Opportunities to Use Energy Efficiency and Demand Flexibility to Reduce Data Center Energy Use and Peak Demand

Steven Nadel

October 2025
White Paper

ACEEE

About ACEEE

The **American Council for an Energy-Efficient Economy** (ACEEE), a nonprofit research organization, develops policies to reduce energy waste and combat climate change. Its independent analysis advances investments, programs, and behaviors that use energy more effectively and help build an equitable clean energy future.

About the author

Steven Nadel has been ACEEE's executive director since 2001. He has worked in the energy efficiency field for more than 40 years and has over 200 publications. His research interests include decarbonization strategies for the buildings, industrial, and transportation sectors; federal, state, and local energy and climate change policy; utility-sector energy efficiency programs and policies; and appliance and equipment efficiency standards. Steve earned a master of science in energy management from the New York Institute of Technology and a master of arts in environmental studies and a bachelor of arts in government from Wesleyan University.

Acknowledgments

This report was made possible through the generous support of the Tilia Fund and internal ACEEE resources. The author gratefully acknowledges external reviewers, internal reviewers, colleagues, and sponsors who supported this report. External expert reviewers included Eric Masanet from the University of California Santa Barbara, Jon Koomey from Koomey Analytics, Arman Shahabi from Berkeley Lab, Micah Sweeney from EPRI, and Hallie Carrao and David Culler from Google. Internal reviewers included Neal Elliott, Matt Malinowski, Mark Kresowik, and Mariel Wolfson. External review and support do not imply affiliation or endorsement. Last, we would like to thank the following members of ACEEE's editorial and communications team: Ethan Taylor, Kate Doughty, Mariel Wolfson, Mark Rodeffer, Nick Roper, Phoebe Spanier, and Roxanna Usher for their help in launching this report.

Suggested citation

Nadel, Steven. 2025. *Opportunities to Use Energy Efficiency and Demand Flexibility to Reduce Data Center Energy Use and Peak Demand*. Washington, DC: ACEEE. https://www.aceee.org/white-paper/2025/10/opportunities-use-energy-efficiency-and-demand-flexibility-reduce-data-center.

Data and licensing information

Contents

Key takeaways

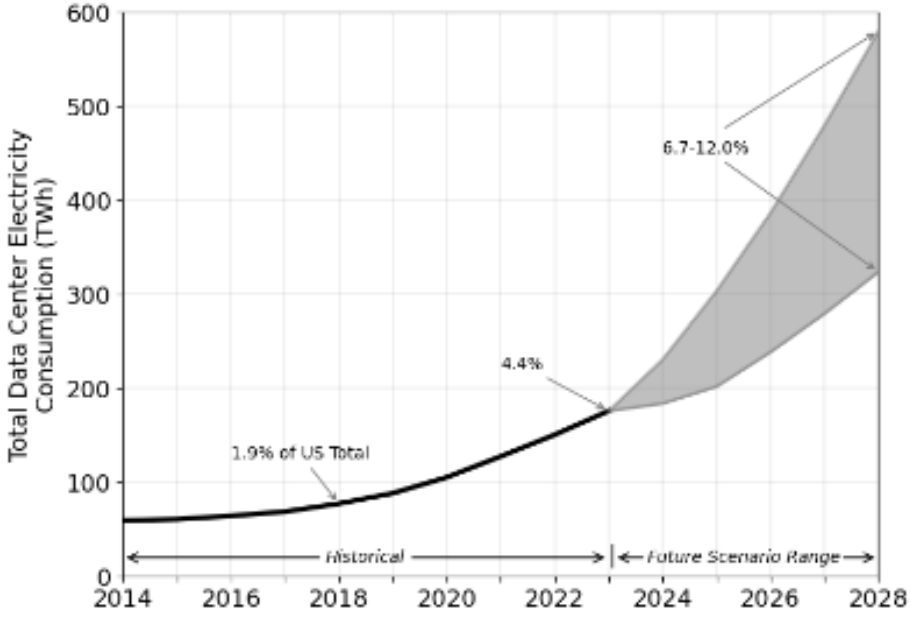
- Data centers are often the largest source of load growth for utilities, according to projections through 2030, and they are contributing to rising electricity costs. Data centers accounted for about 4% of U.S. electricity sales in 2023 (latest available), with projections ranging from 6.7–12.0% in 2028. Up to 70% of the projected growth in data center loads is expected to be hyperscale (deploying services at massive scale, including AI).
- Many new data centers have been proposed, but most data center developers seek permitting and interconnection for many more sites than they expect to build. Several experts estimate only 10–20% of proposed data centers will actually be built; others estimate a higher percentage. Forecasting of data center power demand is improving, but there is a need for additional improvement; power system planners would be wise to carefully review proposals for new data centers and apply probabilities to each of these proposals.
- Achieving higher levels of energy efficiency in data centers is critical for preventing power outages and constraining rising power costs. Opportunities to improve efficiency include developing and using more efficient chips in servers, improving software and algorithms, adding heat recovery, and improving cooling and electric systems.
- Data center energy use roughly doubled from 2000–2010 but only rose modestly from 2010–2020, due to improved efficiency of IT devices, greater use of virtualization software, and movement of more computations to large data centers with very efficient cooling systems. This pattern could potentially repeat in the 2030s if the energy efficiency opportunity is fully realized.
- Complementing energy efficiency are opportunities to reduce loads during periods of peak electric demand using demand flexibility. A study from Duke University estimates that curtailing data center loads for just 0.25% of their uptime would free up enough capacity to accommodate 76 GW of new load (a large nuclear plant is about 1 GW). As one example, a test of a new software platform reduced the peak power consumption of an Oracle data center by 25% during peak grid demand hours.
- Data center efficiency is commonly measured with a metric called PUE, which looks at auxiliary systems such as cooling but not the server part of data centers. This has helped to drive substantial improvements in data center cooling and power system efficiency, but new metrics are needed to better capture all data center energy use including computations, particularly for AI.
- Utilities and states could implement programs and policies to help spur efficiency and demand flexibility progress. Potential opportunities include efficiency and flexibility targets (mandatory, voluntary, or tied to electric rates) as well as energy savings incentive programs.

Introduction

After several decades of very little electric load growth in the United States, electricity use and demand has again started to increase. Data centers are a significant part of this growth, particularly the large data centers used for artificial intelligence (AI) (e.g., EPRI 2024a; Shehabi et al. 2024). Other substantial sources of load growth include large new and expanded manufacturing facilities, increased use of air-conditioning, and electrification of vehicles, heat, and industrial processes (Koomey, Das, and Schmidt 2025). Many new data centers have been proposed and are expected to contribute to substantial load growth over the coming decade. This growth is uneven, with some regions seeing extensive growth (e.g., Virginia and Georgia (Koomey, Schmidt, and Das 2025)) while other regions are experiencing much lower growth. In some regions, data centers are the largest source of load growth for utilities through 2030, including in Virginia (Dominion Energy 2024), Arizona (APS 2023), and American Electric Power in Indiana and Ohio (the latter forecasting nearly 18 GW of data center demand by 2030; St. John 2025a). These new loads are already increasing power costs paid by all customers (e.g., Monitoring Analytics 2025), with further cost increases expected (Wade et al. 2025).[1]

Improving energy efficiency and enabling demand flexibility are critical ways to reduce demand growth, lowering the amount of infrastructure needed and the costs to build this infrastructure. Several analysts have warned about the risks of overbuilding and leaving stranded assets (Lovins 2025; St. John 2025a; McKinsey & Company 2025).

Projections of data center energy use in coming years are uncertain. Probably the most widely used estimate is one published by Berkeley Lab in December 2024 that found that in 2028, data centers could account for 6.7–12.0% of U.S. electricity consumption, up from about 4.4% in 2023 (Shehabi et al. 2024). These estimates are illustrated in figure 1.



[1] We return to the issue of data centers and power costs later in this paper.

Figure 1. LBL estimates of data center energy consumption by year. Source: Shehabi et al. 2024.

Uncertainties in these estimates include the number and size of data centers that are ultimately built, how heavily used these data centers are, the extent of underlying growth in demand for AI and other data center services, and data center and software efficiency. This paper builds on recent ACEEE work on AI (e.g., Wang and Assadi 2025) and focuses on opportunities to use energy efficiency to reduce growth in data center energy use. We also discuss opportunities to reduce peak energy demand through load shifting, storage, and other demand flexibility strategies. Near the end of this paper, we discuss future scenarios for data center energy use, including the impact of energy efficiency.

This paper was developed to help staff at utilities and public agencies who are trying to understand electricity demand growth: how much growth is likely, and what opportunities exist to bring energy demand and supply into balance, such as through programs and policies. The recommendations in this paper can help prevent overbuilding of electric infrastructure and keep consumer and business electricity prices from rising too rapidly. The recommendations can also assist energy efficiency program implementers design efficiency opportunities for this large and growing load. This paper may also be of use to people in the data center industry as they consider all options.

Past Is prologue?

To show that efficiency opportunities are real and significant, we consider data center energy use over the 2010–2018 period. Masanet et al. (2020) examined global data center energy use in 2010 and 2018 in an article published in *Science*, and found that over this period, data center energy use increased 6%, while compute instances increased 550%. They looked at energy use per compute instance and found that this decreased 20% per year over this period. Masanet and Lei (2020) identify three efficiency effects that kept the energy use of conventional data centers[2] in check:

1. The energy efficiency of information technology (IT)devices—and servers and storage drives in particular—improved substantially due to steady technological progress by IT manufacturers.
2. Greater use of server virtualization software, which enables multiple applications to run on a single server, significantly reduced the energy intensity of each hosted application.
3. Most compute instances migrated to large cloud- and hyperscale-class data centers, which utilize ultra-efficient cooling systems (among other important efficiency practices) to minimize energy use.

The 2010–2018 period was before AI became common, but a similar trend could happen after AI becomes well established, just as happened in the 2010s after large data centers became well established.

Data center types

Before discussing opportunities to improve data center efficiency, it is useful to discuss the four main types of data centers, because efficiency opportunities can vary by type. The four main types are hyperscale, colocation, enterprise, and edge. Hyperscale facilities are owned by companies such as

[2] They looked at the types of data centers that predominated before 2018. They did not look at AI data centers.

Google, Microsoft, Meta, and Amazon that deploy Internet services and platforms at massive scale. Colocation companies provide wholesale and retail colocation leasing, typically deploying large and very large data centers. Many companies will share server space and capacity in a colocation facility. Enterprise sites are owned and operated by firms such as banks and phone companies for their internal operations. Their advantage is security and control, but they may have low utilization rates, rigid designs, and long upgrade cycles. Some enterprises are moving less sensitive data to colocation sites. Edge centers serve local Internet of things providers such as Netflix and 5G cell services. While many facilities fit clearly in one category, the lines blur as, for example, hyperscalers will sometimes lease colocation space and colocation providers experiment with grid-edge deployment (Global Data Center Hub 2025).

