference currently available for tracking AI server shipments. We did not incorporate data prior to 2022, as AI server deployment began accelerating in that year, and earlier shipment volumes were both minor and sparsely documented. This comparison provides empirical support for the validity of our CoWoS-based estimation methodology.

2). Historical methodology application: While CoWoS-based estimation has not previously been applied to forecast server shipments, given that CoWoS is a newly emerging bottleneck specific to high-performance AI chips, the use of bottleneck manufacturing steps to estimate supply chain capacity for electrified sectors is well established. For example, a 2024 Nature Communications study ⁷⁰ projected electric vehicle production capacity based on critical mineral bottlenecks, and another recent research ⁷¹ has identified key manufacturing constraints to forecast battery production growth, summarizing previous studies. These studies demonstrate the validity of bottleneck-based estimation strategies for capturing end-to-end supply chain capacity, an approach that we adopt and adapt for AI server projections in this work.

3). Cross-study consistency: As shown in Table 3.1, our projections estimate AI server energy consumption in the range of 204–325 TWh by 2028 and 220–532 TWh by 2030. These values align well with those reported by LBNL, which projects over 150 TWh to more than 300 TWh of AI-related data center consumption in the U.S. by 2028, and with the IEA’s base-case U.S. data center estimate of 426 TWh by 2030. While the specific profiles of their projection are not provided, the target year value ranges show a good match. Given that our study was developed in parallel with these reports, this consistency reinforces the credibility of our projection scope, despite our reliance on a CoWoS-based estimation approach rather than proprietary commercial datasets.

4). Sensitivity analysis: We have conducted a parameter-based sensitivity analysis to evaluate the influence of key drivers in the CoWoS-based projection model, thereby capturing uncertainty

bounds and validating model resilience. The sensitivity analysis result is shown in Figure 6 of the manuscript, which proves the robustness of the projection with variations of key driving factors.

Table S4 Comparison of our study with existing bottom-up methodologies.

Methodology	Historical projection		Current projections include the AI surge		
	LBNL's 2016 report ⁶⁶	LBNL's following efforts ^{67,68}	LBNL's 2024 report ²⁶	IEA's 2025 report ⁶⁵	Our study
AI Server Shipment Estimation	IDC commercial data (undisclosed)	IDC commercial data (undisclosed)	IDC commercial data (undisclosed)	IDC commercial data (undisclosed)	CoWoS-based estimation method
AI Server Electricity Usage Estimation	Utilization-based linear model ⁶⁹	Utilization-based linear model ⁶⁹	Utilization-based linear model ⁶⁹	Utilization-based linear model ⁶⁹	Utilization-based linear model ⁶⁹
PUE/WUE Estimation	Statistical Value	Statistical Value	2022 Statistical and thermodynamic model ³⁶	2022 Statistical and thermodynamic model ³⁶	2022 Statistical and thermodynamic model ³⁶
Grid Factor Estimation for Scope 2 Carbon and Water Footprints	Not included	EIA data	N/A	Brief discussions based on current EIA data	ReEDS model for dynamic simulations
AI Energy Projections	Not included	Not included	Over 150 TWh to over 300 TWh of U.S. AI servers by 2028	Base case: 426 TWh for the U.S. data centers by 2030	204 – 325 TWh by 2028 220 – 532 TWh by 2030
Specified AI Environmental Impact	Not included	Not included	Not included	Not included	Annual 731 to 1125 million m ³ water footprints and 24 to 44 Mt CO ₂ -eq between 2024 and 2030

S6. Nomenclature

Sets

I	Set of states indexed by i .
J	Set of grid cells indexed by j .
K	Set of balancing areas indexed by k .
M	Set of energy resources indexed by m .
T	Set of time indexed by t .
X	Set of model inputs indexed by x .
Θ	Set of model parameters indexed by θ .

