UNCERTAINTY AND UPWARD BIAS ARE INHERENT IN DATA CENTER ELECTRICITY DEMAND PROJECTIONS

Prepared for

Southern Environmental Law Center

by

London Economics International LLC

717 Atlantic Ave, Suite 1A

Boston, MA 02111

July 7, 2025

Uncertainty and upward bias are inherent in data center electricity demand projections

prepared for Southern Environmental Law Center by London Economics International LLC

July 7, 2025

London Economics International LLC ("LEI") was engaged by the Southern Environmental Law Center ("SELC") to evaluate current projections of growth in power demand from a particular type of large load electricity customer: data centers. Data centers are large users of electricity, and projections of their growth are driving utility plans for system expansions. If data center growth does not materialize at the expected levels, the cost of system expansions by vertically integrated utilities would likely fall on other ratepayers.

Utilities, like any other business sector, must make investment decisions under uncertainties about the future. In the case of uncertainty over future load, a particular problem is that data centers are large consumers, so the same general forecast error applied to a large load will be more impactful on the success or failure of investment plans. Based on the analysis documented in this report, uncertainties inherent in outlooks for data center electricity demand reflect a bias to overstating future demand. For example, data center developers have incentives to duplicate requests in different jurisdictions for electric interconnection for the same facility, and industry reports indicate this has been ongoing. In addition, because of the US-wide and even global nature of the options for data center developers to site their facilities, it is difficult to determine in which jurisdictions the projected growth of data center electricity demand would meet, exceed, or fall short of such projections.

LEI's analysis also shows that projections of electric power demand by data centers in the United States exceed the capability of global chip manufacturers to supply the semiconductor chips that data centers need, given the demand from other locations globally. This is further evidence that data center developers are submitting more requests for service for new facilities in the United States than they plan to build.

The uncertainty and upward bias in data center electricity demand projections create the risk that new energy infrastructure (including electricity generation and transmission, and gas pipeline capacity) proposed to meet the needs of these large load customers could become underutilized and lead to higher costs for other customers than what they would have paid if the large load materialized.

Table of contents

London Economics International LLC
717 Atlantic Avenue, Suite 1A
Boston, MA 02111
www.londoneconomics.com

contact:
Marie Fagan/Victor Chung
617-933-7205
marie@londoneconomics.com

Table of figures

Acronyms

AI	Artificial intelligence
AMD	Advanced Micro Devices, Inc.
BA	Balancing authority
CAGR	Compounded annual growth rate
CEC	California Energy Commission
CSA	Customer service agreement
EIA	Energy Information Administration
ERCOT	Electric Reliability Corporation of Texas
ESA	Electricity service agreement
FERC	Federal Energy Regulatory Commission
FLOPS	Floating point operations per second
FLOPS/W	Floating point operations per second per watt
GPC	Georgia Power Company
GPSC	Georgia Public Service Commission
GPU	Graphic processing unit
GW	Gigawatt
GWh	Gigawatt hour
IBM	International Business Machines Corporation
IEA	International Energy Agency
Intel	Intel Corporation
IRP	Integrated Resource Plan
ISO	Independent system operator
ISO-NE	ISO New England
IT	Information technology
LBNL	Lawrence Berkeley National Laboratory
LEI	London Economics International, LLC
MISO	Midcontinent Independent System Operator
mm2	Square millimeter
MW	Megawatt
MWh	Megawatt hour
NPCC	Northwest Power and Conservation Council
NTP	Notice to proceed
NYISO	New York Independent System Operator
PJM	PJM Interconnection, LLC
PUE	Power usage effectiveness
ROE	Return on equity
RTO	Regional transmission organization
SELC	Southern Environmental Law Center
SPP	Southwest Power Pool

SXM	Server PCI Express Module
TDP	Total design power
TOU	Time-of-use
US	United States
W	Watt

1 Executive summary

Recent growth in electricity demand from data centers has been rapid in the United States and other regions throughout the world. A data center can require electric service on a much larger scale compared to other commercial or industrial customers. Though in the past a typical average data center had a load of 5-10 megawatts ("MW"), hyperscale data centers (defined in more detail later in this report) have loads of 100 MW or more, and even a 200 MW facility is now considered typical.[1] As a point of comparison, 200 MW is enough capacity to power over 109,000 households on a peak summer day in the state of Georgia.[2] The large scale of demand represented by even one new data center customer in the footprint of a vertically integrated utility makes it important for the utility to carefully evaluate the drivers and assumptions of such projected demand growth and the risks around such growth forecasts. Federal and state regulators will similarly want to ensure that planning processes and authorizations recognize inherent risks and uncertainties in electric demand growth projections.

As demonstrated in this report, forecasts of electricity demand stemming from data center growth are beset with uncertainty—the extent and pace of data center demand growth remains highly uncertain, and the specific locations even more difficult to predict. For the reasons outlined in this report, LEI believes that not all the electricity demand growth associated with new data centers projected for the United States or for any given individual jurisdiction in the United States will necessarily materialize.

The potential for overestimating electricity demand from new data centers has important implications for evaluating utility plans for generation and transmission capacity expansion, and in some cases related proposals for new interstate gas pipeline infrastructure to serve new gas-fired electricity generation. Currently, many utilities are projecting substantial growth in electricity demand from potential data centers within their territories. These demand projections in turn can drive a significant portion of the utilities' generation capacity expansion plans, and, in some cases, prompt a projected need for additional firm transportation capacity on interstate gas pipelines. However, if the projected data center load fails to fully materialize, then the cost of these assets — whether power plants or pipeline infrastructure—would be borne by other utility customers.

[1] International Energy Agency ("IEA"). "What the data centre and AI boom could mean for the energy sector." October 18, 2024. <https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector>; and McKinsey & Company. "AI power: Expanding data center capacity to meet growing demand." October 29, 2024. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand>

[2] LEI calculated the number of households based on the total number of residential customers for Georgia Power Company ("GPC") in 2023 and the total GWh consumption by GPC residential customers at a recent peak month, July 2023. This results in an average hourly peak for the month, not the peak hour for the month. (Source: Georgia Public Service Commission ("GPSC") Docket No. 56002. GPC. 2025 IRP Technical Appendix Volume 1: B2025 Load and Energy Forecast. January 2025. <https://psc.ga.gov/search/facts-document/?documentId=221233)>

1.1 Organization of the report

With this report, LEI examines the drivers of data center electricity demand in the United States using a two-part approach. First, in Section 2, LEI examines the uncertainties and incentives that impact forecasts of electricity demand growth from data centers. Second, in Section 3, LEI applies a critical lens to the projected growth of data center load in the United States by comparing aggregated US electricity demand projections from select jurisdictions against global semiconductor chip supply projections to discover whether US data center electricity demand growth could credibly be served by the projected global supply of semiconductor chips, a key component to data center operations. Third, Section 4 discusses implications for utility ratepayers; Section 5 concludes the report. Appendix A provides the reader with a basic understanding of data centers. Appendix B provides an index of works relied upon and cited in this report.

1.2 LEI finds uncertainty and inherent bias in data center load outlooks

Based on its analysis, LEI concludes that data center electricity demand projections remain highly uncertain and currently reflect a bias to overestimating growth in the number of data centers that will be built, and therefore also overestimating future electricity demand. In addition, because of the US-wide and even global nature of the options for data center developers to site plants, it is challenging to determine in which jurisdictions the data center-induced electricity demand growth would meet, exceed, or fall short of projections. This uncertainty creates substantial risks that new energy infrastructure proposed to serve large load customers may never become fully utilized as intended and could lead to higher costs for existing ratepayers.

LEI's review of the drivers of data center electric demand projections (detailed in Section 2) shows that:

1) **Data center developers have incentives to duplicate requests for electric service, and evidence shows they are doing so.** This reflects the variety of locations where a data center can choose to locate, in the United States and across the world; the incentive to duplicate requests and evidence of duplication of requests; and recent attrition in data center load based on announcements from such customers;
2) **Vertically integrated electric utilities do not have an incentive to be very skeptical of requests.** A vertically integrated utility under traditional cost-of-service regulation expects to benefit financially from making investments that expand its assets and contribute to increases in rate base; and
3) **Most data center interconnection requests do not reflect large financial commitments from potential customers.** Utilities perform some level of probabilistic analysis of demand requests, and some have adopted contracting terms to reduce risk of shifting of costs to existing customers, which can also help to reduce the incentive of data centers to submit multiple requests for interconnection, so that even if the utility adds new system assets which are underutilized for a time, the contract with the data center may prevent cost shifting to other ratepayers. However, many data center interconnection requests driving current utility load forecasts are at the stage of the interconnection process at which no contracts have yet been signed, so financial commitments are minimal;

4) **Other drivers result in over-forecasting of data center electricity demand.** Bottlenecks in electric generation, transmission, and distribution equipment are slowing the pace of electric sector infrastructure growth that would serve new data centers. If the generation cannot be built in time, the new load must be deferred or move to another location and will grow more slowly than current data center electricity demand projections would imply. Local opposition to data centers is growing and has resulted in delays and cancellations. At the state level, the cost of tax incentives for data centers is increasing as data center development booms. Finally, demand flexibility from future large load customers may reduce the need for new generation capacity.[3] Data centers often incorporate on-site generation and, where appropriate, the capability to shift operations across multiple facilities in far-flung regions. This supports flexibility in hourly consumption of electricity, moderating electricity demand from new data center customers, especially if the customer is on a tariff which incentivizes reduction in electricity demanded from the utility during periods of scarcity and/or high peak demand.

LEI's quantitative analysis (detailed in Section 3) compares the data center load projections of US jurisdictions with the projected physical supply of global chip manufacturers. If total projections for data center electricity demand exceed the number of semiconductor chips that global manufacturing can supply over the same period, then it is likely that not all the announced or planned data centers incorporated into the US electricity demand forecasts could be built. LEI finds that this is, in fact, the case. For all the data centers announced in the United States for 2025 through 2030 to go forward, it would require 90% of incremental global chip supply for that period be directed to the United States market. This is unrealistic, as the United States currently accounts for less than 50% of global chip demand, and other regions in the world are expanding demand for chips.

[3] Norris, T. H., T. Profeta, D. Patino-Echeverri, and A. Cowie-Haskell. 2025. *Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems.* NI R 25-01. Durham, NC: Nicholas Institute for Energy, Environment & Sustainability, Duke University. <https://nicholasinstitute.duke.edu/publications/rethinking-load-growth>

2 Uncertainties beset data center electric load forecasts

LEI finds that there are many uncertainties based on demand-side factors (on the part of data center developers) and on supply-side factors (on the part of generation-owning electric utilities) that impact the outlook for data center demand growth. These are discussed below, and each uncertainty currently points to the likelihood that industry observers, utilities, and system operators will over-forecast rather than under-forecast growth of data center electricity demand because the potential errors all point in the same direction.

2.1 The data center development business model leads to over-forecasting of electricity demand growth potential

As detailed below, data center developers have a variety of choices as to where to site their facilities and they have incentives to submit multiple requests for service to different utilities. Therefore, one would expect attrition relative to initial requests, and this has recently become evident.

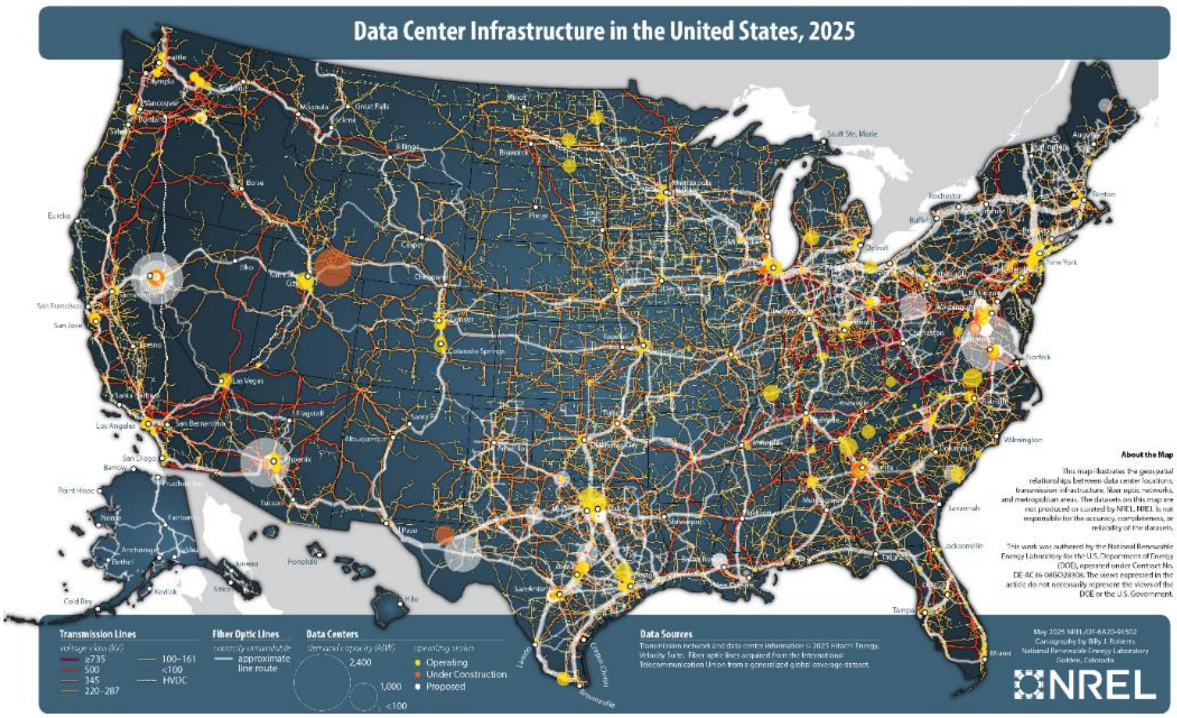
2.1.1 Data center developers have many choices as to where to locate

The design of a data center may be for a specific activity and therefore varies depending on the specific needs of the tenants. But, at its most basic, data center developers must consider certain essential requirements: availability of affordable land, accessibility to fiber capacity, adequate latency (the time delay in data transmission and processing within a network or system), water for cooling (if the design uses water cooling), and high quality and cost-effective electricity service. There are many locations in the United States that meet these criteria and are home to a large and growing data center sector (see Figure 1).