An estimate of the number of servers in each class by year is shown in figure 2. AI processes are primarily carried out in hyperscale facilities, whereas colocation and enterprise facilities are the more traditional data center types. McKinsey & Company (2025) estimates about 70% of data center power demand growth from 2025–2030 will be at hyperscale.[3]

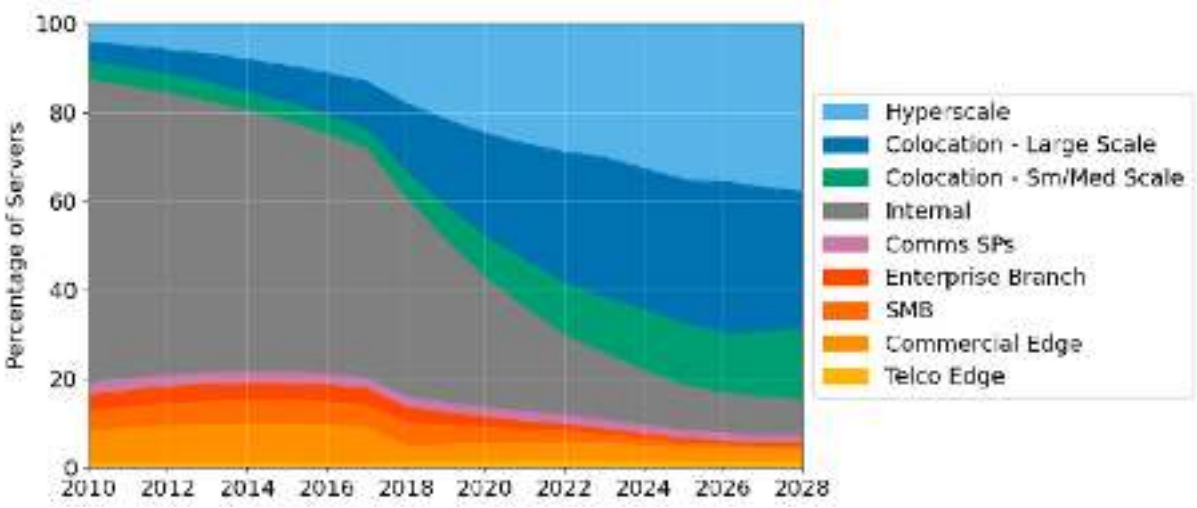


Figure 2. Estimated percentage of servers by data center type in the United States. Internal, comms service providers (SPs), and small/medium businesses (SMB) are part of enterprise. Source: Shehabi et al. 2024.

Opportunities for continued improvements in data center energy efficiency

Data center energy uses

To frame data center energy efficiency opportunities, it is useful to first understand where data centers use energy. A ballpark estimate is provided in figure 3. On average, about 80% of energy use is for

[3] The McKinsey estimate is for peak demand, while figure 2 is for number of servers. AI servers, such as those used in hyperscale facilities, can be particularly power intensive.

servers and other information technologies (IT) assets, about 15% is for cooling, and 5% for electrical equipment and lighting.

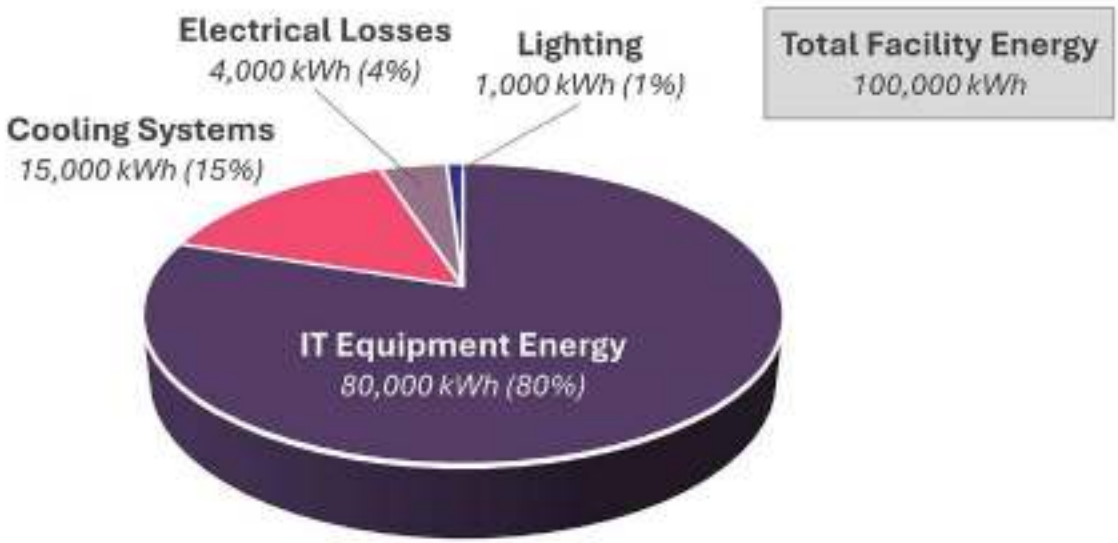


Figure 3. Distribution of data center energy consumption by component. This example is illustrative and assumes a PUE (defined later in this paper) of 1.25, a typical value for a very large data center. Source: Zhang 2024.

Summary of major efficiency opportunities

Based on this distribution, in the sections below we summarize opportunities to improve energy efficiency of servers and chips, software and algorithms, cooling, and heat recovery.[4] An appendix at the end of this paper discusses these opportunities in more detail.

Servers and Chips: Severs are where the computing occurs. In data centers, many servers are installed in vertical racks as shown in figure 4. Each server contains many chips, which is where much of the work takes place. Servers and the chips they contain are purchased when a data center is built, and also when existing servers need to be replaced. Servers and chips have been steadily improving in capacity and energy efficiency. Rahman at Epoch AI (2024) examined recent trends and found that leading chips are becoming about 40% more efficient each year (as measured by FLOPs per watt).[5] Contributing to this trend are not just improvements in general types of chips (CPUs and GPUs[6]) but also development of specialized chips for specific applications. It should be noted that the vast majority of these efficiency gains are channeled into additional computing power, allowing more computing per unit of energy, although potentially some of these efficiency gains could be used to reduce energy use.[7]

[4] This paper gets technical in places. For a more basic introduction, see Clean Energy Transition Institute (2025).

[5] FLOPs are floating point operations per second, a metric commonly used in the IT industry.

[6] CPU is a central processing unit, such as that contained in a typical personal computer. GPU is a graphic processing unit optimized for multiplication operations and is used in gaming computers and also for generative AI and cryptocurrency mining.

[7] This phenomenon is sometimes labeled “rebound” or “Jevons paradox”—when something becomes easier and less expensive, demand for the service can increase.

Figure 4. Inside of a data center, showing vertical racks with each rack containing many servers. Source: UL Solutions.

Software and algorithms: A variety of software is commonly used by large data centers to optimize operations. In recent years, virtualization of network services like firewalls and routers has become common, decoupling them from dedicated hardware and allowing operation of only as many servers as are needed at a given moment.

Another approach for reducing energy use is to write software code in ways that minimize calculations and optimize algorithms (Datacenters.com 2023). This is particularly important for AI software. AI models can be divided into two primary aspects: model training and model inference (applying the models in response to inquiries). Model training requires large numbers of high-power servers. As of 2025, training appears to account for 10–20% of AI energy use, with the remainder used for inference (Craddock 2025). Some models are developed more efficiently than other models, the classic case being Deep Seek, which had access to only a limited number of high-power chips and therefore had to construct their model very efficiently. Reportedly, Deep Seek consumed 90% less power while delivering similar capabilities to other AI models (Esram and Assadi 2025). This drastic reduction in hardware and energy usage was achieved through model compression (making models more compact and easier to run), highly efficient algorithms, and refined workload distribution—essentially, integrating software innovations with cost-effective hardware to maximize performance per watt (Esram and Assadi 2025). Furthermore, the Chinese firm z.ai claims to be even more efficient by using a model that automatically breaks tasks into subtasks in order to complete them more efficiently and accurately (Cheng 2025).

Another new frontier for software is employing AI to further improve operations. For example, Huawei used AI to analyze historical data, employ machine learning algorithms and statistical models, and forecast energy demands so systems can be optimized for these anticipated demands. They achieved an 8% reduction in energy consumption (Digital Reality 2025).

Leiserson et al. (2020) analyze algorithms and provide several examples of major improvements. Based on these examples they conclude that improvements in algorithms depend on human ingenuity and as a result occur unevenly and sporadically but with large efficiency gains possible (on the order of the

annual improvement in chips). Considering multiple software approaches, Patterson et al. (2022) estimate that if the whole machine learning field adopts best practices, by 2030 total carbon emissions from training will decline.

Inference energy use is growing as more and more users harness AI models. As noted above, energy used for inference substantially surpasses energy used for model training. Current models vary by about a factor of 100 on how much energy they use to respond to a medium-sized query. Jegham et al. (2025) tested 28 AI models, and for a specific medium-sized task the median model used 2.7 watt-hours while the best model used 0.3 watt-hours and the worst used 29 watt-hours. This test did not look at the quality of the model response, but the wide dispersion of consumption for the same query indicates an opportunity to learn from the best models and reduce energy use per query. Jegham et al. also looked at smaller and larger queries, and the results for these showed similar dispersion in energy use per query.

Cooling: Keeping servers and other equipment from overheating is critical to reliable operation and is the second largest energy consumer in data centers, accounting for approximately 15% of data center energy use (figure 3). Data centers have typically employed high-efficiency water-to-air chillers[8] as their primary cooling system, as these systems are higher efficiency than air-to-air systems. But water-to-air systems use a large amount of water—as much as 5 million gallons a day. In some regions with a lot of data centers (e.g., northern Virginia), water supplies are tightening (Mulkey 2024).

Traditionally, data centers have used sophisticated engineering and controls to maximize cooling performance. But two major developments provide the potential for substantial additional cooling energy savings: use of AI and liquid cooling.

AI allows optimization of cooling systems in ways that go well beyond prevailing practices. For example, Google employed its DeepMind research staff to analyze historical data on cooling system parameters and performance to train an ensemble of neural networks on the performance of specific cooling systems (e.g., system x at site y). In a case study, they document using the AI models they developed to optimize cooling system performance, reducing cooling energy use by 40% (Evans and Gao 2016).

Liquids are a much more efficient heat carrier than air and there is growing interest in using direct liquid cooling. Several approaches can be used. Figure 5 illustrates two of these: *direct-to-chip* and *immersion*. Direct-to-chip delivers a liquid (usually water) to the CPU or GPU, with a cold plate in between so the electronics are never in contact with the fluid. Claman (cited in Hutchinson 2024) estimates that direct-to-chip cooling reduces cooling energy use nearly 20%. This approach is illustrated on the left side of figure 5. With immersion cooling, all server components are submerged in a tank of nonconductive liquid coolant. This dielectric fluid absorbs and dissipates heat, carrying the warmed fluid away from the components and into a cooling system. Immersion cooling can reportedly reduce cooling energy use by 30% or more (Vincent 2025). For the highest-power chips, such as Nvidia Blackwell, immersion cooling will be essential. This approach is illustrated on the right side of figure 5.

[8] These cool air and transfer the heat to water circulating in the chiller. In many data centers groundwater is used.

Figure 5. Illustrations of liquid cooling techniques. Direct-to-chip is on the left (where cold plates sit atop the chip to remove heat) and immersion on the right. Sources: Vertiv and Tech News World.