Symbols

a	Start time of the exponential shock
b	Reabsorption rate of the exponential shock
c	Shock intensity of the exponential shock
C	Coefficient
CE	Carbon emission
Cap	Accumulative Capacity
CF	Grid Carbon factor
E	Energy consumption
F	Function
p	Innovation rate of the Bass model
P	Power usage
q	Imitation rate of the Bass model
m	Market potential of the bass model
n	Number of time interval
N	Number of grid cells
NCE	Net carbon emission
NG	Net power generation
PUE	Power usage effectiveness
R	Revenue
t	Time index
t^*	Peak sale time
T	<i>Lifetime of AI servers</i>
u	Utilization level of servers
W	Water footprint
WF	Grid water factor
WUE	Water usage effectiveness
x	Inputs of the PUE and WUE estimation model
z	Accumulative adopters of the Bass model
θ	Model parameters
χ	Intervention function

Subscripts

AI	AI server
i	State index
ITR	Idle power to rated power ratio
j	Grid cell index
k	Balancing area index
L	Lighting system
$Loss$	Loss of electricity/water
m	Grid energy resource index

<i>max</i>	Maximum value
<i>min</i>	Minimum value
<i>MTR</i>	Maximum power to rated power ratio
<i>Rate</i>	<i>Rated value</i>
<i>RTC</i>	Revenue to capacity
<i>t</i>	Time index

Supplementary References

- 1 Monica Chen, H. J. S. D. A. in *TSMC to see CoWoS production capacity reach 60,000 wafers in 2025* (<https://www.digitimes.com>, 2024).
- 2 NIKKEI Asia. in *TSMC explores radical new chip packaging approach to feed AI boom* (<https://asia.nikkei.com>, 2024).
- 3 Digitimes Asia. in *TSMC adjusts CoWoS capacity plans amid Trump 2.0 uncertainty* (<https://www.digitimes.com>, 2024).
- 4 Digitimes Asia. in *Nvidia secures 60% of TSMC's doubled CoWoS capacity for 2025* (<https://www.digitimes.com>, 2024).
- 5 TrendForce. in *TSMC Reportedly Sensing Increased Orders Again, CoWoS Production Capacity Surges* (<https://www.trendforce.com/news/>, 2024).
- 6 TrendForce. in *TSMC's Advanced Packaging Sees Surge with Rush Orders from NVIDIA, AMD, Amazon* (<https://www.trendforce.com>, 2023).
- 7 TrendForce. in *Global AI Server Demand Surge Expected to Drive 2024 Market Value to US\$187 Billion; Represents 65% of Server Market, Says TrendForce* (<https://www.trendforce.com>, 2024).
- 8 Shah, A. in *Nvidia Shipped 3.76 Million Data-center GPUs in 2023, According to Study* (<https://www.hpcwire.com>, 2024).
- 9 Nvidia. in *DGX Systems : Built for the Unique Demands of AI* (<https://www.nvidia.com/en-gb/data-center/dgx-systems/>, 2024).
- 10 Nvidia. in *NVIDIA DGX Platform* (<https://docs.nvidia.com/dgx/>, 2024).
- 11 TrendForce. in *TSMC Boosts Investment in Advanced Packaging with NTD 500 Billion Plan to Build Six Plants in Chiayi Science Park* (<https://www.trendforce.com>, 2024).
- 12 TSMC. in *3DFabricTM: Advanced packaging technologies and design ecosystem collaboration* (https://chipletsummit.com/proceeding_files/a0q5f0000044zma/20240206_PreConF_Paris.PDF, 2024).
- 13 Garreffa, A. in *NVIDIA's next-gen Rubin, Rubin Ultra, Blackwell Ultra AI GPUs: also supercharged Vera CPUs* (<https://www.tweaktown.com>, 2024).
- 14 Nvidia. NVIDIA DGX H100 User Guide. (<https://docs.nvidia.com/dgx/dgXH100-user-guide/dgXH100-user-guide.pdf>, 2023).
- 15 Benaich, N. & Air Street Capital Team. State of AI Report 2022. (<https://www.stateof.ai/>, 2022).
- 16 Benaich, N. & Air Street Capital Team. State of AI Report 2023. (<https://www.stateof.ai/>, 2023).
- 17 Benaich, N. & Air Street Capital Team. State of AI Report 2024. (<https://www.stateof.ai/>, 2024).
- 18 Wu, C.-J. *et al.* Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* **4**, 795-813 (2022).
- 19 Li, B. *et al.* Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems. in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1-15.
- 20 Fuchs, H. *et al.* Characteristics and energy use of volume servers in the united states. (Energy Analysis and Environmental Impacts Division, Energy Efficiency Studies Department, <https://ees.lbl.gov/publications/characteristics-and-energy-use-volume>, 2023).