Data center developers are not only interested in the United States. The International Energy Agency ("IEA") estimated that in 2024, US data center total installed capacity amounted to 42 gigawatts ("GW"), while other countries' data center total installed capacity amounted to 55 GW.[4] Edge data centers (data centers which serve low-latency applications like streaming services, and real-time content delivery) need to be close to their customers, so there will be more data centers built around urban centers and areas of high commercial activity around the world.

[4] IEA. *Energy and AI.* April 2025. P. 259. Table A.2. <https://www.iea.org/reports/energy-and-ai> Note that IEA's "installed capacity" refers to the maximum theoretical capacity the facility can support when fully populated with information technology ("IT") equipment and operating at its design limits. This is higher than its "installed IT capacity" which reflects redundancy requirements, operational safety margins, or partial buildouts.

Figure 1. Data centers in the United States as of May 2025



Source: NREL. "Data Center Infrastructure in the United States, 2025." May 2025. <https://docs.nrel.gov/docs/gen/fy25/94502.jpg>

Developers want flexibility to choose between different sites because the facilities can be developed quickly compared to the time it typically takes to develop and build energy generation, transmission, and distribution infrastructure. Data centers are reported to take two to three years to design, permit, and build, though some potential customers are looking to build large data centers in as little as six to nine months.[5] In contrast, the planning, approval, and construction of electric infrastructure such as generation, transmission, and distribution typically involves a rigorous and interdependent process (to ensure reliability, resilience, and system integrity) which can take longer. For example, the median time for clean energy resources in the

[5] Levitt, Ben. "AI and Energy, the Big Picture." S&P Global. December 2024. <https://www.spglobal.com/en/research-insights/special-reports/look-forward/ai-and-energy>; and Bain and Company. "Utilities Must Reinvent Themselves to Harness the AI-Driven Data Center Boom." <https://www.bain.com/insights/utilities-must-reinvent-themselves-to-harness-the-ai-driven-data-center-boom>

United States to move from transmission interconnection request to commercial operation was five years for projects which entered commercial operations in 2023.[6]

2.1.2 Submitting duplicate requests is a low-barrier, low-cost, and low-risk strategy despite efforts to secure commitments

To serve a large load customer, a utility's large load tariff is often paired with a bilateral contract referred to as a customer service agreement ("CSA") or electricity service agreement ("ESA"). These contractual agreements specify a variety of terms and conditions, in addition to the rates that will be applicable under the terms and conditions of the tariffs. A customer who has signed such a contract has "skin in the game" (i.e., some financial exposure) if the contract includes terms that commit the customer to pay for costs at the contracted level of demand requested and/or penalize the customer for not using the level of demand that it initially requested. This creates an incentive for the customer not to overstate the service it will need and supports a more accurate load forecast by the utility.

However, these agreements are only as strong or as weak as the terms of the CSA or ESA, which can vary widely:

- **Minimum demand requirements**: A number of utilities establish demand charges based on a minimum percentage of the customer's contracted capacity. This is essentially a take-or-pay arrangement. The intent is to prevent a customer from requesting more service than it plans to use. This minimum billing demand ranges widely, from 50% to 85% of contracted capacity across several jurisdictions examined by LEI.[7]
- **Contract period:** The length of time the new large customer must commit to its minimum billed demand varies also across jurisdictions, from as little as one year to proposals ranging from 8 to 30 years. A longer contract term increases the cost of walking away from an ESA (and discourages the customer from requesting more service than it will use) because the customer commits to paying its minimum demand charges for longer.
- **Capacity exceedance penalties**: Several jurisdictions include penalties for exceeding the contracted capacity. Such penalties provide a disincentive for a large customer to use more

---

[6] Silverman, A., Wendling, Dr. Z.A., Rizal, K., Samant, D. *Outlook for Pending Generation in the PJM Interconnection Queue.* Columbia University CGEP. May 2024. P. 9. <https://www.energypolicy.columbia.edu/wp-content/uploads/2024/05/PJM-Interconnection-CGEP_Report_042924-2.pdf>

[7] Sources: Dominion Energy: "Schedule 10 Large General Service." Virginia State Corporation Commission. December 20, 2022; "Terms and Conditions Section XXII. Electric Line Extensions and Installations." Virginia State Corporation Commission. March 8, 2024; and "Terms and Conditions Section IX. Deposits." Virginia State Corporation Commission. December 12, 2013; Georgia Power Company: "Customer Renewable Supply Procurement Schedule CRSP-1." Georgia Public Service Commission. October 2023; "Rules Regulations and Rate Schedules for Electric Service," Georgia Public Service Commission. January 2023; "Power and Light Large Schedule PLL-18." Georgia Public Service Commission. April 2025; and GPSC Docket No. 44280. GPC. *Rules & Regulations Update 12-11-2024*. December 2024. <https://psc.ga.gov/search/facts-document/?documentId=220667>'; Indiana Michigan Power Company: "Submission of Unopposed Settlement Agreement and Unopposed Motion for Acceptance of Out of Time Filing." Indiana Utility Regulatory Commission Cause No. 46097. November 22, 2024.

capacity than it has contracted for; or, alternatively, to contract for less capacity than it expects to use.

- **System allocation fee:** A system allocation fee is an upfront deposit required to secure a given level of system (generation and transmission) capacity. This is currently not widely used, though Portland General Electric proposed such a fee in Oregon Public Utility Commission Docket UE-430.

The effectiveness of the terms described above will depend on the specific levels of demand requirements, contract periods, and other details. Nevertheless, even with contract terms that are meant to winnow potential data center requests, such contracts are typically signed **after** a request to interconnect or an announcement of a new data center facility is released. Of course, the developer is aware of the terms of the ESA it will have to sign, but the utility does not require a customer to sign the ESA at the front end. The developer will wait to hear from the utility on what the costs of any necessary new substation expansions would be and the expected in-service date, and only then does the developer decide if it will sign an ESA.

Therefore, in the pre-ESA early stages of the interconnection process, the cost is low to submit multiple requests. A former commissioner at the Texas Public Utility Commission noted that the cost of putting forward an initial request for service is typically minimal: "[t]he phantom load problem arises because the cost of getting in a queue is lower than the weighted likelihood that they'll want to use their position. When it's cheaper to buy a queue position than not to use your queue position, you'll buy queue positions all day long."[8] In addition, potential customers who have requested service or simply announced tentative plans to build a data center (but do not have a significant financial commitment—i.e., have not signed a CSA or ESA), do not have the same level of financial obligation as a customer who has executed a CSA or ESA. Nevertheless, because of the longer timeframe required to plan for and build electricity infrastructure, such early-stage customers (which are often more numerous than the data center customers who have made financial commitments through an ESA) impact utility projections of future electricity load. Such early-stage requests are considered in utility load forecasts, which are then used for generation, transmission and distribution system planning. LEI has observed that utilities increasingly apply weighting factors to early-stage projects, to reflect a lower probability of an early-stage request materializing compared to the probability of a project with an executed CSA or ESA in place. However, with a very short history of such potential customers and their ultimate attrition, utilities must rely on limited information to create such weighting factors.

Therefore, even assuming many data centers are built in the United States (rather than elsewhere in the world), unless a new customer signs a contract, and the contract commits the customer to a specific level of demand and a material financial commitment, there is no guarantee that the data center would be built in any given utility's territory versus somewhere else.

[8] Martucci, Brian. "A fraction of proposed data centers will get built. Utilities are wising up." Utility Dive. May 15, 2025. <https://www.utilitydive.com/news/a-fraction-of-proposed-data-centers-will-get-built-utilities-are-wising-up/748214>

2.1.3 Data center developers submit redundant requests for electric service across multiple jurisdictions

The timing mismatch between data center development and the comparatively slower pace of energy infrastructure buildout, coupled with the typically low cost of submitting a request for service to a utility, creates a strong incentive for potential data center customers to submit multiple, redundant requests for service across several utility jurisdictions.[9] This approach gives developers optionality for negotiating and selecting a site that meets their needs, including timetable for achieving commercial operations. Industry experts have observed that "[D]ata center developers consider multiple states as possible locations for data centers, and they query multiple utilities simultaneously for electricity rates and incentives prior to making a final selection."[10] Meta's former energy strategy director remarked that technology companies themselves are "getting the same project bid into them multiple times," making it difficult to distinguish viable projects from speculative ones.[11]

This dynamic is increasingly evident, including in regional planning processes. As parties to a recent Federal Energy Regulatory Commission ("FERC") docket noted "… PJM has no way to cross-check whether a data center in, for example, Exelon's service territory has also made the same proposal in Dominion's territory, and both proposals end up in PJM's forecast even though only one will be built. … [I]n a recent presentation at the Pennsylvania Environmental Law Forum, PJM's own Senior Manager of Government Services, Stephen Bennett, stated that data center companies 'are pitching the same data centers in different locations.'"[12]

A former Google senior director of software engineering said there are "five to 10 times more interconnection requests than data centers actually being built."[13] Microsoft expressed concerns that "over-forecasting demand from data centers could lead to procuring excessive carbon-intensive generation," and recommended that the Georgia Public Service Commission ("GPSC") only approve near-term resource planning decisions in Georgia Power Company's 2023

---

[9] Giacobone, Bianca March 2025. "Phantom data centers are flooding the load queue." March 26, 2025. <https://www.latitudemedia.com/news/phantom-data-centers-are-flooding-the-load-queue/>

[10] Koomey, Jonathan (Koomey Analytics), Schmidt, Zachary (Koomey Analytics), and Das, Tania (Bipartisan Policy Center). *Electricity Demand Growth and Data Centers: A Guide for the Perplexed*. February 2025. P. 10. </https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2025/02/BPC-Report-Electricity-Demand-Growth-and-Data-Centers-A-Guide-for-the-Perplexed.pdf>

[11] Giacobone, Bianca March 2025. "Phantom data centers are flooding the load queue." March 26, 2025. https://www.latitudemedia.com/news/phantom-data-centers-are-flooding-the-load-queue/.

[12] FERC Docket No. EL25-49-000. Public Interest Organizations. *Comments of Public Interest Organizations in response to PJM Interconnection, L.L.C.'s 03/24/2025 Answer to FERC's 02/20/2025 Order under EL25-49.* April 23, 2025. P. 16-17. <https://elibrary.ferc.gov/eLibrary/filedownload?fileid=9D23F7BC-5484-CAA9-9030-96645FF00000>

[13] Martucci, Brian. "A fraction of proposed data centers will get built. Utilities are wising up." *Utility Dive*. May 15, 2025. <https://www.utilitydive.com/news/a-fraction-of-proposed-data-centers-will-get-built-utilities-are-wising-up/748214/>

Integrated Resource Plan ("IRP") Update based primarily on "known, mature projects that have made firm commitments to Georgia Power."[14]

The problem reverberates beyond electric utilities. Because of concerns over potential duplication of requests, executives in the US natural gas industry have publicly voiced skepticism for growth in gas demand from data center electric power customers. The Vice President of New Ventures for pipeline company The Williams Companies ("Williams") noted at an industry event: "… if you look at how these [data center] projects are coming into different organizations, there is double and triple [counting] … it is the same project because you have different players that are developing pieces."[15] "It's creating a lot of problems for these regulators and utilities because how do you differentiate between a real project and a fake project?" the president of a shale gas producer remarked at the same event, adding that he expects only 10% of data center projects that have been announced to be built.[16]

2.1.4 Expected attrition is in evidence

With multiple requests for service and the implied intent of developers to ultimately build only a subset of the proposed data centers, some attrition among large load projects should be expected (and therefore warrant downward adjustment to current electricity demand growth prospects). There is already evidence of this attrition. A recent industry report noted that Microsoft put on hold three data centers in Ohio (a total of $1 billion in spending), as well as portions of a data center in Wisconsin.[17] Microsoft is reported to be slowing down or pausing projects, including those for which it has "secured energy and all necessary approvals"[18] though this distinction does not necessarily apply to the Ohio or Wisconsin projects (in other words, Microsoft did not report whether or not the Ohio and Wisconsin projects had already secured energy and necessary approvals). The President of Microsoft Cloud Operations noted in April 2025, "[b]y nature, any significant new endeavor at this size and scale requires agility and refinement as we learn and grow with our customers. What this means is that we are slowing or pausing some early-stage projects. While we may strategically pace our plans, we will continue to grow strongly and

---

[14] GPSC Docket No. 55378. *Comments on Georgia Power's 2023 Integrated Resource Plan Update*. Microsoft. Docket No. 55378 at 1, 4. Apr. 1, 2024. <https://psc.ga.gov/search/facts-document/?documentId=218199>

[15] Energy Intelligence. "US Gas Companies Temper Data Center Demand Expectations." *Natural Gas Week*, Vol. 41, No. 11. March 14, 2025.

[16] Ibid.

[17] Mannion, Annemarie. "Microsoft Hits Pause Button on $1B in Data Centers in Ohio." *Engineering News-Record*. April 10, 2025. <https://www.enr.com/articles/60572-microsoft-hits-pause-button-on-1b-in-data-centers-in-ohio#:~:text=By%20Annemarie%20Mannion,contractors%20had%20not%20been%20announced>

[18] Patel, Dylan; Jeremie Eliahou Ontiveros; Maya Barkin. "Microsoft's Datacenter Freeze - 1.5GW Self-Build Slowdown & Lease Cancellation Misconceptions." *SemiAnalysis*. April 28, 2025. <https://semianalysis.com/2025/04/28/microsofts-datacenter-freeze/>

allocate investments that stay aligned with business priorities and customer demand."[19] This statement seems intended to provide Microsoft with the flexibility to pause investment as and where needed, without alarming its shareholders.

Data center developers face their own uncertainties and may be evaluating changes in their market environment, some of which may be contributing to a slowdown in commitments. In addition to long waits for interconnection, these include, but are not limited to, uncertainty about the impact of future artificial intelligence ("AI") models that are more energy-efficient and cheaper to train and run than existing AI models; the US administration's proposed $500-billion AI-powered Stargate project; and concerns over tariffs and the cost of inputs.[20]

2.2 Utilities and shareholders benefit from strong load growth

Although utilities have some processes in place to reduce the number of "phantom load" requests by data centers—by assigning probabilities to requests for service and requiring varying levels of monetary commitments in CSAs and ESAs— a vertically integrated utility under traditional cost-of-service regulation expects to benefit financially from making investments that expand its assets and contribute to increases in rate base. As detailed below, investments in assets to support service to new customers will increase rate base, which in turn increases allowed returns under a cost-of-service regulatory model. The utility and its shareholders therefore benefit from an expanded rate base, with only some of the risk of under-utilization of assets borne by shareholders (if the investment is deemed imprudent), and the rest of the risk borne by the utility's customers.