Heat recovery: Data centers produce a lot of heat, which can potentially be used in other applications. The challenge is that this is often low-grade heat—typical temperatures of data center waste heat are 25–35°C (77–95°F) for air-cooled systems. But with water and refrigerant cooling, temperatures are generally higher: 50–60°C (122–140°F) for water-cooled systems, and up to 90°C (194°F) for two-phase refrigerant-cooled systems (Wang et al. 2024). With the move to the latest high-output chips and liquid cooling, more heat at higher temperature is available. One university reviewer of a draft of this paper thought this could be a game-changer for heat recovery.

This heat can be used for a variety of applications such as

- *District heating.*[9] Israelsen (2025) notes many data center heat recovery projects in Europe as well as one that helps heat an Amazon office in Seattle and a project now being planned for San Jose that will use heat from three data centers to provide district heat for 4,100 residences.
- *Greenhouses and other farming applications such as dairy*. As an illustration, a recent report specifically looked at opportunities to locate greenhouses near data centers in Virginia (Resource Innovation Institute 2025).
- *Industrial processes*, particularly low temperature processes such as food processing and pharmaceuticals
- *Desalinization for potable water production*
- *Other applications such as apartments or industrial processes* that use hot water

In many cases the low-grade heat from the data center needs to be raised in temperature, such as with a heat pump. That said, the data center exhaust is much warmer than the outside air heat pumps often use, and thus the heat recovery improves heat pump performance.

Heat recovery is easier and less expensive when the data center is designed with heat recovery in mind (Israelsen 2025). Such designs are possible for new data centers but are very difficult for most existing data centers.

[9] District heating systems have a central heating plant to heat water and distribute this hot water to provide heat to individual buildings in a neighborhood. These systems are common in Europe and China but there are only a few such systems in the United States.

To minimize heat loss, heat reuse sites should be near the data center. Increasingly, data centers are being located in rural areas; such projects should consider colocating factories or greenhouses nearby.

Data centers generally produce a lot more heat in summer than winter (since outdoor air can help with cooling in the winter). This factor needs to be taken into account when matching data centers to potential heat reuse sites.

A note on efficiency metrics for data centers

Data center efficiency has often been rated with a metric called *power utilization effectiveness* (PUE). PUE was introduced in 2006 and has become the most commonly used metric for reporting the energy efficiency of data centers. PUE was originally developed by a consortium called The Green Grid but was then revised and published in 2016 as a global standard under ISO/IEC (ISO 2016).

PUE = Total facility energy/IT equipment energy

Essentially, PUE is a measure of the efficiency of cooling and other auxiliary loads, since IT equipment energy is part of both the numerator and denominator. The ideal PUE is 1.0, which means no additional energy use beyond IT equipment.

Data center software commonly reports PUE in real time and often seeks to minimize PUE (e.g., see figure 6). According to the Uptime Institute (2025), globally the average PUE in 2024 was 1.56. Google has perhaps the lowest PUE in the data center industry, reporting an average annual PUE of 1.09 across all their large-scale data centers (once they reach stable operations), averaged across the year, including all sources of overhead. Their best data center had a PUE of 1.04 (Google 2025). One source published in 2024 (using 2022 or 2023 data) lists PUEs of 1.18–1.46 for several colocators and 1.08–1.22 for several hyperscalers (Zhang 2024).

While PUE is useful and has helped drive cooling and other improvements, it does not evaluate or capture server energy use or effectiveness. Many observers have called for an improved efficiency metric based on some measure of energy use (kWh) per unit of work done (e.g., Trueman 2024; Esram and Assadi 2025). There is not yet consensus on what metric to use for work done. Following are three concepts that have been developed:

1. Google researchers (Schneider et al. 2025) report on work to develop a new metric called compute carbon intensity (CCI). CCI is electricity emissions (based on electricity use and emissions per kWh) and performance (FLOPS). For now, they use data on measured FLOPS, but they are interested in future improvements to include only useful and not wasted work. They provide data on how a new tensor processing unit released in December 2024 improves CCI by a factor of three relative to the previous generation.
2. The Green Grid, a group of researchers and industry experts, developed the original PUE criteria, as well as a companion water use efficiency (WUE) criteria. In 2025 they released two new metrics: data center resource effectiveness (DCRE) and IT work capacity (ITWC). DCRE integrates energy effectiveness, water usage, and other factors. ITWC refers to server central processing unit (CPU) performance metrics that measure a server's processing capabilities. Green Grid has indicated they hope to add ITWC to DCRE in the future, which potentially could provide a path

to include server performance in an integrated metric (for more information, see the Green Grid Library[10]).

3. In 2011 several Intel and National Laboratory researchers proposed two new metrics: IT-power usage effectiveness (ITUE) and total-power usage effectiveness (TUE) (Patterson et al. 2013). ITUE is a PUE-like metric for the IT components. It is the ratio of total energy use into the IT equipment (both compute and support energy) to total energy use of the compute components. Like PUE, it will be greater than one. It does not evaluate or capture computing efficiency but only inefficiencies in the IT system. TUE is PUE times ITUE. Shehabi et al. (2024) note that these metrics "have seen little uptake" but also note that "computation efficiency alone provides little insight into energy efficiency opportunities without understanding how those computations are applied to different workloads."

In our view, the Google work noted in point 1 above seems particularly promising.

How much efficiency savings are available?

It is difficult to estimate how much these different efficiency opportunities can reduce overall data center energy use. Some of these opportunities are already being used in the newest data centers, such as state-of-the-art chips and cooling systems and some software improvements. For example, Google estimates that over a recent 12-month period, the energy and total carbon footprint of the median Gemini Apps text prompt dropped by 33x and 44x, respectively, all while delivering higher-quality responses (Ellsworth et al. 2025). But their energy use did not go down by 33x: Much of this efficiency gain went into increased computing. Thus, a key question for the future is how much of the efficiency improvement will go to energy savings versus how much will enable increased computing? The vast majority is likely to go to increased computing due to rebound.

The one published estimate we found about potential energy savings from energy efficiency is by the International Energy Agency (IEA). They published several scenarios of possible future global energy demand by data centers including a base case and a high-efficiency case, with the latter case having about 15% lower energy use (IEA 2025). Their scenarios are shown in figure 8.

[10] www.thegreengrid.org/resources/library-and-tools provides many useful resources.

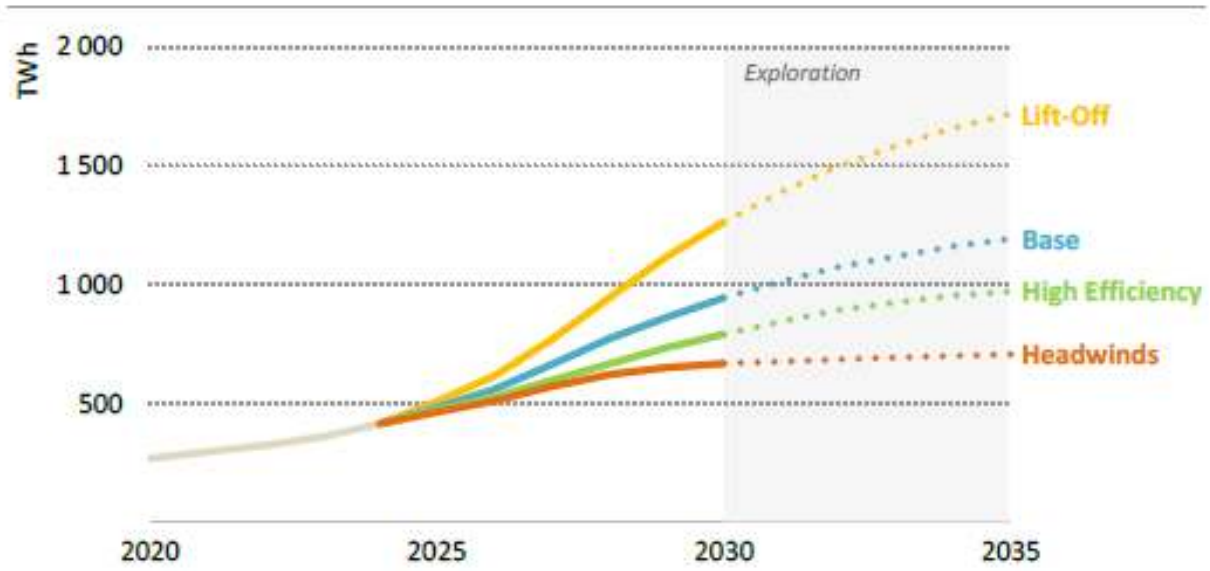


Figure 8. Global data center electricity consumption by IEA sensitivity case. Source: IEA 2025.

Another input on the question of how much efficiency is available is provided by researchers at Berkeley Lab (Shehabi et al. 2024) who published a range of potential U.S. data center energy use, as shown previously in figure 1. The lower bound is just over half the energy use of the upper bound. The range is due to efficiency variations for cooling as well as variations in equipment assumptions based on the range of processing unit shipments and the average operational power and operational time of servers. It is unclear how much of the range is due to efficiency versus differences in other equipment assumptions, but clearly efficiency plays a significant role.

Potentially, much greater efficiency savings are possible than 15%. For example, as discussed earlier, chips are getting 40% more efficient each year. Much of these savings will go to increased computations. If only 10% goes to energy savings, after five years of chip efficiency savings, server energy savings will total about 20%.[11] Similar efficiency improvements could be available from improved software and algorithms as Leiserson et al. (2020) suggest, also discussed earlier. Together, these improvements could drive energy demand down by about 40%, similar to levels shown in the lowest energy use scenario in figure 8.[12] Clearly, large energy efficiency opportunities are available, but it is very unclear how much of these gains will go toward energy savings and not to more computations.

Demand flexibility opportunities

Prior sections discuss the many opportunities to improve the efficiency of data centers and reduce annual energy use. But there are also opportunities to shift load away from periods of peak demand, helping to reduce the need for new generation and other grid resources that serve data centers.

In early 2025 the Nicholas Institute at Duke University published a study looking at available power for data centers and other new loads (Norris et al. 2025). They examined 22 of the largest power balancing

[11] 4%/year (10% of 40% per year) times five years is about 20%

[12] This lowest energy use scenario is labeled “headwinds,” but high use of efficiency measures would produce a similar result.

authorities (e.g., Regional Transmission Organizations) serving 95% of U.S. peak load. They developed a first-order estimate of the gigawatts of new load that could be added in each balancing authority before total load exceeds what system planners are prepared to serve, provided the new load can be temporarily curtailed as needed. This serves as a proxy for the system's ability to integrate new load, which they term *curtailment-enabled headroom*. They find

- 76 GW of new load—equivalent to 10% of the nation's current aggregate peak demand—could be integrated with an average annual load curtailment rate of 0.25% (i.e., if new loads can be curtailed for 0.25% of their uptime). An illustration is provided in figure 9.
- 98 GW of new load could be integrated at an average annual load curtailment rate of 0.5%, and 126 GW at a rate of 1.0%.
- The number of hours during which curtailment of new loads would be necessary per year is comparable to those of existing U.S. demand flexibility programs.
- The average duration of load curtailment (i.e., the length of time the new load is curtailed during curtailment events) would be relatively short, at 1.7 hours when average annual load curtailment is limited to 0.25%, 2.1 hours at a 0.5% limit, and 2.5 hours at a 1.0% limit.
- Nearly 90% of hours during which load curtailment is required retain at least half of the new load (i.e., less than 50% curtailment of the new load is required).