- 21 Shilov, A. in *Datacenter GPU service life can be surprisingly short — only one to three years is expected according to unnamed Google architect* (<https://www.tomshardware.com>, 2024).
- 22 Kung, F. in *See Generative AI's Impact on the AI Server Market to 2025* (https://files.futurememorystorage.com/proceedings/2024/20240808_BMKT-301-1_KUNG.pdf, 2024).
- 23 Liu, M. L. & Kung, F. in *Global AI Server Shipments Forecasted to Increase 40% in 2023 amid Rising AI Demand, Says TrendForce* (<https://www.trendforce.com/presscenter/news>, 2023).
- 24 Nvidia. in *NVIDIA AI Enterprise: Sizing Guide* (<https://docs.nvidia.com/ai-enterprise/sizing-guide/latest/>, 2024).
- 25 Nvidia. in *Data Center Products* (<https://www.nvidia.com/en-us/data-center/products>, 2024).
- 26 Shehabi, A. et al. 2024 United States Data Center Energy Usage Report. (<https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report>, 2024).
- 27 Synergy Research Group. in *Hyperscale Data Center Capacity to Almost Triple in Next Six Years, Driven by AI* (<https://www.srgresearch.com/articles>, 2023).
- 28 Google. in *Discover our data center locations* (<https://www.google.com/about/datacenters/locations/>, 2024).
- 29 Meta. in *Meta's U.S. data center fleet* (<https://datacenters.atmeta.com/us-locations/>, 2024).
- 30 Microsoft. in *Microsoft Datacenters* (<https://datacenters.microsoft.com/globe/explore/>, 2024).
- 31 AWS. in *AWS Global Infrastructure* (<https://aws.amazon.com/about-aws/global-infrastructure/>, 2024).
- 32 Baxtel. in *United States Data Center Market* (Baxtel, <https://baxtel.com/data-center/united-states>, 2024).
- 33 Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology* **33**, 364-376 (2015).
- 34 Wei, R. et al. Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific reports* **8**, 663 (2018).
- 35 Jan Remund, S. M., Michael Schmutz and Pascal Graf. in *Global Meteorological Database for Engineers, Planners and Education* (Meteonorm, <https://meteonorm.com/>, 2020).
- 36 Lei, N. & Masanet, E. Climate-and technology-specific PUE and WUE estimations for US data centers using a hybrid statistical and thermodynamics-based approach. *Resources, Conservation and Recycling* **182**, 106323 (2022).
- 37 Lei, N. & Masanet, E. Statistical analysis for predicting location-specific data center PUE and its improvement potential. *Energy* **201**, 117556 (2020).
- 38 Minnesota Mining and Manufacturing Company. 3M Specialty Fluids. (https://www.3m.com/3M/en_US/p/c/electronics-components/specialty-fluids/, 2024).
- 39 Quirk, D., Davidson, T. & Roger Schmidt PHD, P. ASHRAE's Data Center Thermal Guidelines. *ASHRAE Journal* **64**, 24-29 (2022).
- 40 Li, X., Xu, Z., Liu, S., Zhang, X. & Sun, H. Server performance optimization for single-phase immersion cooling data center. *Applied Thermal Engineering* **224**, 120080 (2023).
- 41 Agonafer, D., Bansode, P., Saini, S., Gullbrand, J. & Gupta, A. Single Phase Immersion Cooling for Hyper Scale Data Centers: Challenges and Opportunities. in *Heat Transfer Summer Conference*. V001T016A008 (American Society of Mechanical Engineers).
- 42 Kong, R. et al. Enhancing data center cooling efficiency and ability: a comprehensive review of direct liquid cooling technologies. *Energy* **308**, 132846 (2024).
- 43 Patel, P. et al. Characterizing Power Management Opportunities for LLMs in the Cloud. in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*. 207-222.
- 44 Mohan, J., Phanishayee, A., Kulkarni, J. & Chidambaram, V. Looking beyond GPUs for DNN scheduling on Multi-Tenant clusters. in *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. 579-596.