2.2.1 Utility investments are ultimately paid for by customers

A public utility has an obligation to serve customers in its territory but cannot jeopardize the reliability of the service it provides to existing customers to take on new customers. Therefore, the utility periodically performs forward-looking load studies, which must incorporate assumptions about the interest of new potential customers, and the load growth which will ultimately materialize. A vertically integrated electric utility (which owns generation, transmission, and distribution assets) will plan capacity expansions of its system to meet this expected load. If the utility is regulated under a cost-of-service model, and the investment is deemed prudent based on available information, then the utility will be allowed to recover the costs of such new infrastructure in its rates. The costs will include a regulator-approved return on equity ("ROE") on the rate base, which contributes to the profits that the utility and its

---

[19] Noelle Walsh, President Microsoft Cloud Operations and Innovation | Nouryon, Non-Executive Director. <https://www.linkedin.com/posts/noelle-walsh-b29356108_microsoftcloud-datacenters-activity-7315439628562423808-W67e/>

[20] Williams, Kevin. "AI data center boom isn't going bust but the 'pause' is trending at big tech companies." CNBC. April 27, 2025. <https://www.cnbc.com/2025/04/27/ai-data-center-boom-isnt-going-bust-but-the-pause-is-trending.html>

shareholders earn. The cost of the investments as well as the ROE are paid for by the utility's customers.

While the utility expects to benefit financially from the investments it must make to serve new customers, it still faces some investment risk. This risk to the utility may come in the form of prudency reviews and/or performance incentives or multi-year plans whereby a utility will face regulatory lag and may not be able to adjust its rate base immediately. These practical implications temper the zeal to build assets too quickly, though the utility still has a fundamental motivation to expand investment in its system.

On balance, then, the utility does not have a strong incentive to be skeptical regarding forecasted electricity load growth (whether from data centers or other types of customers) given the opportunity to expand its rate base. Nevertheless, utilities often perform analyses in which they adjust the total requests for service of prospective large load customers such as data centers for the possibility that some load will not materialize. However, experience with attrition of large load customers is still limited, so difficult for a utility to predict.

2.2.2 Risk of over-forecasting may impact ratepayers more than the utility

Large incremental capital expenditures for large load customers can impact existing ratepayers as a result of several sources of risk:

- **The first risk stems from the lack of spare generation capacity and/or transmission system headroom in some utility systems in the United States.** Such utility systems would need new investment to serve new large loads such as data centers. New investment in system generation and/or other network resources (transmission) would cause rates to rise for all customers (but for the new large load, an investment in transmission or generation could be deferred for some time and therefore current customers would not have seen their rates increase). If a system has spare generation and transmission capacity, it can add new customers without a large increase in rates (all else equal), but as noted above, some utility systems do not have spare capacity. Under some circumstances a new large load customer could benefit other ratepayers, if it materializes and is sustained. Large loads increase energy consumption and demand over which to recover certain fixed costs. In addition, large load customers can create economies of scale for the electric system, because the utility may be able to use larger transformers with a lower cost per unit of capacity. As the new infrastructure investment made to increase capacity to serve a large load customer comes online, the average cost to serve customers can decline both because the larger investments cost less per unit, and fixed operating costs can be spread among a larger base of customers. So, it is possible that large loads can, under certain circumstances, lead to a lower dollar rate per unit of energy consumed or per unit of peak demand. However, these benefits are contingent on the load materializing and remaining on the system long enough to offset the initial investment cost, and on the relative cost of the new investment(s) versus the embedded cost of existing infrastructure.
- **The second risk stems from the uncertain permanence of new large loads.** Some data centers do not require as much complex fixed investment as, for example, an auto

manufacturer, so not only can a data center be built quickly, it can also relocate quickly. If the data center leaves the system, and the system investments related to new generation plants and transmission lines were triggered by that data center's load, the recovery of the fixed costs of those new assets will be shifted to remaining customers. This also happens if, for example, population declines and/or households migrate—but because data centers are such large users of electricity, the loss of single data center customer will have a disproportionate impact.

- **A third risk is the shorter expected life span of a data center as compared to the operational life and typical cost recovery period for electricity infrastructure.** Even if the data center does materialize as proposed and does not re-locate, data centers have shorter operational lifespans—approximately 15 years[21]—compared to the energy facilities built to serve them. The economic life of power assets is about 30 to 40 years for gas generation, 30 years for transmission and distribution facilities, and 35 years for natural gas pipelines.[22] Assuming data center facilities are built and the electricity demand materializes as planned, the data center customer may not be part of the system for the entire lifetime of the transmission and generation assets that were built by the utility to provide service. This mismatch in infrastructure lifetimes versus economic life of customer facilities means that some burden may be shifted to other customers.

2.3 Other factors indicate that growth in data centers may be weaker than implied by current announcements

Apart from evidence of multiple requests for service, there are other factors that suggest that electricity demand from data centers will fall well below the levels implied based on review of requests for service or announced plans of data centers.

2.3.1 Equipment bottlenecks limit the pace of development of gas-fired generation resources

Tight supplies of equipment for the next few years could become a bottleneck to serving potential new load from data centers, particularly with natural gas-fired generation. Siemens and GE Vernova, two industry leaders in gas turbine manufacturing, are seeing lead times increase for

---

21 Data Center Knowledge. "Hyperscalers in 2024: Where Next for the World's Biggest Data Center Operators?" February 28, 2024. <https://www.datacenterknowledge.com/hyperscalers/hyperscalers-in-2024-where-next-for-the-world-s-biggest-data-center-operators->

22 Gas generation: Florida Power & Light, Power Generation Division. "Depreciation Analysis for Power Generation." February 2016. <https://www.floridapsc.com/library/filings/2016/07554-2016/Support/OPCs%201st-38-Attachment%203.pdf>; Transmission and Distribution: Thomson Reuters Onvio. MACRS asset life table. Item 49.14, Electric Utility Transmission & Distribution. <https://onvio.us/ua/help/us-en/staff/fixed-assets/depreciation/macrs-asset-life-table.htm>; Pipelines: Thomson Reuters Onvio. MACRS asset life table. Item, 49.21, Gas Utility Distribution Facilities <https://onvio.us/ua/help/us-en/staff/fixed-assets/depreciation/macrs-asset-life-table.htm>

turbine delivery; the lead time for a 200 MW Siemens turbine has increased to three years,[23] and GE Vernova has a company backlog through 2028 for delivery times on orders.[24] Unless they have already put orders in, utilities and independent power producers may face not only delays but also higher prices in bringing new gas-fired power plants online in the next few years. The CEO of NextEra Energy (one of the largest power generation companies in the United States) recently noted that "We built our last gas-fired facility in 2022, at $785 [per kilowatt ("kW")]. If we wanted to build that same gas-fired combined cycle unit today [the cost would be] $2,400/kW," continuing, "[t]he cost of gas-fired generation has gone up three-fold."[25] Timing is also an issue, as he added, "[w]hen you look at gas as a solution…you're really looking at 2030 or later." Industry observers also report an average lead time of three years for delivery of large transformers (a key part of many transmission and distribution projects); and in some cases, lead times as long as five years.[26]

Until the utility places orders for necessary equipment it needs to procure and a notice to proceed ("NTP") on construction of substation or other required facilities is authorized, there is no clear timeline of when the utility can serve the new load. If the receipt of necessary transmission, distribution, or generation equipment is delayed, then bringing on new customers will, naturally, have to be deferred.

2.3.2 Some state authorities are re-examining tax incentives

Most states offer tax incentives to attract data centers. This is intended to boost economic development driven by the employment opportunities at data centers, and the indirect impacts on the local economy for services that support data centers and their workers. However, a recent study found that in some cases the cost to the state treasury exceeded the benefits to the state's economy, and some states have begun to re-examine the value of such tax incentives.[27] In Georgia, the state legislature passed a bill in 2024 halting tax breaks to data centers for two years, though

[23] Malik, Naureen S. "Gas Power Won't Provide an Easy Fix for AI Boom." *Bloomberg*. January 8, 2025. <https://www.bloomberg.com/news/newsletters/2025-01-08/gas-power-won-t-provide-an-easy-fix-for-ai-boom>

[24] Casey, Simon. "GE Vernova CEO Sees Order Backlog Stretching Into 2028." *Bloomberg*. March 11, 2025. <https://www.bloomberg.com/news/articles/2025-03-11/ge-vernova-ceo-sees-order-backlog-stretching-into-2028?utm_source=chatgpt.com&embedded-checkout=true>

[25] Cunningham, Nicholas. "Costs to build gas plants triple, says CEO of NextEra Energy." Gas Outlook. March 25, 2025. <https://gasoutlook.com/analysis/costs-to-build-gas-plants-triple-says-ceo-of-nextera-energy/>

[26] Seiple, Chris. "Gridlock: The demand dilemma facing the US power industry." *Wood Mackenzie*. October 2024. <https://www.woodmac.com/horizons/gridlock-demand-dilemma-facing-us-power-industry/>

[27] LeRoy, Greg and Tarczynska, Kasia. *Cloudy with a loss of Spending Control*: *How Data Centers Are Endangering State Budgets*. goodjobsfirst.org. April 2025. </https://goodjobsfirst.org/wp-content/uploads/2025/04/Cloudy-with-a-Loss-of-Spending-Control-How-Data-Centers-Are-Endangering-State-Budgets.pdf>

the bill was later vetoed by the Governor. The Governor's Office of Planning and Budget expects the tax breaks to cost the state $327 million in 2025.[28]

As the data center industry grows, so does the cost of providing tax incentives. In Texas, the sales tax exemption program for data centers cost the state $157 million in 2023; the state estimates the cost will increase by an order of magnitude, to $1 billion in 2025.[29]

2.3.3 Local opposition has derailed and/or delayed data center projects

Industry observers report that a large number of proposed data centers have been cancelled or delayed, owing to local opposition.[30] Residents object to local impacts such as noise, and fears over water resource depletion and pressure on electricity rates. Local authorities may deny permits because of such concerns.

2.3.4 Data center demand flexibility may reduce peak demand

Data centers serve a variety of users and applications. Certain types of data centers can temporarily reduce their energy consumption and may do so to avoid high prices during periods of surging demand, similar to how some other industrial customers of electricity act under demand response programs. Some data centers have flexibility in their operations and can provide demand response. They can lower energy consumption from the grid by reverting to on-site generation, and/or shifting computational work to other locations, or even reducing operations.[31] These strategies may be limited, however. On-site generation is subject to local environmental laws, where diesel (which is used in back-up generation) may only run a few hundred or less hours per year because of emissions limitations. In co-tenant sites, leases often have pass-through clauses where utility costs are passed on to tenant, so that the landlord has limited incentive to be part of a utility demand response program.

To incentivize a data center customer to provide a demand response when needed, a utility can offer it a tariff tailored to reflect the value of such flexibility, and/or it can offer a time-of-use

---

[28] Chow, Andrew R. "Why Tax Breaks for Data Centers Could Backfire on States." April 25, 2025. <https://time.com/7280058/data-centers-tax-breaks-ai/>

[29] LeRoy, Greg and Tarczynska, Kasia. *Cloudy with a loss of Spending Control*: *How Data Centers Are Endangering State Budgets*. goodjobsfirst.org. April 2025. </https://goodjobsfirst.org/wp-content/uploads/2025/04/Cloudy-with-a-Loss-of-Spending-Control-How-Data-Centers-Are-Endangering-State-Budgets.pdf>

[30] Eddy, Nathan. "Local Opposition Hinders More Data Center Construction Projects." Data Center Knowledge. May 15, 2025. <https://www.datacenterknowledge.com/regulations/local-opposition-hinders-more-data-center-construction-projects>

[31] See for example: Bloomberg News. "How 'Load Shifting' May Help Improve Data Center Sustainability." *Data Center Knowledge*. February 26, 2024. <https://www.datacenterknowledge.com/sustainability/how-load-shifting-may-help-improve-data-center-sustainability>; and Judge, Peter. "Google tests system to cut data center power use during grid problems." *Data Center Dynamics*. October 5, 2023. <https://www.datacenterdynamics.com/en/news/google-tests-system-to-cut-data-center-power-use-during-grid-problems/>

("TOU") tariff which automatically incentivizes shifting load away from peak hours. This flexibility could reduce the level of expansion of generation resources a utility would need to meet load growth.

3 US data center load growth projections cannot be met by global chip supply

The data center boom in the United States is part of a larger global phenomenon, but that global phenomenon will ultimately be constrained by the supply of intrinsic components that data centers require, most critically semiconductor chips. By comparing the load projections from a set of US jurisdictions with available data (covering 77% of US electricity generation and demand) with the physical supply constraints of global chip manufacturers, LEI sought to assess the credibility of the total US-wide data center demand outlook. LEI hypothesized that if the sum of data center demand projections for these jurisdictions of the United States electricity market exceeds the global capacity to manufacture semiconductor chips over the same period, then it is likely that not all the announced or planned data centers incorporated into those US electricity demand forecasts could be built. If not all announced or planned data centers in the US jurisdictions surveyed could be built, then some of the implied attrition would impact specific states.

LEI developed an approach to examining the reasonableness of data center forecasts by working backward from projections for data center electricity demand to the implied need for semiconductor chips, the primary electricity-using component in a data center. The purpose of this analysis is not to forecast the number of AI-related chips ordered or chip production capacity in the future. Instead, it is meant to be a sanity check on the scale of projected data center growth and by extension the forecast of the aggregated electricity demand growth reported by the US regional transmission organizations ("RTOs"), independent system operators ("ISOs"), and balancing authorities ("BAs"). In other words, if the aggregated US RTOs', ISOs', and BAs' data center demand forecasts were to materialize, how many semiconductor chips would this require, and is it reasonable to expect the needed chip-making capacity would materialize and be available to meet that forecasted demand?

As explained below, LEI found evidence that the high-level forecast is not credible. LEI found that, even if global AI semiconductor chip manufacturing were to grow at an average rate of 10.7% annually (significantly faster rate than the 6.1% growth rate over the past decade)[32] it could only satisfy an incremental 63 GW of data center-related demand ***globally*** over the six years from 2025 to 2030. The implied RTO/ISO/BA demand projection, detailed below, is for ***US-only*** data center demand growth of 57 GW over the six years from 2025 to 2030. This would amount to more than 90% of the new total global manufacturing capability being earmarked for US data centers. In other words, for the forecasts to hold water, this would mean the US would require more than 90% of the world's new supply of semiconductor chips from 2025 through 2030.