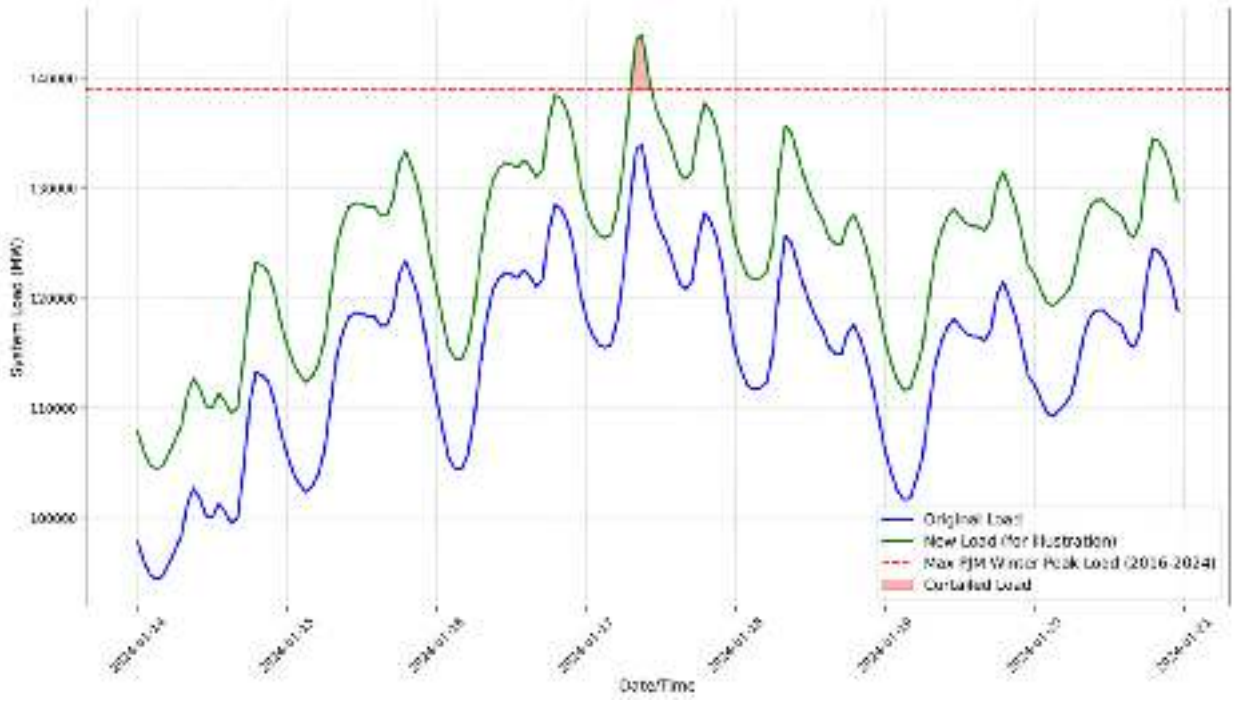


Figure 9. Illustration of recommended load reduction (shaded in orange) in study by Duke University researchers. This example is for the area served by PJM. Source: Norris et al. 2025.

To help address this opportunity, the Electric Power Research Institute (EPRI) has formed a new initiative—DCFlex—to explore how data centers can support the electric grid, enable better asset utilization, and support the clean energy transition. The initiative's founding members include Compass Datacenters, Constellation Energy, Duke Energy, the Electric Reliability Council of Texas (ERCOT), Google, Meta, New York Power Authority, NRG Energy, Nvidia, Pacific Gas and Electric Company (PG&E), PJM

Interconnection, Portland General Electric, QTS Data Centers, Southern Company, and Vistra Corp (EPRI 2024b). In June 2025, they announced three initial projects:

1. Nvidia and Oracle are partnering with Emerald AI to use Emerald's new Conductor platform that "enables AI data centers to flexibly adjust their power consumption from the electricity grid on demand, orchestrating the computing loads of inference, training, and fine-tuning AI models across a network of data centers to bolster the power grid's reliability while meeting strict compute performance standards" (Emerald AI 2025). The Emerald platform continuously profiles jobs across multiple dimensions—including computational flexibility, time sensitivity, and performance tolerance—to conduct dynamic load balancing. When an incoming signal from the grid requires power reduction, the platform models thousands of optimization scenarios in seconds, predicting the effect of each on both power draw and AI performance metrics. Then, rather than shutting down entire data halls, the platform can slow down less critical tasks, pause batch processing that is not time sensitive, and reschedule flexible workloads. In an initial test, an Oracle data center in Phoenix reduced its power consumption by 25% during peak grid demand hours—while claiming it was also maintaining AI compute quality. It did so by choreographing clusters of Nvidia GPUs in real time, on a task-by-task basis (Allsup 2025).
2. In North Carolina, Google will work directly with the local utility, Duke Energy, to temporarily reduce power consumption on an event basis. This approach involves a data center offloading its computing tasks to other facilities or pushing those tasks to a different time frame to reduce the overall load on the grid at a given time (Tilton 2025).
3. At a site in Paris, France, data center operator Data4 will work with Schneider Electric and RTE, France's transmission system operator, to explore how an uninterruptable power supply (UPS) system can be used to power through voltage and frequency issues. Currently, voltage drops and other grid issues can trigger data centers going offline to protect their computers from damage (Tilton 2025).

In another recent development, Google announced the first known contracts between a hyperscaler and U.S. utilities around AI data center flexibility. This is the first public case of U.S. data center flexibility being incorporated into the utility planning realm, instead of limited to operations. Google is using demand flexibility to leverage flexible workloads to respond to periods of high stress on the local power grid. This is done by temporarily reducing the availability of a subset of machine learning hardware. The flexibility is time limited, available for only "certain hours or times of the year." Google has signed long-term contracts with Indiana-Michigan Power and the Tennessee Power Authority that make use of this capability (Terrell 2025). A key motivation for Google appears to be getting access to power in exchange for agreeing to constrain power use during a limited number of critical hours. As noted by Google: "We can't do it everywhere. Some of our loads can't be curtailed." [But where we are able to do it], "there's value to being able to secure capacity without having to wait for new infrastructure" (St. John 2025b).

Following are a few other ways to reduce data center peak loads:

- Using the dynamic voltage and frequency scaling reduces the electrical frequency in GPUs and some other equipment, slowing them down and reducing energy use. This approach, which is sometimes called *de-clocking*, has been used to reduce loads during periods with reduced compute needs, but it may not work in certain performance sensitive applications such as AI. We are not aware of it yet being used for demand flexibility.
- Use of batteries or on-site generation can help glide through peak periods. Versus, a data center developer, has worked with the National Renewable Energy Laboratory (NREL) to test the

concept, mixing sophisticated controls with utility-scale batteries (e.g., on the order of 70 MW per site) (Giacobone 2025b).

- Cool storage could help to reduce cooling demand during peak hours. Such systems have been used in many commercial buildings. Nun (2023) discusses this option for data centers.

In addition to addressing peak electric loads, Google is also managing loads to increase use of clean energy by shifting loads between data centers to take advantage of available clean energy (Koningstein 2021).

Demand flexibility for data centers is a rapidly emerging area that is likely to grow substantially in the coming months and years. The success of demand response growth will depend on the development of the right incentives and market constructs to meet data center constraints while capitalizing on the demand reduction opportunity. The various projects discussed above show that some load reductions and shifting are possible through strategic partnerships and planning. However, load reductions can be costly and disruptive, so demand response disruptions should have guardrails that hyperscalers can count on and plan around. Load shifting without clear guardrails has been met with opposition from the data center industry (Zeitlin 2025). Demand flexibility should be considered by utility planners as part of a solution set to help make decisions on how to best meet load growth affordably and efficiently.

Predicting and managing future data center energy demands

There is great uncertainty as to the amount of power data centers will need in the future. Figure 1 earlier in this paper shows an estimate that data centers will account for approximately 6.7–12.0% of U.S. electricity consumption in 2028. The researchers note that the amount of energy efficiency adopted is one contributor to this range (Shehabi et al. 2024). Therefore, the more that data center developers and owners pursue energy efficiency, the more likely that electricity demand will be at the low end of this range.

Likewise, global scenarios prepared by IEA and summarized in figure 8 show a factor of three difference in 2035 energy use between the lowest and highest demand scenarios. With an additional seven years of uncertainty, the range grows.

High-end forecasts essentially assume a substantial rebound effect: More efficient and cheaper AI spurs increased demand for AI and data centers. This in turn implicitly assumes that a large number of recently proposed data centers would in fact be built. While possible, in our view this is unlikely. In recent years, while many data centers have been proposed. Most data center developers seek permitting and interconnection for many more sites than they expect to build, allowing them to see which projects can overcome obstacles and which cannot. For example, the chief executive officer (CEO) of one major utility (Public Service Electric & Gas in New Jersey) has estimated only 10–20% of proposed data centers will actually be built (Howland 2025). Some data center experts have made similar estimates and even some hyperscalers have commented about over forecasts (Martucci 2025a; St. John 2025a). Lovins (2025) suggests a “small fraction” will actually get built.[13] An expert cited by Martucci (2025a) noted that even large companies such as Microsoft, Meta, Amazon, and Google propose several

[13] The figure noted in the paper is approximately 18%.

times more projects than they are likely to need due to uncertainties around power availability at a given site, while less sophisticated developers abandon proposed projects at an even higher rate. Reviewers from a university and a research institute who reviewed a draft of this paper thought that a 10–20% completion rate is too low but did not provide an alternative figure. For the high-end scenario to apply, the number of proposed data centers that are actually built would have to be substantially higher than this range. Also, even when built, many data centers do not operate at full capacity (Norris 2025).

It should also be noted that AI developers are often in a hurry to build their data centers because they are looking for power within 18 months. Utilities and other power plant developers are rarely able to move that quickly, resulting in creative ways to seek power quickly, such as restarting closed nuclear power plants and building data centers inside the fence-line of power plants to avoid some utility interconnection and other regulations.

As noted above, widespread use of efficiency improvements would drive power demand toward the lower end of the range shown in figure 8. But a variety of headwinds could also drive demand lower. Among the potential headwinds mentioned in recent (2025) publications include the following:

- Concerns that improvements in AI are slowing down; for example, several observers have suggested that the improvement from ChatGPT v4 to v5 was smaller than were the improvements from v2 to v3 and from v3 to v4 (Newport 2025)
- Concerns about whether profits from AI will continue to justify the massive investments being made (Lehr 2025)
- Estimates that there are unlikely to be enough chips to serve high forecasts of U.S. AI power consumption (Walton 2025)
- Questions about demand for AI given issues about the accuracy of responses provided by AI (Koomey, Das, and Schmidt 2025)
- Concerns that the high-quality data needed to train models has mostly been tapped, and that limits on the amount of additional high-quality data will limit what AI can do, thus limiting demand for AI (Chakrabarti 2025)
- Questions about the amount of power that can be supplied (Krugman 2025)

These concerns could perhaps be overcome, but they do place obstacles in the path of rapid AI growth.