- 45 Hu, Q., Sun, P., Yan, S., Wen, Y. & Zhang, T. Characterization and prediction of deep learning workloads in large-scale gpu datacenters. in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1-15.
- 46 Fung, B. & Beran, L. Achieving Sustainable Data Center Growth. (<https://www.delloro.com/wp-content/uploads/2022/10/DellOro-Achieving-Sustainable-Data-Center-Growth.pdf>, 2022).
- 47 Mordor Intelligence. Data Center Cooling Market Size & Share Analysis - Growth Trends & Forecasts (2024 - 2029). (<https://www.mordorintelligence.com/industry-reports/global-data-center-cooling-market-industry>, 2023).
- 48 Mordor Intelligence. Immersion Cooling Market In Data Centers Market Size. (<https://www.mordorintelligence.com/industry-reports/immersion-cooling-market-in-data-centers/market-size>, 2023).
- 49 Markets and Markets. Immersion Cooling Market by Type, Application, Cooling Fluid, Component, and Region - Global Forecast to 2031. (<https://www.marketsandmarkets.com/Market-Reports/immersion-cooling-market-107040948.html>, 2023).
- 50 Markets and Markets. Data Center Market by Solution, Service, Type of Cooling, Data Center Type, Industry, & Geography - Global Forecast to 2030. (<https://www.marketsandmarkets.com/Market-Reports/data-center-cooling-solutions-market-1038.html>, 2024).
- 51 Global Market Insights. Data Center Immersion Cooling Market - By Component, By Cooling Technique, By Cooling Fluid , By Organization Size, By End-Use, Forecast 2024 – 2032. (<https://www.gminsights.com/industry-analysis/data-center-immersion-cooling-market>, 2024).
- 52 Global Market Insights. Data Center Cooling Market - By Component, By Cooling Technique, By End Use, By Data Center Size, Forecast 2024 – 2032. (<https://www.gminsights.com/industry-analysis/data-center-cooling-market>, 2024).
- 53 Maximize Market Research. Data Center Cooling Market: Global Industry Analysis and Forecast (2023-2029). (<https://www.maximizemarketresearch.com/market-report/global-data-center-cooling-market/23969/>, 2023).
- 54 Maximize Market Research. Data Center Liquid Immersion Cooling Market: Global Industry Analysis and Forecast (2024-2030). (<https://www.maximizemarketresearch.com/market-report/global-data-center-liquid-immersion-cooling/107348/>, 2024).
- 55 Ho, J. *et al.* Regional Energy Deployment System (ReEDS) Model Documentation (Version 2020). (National Renewable Energy Lab.(NREL), Golden, CO (United States), <https://docs.nrel.gov/docs/fy21osti/78195.pdf>, 2021).
- 56 NREL. in 2023 *Electricity ATB Technologies and Data Overview* (<https://atb.nrel.gov/electricity/2023/index>, 2023).
- 57 Grubert, E. A. Water consumption from hydroelectricity in the United States. *Advances in Water Resources* **96**, 88-94 (2016).
- 58 EIA. in *Historical State Data* (<https://www.eia.gov/electricity/data/state/>, 2024).
- 59 EIA. in *EIA Hourly Electric Grid Monitor* (https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/US48/US48, 2024).
- 60 Siddik, M. A. B., Shehabi, A., Rao, P. & Marston, L. T. Spatially and temporally detailed water and carbon footprints of US Electricity generation and use. *Water Resources Research* **60**, e2024WR038350 (2024).
- 61 EIA. in *Thermoelectric cooling water data* (<https://www.eia.gov/electricity/data/water/>, 2024).
- 62 EIA. in *Form EIA-923 detailed data with previous form data (EIA-906/920)* (<https://www.eia.gov/electricity/data/eia923/>, 2024).
- 63 EIA. in *Data and methodology changes* (<https://www.eia.gov/state/seds/seds-data-changes.php?sid=US#2021>, 2023).
- 64 National Renewable Energy Laboratory. in *ReEDS 2.0* (<https://nrel.github.io/ReEDS-2.0>, 2024).
- 65 International Energy Agency. Energy and AI. (IEA, <https://www.iea.org/reports/energy-and-ai>, 2025).
- 66 Shehabi, A. *et al.* United states data center energy usage report. (2016).

- 67 Siddik, M. A. B., Shehabi, A. & Marston, L. The environmental footprint of data centers in the United States. *Environmental Research Letters* **16**, 064017 (2021).
- 68 Shehabi, A., Smith, S. J., Masanet, E. & Koomey, J. Data center growth in the United States: decoupling the demand for services from electricity use. *Environmental Research Letters* **13**, 124030 (2018).
- 69 Fan, X., Weber, W.-D. & Barroso, L. A. Power provisioning for a warehouse-sized computer. *ACM SIGARCH computer architecture news* **35**, 13-23 (2007).
- 70 Woodley, L. *et al.* Climate impacts of critical mineral supply chain bottlenecks for electric vehicle deployment. *Nature Communications* **15**, 6813 (2024).
- 71 Sihvonen, V., Grönman, A. & Honkapuro, S. Techno-socio-economic bottlenecks in increasing battery capacity for supporting the energy transition. *Renewable and Sustainable Energy Reviews* **210**, 115259 (2025).