Such a scenario is unrealistic. The United States currently accounts for slightly less than 50% of global semiconductor chip demand, and other regions in the world are also projecting strong

[32] Semiconductor Industry Association. "Strengthening the U.S. Semiconductor Supply Chain." May 2024. P. 2. Figure 1. <https://www.semiconductors.org/wp-content/uploads/2024/05/Emerging-Resilience-in-the-Semiconductor-Supply-Chain_SIA-Summary.pdf>

demand for semiconductor chips to support their region's growth in data centers.[33] LEI concludes that over-forecasting data center growth and therefore the associated electricity demand is significant at the national level in the United States.

As described in this section, LEI focused its analysis on AI semiconductor chips. While most current data centers in the US are not used for AI applications, most future data centers, especially those with high electricity demand, will be used for AI.[34]

3.1 LEI's approach compares demand outlooks to the global supply of chips

LEI's approach worked backward from projections for data center electricity demand to the implied need for semiconductor chips—the fundamental elements in the data center supply chain. LEI's approach followed a four-step process (see Figure 2):

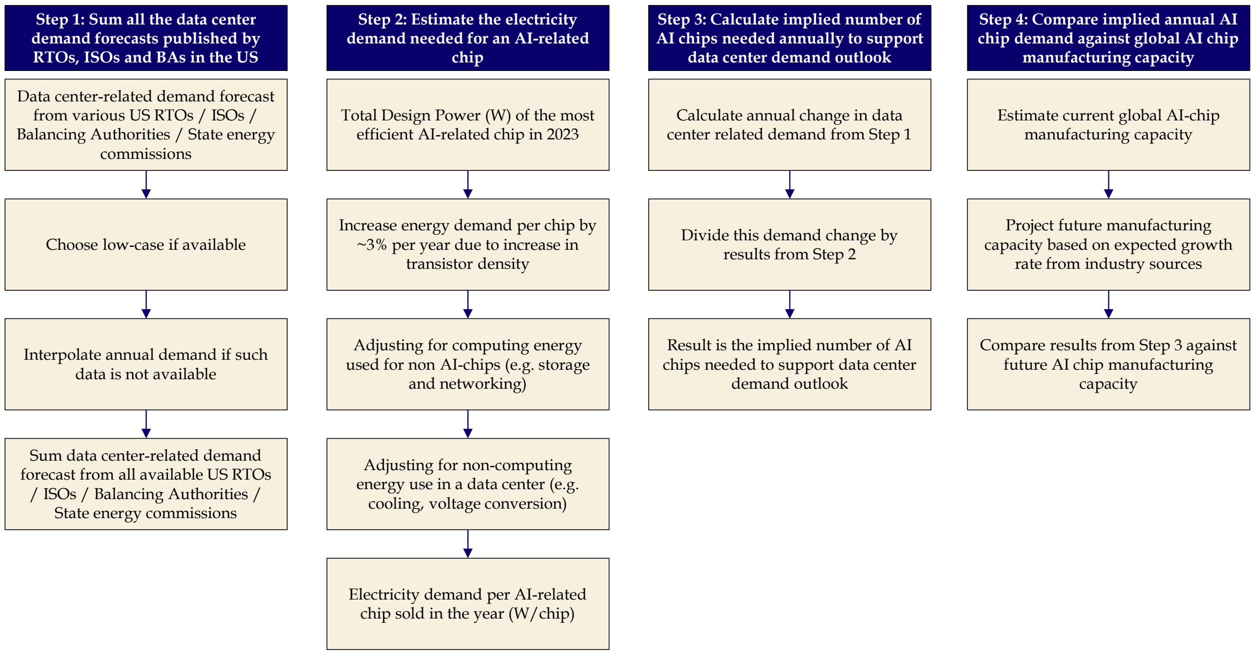
- First, LEI summed the data center demand forecasts published by the RTOs, ISOs, and BAs in the United States that could be obtained. LEI took a conservative approach to this, meaning that LEI referred to electricity demand growth based on the slower rather than faster end of the range offered by each jurisdiction, if the jurisdiction provided a range of electricity demand growth forecasts rather than a single forecast. Slower growth is more conservative in that it is more likely to be met by the global supply of semiconductor chips and less likely to be constricted by semiconductor chip shortages.
- Second, LEI estimated the electricity needed to power an AI-related chip. LEI also accounted for energy used for non-computing activities, like cooling (especially important in warm climates such as in the southeastern United States). LEI made assumptions as to the use of energy for non-computing activities, and the rate of efficiency improvement of semiconductor chips. Efficiency and computing capabilities are discussed in more detail in Appendix A.
- Third, LEI calculated the implied number of semiconductor chips needed annually based on the data center demand outlook from Step 1.
- Fourth, LEI compared this annual demand for semiconductor chips to global chip manufacturing capacity (including expansion of that capacity). Where LEI had to make assumptions, for example, as to the rate of annual expansion of chip manufacturing capacity, LEI made conservative assumptions—in this case, assuming that chip manufacturing growth would be rapid (to allow fast growth in data centers).

---

[33] IEA. *Energy and AI*. April 2025. P. 259. Table A.2. <https://www.iea.org/reports/energy-and-ai>

[34] See, for example, 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. P. 31.

Figure 2. LEI's approach to examining reasonableness of data center load forecast by selected US RTOs, ISOs, and utilities



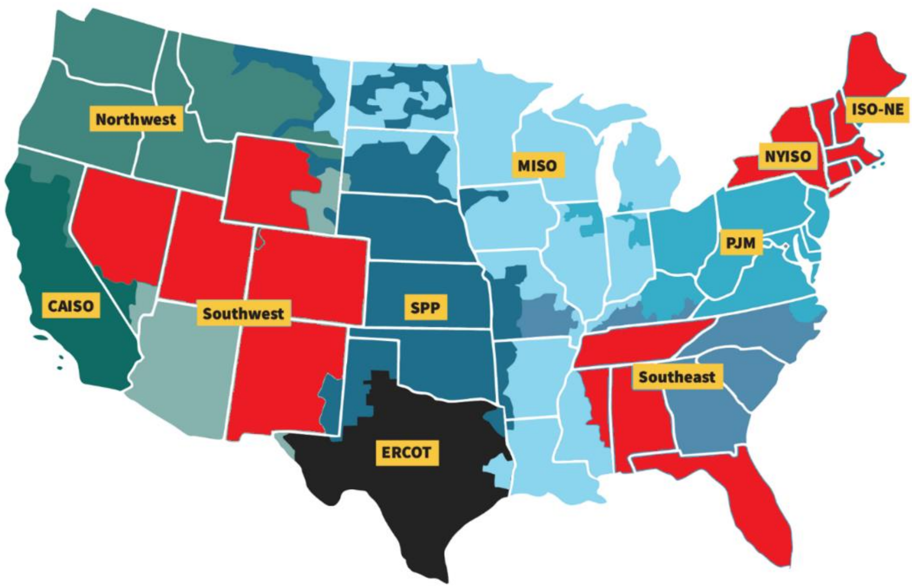
The following sections detail each of the steps of LEI's methodology.

3.1.1 Step 1: Aggregate available load forecasts across US RTOs, ISO, and BAs

First, LEI summed all the available data center demand forecasts published by the RTOs, ISOs, and BAs in the United States that LEI was able to obtain from public sources. Data was available for Midcontinent Independent System Operator ("MISO"), Electric Reliability Corporation of Texas ("ERCOT"), Southwest Power Pool ("SPP"), California Energy Commission ("CEC"), the Northwest Power and Conservation Council ("NPCC"), Arizona Public Service, PJM (AEP, APS, PSEG, and Dominion), North Carolina and South Caroline (Duke Energy Carolinas and Duke Energy Progress), and Georgia Power Company ("GPC"). LEI was unable to obtain publicly available forecasts for several jurisdictions (New England, New York, parts of the Southeast United States, and parts of the Western United States) (see Figure 3). LEI's aggregation accounts for 77% of net summer electric capacity in the United States.[35]

[35] Based on data from the Energy Information Administration, LEI's jurisdictions omit 23.49% of US net summer capacity (MW) for 2023; they omit 23.55% of US consumption (MWh) for 2023.

Figure 3. US RTOs and ISOs—jurisdictions in red are not covered in LEI's analysis



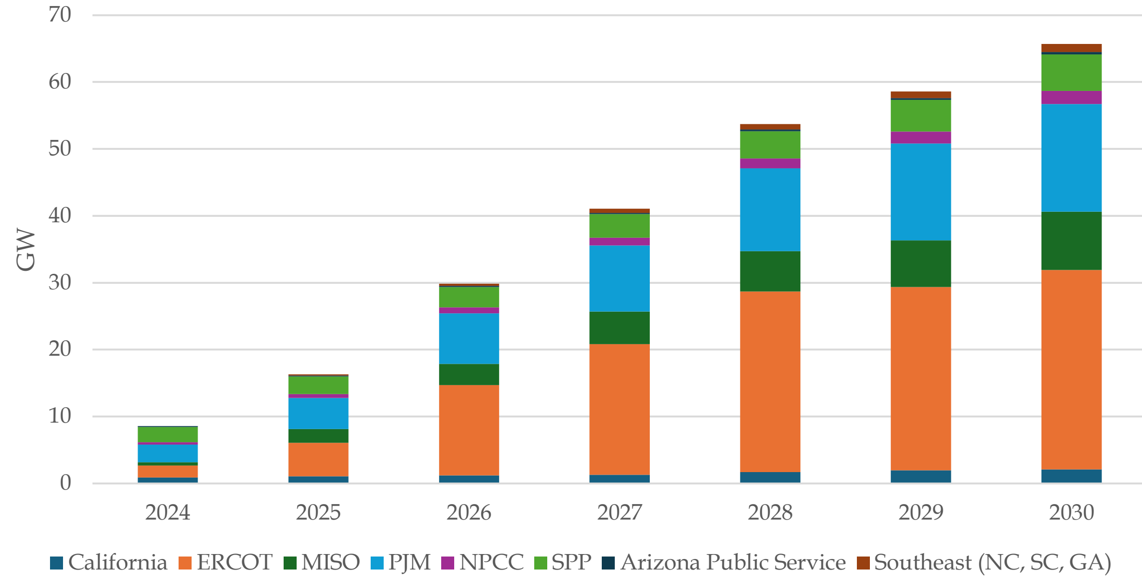
Note: LEI was not able to obtain specific data center-related demand forecast in the latest published integrated resource plans or load forecasts in the jurisdictions colored in red. Therefore, the red colored areas, which represent approximately 23% of US electricity capacity, are not included in LEI's data center-related demand calculations. The other colors indicate RTO or BA boundaries.

The forecasts from the jurisdictions differed in issue dates and in periodicity. If a forecast was not available for every year, LEI interpolated using the cumulative average growth rate ("CAGR") between two years. If the forecasts had an outlook timeframe shorter than required, LEI assumed data center-related load growth would be 50% of the change in demand (in GW terms) between the last two years of the available forecasts. For example, if an electricity demand forecast ended in 2029, with data center-related demand being 7 GW in 2028 and 8 GW in 2029, then LEI extended the forecast to 2030 by adding 50% of the growth between 2028 and 2029 (i.e., 50% of 1 GW which is equal to 0.5 GW) to the projected data center-related electricity demand in 2029 (8 GW), resulting in a 2030 projection of 8.5 GW. This approach is conservative because it uses a slower growth rate than in the prior years forecasted by the jurisdiction.

Applying this approach to US data center-related demand to the 77% of the US electricity sector with available data results in an electricity demand of somewhat less than 9 GW in 2024. This is the total aggregation based on LEI's methodology described in Step 1, and because it includes only 77% of the US sector, it is a conservative (low) estimate for the US overall.

Using the same methodology for the ensuing years results in electricity demand of 65 GW by 2030 (see Figure 4).

Figure 4. RTO/ISO/BA data center-driven electricity demand outlook*



*US RTOs and utilities with data available as noted in the text—this covers 77% of the US power sector by capacity and consumption.

Note: Forecasts for California, MISO and NPCC are low cases. NPCC's forecast is only available until 2029, therefore LEI assumed 2030 demand would be 2029 demand plus 50% of the increment of data center-related demand growth from 2028 to 2029.

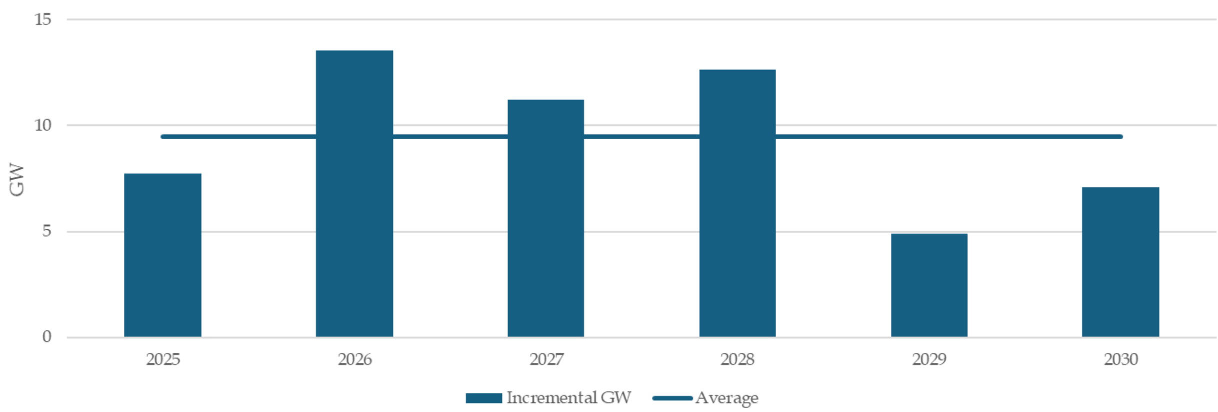
For some utilities, although the demand outlook may be small compared to other regions, the share of forecasted data center demand could be large compared with electricity demanded by existing customers. This adds risk to the smaller utilities' customers, because a potential buildout of new infrastructure to meet demand (some of which might not materialize) has fewer customer billable units over which to spread the cost. All else equal, risk to customers will be higher in jurisdictions with vertically integrated utilities where investments in generation assets are also paid for by customers. By contrast, in some deregulated wholesale electricity markets, some of the risk of not recovering capital expenditures stays with the generation owners (rather than being backstopped by the utility's customers).