Given these uncertainties, power system planners would be wise to carefully review proposals for new data centers and apply probabilities to each of these proposals. Koomey, Das, and Schmidt (2025) find that many “utilities are collecting better data, tightening criteria about how to ‘count’ projects in the pipeline, and assigning probabilities to projects at different stages of development. These changes are welcome and should help reduce uncertainty in forecasts going forward.”

At the national level, bottoms up models based on the number of servers and their use have been much more accurate than top-down models based on proposals (Masanet, Lei, and Koomey 2024). Shehabi et al. (2025), the source of figure 1, is a good example of such a model. But it is probably not possible to do a bottoms-up model at the local level, since data can move across the country at the speed of light and data centers are not spread evenly across the country, but instead tend to be concentrated in areas with favorable conditions (power availability and price, water availability for cooling, etc.). Various proposals have been made to more easily figure out which proposed data centers get built; for a discussion of some proposals, see Martucci (2025a) and St. John (2025a).

The cost to electric consumers of data center growth

High load growth has a cost as additional electrical generation, transmission, and distribution investments need to be made. Typically, these costs are paid by all customers and thus can result in higher electric rates for everyone. To provide just two examples:

- The Independent Market Monitor for PJM (the system operator serving 13 states in the Mid-Atlantic and Midwest) found that "data center growth is the primary reason for recent and expected market capacity conditions, including total forecast load growth, the tight supply and demand balance, and high prices. But for data center growth, both actual and forecast, the PJM Capacity Market would not have seen the tight supply demand conditions, the high prices observed in the BRA [Base Residual Auction] for 2025/2026 or the high prices expected for the 2026/2027 and subsequent capacity auctions. Holding aside all the other issues raised [in] this report, data center load by itself resulted in an increase in the 2025/2026 BRA revenues of $9,332,103,858 or 174.3 percent" (Monitoring Analytics 2025).
- A study by researchers at Carnegie Mellon and North Carolina State Universities found that data center and cryptocurrency mining growth through 2030 could increase average U.S. electricity generation costs by 8%, with increases of more than 25% in Central and Northern Virginia (Wade et al. 2025).

To address these costs, a number of states are adopting and considering special rules for data centers and other large new loads, to better ensure that the costs of power for data centers are not passed on to other customers such as households and other businesses. Berkeley Lab has published a technical brief describing some of these efforts (Satchwell et al. 2025) and the Smart Electric Power Alliance (SEPA) has a database on many of these efforts (SEPA 2025).

Energy efficiency and demand flexibility could be a complementary approach to help reduce these costs and rate increases.

Potential roles for policies and programs

Given the large potential for energy efficiency and demand flexibility savings, as well as the growing cost of power and questions about power availability, the data center industry is doing a lot to improve data center efficiency. Programs and policies implemented by utilities and states could help increase these efforts. Potential opportunities include efficiency targets (mandatory, voluntary, or tied to electric rates) as well as energy savings incentive programs.

Efficiency targets

Several initiatives are already underway, primarily in Europe, as summarized by Energize Capital (2025):

- *European Green Deal & Energy Efficiency Deal*: This set of initiatives that aims to make Europe climate neutral by 2050 includes a number of requirements for data centers. Starting May 2024, data centers (>500 kW) are required to report a range of information and key performance indicators. These encompass energy consumption (kWh), power utilization (kW), and renewable energy sources, among others.
- *German Energy Efficiency Act*: Requires all data centers that start operations before July 2026 to sustain a PUE of 1.5 or lower by 2027, and eventually 1.3 PUE or lower by 2030

- *Climate Neutral Data Centre Pact*: ~100 data center operators have signed or are already certified under this pact. Objectives include (by 2025)
    - New data centers operating at full capacity in cool climates will meet an annual PUE target of 1.3; in warm climates the target is 1.4.
    - Data center electricity demand will be matched by 75% renewable energy or other carbon-free energy, matched on an hourly basis.

U.S. states could potentially set similar targets as mandatory requirements in order to obtain commitments for large amounts of electric power. States and/or utilities could also set such targets as voluntary objectives or as requirements to be eligible for favorable electric rates. Participation in demand flexibility programs and even demand flexibility targets could also be conditions for special rates.

In the near term, efficiency targets could be in terms of PUE, but given some recent PUEs of less than 1.1, "free riders" (customers who would have made the efficiency improvements even if there were not a program) are likely to be high. For incentives to be workable, a broader efficiency metric in terms of kWh/FLOP will probably be needed.

Sometimes, a set of targets or other steps can be negotiated with data center developers in exchange for access to power. The Google deals with Indiana-Michigan and the Tennessee Valley Authority (TVA) (discussed earlier) are examples, as is legislation passed in Texas, for which the industry had substantial input (Martucci 2025b).

Incentive programs

The data center industry is well capitalized and has incentives to invest in energy efficiency to better assure power availability, reduce costs, and improve its image. Thus, incentives should probably be used sparingly. This said, opportunities might include the following:

- Demand flexibility programs can reduce data center peak electric load loads, including by promoting use of Emerald AI and similar software. However, data centers need to meet reliability requirements and have operational constraints, so they can only sometimes shift loads. Programs may need to be optimized for data center participation (e.g., in terms of advance notice, limits on how often shifts are requested) instead of assuming that traditional programs will work for data centers.
- Incentives can be introduced for reducing PUE below specified targets. Particularly for new construction, any PUE targets should be set fairly low. The incentives could be rebates based on output and PUE, or they could take the form of rate reductions according to a defined schedule in a data center rate (increasingly, utilities are developing special rates for large data centers—Satchwell et al. 2025; SEPA 2025). A new, broader efficiency metric, once developed, could be used instead of PUE.
- Particular emerging technologies, such as improved energy storage, could be incentivized.

Another program option is for data centers to help pay for load-reducing retrofits for other customers, freeing up capacity to serve data centers. As an example, Google is starting to fund community organizations working on weatherization and solar for low-income homes (Solstice Power Technologies 2024). And Rewiring America has proposed that in order to free up capacity for data centers, hyperscalers should cost share conversion of electric resistance heat in homes to heat pumps and also cost share installation of home solar and battery storage systems (Wyant, Verma, and Kanj 2025).

Data and metrics, research and development (R&D)

Data on data center energy consumption and other parameters are provided by a few firms such as Google, but many firms keep the data internal. Given their large impact on electric grids, governments should consider establishing explicit data center reporting requirements. Such requirements are in place in Europe. To preserve confidentiality, data from multiple firms could be aggregated so that firm-specific proprietary data are not divulged.

Industry and government, such as the national laboratories, should take the lead in developing a broader data center efficiency metric, such as kWh/FLOP. The Google CCI metric may be a good place to start.

While the data center industry is well capitalized and does not need much R&D support, one potentially productive research area is understanding differences between models in energy use and accuracy, and using this to develop guidelines for model energy use, helping to improve all models, but particularly ones with high energy use.

Conclusions

Data centers are contributing to substantial load growth in many regions of the U.S. Data centers accounted for about 4% of U.S. electricity sales in 2023, with projections ranging from 6.7–12.0% in 2028. Data centers are often the largest source of load growth for utilities through 2030; achieving higher levels of energy efficiency is critical for keeping the lights on and constraining rising power costs.

Data centers have improved dramatically in energy efficiency in recent years, but many large opportunities remain to improve even further. For example, the chips in servers have been improving in efficiency (computations per kWh) at a rate of 40% per year; if just 10% of these efficiency gains were channeled into energy savings and not more computations, after five years energy use would be reduced by 20%. A similar scale of available savings exists from improved software and algorithms. Improved cooling and electric systems can provide additional savings.

Data center energy use roughly doubled from 2000–2010 but only rose modestly from 2010–2020 due to improved efficiency of IT devices, greater use of virtualization software, and movement of more computations to large data centers with very efficient cooling systems. As discussed earlier, Patterson et al. (2022) estimate that if the whole machine learning field adopts best practices, by 2030 total carbon emissions from training will decline. AI is becoming established in the 2020s, with many new data centers built to handle this growth; concurrent with and after this growth, we can hopefully enter a phase of optimizing AI and data center efficiency, applying the many techniques discussed in this paper and leveling off energy use while computations continue to grow, as happened in the 2010s. If the energy efficiency opportunity is fully realized, data center energy consumption can be brought toward the low end of recent power demand forecasts, such as those from Berkeley Lab (Shehabi et al. 2024) and IEA (2025).

Complementing energy efficiency are opportunities to reduce loads during periods of peak electric demand using demand flexibility. Duke University researchers estimate that curtailing data center loads for just 0.25% of their maximum uptime would free up enough capacity to accommodate 76 GW of new load. As one example, a test of a new software platform reduced the peak power consumption of an Oracle data center by 25% during peak grid demand hours.

Data center efficiency is commonly measured with a metric called PUE, which looks at just auxiliary systems such as cooling but not the server part of data centers. This has helped to drive substantial

improvements in data center cooling and power system efficiency, but new metrics are needed to better capture all data center energy use relative to the computations, in order to make metric-based programs and targets possible. We find the Google CCI metric discussed earlier a promising start as it is essentially energy use per unit of computation (e.g., kWh per FLOP), plus a factor for emissions.

Many new data centers have been proposed but most data center developers seek permitting and interconnection for many more sites than they expect to build. Several experts estimate only 10–20% of proposed data centers will actually be built; others estimate a higher percentage. Forecasting of data center power demand is improving but needs further improvement in order to support energy planning and minimizing the impact of data centers on electricity costs.

The data center industry is doing a lot to improve data center efficiency. Utilities and states could implement programs and policies to help accelerate these efforts. Potential opportunities include efficiency targets (mandatory, voluntary, or tied to electric rates) as well as energy savings incentive programs.

Recommended next steps

Given our findings, we recommend several immediate steps for policymakers and those interested in using energy efficiency and demand flexibility to help reduce data center energy use and the costs to electric consumers from growing data center loads.

1. Forecasts of future loads, while improving, continue to need attention. Proposed new data centers need to demonstrate that proposals are real through application fees, making long-term power purchase commitments, and other steps. Utilities should assign probabilities to each proposal to come up with probability-weighted forecasts.
2. Utilities and regulators should establish special rate classes for data centers and other large new loads, with the details designed to ensure these customers fully cover their costs of service without subsidies from other customers. An example of such a rate is one recently approved in Ohio (Skidmore 2025). Such rates should include discounts or other inducements for loads that can be flexed during critical peak periods.
3. Data center owners and developers should accelerate efforts to improve data center energy efficiency. Software and analytics are particularly attractive opportunities, complementing ongoing improvements in chips and servers. Improved efficiency will reduce data center power loads together with the emissions associated with those loads.
4. Utilities and data center operators should continue to experiment with approaches to flex load during peak demand periods. Appropriate approaches need to be developed for different data center tasks, so that operators have a range of options they can tap during critical periods. Utilities need to recognize that data centers have unique attributes that mean they will often need load flexibility approaches that are different from other types of large customers.
5. Targets and other policies now being used in Europe should be studied to see how well they work and whether they might be appropriate for the United States.
6. Data center operators, standards bodies, and national laboratories should develop, test, and finalize an efficiency metric for data centers that looks at all the energy used in the data center, including computation, and not just data center auxiliary systems.