In terms of annual increments to projected load growth (the conservative outlook shown in Figure 4 above), the growth from 2025 to 2030 is uneven but averages 9.5 GW per year (see Figure 5). Again, this is a conservative (low) estimate for the US overall, because it only includes 77% of the electricity sector, and because it is based on the low-end forecasts provided by the RTOs, ISOs, and BAs, as noted previously.

London Economics International LLC
717 Atlantic Avenue, Suite 1A
Boston, MA 02111
www.londoneconomics.com

contact:
Marie Fagan/Victor Chung
617-933-7205
marie@londoneconomics.com

Figure 5. Total annual incremental electricity demand implied by RTO/ISO/BA US data center outlook



To reiterate, LEI ensured its aggregation of US demand forecasts is conservative because:

- LEI used the lowest data center demand estimate from the forecasting entity if multiple outlooks were available (this applied to California, MISO, and NPCC);
- US coverage is not complete. The available data covers 77% of the market as noted above, because not all utilities or BAs report electricity demand growth associated with data centers. Data center-related demand forecasts for Colorado, Florida, Utah, New Mexico, the area served by Tennessee Valley Authority (including Tennessee and parts of Alabama, Georgia, Kentucky, and Mississippi), New York state (for New York Independent System Operator ("NYISO")), and the six states covered by ISO-NE, were not publicly available at the time LEI performed this analysis; and
- Not all utilities within regions that LEI included provided data center-related load growth forecasts. For example, the parts of Arizona that are not covered by Arizona Public Service are not included. Data centers may be developed in these regions in the future,[36] which will create demand for AI-related chips that are not included in LEI's total. Therefore, LEI's analysis is conservative in relation to the hypothesis being tested.

3.1.2 Step 2: Estimate the electricity demand per AI chip and other uses

The data center energy demand forecasted by the US utilities, RTOs, and ISOs comes from expectations that new data centers will be built, which implies that new AI-related semiconductor

[36] For example, under the provisions of Tennessee Code Ann. Section 67-6-206(c), qualified data centers may purchase electricity at a reduced rate of sales tax, and under Section 67-6-102, qualified data centers are also eligible for sales tax exemption on certain computers and devices. This is intended to incentivize data centers to locate in Tennessee and would create data center-related demand for Tennessee Valley Authority that is not included in LEI's analysis.

chips must be purchased and installed. Fundamentally, the world must therefore be able to produce sufficient AI semiconductor chips for this forecasted demand to materialize.

To calculate the number of AI semiconductor chips needed to meet the energy demand forecast, LEI first quantified the energy need of AI semiconductor chips. To estimate energy use per AI semiconductor chip, LEI made the conservative assumption that a falling number of AI semiconductor chips will be needed over time to drive the same MWs of data center-related demand. While the energy efficiency of AI-related semiconductor chips measured in floating point operations per second per watt ("FLOPS/W") has improved over the past decade, each AI-related semiconductor chip consumes more energy because the density of transistors has increased. From the first generation of AI-related semiconductor chips in 2016[37] to the most recent generation in 2024,[38] the total design power ("TDP") (in watts ("W")) per square millimeter ("mm2") of silicon has increased at an annual rate of 3.04% (from 0.49W per mm2 to 0.63W/mm2). Assuming the trend of increasing energy consumption per chip size continues over the forecast period and holding each AI-related semiconductor chip's silicon size the same,[39] the same MW of data center-driven demand growth over time will require fewer AI-related semiconductor chips.

LEI assumed that 41% of demand for energy from data centers is consumed for non-computing activities, and 59% is used for computing. This is based on the low end of Lawrence Berkeley National Laboratory's ("LBNL") estimation of 59% to 76%,[40] while the remaining energy would be consumed for other equipment such as data storage and networking. LEI's assumption is conservative because the smaller share of energy used for AI-related computing means a smaller implied number of chips would be required per MW of data-center related load.

After applying these calculations, LEI arrived at an energy consumption per AI-related semiconductor chip that starts at a little over 1,500 W in 2025 and rises to over 1,750 W by 2030 (see Figure 6). The increase is driven by an increasing density of transistors per surface area of a chip, which is discussed in more detail in Appendix A.

---

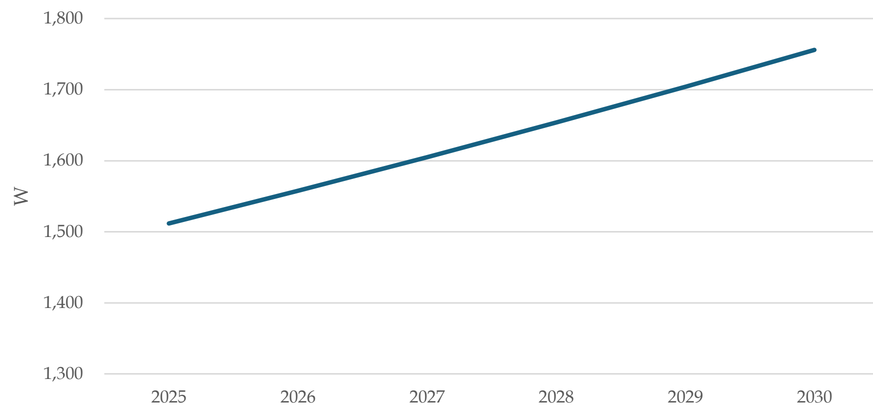
[37] Nvidia's P100 DGXS, launched in April 2016. <https://www.techpowerup.com/gpu-specs/tesla-p100-dgxs.c3285>

[38] Nvidia's B200 SXM, launched in 2024. <https://www.techpowerup.com/gpu-specs/b200-sxm-192-gb.c4210>

[39] The global supply of semiconductors is measured by the number of specific sized wafers that can be produced over a period of time. This means the size of the AI-related chip does not impact the overall limitation on chip output – if the AI-related chip is bigger, it means a smaller number of chips can be produced over a period of time, and vice versa.

[40] 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. P. 53. Figure 5.6.

Figure 6. Estimated energy demand (W) per AI-related chip

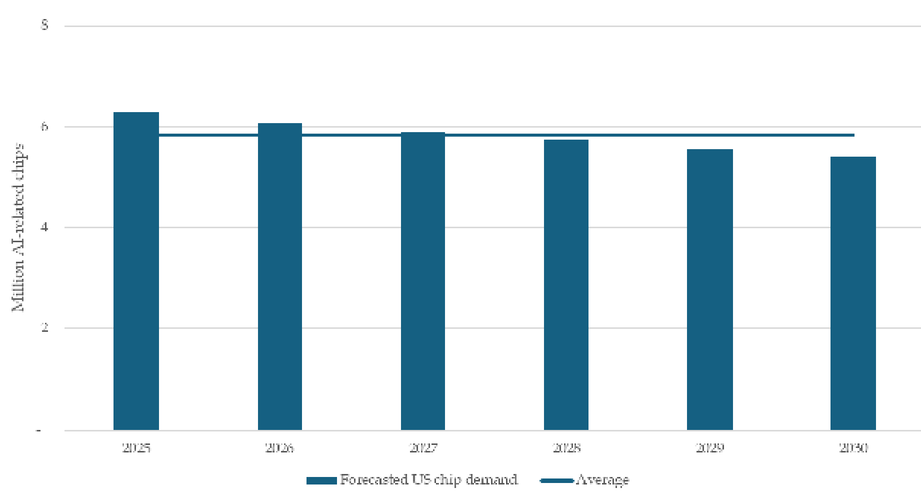


Note: Adjusted for power usage efficiency ("PUE") of 1.2. PUE refers to the share of energy consumption for computing versus other data center usage, and is discussed in detail in Appendix A.
Source: LEI analysis.

3.1.3 Step 3: Estimate the number of chips implied by aggregated data center demand forecasts

Dividing the annual data center-driven load growth from Step 1 by the energy demand impact of an AI-related semiconductor chip results in the number of new semiconductor chips which need to be installed in the US each year if the data center electricity demand outlook for RTO/ISO/BAs is to be achieved. This demand averages 5.8 million chips per year for the outlook period (see Figure 7).

Figure 7. Total annual incremental AI chip demand implied by RTO/ISO/BA US data center outlook



3.1.4 Step 4: Estimate global production capacity based on industry growth

Will the world have enough production capacity for AI-related chips to meet this demand?

LEI began by estimating the baseline annual production capacity by examining the largest individual chip producers. Nvidia held a 98% market share in 2023, shipping 3.76 million units of data-center graphic processing units ("GPUs"). Including the other two main suppliers, Advanced Micro Devices ("AMD") and Intel Corporation ("Intel"), data center-related chip shipments in 2023 totaled 3.85 million.[41] Note that this 3.85 million is already much less than the projected need of the RTO/ISO/BA forecast for over 6 million in 2025—which indicates that even near-term outlooks for data center demand are too high.

However, the chip manufacturing industry has been growing and LEI examined its prospects for further growth. Industry observers have noted that the fast growth seen in 2024 and 2025 (over 15% annually) is likely to slow somewhat for the longer term:

- Deloitte reported that the semiconductor industry had a robust 2024 with growth of about 19% in sales revenue, and it expects this to slow somewhat for 2025, with sales revenue growth projected at about 11%.[42] PricewaterhouseCoopers forecasts that the logic product sector (used for AI and cloud computing) of the global semiconductor industry would grow by a CAGR of 6.28% from 2024 to 2030.[43]
- The Semiconductor Industry Association projects that global semiconductor capacity (as measured by wafer starts per month) will increase by 108% from 2022 to 2032, implying a 7.6% CAGR.[44]
- IDC, a consultancy, forecasts growth in worldwide foundry (semiconductor chip manufacturing facility) market revenue of about 10% from 2026 through 2028.[45]

---

41 Shah, Agam. *Nvidia Shipped 3.76 Million Data-center GPUs in 2023, According to Study*. June 10, 2024. <https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/>

42 Kusters, J., Bhattacharjee, D., Bish, J., Nicholas, J.T., Stewart, D., Ramachandran, K. "2025 global semiconductor industry outlook." Deloitte. February 4, 2025. <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-telecom-outlooks/semiconductor-industry-outlook.html>

43 PricewaterhouseCoopers, *State of the Semiconductor Industry*. November 28, 2024. P. 3. Exhibit 1. <https://www.pwc.com/gx/en/industries/technology/state-of-the-semiconductor-industry-report.pdf>

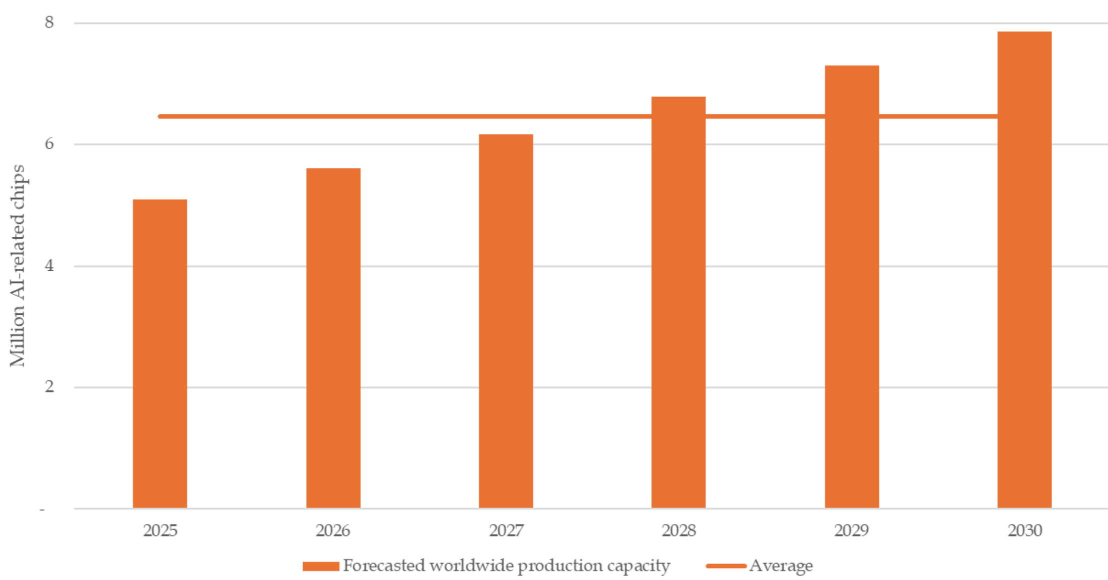
44 Semiconductor Industry Association. "Strengthening the U.S. Semiconductor Supply Chain." May 2024. <https://www.semiconductors.org/wp-content/uploads/2024/05/Emerging-Resilience-in-the-Semiconductor-Supply-Chain_SIA-Summary.pdf>

45 IDC. "Global Semiconductor Market to Grow by 15% in 2025, Driven by AI." December 12, 2024. <https://my.idc.com/getdoc.jsp?containerId=prAP52837624>

- Another industry source projects that the global market for silicon wafers was valued at $17.27 billion in 2023, and would reach $22.1 billion by 2030, a CAGR of 5.77%.[46]

To be conservative (i.e., to project higher AI chip production capacity), LEI applied high end industry estimates and forecasts: a 19% growth rate applied to the 3.85 million AI-related semiconductor chips shipped in 2023, to estimate worldwide manufacturing capacity in 2024,[47] followed by 11% growth for 2025, 10% growth for 2026 to 2028, and 7.6% thereafter. This results in world-wide production capacity of reach 7.9 million AI-related semiconductor chips by 2030 (an average growth rate of 10.7% from 2023 to 2030) (see Figure 8).

Figure 8. LEI outlook for AI-related global chip capacity



3.2 Findings

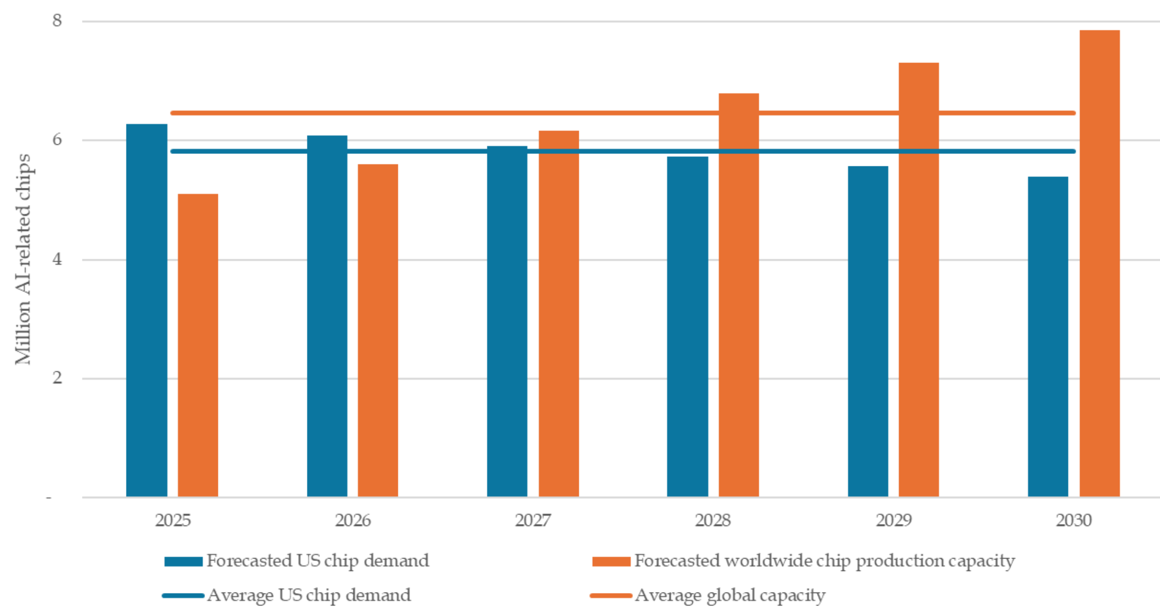
The bottom line is derived by comparing total global capacity for AI-related semiconductor chips (which reaches 7.9 million chips by 2030 as noted above) with the implied growth in chip demand from the RTO/ISO/BA outlook. Over the six-year period from 2025 to 2030, US demand for

[46] Semiconductor Insight. "Silicon Wafer Market Size, Share, Trends, Market Growth and Business Strategies 2025-2032." (Undated). <https://semiconductorinsight.com/report/silicon-wafer-market/>

[47] Morgan Stanley research also estimated that AI-related chip sales from Nvidia to be 4 million in 2024 (Source: Morgan Stanley Research. "AI Supply Chain – The Latest about NVDA GB200 Superchip." P. 10. Exhibit 11. May 13, 2024). Note that, for conservativeness, LEI assumed that most of the capacity expansion would happen by 2025, reflecting the commissioning of the Taiwan Semiconductor Manufacturing Company's Arizona Fab (Source: Taiwan Semiconductor Manufacturing Company Limited. *TSMC Arizona.* <https://www.tsmc.com/static/abouttsmcaz/index.htm>).

semiconductor chips would account for 90% of global capacity if all the projected US data center demand were to materialize (see Figure 9, blue line and orange line).

Figure 9. Implied incremental AI-related chips needed in the US vs forecasted worldwide AI-related chip production capacity



It is not realistic to expect that the United States will be able to acquire 90% of the global supply of chips over the next six years. Global chip manufacturing capacity will be sought after by customers around the globe, not just in the United States, and not just in the US jurisdictions surveyed. As noted previously in Section 2.1.1, the IEA estimated that in 2024, US data center installed capacity amounted to 42 GW, while other countries' data center installed capacity amounted to approximately 55 GW, so that US data center electricity installed capacity accounted for less than half of the world's total.[48] Nvidia (the world's largest chip maker) reported that revenue from sales to customers outside of the United States accounted for 69%, 56%, and 53% of total revenue for fiscal years 2023, 2024, and 2025 respectively."[49]

---

[48] IEA. *Energy and AI*. April 2025. P. 259. Table A.2. <https://www.iea.org/reports/energy-and-ai>

[49] Nvidia. *Form 10-K. Annual Report for fiscal year ended January 26, 2025*. P. 79.

There is demand for AI-related semiconductor chips in other parts of the world, such as Europe (e.g. France[50] and Ireland[51]), Canada,[52] China, East/Southeast Asian countries such as Japan, South Korea, Singapore and Malaysia; and the Middle East.[53] These regions and countries are competing for the global supply of chips. Recent talks between the US administration and Saudi Arabia indicate the kingdom's expectations for strong growth in data centers and associated demand for chips and plans for other Middle Eastern nations for purchase of chips from the United States (as well as plans to invest in data centers in the United States).[54]

This implies that the total data center electricity demand growth projections over the next five years for the US are unrealistic compared to the chip manufacturing industry's own projections for how fast it can ramp up production. It is unlikely that sufficient equipment could be manufactured globally to create such electricity demand unless a large (perhaps government-funded) project was put into place to do so. Hence, it is not reasonable to believe that data centers in the United States will procure almost all the output of production capacities for AI-related semiconductor chips from the entire world over a six-year period.

As noted in Section 2.1 of this report, data center requests for service across the United States have some level of duplication. Even ignoring the possibility of duplication, LEI's analysis of chip supply suggests that some data center developers will be unable to build as intended in the next few years and therefore not all the electricity demand growth from data centers reflected in the aggregate US outlook will be realized.

---

[50] *France 24*. "Microsoft, Amazon to invest billions in French tech." December 5, 2024. <https://www.france24.com/en/live-news/20240512-amazon-plans-to-invest-1-2-bn-euros-in-france-macron-s-office?ref=frenchtechjournal.com>

[51] Burns, John. "Up to 82 data centres operating in Republic." *Irish Times.* June 17, 2023. <https://www.irishtimes.com/business/2023/06/17/up-to-82-data-centres-operating-in-republic/>

[52] Microsoft. "Microsoft Expands Digital Footprint In Quebec With Usd$500 Million Investment In Infrastructure And Skilling Initiatives." November 2023. <https://news.microsoft.com/en-ca/2023/11/22/microsoft-expands-digital-footprint-in-quebec-with-usd500-million-investment-in-infrastructure-and-skilling-initiatives/>

[53] Countries in the Middle East, including Saudi Arabia and UAE, have plans to attract investments from technology companies such as Google and Microsoft to develop data centers in their region. For example, see: MIT Sloan Management Review. "How the Middle East is Emerging as a Data Center Powerhouse Amid Booming AI Demand." October 2024. <https://www.mitsloanme.com/article/how-the-middle-east-is-emerging-as-a-data-center-powerhouse-amid-booming-ai-demand/>

[54] Gunia, Amy. "Will 'massive' Gulf deals cement the US lead in the race for global AI dominance?" CNN. May 22, 2025. <https://www.cnn.com/2025/05/22/business/trump-gulf-deals-global-ai-race-hnk-spc-intl>

4 Uncertainty around electricity demand from data center growth poses risks for ratepayers

As LEI examined in detail in this report, future growth of data centers is highly uncertain, especially projections for growth in any given location. As detailed, there are a number of drivers of this uncertainty, and they all point to upside bias in demand outlooks.

The uncertainty surrounding data centers impacts utility load projections because data centers are such large customers. As such, they can create substantial risk for other customers if risks are left unmitigated. As discussed earlier in this report, vertically integrated utilities build infrastructure and pass along costs to customers; such costs could include new system assets (generation and transmission capacity), as well as related costs such as the cost of firm transportation on natural gas pipelines intended to serve new gas-fired generating plants. If the expected electric demand does not appear, the costs incurred to prepare the system to serve the demand would typically be re-allocated to the remaining customers, and if the expected electric demand is large compared to the demand of the existing customer base, the impact could be substantial. Even if the new large load materializes, it may not be on the system long enough to cover the costs of system assets which it caused to be built to serve it. The loss of any kind of customer (including residential customers if population migrates, or an industrial customer who shuts down a factory) would impact the remining customers, but with data centers, the risk posed by a single customer is amplified because data centers can be such large loads.

If a utility takes a more conservative approach and underestimates future load, then it will have to tell some new customers that they must wait for service. This may not be ideal for the utility, because it may mean that these customers may move their plans to other regions, causing the utility to forgo expansion of its customer base and associated rate base growth. But when the scale of investment needed is in the hundreds of millions of dollars, utilities and their regulators will want to avoid basing expansion plans on speculative requests.

In short, over-forecasting based on speculative data center load can expose existing customers to potentially unnecessary and avoidable costs. An approach that prioritizes verified commitments, investment planning to meet demonstrated need, and ensuring ESA or other contract and tariff terms require financial commitments from potential data center customers would help reduce risk to existing customers.

5 Conclusion

The purpose of this analysis was to examine the projections of data center electric load growth in the United States. These projections are beset by uncertainty, and the factors driving the uncertainty (the incentive for duplication of requests by data centers, evidence of attrition in data center announcements, utility investment incentives, long lead times for new generation equipment, the typical timeline for adding infrastructure to the grid, and limits to global semiconductor chip supplies) all point to the greater risk of over-forecasting demand than under-forecasting demand. This bias may prove costly to existing customers, especially in regions where vertically integrated utilities own generation and pass costs to customers under cost-of-service regulation and the incremental cost of investment far exceeds the average embedded cost.

LEI's approach to this analysis was two-pronged:

First, LEI examined factors on the demand side (on the part of data center developers) and the supply side (on the part of utilities) which impact forecasts of data center electricity load. LEI found that these factors not only contribute to uncertainty in load forecasts, but they also add upside bias to such forecasts.

Second, LEI performed a reality check on data center forecasts for the United States by comparing them to the global capacity to manufacture the number of chips required to support such forecasts. LEI found that global chip manufacturing is likely to be far short of what would be needed to supply the US demand forecast based on LEI's tally of RTO/ISO/BA data center projections.

LEI concludes that the risk to the overall outlook for data center demand at the US level contains upside bias: it is more likely that the outlook implied by the RTO/ISO/BA tally is higher than what will materialize in the time frame of the forecast. If this happens and utilities have invested in incremental generation and transmission based on unrealistic projections of data center demand, costs will increase for other utility customers.

Appendix A: The basics of data center electricity demand

A data center is a physical facility composed of computing and data storage equipment, and software. At a high level, its three main components are:

- **Network infrastructure** to connect servers (physical and virtual) that perform data center services and provide storage internally in the data center, and to provide external connectivity to end-user locations;
- **Storage infrastructure** to hold data; and
- **Computing resources** to provide the processing, memory, local storage, and network connectivity that drive applications.[55]

The data center industry is shifting towards larger companies and larger data centers known as hyperscale centers, which provide cloud computing and data management services to organizations that require extensive infrastructure for large-scale data processing and storage. In 2024 about 80% of North American demand for data centers was driven by hyperscalers.[56] Not only is the size of an individual hyperscale data center much larger, but the companies that use them are huge—large cloud service providers include Amazon Web Services, Google Cloud, Microsoft Azure, International Business Machines Corporation ("IBM") Cloud, Baidu, and Alibaba Cloud.[57] By 2023, hyperscale centers accounted for about 95% of computing instances (a measure of computing resources), meaning that these hyperscale centers now dominate the data center industry (see dotted line in Figure 10). The computing power needed for AI workloads is the driver of the growth in the size of data centers and a factor in the ongoing expansion of the data center industry.[58]

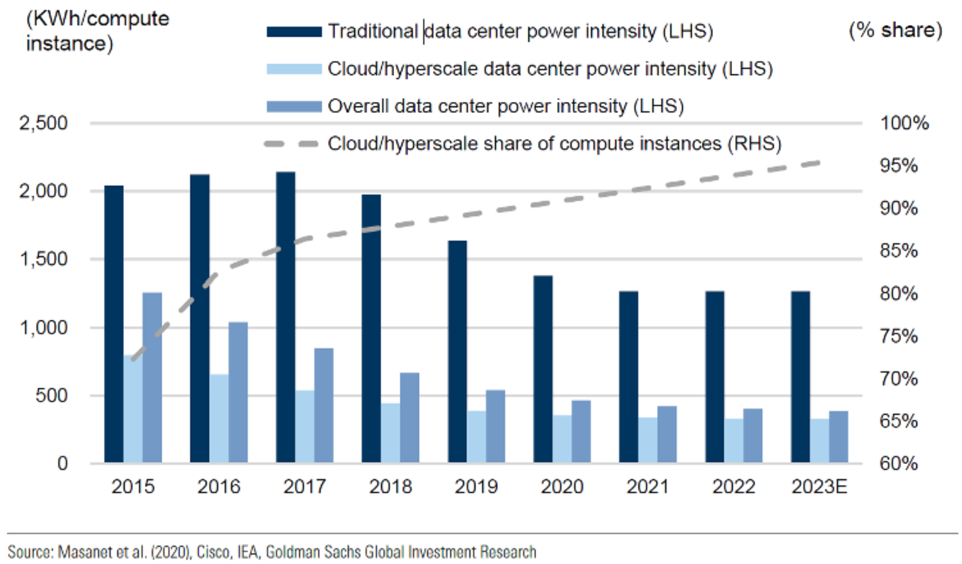
[55] Cisco. "What Is a Data Center?" <https://www.cisco.com/c/en/us/solutions/data-center-virtualization/what-is-a-data-center.html>

[56] Collier's. "Accessing Power and Capital: The Future of Data Center Development 2024." P. 3. <https://www.colliers.com/en/research/dmc-nrep-uscm-usdc-colliers-data-center-white-paper-2024 >

[57] New England States Committee on Electricity. "Data Centers and the Power System: A Primer." Soring 2024. P. 6. <https://nescoe.com/resource-center/data-centers-primer/>

[58] McKinsey & Company. "AI power: Expanding data center capacity to meet growing demand." October 29, 2024. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand>

Figure 10. Hyperscale share of data center computing growth over time, and energy efficiency (global estimates)



Source: Goldman Sachs. "AI, data centers and the coming US power demand surge." April 28. 2024. P. 13. Exhibit 10. <https://www.goldmansachs.com/pdfs/insights/pages/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf>

Note: A compute instance is a measure of computing resources (independent of the hardware used). In cloud computing, a compute instance is a server resource provided by a third-party cloud service and supported by the hardware in the data center (Source: Amazon Web Services. *What is an Instance In Cloud Computing?* <https://aws.amazon.com/what-is/cloud-instances/>).

Hyperscale data centers are more energy efficient than smaller data centers in the sense that a larger share of the energy used by hyperscale data centers goes into actual computing work. This is referred to as power usage effectiveness ("PUE"). As of 2023, the estimated average PUE for all data centers in the United States was 1.4[59] (meaning 40% of energy is not used for computing), while AI-specialized data centers had an average PUE of 1.14.[60] Future improvements in this kind of efficiency might be limited because the most efficient data centers are already close to a PUE of 1, which is the lowest possible. However, computation efficiency can still improve.