Appendix: More detailed description of some of the major efficiency opportunities

Servers and chips

Servers and the chips they contain are purchased when a data center is built, and also when existing servers need to be replaced. In data centers, servers historically are replaced after 3–5 years before failures increase in frequency, although with good maintenance, servers can be used for 8 years or even longer (Infiniti undated). Recently, hyperscalers have lengthened server lifetime to reduce capital costs—major hyperscalers are now reporting 5–6 years (Brightmore 2024). While servers and chips do not present efficiency retrofit opportunities, efficiency can and should be a very important factor when servers, including replacement servers, are purchased.

Servers and chips have been steadily improving in capacity and energy efficiency. Rahman at Epoch AI (2024) examined recent trends and found that leading chips are becoming about 40% more efficient each year (as measured by FLOPs per Watt).[14] His findings are illustrated in figure A1.

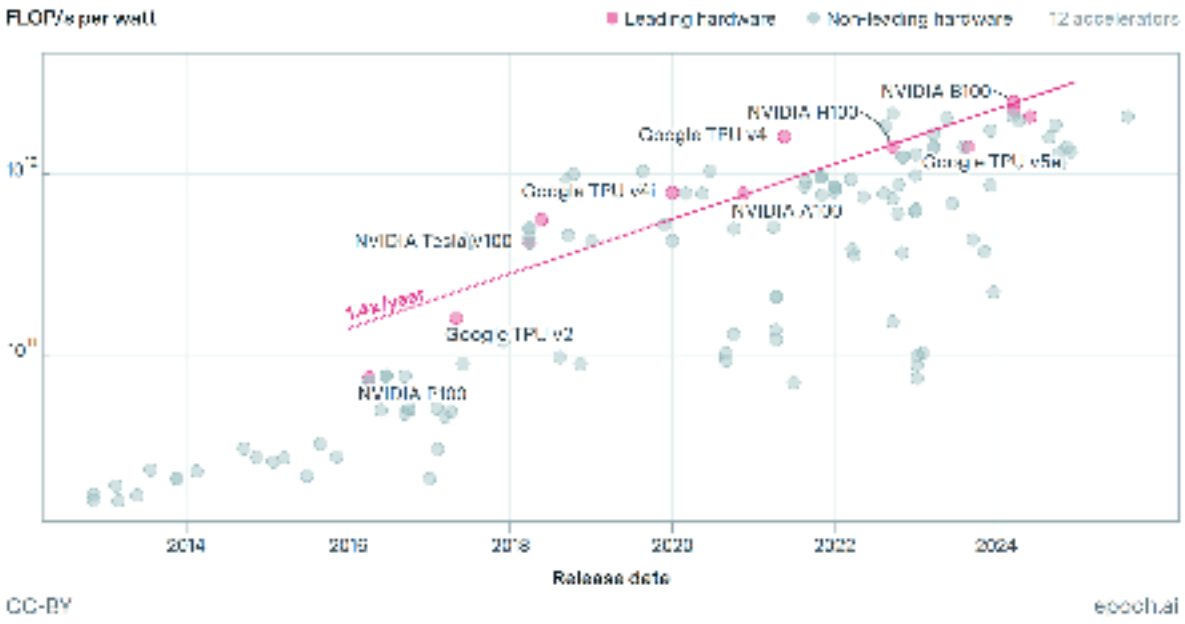


Figure A1. Energy efficiency (FLOPs per watt) of leading chips. Source: Rahman 2024.

Contributing to this trend are not just improvements in broad categories of chips (CPUs and GPUs[15]) but also development of specialized chips for specific applications such as tensor cores (better optimized for the matrix operations used with AI), Data processing units (DPUs, essentially mini-servers on a chip that

[14] FLOPs are floating point operations per second, a metric commonly used in the IT industry.

[15] CPU is a central processing unit, such as contained in a typical personal computer. GPU is a graphic processing unit optimized for multiplication operations and is used in gaming computers and also for generative AI and cryptocurrency mining.

take on data-intensive functions), neural processing units (NPUs, which are optimized for neural networks), flexible and reconfigurable chips, three-dimensional (3D) chips to improve performance without increasing footprint, chiplet architecture where modular processing units are combined to optimize specific tasks, multi-instance technology that allows partitioning a single GPU into several units for more efficient resource utilization, and on-chip memory that reduces time to save and retrieve information. These approaches are sometimes called *accelerated computing*. Given these many types of specialized chips, which are designed for AI applications, there is growing work to optimize AI computations by routing specific tasks to specific types of chips that are optimized for those tasks, better integrating hardware (chips) and software.

It should be noted that the vast majority of these efficiency gains are channeled into additional computing power, allowing more computing per unit of energy, although some of these efficiency gains could potentially be used to reduce energy use.

Software and algorithms

A variety of software is commonly used by large data centers to optimize operations, often under the heading data center infrastructure management (DCIM) software, which is sold by several major vendors. This software is widely used in conventional (non-AI) large data centers, but there may be some remaining opportunities in smaller data centers. DCIM mainly helps optimize infrastructure and does not give visibility into the actual computing. This software typically includes

- Virtualization of network services like firewalls, routers, and load balancers, decoupling them from dedicated hardware and allowing operation of only as many servers as are needed at a given moment; with virtualization, more than one application can be run on a server and a single application to be run on multiple servers, allowing server capacity to be optimized
- Tracking essential metrics such as power consumption, temperature, humidity, and equipment health, enabling data center administrators to identify and address problems, including addressing potential issues before they affect performance or cause downtime
- Automation of repetitive tasks such as server provisioning, monitoring, incident response, and power and server capacity management

A newer approach called containerization is taking virtualization to the next level. Containerization is a software deployment process that bundles an application's code with all the files and libraries it needs for running on any infrastructure. Containers enable fine-grained control over resource allocation, enabling data centers to allocate only the necessary CPU, memory, and storage resources for each application or service. This efficient resource utilization translates into reduced energy consumption, as unnecessary resources are not allocated or wasted. Containerization also enables rapid deployment, scaling, and migration of applications, leading to improved agility and optimized utilization of data center resources (Perez, Porter, and Narasimhan 2023).

Another approach for reducing energy use is to write software code in ways that minimize calculations and that optimize algorithms (Datacenters.com 2023). This is what Deep Seek did in developing their AI models with access to only a limited number of high-power chips (Esram and Assadi 2025). Using smaller models for specific jobs can also be an important tool for improving processing efficiency (Patterson et al. 2022). A Chinese company called z.ai claims to be even more efficient than Deep Seek by using a model that automatically breaks tasks into subtasks in order to complete them more efficiently and accurately (Cheng 2025). This can be a very large source of efficiency gains, particularly for AI, as discussed in the section just below.

Another new frontier for software is employing AI to further improve operations. For example, Huawei used AI to analyze historical data, employ machine learning algorithms and statistical models, and forecast energy demands so systems can be optimized for these anticipated demands. They achieved an 8% reduction in energy consumption (Digital Reality 2025). This source also includes examples from Meta and Microsoft but does not provide data on energy savings from these applications.

Efficiency of AI models

A subcategory of software efficiency is the software used to establish and use AI models. Models can be divided into two primary aspects—model training and model inference (applying the models in response to inquiries). Model training requires large numbers of high-power servers. As of 2025, training appears to account for 10–20% of AI energy use, with the remainder used for inference (Craddock 2025). Some models are developed more efficiently than other models, the classic case being Deep Seek. Reportedly, Deep Seek consumed 90% less power while delivering similar capabilities. This drastic reduction in hardware and energy usage was achieved through model compression (making models more compact and easier to run), highly efficient algorithms, and refined workload distribution—essentially integrating software innovations with cost-effective hardware to maximize performance per watt (Esram and Assadi 2025). And z.ai claims to be even more efficient, as discussed above.

Craddock (2025) suggests that "[t]he next frontier in AI isn't just about making models bigger—it's about making them smarter and more efficient." He goes on to list a variety of techniques that can be used:

- Model distillation: Compressing larger models into smaller, more efficient versions while preserving core functionality
- Sparse architecture design: Implementing attention mechanisms that focus computational resources only where needed
- Quantization techniques: Reducing model precision without significant performance degradation
- Neural architecture search (NAS): Automated discovery of optimal model structures for specific tasks
- Modular design patterns: Creating reusable, specialized components that can be combined efficiently
- Parameter-efficient fine-tuning (PEFT) techniques reducing resource requirements by up to 95%
- Mixture-of-experts architectures enabling dynamic resource allocation
- Task-specific pruning strategies eliminating redundant computational paths
- Energy-aware architecture design incorporating power consumption as a design constraint
- Adaptive computation time mechanisms allowing models to vary computational effort based on input complexity

He suggests that "[t]he implementation of efficient model architectures must be approached holistically, considering both immediate resource savings and potential rebound effects… [S]uccessful deployment requires careful consideration of the entire AI lifecycle, from development through to deployment and maintenance. This includes establishing clear metrics for efficiency, regular monitoring of resource usage patterns, and mechanisms to prevent efficiency gains from simply enabling more extensive model deployment without strategic justification." He notes that a "senior technical advisor to government AI

initiatives" told him that government departments have reduced their AI infrastructure costs by 60% through the implementation of efficient architectures, while actually improving model performance in their specific use cases.

Leiserson et al. (2020) hone in on algorithms and provide several examples of major improvements in algorithms. Based on these examples, they conclude that improvements in algorithms depend on human ingenuity and as a result occur unevenly and sporadically but with large efficiency gains possible (on the order of annual improvement in chips).

Some recent articles discuss technical ways to implement some of these strategies, including "accurate quantized training," "speculative decoding," and "pareto governors for energy-optimal computing" (Lew and Zhang 2023; Leviathan et al. 2024; Sen and Wood 2027). We mention these for those who want specific examples but note that the underlying papers are very technical.

Considering all of these approaches, Patterson et al. (2022) estimate that if the whole machine learning field adopts best practices, by 2030 total carbon emissions from training will decline.

Inference energy use is growing as more and more users harness AI models. As noted above, energy used for inference substantially surpasses energy for model generation. Current models vary by about a factor of 100 on how much energy they use to respond to a medium-size query. Jegham et al. (2025) tested 28 AI models, and for a specific medium-sized task the median model used 2.7 watt-hours, while the best model used 0.3 watt-hours and the worst used 29 watt-hours. This test didn't look at the quality of the model response, but the wide dispersion of consumption for the same query indicates an opportunity to learn from the best models and reduce energy use per query. Jegham et al. also looked at smaller and larger queries and the results for these showed similar dispersion in energy use per query.

Finally, it should be noted that there are opportunities to co-design software and hardware to leverage synergies and maximize energy efficiency and performance (IEA 2025).

Cooling

Keeping servers and other equipment from overheating is critical to reliable operation and the second largest energy consumer in data centers, accounting for approximately 15% of data center energy use (figure 3). Data centers have typically employed high-efficiency water-to-air chillers[16] as their primary cooling system, as these systems are higher efficiency than air-to-air systems. But water-to-air systems use a large amount of water—as much as 5 million gallons a day. In some regions with a lot of data centers (e.g., northern Virginia), water supplies are tightening (Mulkey 2024).