5.1 Data centers support a variety of businesses, some with volatile markets

There are three overarching categories of business which data centers support:

[59] 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. P. 47.

[60] Ibid. P. 47. Figure 4.5.

- **Cloud computing data centers** support a broad spectrum of information technology ("IT") services that involve storing and processing data;
- **AI data centers** (including for training and inference) are specifically designed to handle the computationally intensive tasks of artificial intelligence and machine learning, requiring specialized and advanced equipment; and
- **Crypto mining centers** have the narrowest function: they validate blockchain transactions and generate new cryptocurrency units.

Cloud computing and AI data centers are expected to have more stable operations and earnings compared to crypto mining centers activities and earnings which are more volatile, rising and falling with the value of cryptocurrencies and energy prices.[61] The EIA estimated that electricity demand associated with cryptocurrency mining operations in the United States likely accounted for 0.6% to 2.3% of US electricity consumption by January 2024, but EIA has not updated these numbers.[62] These ranges compare to the 4.4% energy use estimated for data centers overall for 2023.[63]

5.2 Computational and design efficiency can impact electricity demand from data centers

The energy efficiency of data centers is an important factor in driving future data center load growth.

There are two measures of efficiency for data centers. First, as noted previously, PUE measures the ratio of power (MW of electricity) used by the data center as a whole and the power used by the computing equipment. Power used by non-computing equipment can include cooling, power transformation (from high voltage to low voltage), and conversion (from alternating current to direct current). The PUE of a data center depends on factors including the data center's design, the capacity factor of the data center, location of the data center (which reflects the ambient temperature and impacts the efficiency of cooling equipment), among others.

Over the past decade, the PUE of hyperscale data centers owned by large companies has improved. For example, Google reports its improvement in PUE since 2008, when the PUE for all large-scale Google data centers was above 1.2. Google's PUE improved over time, falling to 1.10

---

[61] Cloudnium LLC. *Understanding the Differences Between Data Centers and Crypto Mining Facilities*. July 12, 2024. <https://cloudnium.net/understanding-the-differences-between-data-centers-and-crypto-mining-facilites/>

[62] EIA. "Tracking electricity consumption from U.S. cryptocurrency mining operations." February 1, 2024. <https://www.eia.gov/todayinenergy/detail.php?id=61364#>

[63] 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. P. 52.

by Q3 2024.[64] Similarly, Amazon Web Services announced that it reached a global PUE of 1.15 across its data centers in 2023, while Meta (the parent company of Facebook) has an average PUE of around 1.08.[65] In theory, a perfect data center (where all the energy used goes to computing) would have a PUE of 1. This means there is only very limited room for the PUE of hyperscale data centers to improve further. At best, the PUE of hyperscale data centers owned by large companies can only improve by approximately 10%, as the PUE value cannot drop below 1.

The other measure of data center efficiency is the energy required per unit of computing work. Generally, computing work is measured in floating point operations ("FLOPS"). A FLOP is a calculation done by a computer. The energy efficiency of a data center's computing power can be measured by FLOPS/W, with a higher number indicating greater efficiency. Floating point operations differ by the level of precision: the more precise the calculation, the more inefficient the calculation process.

The FLOPS/W of a data center depends largely on the semiconductor chips used by the data center. In recent years, AI-related work has been typically carried out by GPUs, which are semiconductor chips originally designed for graphics-related purposes, but which were found to be very efficient in doing the type of calculations needed for AI.[66]

5.3 Semiconductor chips have become more efficient, but also more energy dense

The IEA reported that a modern AI-related computer chip uses 99% less electricity to perform the same calculations than a similar chip in 2008, and the efficiency of these chips doubled roughly every two-and-a-half to three years (see Figure 11). This implies an average improvement of 30% in FLOPS/W per year.[67]

---

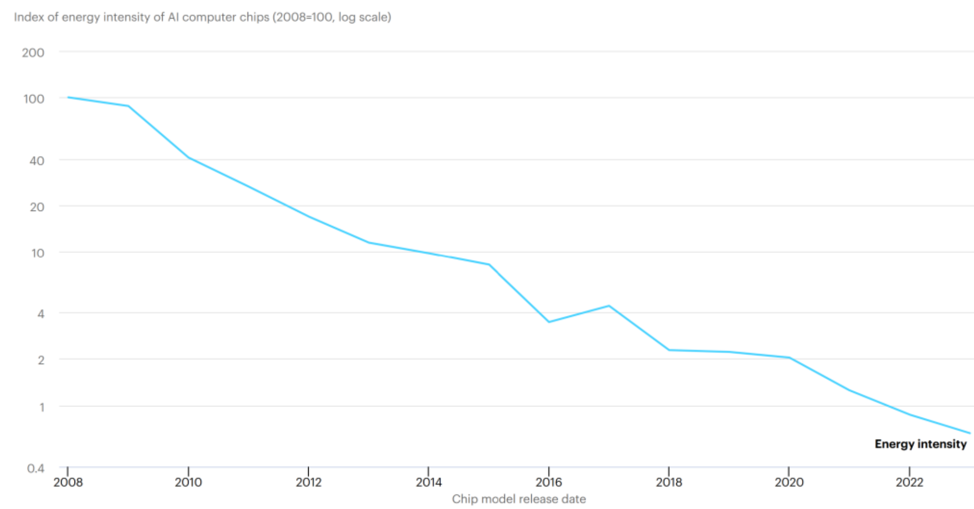
[64] Google Data Centers. "Growing the internet while reducing energy consumption." <https://www.google.com/about/datacenters/efficiency/>

[65] Butler, Georgia. Data Center Dynamics. *AWS global data centers achieved PUE of 1.15 in 2023.* December 4, 2024. <https://www.datacenterdynamics.com/en/news/aws-global-data-centers-achieved-pue-of-115-in-2023/>

[66] Khan, Saif M. and Alexander Mann. *AI Chips: What They Are and Why They Matter*. Center for Security and Emerging Technology. April 2020. P. 20. <cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter>

[67] A 100% improvement over 2.75 years translate to an average of 28.7% improvement every year.

Figure 11. Energy intensity of AI computer chips



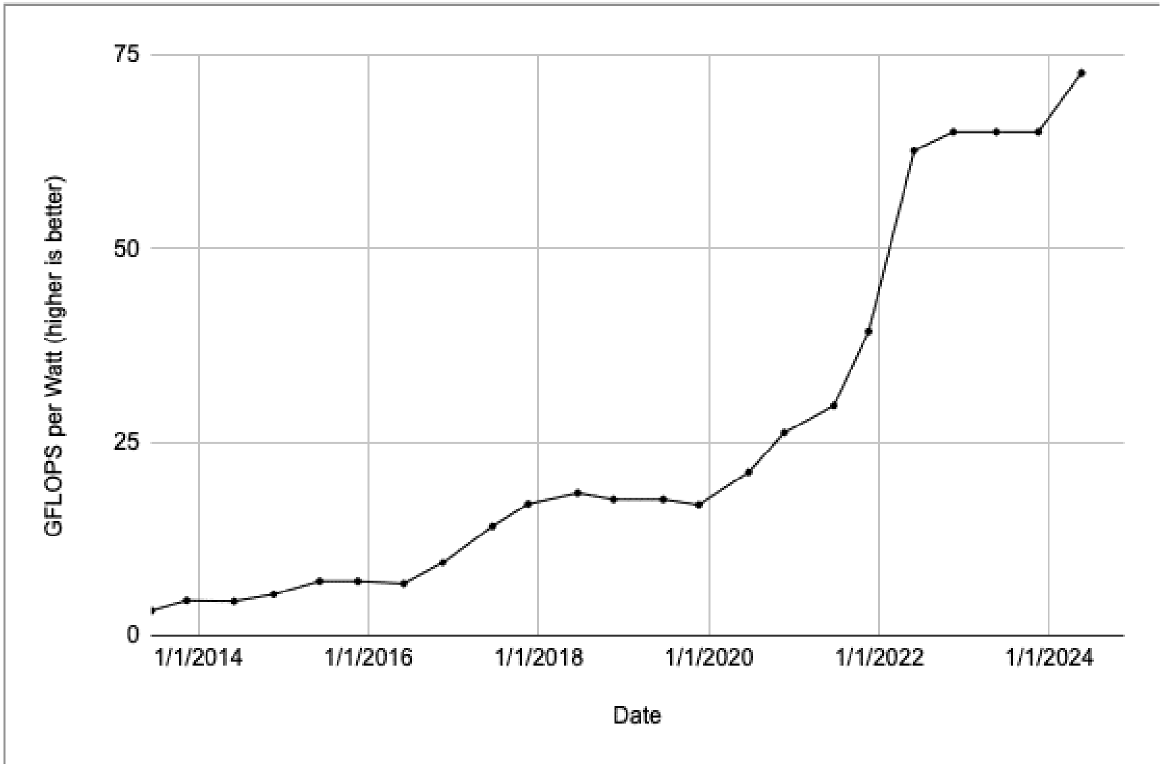
Source: IEA. "What the data centre and AI boom could mean for the energy sector." October 18, 2024. https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector.

Similarly, Nvidia, using data from the top 500 supercomputers in the world, showed that the FLOPS/W of supercomputers improved from approximately 4 giga-FLOPS ("GFLOPS")/W[68] in late 2013, to 74 GFLOPS/W in 2024 (see Figure 12). This also implies an approximately 30% average improvement in energy efficiency per year, similar to the IEA calculation.[69]

[68] A GFLOP means a giga-FLOP, which is approximately a billion FLOPS.

[69] A 18.5-fold improvement over approximately 11 years translate to 30.4% improvement every year.

Figure 12. Energy efficiency gains over time for the most efficient supercomputer



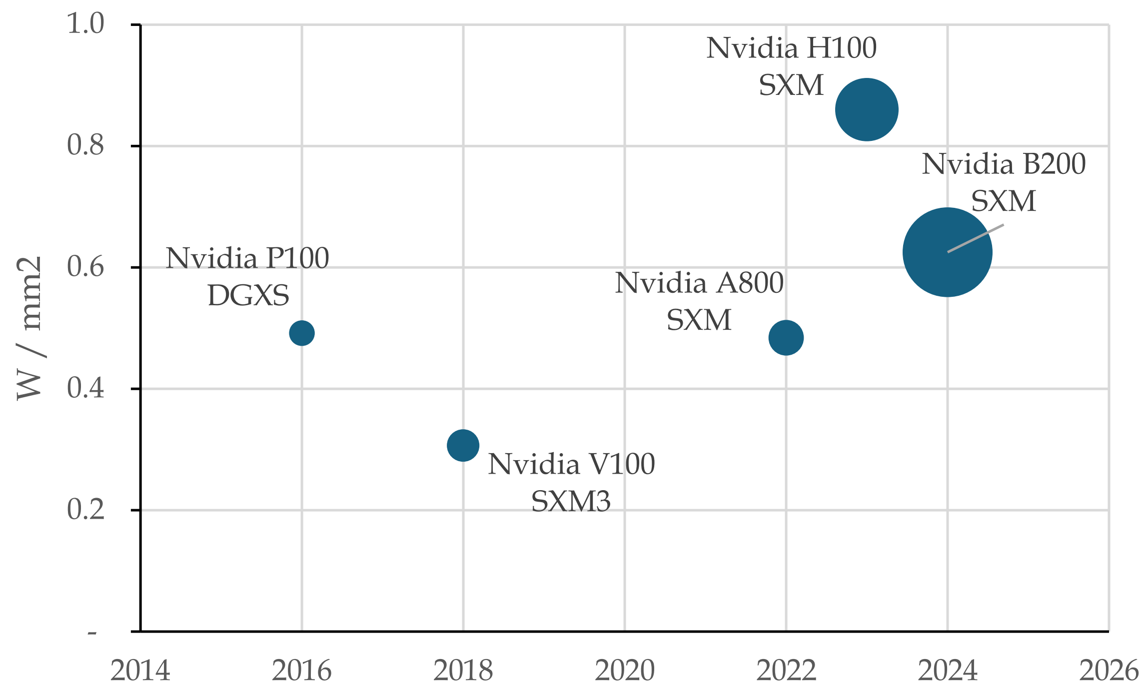
Source: Harris, Dion. *Sustainable Strides: How AI and Accelerated Computing Are Driving Energy Efficiency*. July 22, 2024. Nvidia. <https://blogs.nvidia.com/blog/accelerated-ai-energy-efficiency/>

Note: Data is from Top500.org, an organization which compiles statistics about supercomputers.

Another key driver of data center-related electricity demand growth is that AI-related chips use more energy than conventional chips because AI-related chips have more computing power packed into a single chip. While the FLOPS/W of AI-related chips has improved over time, each new generation of AI-related chips has more FLOPS packed into an individual chip than the previous generation. This has resulted in an overall increase in energy demand per chip. Based on LEI's comparison of five generations of AI-related chips specifically designed for data centers, the energy density of AI-related chips (measured in watt per square millimeter, or W/mm2 of chip size) has increased at an average rate of 3.04% per year (see Figure 13).

While PUE efficiencies are near the maximum, the computational efficiency of AI chips is still improving at a fast clip, even as the electricity demand associated with AI chips has risen. Each chip has more GFLOPS packed in the same chip surface area, so each chip may still demand an increasing level of electricity.

Figure 13. Energy density (W/mm2 of die size) of generations of AI-related chips



Note 1: Size of the dot represents the relative FLOPS (for FP32 calculations) of the generation of AI-chips presented.

Note 2: The "die size" of a chip is the size of the silicon wafer where the circuit is printed, while the energy demand is based on the total design power of the chip.

Note 3: The Nvidia P100 DGXS is chosen as the first AI-related chip in the analysis as it is the first Nvidia GPU that uses a Server PCI Express Module ("SXM"), indicating that this chip is designed specifically for data centers (as opposed to workstation or gaming computer purpose).

Sources: TechPowerUp GPU database, IEEE Spectrum, LEI analysis.

5.4 Innovations in AI show dramatically lower energy demand

The overall energy demand associated with operating generative AI may prove significantly lower than previously expected. The launch of DeepSeek-R1 in early 2025 took the AI and energy industry by surprise—it is an event which might be considered a "wild card." It is a more energy-efficient AI model, and it raises important questions about how much electricity would be needed to support an AI-driven economy.[70] For example, the efficiency of DeepSeek-R1 implies less energy is needed. However, the model is also cheaper to train and use, and this could translate to lower prices charged to AI users; the lower prices could incentivize new and broader applications for AI, which would increase demand for data centers and the electricity to power them. The net effect on electricity demand is still unknown.