Traditionally, data centers have used sophisticated engineering and controls to maximize cooling performance. But two major developments provide the potential for substantial additional cooling energy savings: use of AI and liquid cooling.

AI allows optimization of cooling systems in ways that go well beyond prevailing practices. For example, Google employed its DeepMind research staff to analyze historical data on cooling system parameters and performance to train an ensemble of neural networks on the performance of specific cooling systems (e.g., system x at site y). They also used AI to look at temperature and pressure data and improve predictions of site conditions an hour ahead. They then combined these two models to

[16] These cool air and transfer the heat to water circulating in the chiller. The water is then typically cooled using cooling towers which transfer this heat to the air outside.

optimize cooling system performance, reducing cooling energy use by 40% (Evans and Gao 2016). Figure A2 illustrates a sample control period showing cooling impacts from their machine learning algorithms.

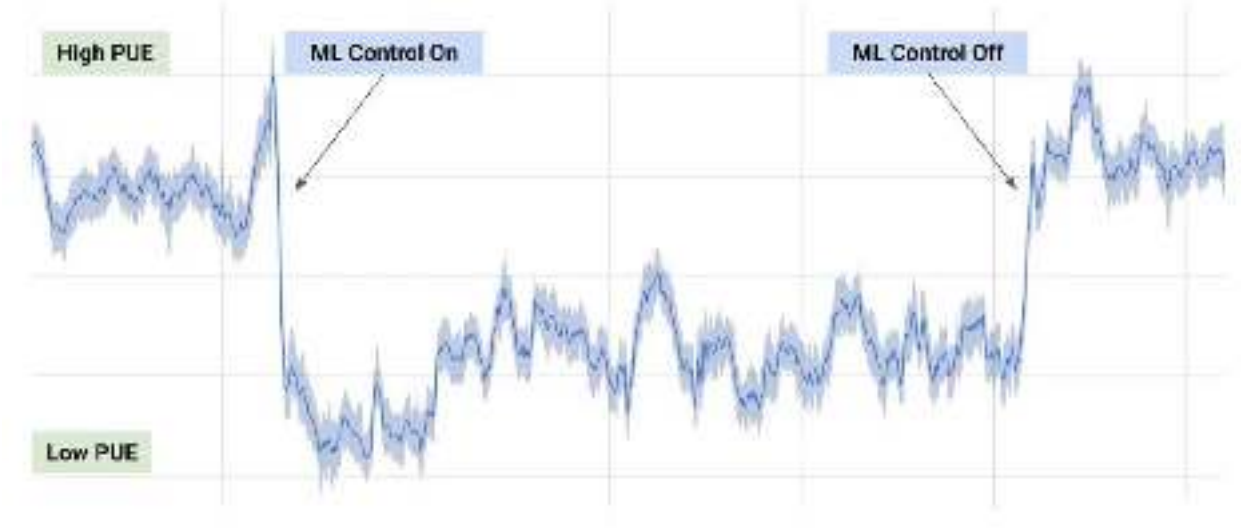


Figure A2. Illustration of how a machine learning (ML) algorithm reduced cooling energy use at Google. PUE is power usage effectiveness, a measure of site energy efficiency. Lower PUE means less energy is used. The y-axis is PUE but specific values are not provided in the article from which we take this figure. Source: Evans and Gao 2016.

In many current data centers, both water-to-air and air-to-air cooling produce cold air, which is blown across server racks to cool them. Liquids are a much more efficient heat carrier than air, and there is growing interest in using direct liquid cooling. Presently two types of systems are being deployed (Trane 2024). These systems are

1. *Rear Door Heat Exchanger.* A coolant-filled exchanger is mounted to the back of the server. An internal fan blows heat out of the server and across the exchanger, cooling it down before it is released into the data center's indoor air environment. Because they can be implemented without major modifications to the existing servers, rear door heat exchangers can be a cost-effective solution for data centers seeking to improve cooling.
2. *Direct-to-Chip*. This method delivers a liquid (usually water) to the CPU or GPU, with a cold plate in between so the electronics are never in contact with the fluid. Fans then blow in chilled air to move the tempered air away from the servers. The data center's chiller still plays a major role with this type of liquid cooling. This approach removes about 70–75% of the heat in the rack; air-based cooling is needed to remove remaining heat (Vertiv 2025). Claman (cited in Hutchinson 2024) estimates that direct-to-chip cooling reduces cooling energy use nearly 20%. This approach is illustrated on the left-hand side of figure 5 in the body of the report.

A third type of liquid cooling system is even more efficient and is now starting to be deployed:

3. *Immersion Cooling*. With this technology, all server components are submerged in a tank of nonconductive liquid coolant. This dielectric fluid absorbs and dissipates heat, carrying the warmed fluid away from the components and into a cooling system. There, the heat is expelled using other cooling methods. Immersion cooling offers high cooling efficiency compared to other liquid technologies. It is also quieter, more energy efficient, and more compact, allowing for more densely packed hardware configurations (Trane 2024). Immersion cooling can reportedly reduce cooling

energy use by 30% or more (Vincent 2025). For the highest power chips, such as Nvidia Blackwell, immersion cooling will be essential. This approach is illustrated on the right-hand side of figure 5. World.

While each type of liquid cooling system offers advantages, they have downsides as well (Trane 2024).

- Initial costs are high due to the need for specialized equipment such as coolant tanks, pumps, and heat exchangers.
- There is the risk of leaks, which can damage equipment and cause outages.[17]
- Liquid cooling components require regular maintenance, including periodic fluid replacement.
- Certain hardware components cannot be used with liquid cooling systems due to their design and/or materials.
- Systems are complex and require careful design, installation, and maintenance, with special attention paid to proper sealing, plumbing, and monitoring.
- Because these are new technologies, employees who are accustomed to maintaining air-cooled heating, ventilating and air-conditioning (HVAC) systems will experience a learning curve.

In addition, we note that immersion cooling systems and some direct to chip systems use "forever chemicals" such as PFAS (per- and polyfluoroalkyl substances), which need to be very carefully managed to avoid health and environmental problems.

Use of liquid cooling is starting to accelerate with companies such as Microsoft, Vertiv, and Evolution generally using liquid cooling in their new data centers (Craske 2025).

Other hardware

Other hardware, such as electric equipment, also contributes to data center energy use. Some potential opportunities to reduce this energy use include

- 48-volt power supply for servers (rather than the traditional 12 volt) to reduce energy losses and improve system efficiency; in tests these have been shown to reduce energy losses by at least 25% (McKinsey& Company 2024)
- 1.6 terabyte ethernet, doubling the speed relative to current typical ethernets and reducing latency and power consumption (Vincent 2025)
- High bandwidth memory, a type of memory that achieves higher data transfer rates than traditional memory by using a 3D stacked architecture; this stacking allows for a wider memory bus and increased data transfer speeds, increasing speed and reducing energy use (Chang and de Costa 2024)
- Direct current electrical infrastructure to reduce AC/DC conversion losses (e.g., see Siemens 2025)

[17] To address this issue, some facilities operate their liquid loops at below atmospheric pressure so that if a leak develops, atmospheric air goes into the loop but liquid does not come out.

- Improved efficiency uninterruptible power supplies; significant progress has been made but more is possible (Shehabi et al. 2024)

Heat recovery

This is discussed in the body of the report.

References

Allsup, M. 2025. "Nvidia and Oracle Tapped This Startup to Flex a Phoenix Data Center." *Latitude Media*. July 1. www.latitudemedia.com/news/nvidia-and-oracle-tapped-this-startup-to-flex-a-phoenix-data-center/.

APS (Arizona Public Service). 2023. *2023 Integrated Resource Plan*. Phoenix, AZ: APS. www.aps.com/-/media/APS/APSCOM-PDFs/About/Our-Company/Doing-business-with-us/Resource-Planning-and-Management/APS_IRP_2023_PUBLIC.pdf?la=en&hash=F601897086C6836F7FD33C5C2F295F47.

Brightmore, D. 2024. "Hyperscalers Lengthen Server Lifespans to Save Billions." 2024. *Interface Media*. March 6. www.interface.media/blog/2024/03/06/hyperscalers-lengthen-server-lifespans-to-save-billions/.

Chakrabarti, K. 2025. "The Bitter Lesson Is Misunderstood." Substack. August 13. https://obviouslywrong.substack.com/p/the-bitter-lesson-is-misunderstood

Chang, W., and R. de Acosta. 2024. "What Is High Bandwidth Memory and Why Is the US Trying to Block China's Access to It?" *CNN Business*. December 8. www.cnn.com/2024/12/08/tech/us-china-hbm-chips-hnk-intl.

Cheng, E. 2025. "China's Latest AI Model Claims to Be Even Cheaper to Use than DeepSeek." *CNBC*. July 28. www.cnbc.com/2025/07/28/chinas-latest-ai-model-claims-to-be-even-cheaper-to-use-than-deepseek.html.

Clean Energy Transition Institute. 2025. "Data Centers, Artificial Intelligence and Energy Use 101." Seattle, WA: Clean Energy Transition Institute. www.cleanenergytransition.org/post/data-centers-artificial-intelligence-and-energy-use-101.

Craddock, M. 2025. "The AI Efficiency Paradox: How Generative AI's Success Could Drive Unsustainable Resource Consumption." *Medium*. January 28. www.medium.com/@mcraddock/the-ai-efficiency-paradox-how-generative-ais-success-could-drive-unsustainable-resource-55b3508448a4.

Craske, B. 2025. "How Are Data Centres Shifting to Zero-Water Cooling Tech?" *Data Centre*. July 22. www.datacentremagazine.com/news/how-are-companies-pioneering-data-centre-zero-water-cooling.

Datacenters.com. 2023. "Green Software: How It Works and Benefits Data Centers." September 12. www.datacenters.com/news/green-software-how-it-works-and-benefits-data-centers.

Digital Reality. 2025. "How AI Can Help Sustainable Data Centres by Revolutionising Energy Efficiency." www.digitalrealty.co.uk/resources/articles/sustainable-data-centre-ai.

Dominion Energy. 2024. *Virginia Electric and Power Company's Report of Its 2024 Integrated Resource Plan*. Richmond, VA : Dominion Energy. www.dominionenergy.com/-/media/content/about/our-company/irp/pdfs/2024-irp-w_o-appendices.pdf.

Elsworth, C., K. Huang, D. Patterson, I. Schneider, R. Sedivy, S. Goodman, B. Townsend, P. Ranganathan, J. Dean, A. Vahdat, B. Gomes, and J. Manyika. 2025. "Measuring the Environmental Impact of Delivering AI at Google Scale. https://services.google.com/fh/files/misc/measuring_the_environmental_impact_of_delivering_ai_at_google_scale.pdf.

Emerald AI. 2025. "Powering the AI Revolution." www.emeraldai.co/.

Energize Capital. 2025. *Data Center Deep Dive: O&M Efficiency*. www.energizecap.com/news-insights/tackling-data-center-efficiency-software-in-operations-maintenance.