---

[70] DeepSeek uses the same kinds of chips as other AI programs.

Wherever data centers do locate, questions remain around how much energy they will require, and what kinds of resources will be necessary to meet that demand. Specifically, trends in AI suggest that gains in energy efficiency and demand response may significantly lessen the electricity requirements of future data centers. On the other hand, lower electricity costs could expand AI into new market applications.

Appendix B: Works cited

Amazon Web Services. *What is an Instance In Cloud Computing?* <https://aws.amazon.com/what-is/cloud-instances/>

Bain and Company. “Utilities Must Reinvent Themselves to Harness the AI-Driven Data Center Boom.” October 2024. <https://www.bain.com/insights/utilities-must-reinvent-themselves-to-harness-the-ai-driven-data-center-boom/>

Bloomberg News. “How ‘Load Shifting’ May Help Improve Data Center Sustainability.” *Data Center Knowledge.* February 26, 2024. <https://www.datacenterknowledge.com/sustainability/how-load-shifting-may-help-improve-data-center-sustainability>.

Burns, John. “Up to 82 data centres operating in Republic.” *Irish Times.* June 17, 2023. <https://www.irishtimes.com/business/2023/06/17/up-to-82-data-centres-operating-in-republic/

Butler, Georgia. Data Center Dynamics. *AWS global data centers achieved PUE of 1.15 in 2023.* December 4, 2024. <https://www.datacenterdynamics.com/en/news/aws-global-data-centers-achieved-pue-of-115-in-2023/>

Casey, Simon. “GE Vernova CEO Sees Order Backlog Stretching Into 2028.” *Bloomberg.* March 11, 2025. <https://www.bloomberg.com/news/articles/2025-03-11/ge-vernova-ceo-sees-order-backlog-stretching-into-2028?utm_source=chatgpt.com&embedded-checkout=true>

Chow, Andrew R. “Why Tax Breaks for Data Centers Could Backfire on States.” April 25, 2025. <https://time.com/7280058/data-centers-tax-breaks-ai/>

Cisco. “What Is a Data Center?” <https://www.cisco.com/c/en/us/solutions/data-center-virtualization/what-is-a-data-center.html>

Cloudnium LLC. *Understanding the Differences Between Data Centers and Crypto Mining Facilities*. July 12, 2024. <https://cloudnium.net/understanding-the-differences-between-data-centers-and-crypto-mining-facilites/>

Collier’s. “Accessing Power and Capital: The Future of Data Center Development 2024.” <https://www.colliers.com/en/research/dmc-nrep-uscm-usdc-colliers-data-center-white-paper-2024?utm_source=google&utm_medium=paidsearch&utm_campaign=2023-colliers-leadgen-reports-nat-datacenter&utm_content=PID0001-v3&utm_term=hyperscale%20data%20centers&gad_source=1&gclid=CjwKCAiA1eO7BhATEiwAm0Ee-BlRD5_fY_rrI31camaws8bQjoF2ql4P796L6-FBwNxOHfyQSiCcgBoCiEAQAvD_BwE>

Cunningham, Nicholas. "Costs to build gas plants triple, says CEO of NextEra Energy." Gas Outlook. March 25, 2025. <https://gasoutlook.com/analysis/costs-to-build-gas-plants-triple-says-ceo-of-nextera-energy/>

Data Center Knowledge. "Hyperscalers in 2024: Where Next for the World's Biggest Data Center Operators?" February 28, 2024. <https://www.datacenterknowledge.com/hyperscalers/hyperscalers-in-2024-where-next-for-the-world-s-biggest-data-center-operators->

Dominion Energy. "Schedule 10 Large General Service." Virginia State Corporation Commission. December 20, 2022.

Dominion Energy. "Terms and Conditions Section IX. Deposits." Virginia State Corporation Commission. December 12, 2013.

Dominion Energy. "Terms and Conditions Section XXII. Electric Line Extensions and Installations." Virginia State Corporation Commission. March 8, 2024.

Eddy, Nathan. "Local Opposition Hinders More Data Center Construction Projects." Data Center Knowledge. May 15, 2025. <https://www.datacenterknowledge.com/regulations/local-opposition-hinders-more-data-center-construction-projects>

EIA. "Tracking electricity consumption from U.S. cryptocurrency mining operations." February 1, 2024. <https://www.eia.gov/todayinenergy/detail.php?id=61364#>

Energy Intelligence. "US Gas Companies Temper Data Center Demand Expectations." *Natural Gas Week* 41.11 (March 14, 2025).

IDC. "Global Semiconductor Market to Grow by 15% in 2025, Driven by AI." December 12, 2024. https://my.idc.com/getdoc.jsp?containerId=prAP52837624

FERC Docket No. EL25-49-000. Public Interest Organizations. *Comments of Public Interest Organizations in response to PJM Interconnection, L.L.C.'s 03/24/2025 Answer to FERC's 02/20/2025 Order under EL25-49*. April 23, 2025. P. 16-17. <https://elibrary.ferc.gov/eLibrary/filedownload?fileid=9D23F7BC-5484-CAA9-9030-96645FF00000>.

France 24. "Microsoft, Amazon to invest billions in French tech." December 5, 2024. <https://www.france24.com/en/live-news/20240512-amazon-plans-to-invest-1-2-bn-euros-in-france-macron-s-office?ref=frenchtechjournal.com>GPC. "Rules Regulations and Rate Schedules for Electric Service." Georgia Public Service Commission. January 2023.

Giacobone, Bianca. "Phantom data centers are flooding the load queue." March 26, 2025. <https://www.latitudemedia.com/news/phantom-data-centers-are-flooding-the-load-queue/>

Goldman Sachs. "AI, data centers and the coming US power demand surge." April 28. 2024. <https://www.goldmansachs.com/pdfs/insights/pages/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf>

Google Data Centers. "Growing the internet while reducing energy consumption." <https://www.google.com/about/datacenters/efficiency/>

GPC. "Customer Renewable Supply Procurement Schedule CRSP-1." Georgia Public Service Commission. October 2023.

GPC. "Power and Light Large Schedule PLL-18." Georgia Public Service Commission. April 2025.

GPC. "Rules Regulations and Rate Schedules for Electric Service," Georgia Public Service Commission. January 2023.

GPSC Docket No. 44280. GPC. *Rules & Regulations Update 12-11-2024*. December 2024. <https://psc.ga.gov/search/facts-document/?documentId=220667>

GPSC Docket No. 55378. *Comments on Georgia Power's 2023 Integrated Resource Plan Update*. Microsoft. Docket No. 55378 at 1, 4. Apr. 1, 2024. <https://psc.ga.gov/search/facts-document/?documentId=218199>

GPSC Docket No. 56002. GPC. "2025 IRP Technical Appendix Volume 1: B2025 Load and Energy Forecast." January 2025. <https://psc.ga.gov/search/facts-document/?documentId=221233>

Gunia, Amy. "Will 'massive' Gulf deals cement the US lead in the race for global AI dominance?" CNN. May 22, 2025. <https://www.cnn.com/2025/05/22/business/trump-gulf-deals-global-ai-race-hnk-spc-intl>

Harris, Dion. *Sustainable Strides: How AI and Accelerated Computing Are Driving Energy Efficiency*. July 22, 2024. Nvidia. <https://blogs.nvidia.com/blog/accelerated-ai-energy-efficiency/>

IEA. "Average lead times to build new electricity grid assets in Europe and the United States, 2010-2021." January 12, 2023. <https://www.iea.org/data-and-statistics/charts/average-lead-times-to-build-new-electricity-grid-assets-in-europe-and-the-united-states-2010-2021>

IEA. "Average power generation construction time (capacity weighted), 2010-2018." April 14, 2019. <https://www.iea.org/data-and-statistics/charts/average-power-generation-construction-time-capacity-weighted-2010-2018>

IEA. *Energy and AI*. April 2025. <https://www.iea.org/reports/energy-and-ai>.

IEA. "What the data centre and AI boom could mean for the energy sector." October 18, 2024. <https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector>

Indiana Michigan Power Company. “Submission of Unopposed Settlement Agreement and Unopposed Motion for Acceptance of Out of Time Filing.” Indiana Utility Regulatory Commission Cause No. 46097. November 22, 2024.

Judge, Peter. “Google tests system to cut data center power use during grid problems.” *Data Center Dynamics*. October 5, 2023. <https://www.datacenterdynamics.com/en/news/google-tests-system-to-cut-data-center-power-use-during-grid-problems/>

Khan, Saif M. and Alexander Mann. *AI Chips: What They Are and Why They Matter*. Center for Security and Emerging Technology. April 2020. <cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter>

Koomey, Jonathan (Koomey Analytics), Schmidt, Zachary (Koomey Analytics), and Das, Tania (Bipartisan Policy Center). *Electricity Demand Growth and Data Centers: A Guide for the Perplexed*. February 2025. <https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2025/02/BPC-Report-Electricity-Demand-Growth-and-Data-Centers-A-Guide-for-the-Perplexed.pdf>.

Kusters, J., Bhattacharjee, D., Bish, J., Nicholas, J.T., Stewart, D., Ramachandran, K. “2025 global semiconductor industry outlook.” Deloitte. February 4, 2025. <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-telecom-outlooks/semiconductor-industry-outlook.html>

Levitt, Ben. “AI and Energy, the Big Picture.” S&P Global. December 2024. <https://www.spglobal.com/en/research-insights/special-reports/look-forward/ai-and-energy>

Malik, Naureen S. “Gas Power Won’t Provide an Easy Fix for AI Boom.” Bloomberg. January 8, 2025. <https://www.bloomberg.com/news/newsletters/2025-01-08/gas-power-won-t-provide-an-easy-fix-for-ai-boom>

Mannion, Annemarie. “Microsoft Hits Pause Button on $1B in Data Centers in Ohio.” *Engineering News-Record*. April 10, 2025. <https://www.enr.com/articles/60572-microsoft-hits-pause-button-on-1b-in-data-centers-in-ohio#:~:text=By%20Annemarie%20Mannion,contractors%20had%20not%20been%20announced>

Martucci, Brian. “A fraction of proposed data centers will get built. Utilities are wising up.” *Utility Dive*. May 15, 2025. <https://www.utilitydive.com/news/a-fraction-of-proposed-data-centers-will-get-built-utilities-are-wising-up/748214/>

McKinsey & Company. “AI power: Expanding data center capacity to meet growing demand.” October 29, 2024. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand>

Microsoft. "Microsoft Expands Digital Footprint In Quebec With Usd$500 Million Investment In Infrastructure And Skilling Initiatives". November 2023. <https://news.microsoft.com/en-ca/2023/11/22/microsoft-expands-digital-footprint-in-quebec-with-usd500-million-investment-in-infrastructure-and-skilling-initiatives/>

MIT Sloan Management Review. "How the Middle East is Emerging as a Data Center Powerhouse Amid Booming AI Demand." October 23, 2024. <https://www.mitsloanme.com/article/how-the-middle-east-is-emerging-as-a-data-center-powerhouse-amid-booming-ai-demand/>

Morgan Stanley Research. "AI Supply Chain – The Latest about NVDA GB200 Superchip." May 13, 2024.

New England States Committee on Electricity. "Data Centers and the Power System: A Primer." Spring 2024. <https://nescoe.com/resource-center/data-centers-primer/>

Noelle Walsh, President Microsoft Cloud Operations and Innovation | Nouryon, Non-Executive Director. <https://www.linkedin.com/posts/noelle-walsh-b29356108_microsoftcloud-datacenters-activity-7315439628562423808-W67e/>

Norris, T. H., T. Profeta, D. Patino-Echeverri, and A. Cowie-Haskell. 2025. "Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems." NI R 25-01. Durham, NC: Nicholas Institute for Energy, Environment & Sustainability, Duke University. <https://nicholasinstitute.duke.edu/publications/rethinking-load-growth>

Novet, Jordan. "Amazon has paused some data center lease commitments, Wells Fargo says." CNBC. April 21, 2025. <https://www.cnbc.com/2025/04/21/amazon-has-paused-some-data-center-lease-commitments-wells-fargo.html>

NREL. "Data Center Infrastructure in the United States, 2025." May 2025. <https://docs.nrel.gov/docs/gen/fy25/94502.jpg>

Nvidia. *Form 10-K. Annual Report for fiscal year ended January 26, 2025.*

Patel, Dylan; Jeremie Eliahou Ontiveros; Maya Barkin. "Microsoft's Datacenter Freeze - 1.5GW Self-Build Slowdown & Lease Cancellation Misconceptions." *SemiAnalysis*. April 28, 2025. <https://semianalysis.com/2025/04/28/microsofts-datacenter-freeze/>

PricewaterhouseCoopers, *State of the Semiconductor Industry*. November 28, 2024. <https://www.pwc.com/gx/en/industries/technology/state-of-the-semiconductor-industry-report.pdf>

Seiple, Chris. "Gridlock: The demand dilemma facing the US power industry." *Wood Mackenzie.* October 2024. <https://www.woodmac.com/horizons/gridlock-demand-dilemma-facing-us-power-industry/>

Semiconductor Industry Association. "Strengthening the U.S. Semiconductor Supply Chain." May 2024. <https://www.semiconductors.org/wp-content/uploads/2024/05/Emerging-Resilience-in-the-Semiconductor-Supply-Chain_SIA-Summary.pdf>

Semiconductor Insight. "Silicon Wafer Market Size, Share, Trends, Market Growth and Business Strategies 2025-2032." (Undated). <https://semiconductorinsight.com/report/silicon-wafer-market/>

Shah, Agam. *Nvidia Shipped 3.76 Million Data-center GPUs in 2023, According to Study*. June 10, 2024. <https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/>

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. Pg 53. Figure 5.6.

Taiwan Semiconductor Manufacturing Company Limited. *TSMC Arizona.* <https://www.tsmc.com/static/abouttsmcaz/index.htm>

TechPowerUp. "NVIDIA B200 SXM 192 GB." <https://www.techpowerup.com/gpu-specs/b200-sxm-192-gb.c4210>

TechPowerUp. "NVIDIA Tesla P100 DGXS." <https://www.techpowerup.com/gpu-specs/tesla-p100-dgxs.c3285>

Williams, Kevin. "AI data center boom isn't going bust but the 'pause' is trending at big tech companies." CNBC. April 27, 2025. <https://www.cnbc.com/2025/04/27/ai-data-center-boom-isnt-going-bust-but-the-pause-is-trending.html>