EPRI (Electric Power Research Institute). 2024a. *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption*. Palo Alto, CA: EPRI. www.epri.com/research/products/3002028905.

_____. 2024b. "EPRI Launches Initiative to Enhance Data Center Flexibility and Grid Reliability." Palo Alto, CA: EPRI.

Esram, N., and C. Assadi. 2025. *Future-Proof AI Data Centers, Grid Reliability, and Affordable Energy: Recommendations for States*. Washington, DC: American Council for an Energy-Efficient Economy. www.aceee.org/white-paper/2025/04/future-proof-ai-data-centers-grid-reliability-and-affordable-energy.

Evans, R., and J. Gao. 2016. "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%." www.deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/.

[likely to drop] FlexGen. Undated. "Accelerating Data Center Interconnection." FlexGen Power Systems. www.flexgen.com/solutions/data-center-solutions.

Giacobone, B. 2025. "Verrus Successfully Demos Its Flexible Data Center Technology." *Latitude Media*. May 15. www.latitudemedia.com/news/verrus-successfully-demos-its-flexible-data-center-technology/.

Global Data Center Hub. 2025. "The 4 Types of Data Centers (and Who Uses Them)." *Global Data Center Hub*. September 5. www.globaldatacenterhub.com/p/the-4-types-of-data-centers-and-who.

Google. 2025. "Growing the Internet While Reducing Energy Consumption." *Google Data Centers*. www.datacenters.google/efficiency/.

Howland, E. 2025. "PSE&G Large Load Pipeline Jumps 47%, to 9.4 GW, Much of It Speculative." *Utility Dive*. August 6. www.utilitydive.com/news/pseg-data-centers-pjm-earnings/756911/.

Hutchinson, E. "Embracing Direct-to-Chip Cooling for an Energy-Efficient AI Era." Intelligent Data Centres. May 3. www.intelligentdatacentres.com/2024/05/03/embracing-direct-to-chip-cooling-for-an-energy-efficient-ai-era/ .

IEA (International Energy Agency). 2025. *Energy and AI*. Paris, France: IEA. www.iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf.

Infiniti. Undated. "Infiniti Guide: How long Do Data Centres Last?" www.infiniti-it.co.uk/news/tech-news/how-long-do-data-centres-last-.

ISO (International Standards Organization). 2016. *ISO/IEC 30134-2:2016 Information Technology—Data Centres—Key Performance Indicators, Part 2: Power Usage Effectiveness (PUE)*. www.iso.org/standard/63451.html.

Israelson, B. 2025. "Are Data Centers the New Heat Source?" *District Energy*. Quarter 3. www.districtenergy-digital.org/districtenergy/library/item/q3_2025/4286169/.

Jegham, N., M. Abdelatti, L. Elmoubarki, and A. Hendawi. 2025. "How Hungary Is AI? Benchmarking Energy, Water and Caron Footprint of LLM Interface." ARXIV. Cornell University. https://arxiv.org/abs/2505.09598.

Koningstein, R. 2021. "We Now Do More Computing Where There's Cleaner Energy." *Google The Keyword*. May 18. www.blog.google/outreach-initiatives/sustainability/carbon-aware-computing-location/.

Koomey, J., T. Das, and Z. Schmidt. 2025. *Electricity Demand Growth and Data Centers: A Guide for the Perplexed*." Washington, DC: Bipartisan Policy Center. www.bipartisanpolicy.org/report/electricity-demand-growth-and-data-centers/.

Krugman, P. 2025. "What Happens If AI Hits an Energy Wall?" *Substack*. August 18. www.paulkrugman.substack.com/p/what-happens-if-ai-hits-an-energy.

.

Lehr, S. 2025. "Companies Are Pouring Billions into A.I. It Has Yet to Pay Off." *New York Times*. August 13. www.nytimes.com/2025/08/13/business/ai-business-payoff-lags.html.

Leiserson, C., N. Thompson, J. Emer, B. Kuszmaul, B. Lampson, D. Sanchez, and T. Schardl. 2020. "There's Plenty of Room at the Top: What Will Drive Computer Performance after Moore's Law?" *Science*. June 5. www.science.org/doi/full/10.1126/science.aam9744.

Leviathan, Y., M. Kalman, and Y. Matias. 2024. "Looking Back at Speculative Decoding." Google Research. December 6. https://research.cs.wisc.edu/multifacet/papers/taco2017_pareto_governors.pdf www.research.google/blog/looking-back-at-speculative-decoding/.

Lew, L., and Y. Zhang. 2023. "Introducing Accurate Quantized Training (AQT) for Accelerated ML Training on TPU v5e." *Google Cloud Blog*. November 8. www.cloud.google.com/blog/products/compute/accurate-quantized-training-aqt-for-tpu-v5e.

Lovins, A. 2025. "Artificial Intelligence Meets Natural Stupidity: Managing the Risks." Stanford University. www.integrative-design-for-radical-energy-efficiency.stanford.edu/sites/extreme_energy_efficiency/files/media/file/data-centersaiel-dr-16-10-may-2025.pdf.

Martucci, B. 2025a. "A Fraction of Proposed Data Centers Will Get Built. Utilities Are Wising Up." *Utility Dive*. May 15. www.utilitydive.com/news/a-fraction-of-proposed-data-centers-will-get-built-utilities-are-wising-up/748214/.

_____. 2025b. "Texas Law Gives Grid Operator Power to Disconnect Data Centers during Crisis." Utility Dive. June 25. www.utilitydive.com/news/texas-law-gives-grid-operator-power-to-disconnect-data-centers-during-crisi/751587/.

Masanet, E., A. Shehabi, N. Lei, S. Smith, and J. Koomey. 2020. "Recalibrating Global Data Center Energy-Use Estimates." *Science* (367): 6481. February 28. www.science.org/doi/10.1126/science.aba3758.

Masanet, E., and N. Lei. 2020. "How Much Energy Do Data Centers Really Use?" Aspen Global Change Institute. www.agci.org/research-reviews/how-much-energy-do-data-centers-really-use.

Masanet, E., N. Lei, and J. Koomey. 2024. "How Will the Electricity Use of AI Data Centers Evolve? To Answer This Question, Energy Analysts Need Better Data." *Research Gate*. June. www.researchgate.net/publication/381152478_How_will_the_electricity_use_of_AI_data_centers_evolve_To_answer_this_question_energy_analysts_need_better_data.

McKinsey & Company. 2025. *The Data Center Balance: How US States Can Navigate the Opportunities and Challenges*. August 8. www.mckinsey.com/industries/public-sector/our-insights/the-data-center-balance-how-us-states-can-navigate-the-opportunities-and-challenges.

_____. 2024. "AI Power: Expanding Data Center Capacity to Meet Growing Demand." October 29. www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand.

Monitoring Analytics. 2025. *Analysis of the 2025/2026 RPM Base Residual Auction, Part G*. www.monitoringanalytics.com/reports/reports/2025/IMM_Analysis_of_the_20252026_RPM_Base_Residual_Auction_Part_G_20250603_Revised.pdf.

Mulkey, S.K. 2024. "The Surging Demand for Data Is Guzzling Virginia's Water." *Grist*. May 8. www.grist.org/technology/surging-demand-data-guzzling-water-ai/#:~:text=But%20the%20gargantuan%20facilities%20do%20more%20than,ubiquitous%2C%20they%27re%20using%20more%20water%20than%20ever.

Norris, T. 2025. "The Puzzle of Low Data Center Utilization Rates." *Power and Policy*. August 7. www.powerpolicy.net/p/the-puzzle-of-low-data-center-utilization.

Nun, Y. 2023. "Ice-Based Cooling in the Data Center: Less Carbon, More Profit." Blog. May 27. www.nostromo.energy/blogs/ice-based-cooling-in-the-data-center-less-carbon-more-profit.

Newport, C. 2025. "What If A.I. Doesn't Get Much Better Than This?" *The New Yorker*. August 12. www.newyorker.com/culture/open-questions/what-if-ai-doesnt-get-much-better-than-this.

Norris, T. 2025. "The Puzzle of Low Data Center Utilization Rates." *Power and Policy*. August 7. www.powerpolicy.net/p/the-puzzle-of-low-data-center-utilization.

Patterson, D., J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. l. Rothchild, D. So, M. Texier, and J. Dean. 2022. "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." Computer 55 (7): 18–28. https://doi.org/10.1109/MC.2022.3148714 .

Patterson, K., S. Poole, C.-H. Hse, D. Maxwell, W. Tschudi, H. Coles, D. Martinez, and N. Bates. 2013. "TUE, A New Energy-Efficiency Metric Applies to ORNLs Jaguar." In: Kunkel, J. M., T. Ludwig, and H. W. Meuer (eds.), *Supercomputing. ISC 2013. Lecture Notes in Computer Science*, vol. 7905. Berlin, Heidelberg: Springer, Berlin. https://doi.org/10.1007/978-3-642-38750-0_28. Also www.datacenters.lbl.gov/sites/default/files/isc13_tuepaper.pdf.

Perez, K., C. Porter, and S. Narasimhan. 2023. "Strategies for Maximizing Data Center Energy Efficiency." *Nvidia Developer*. May 23. www.developer.nvidia.com/blog/strategies-for-maximizing-data-center-energy-efficiency/.

Rahman, R. 2024. "Leading ML Hardware Becomes 40% More Energy-Efficient Each Year." Epoch AI. www.epoch.ai/data-insights/ml-hardware-energy-efficiency.

Resource Innovation Institute. 2025. "Virginia Explores Innovative Path to Economic Growth through Colocation of Data Centers and Food Production Greenhouses." Press Release. September 5. www.resourceinnovation.org/press-release/virginia-explores-innovative-path-to-economic-growth-through-colocation-of-data-centers-and-food-production-greenhouses/.

Satchwell, A., N. Frick, P. Cappers, S. Sergici, R. Hledik, G. Kavlak, and G. Oskar. 2025. *Electricity Rate Designs for Large Loads: Evolving Practices and Opportunities.* Berkeley, CA: Berkeley Lab. https://emp.lbl.gov/publications/electricity-rate-designs-large-loads.

Schneider, I., H. Xu, S. Benecke, D. Patterson, K. Huang, P. Ranganathan, and C. Elsworth. 2025. "Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends." https://arxiv.org/abs/2502.01671 .SEPA (Smart Electric Power Alliance). 2025. "Database of Emerging Large-Load Tariffs (DELTa)." www.sepapower.org/large-load-tariffs-database/.

Shehabi, A., S. J. Smith, A. Hubbard, A. Newkirk, N. Lei, M. A. B. Siddik, B. Holecek, J. Koomey, E. Masanet, and D. Sartor. 2024. *2024 United States Data Center Energy Usage Report*. LBNL-2001637. Berkeley, CA; Berkeley Lab. www.escholarship.org/uc/item/32d6m0d1\.

Siemens. 2025. "DC Applications—A Wide Range of Possible Uses for Direct Current Technologies." March 21. www.support.industry.siemens.com/cs/document/109977881/dc-applications-%E2%80%93-a-wide-range-of-possible-uses-for-direct-current-technologies?dti=0&lc=en-CN.

Skidmore, D. 2025. "Ohio Regulators Approve Settlement Requiring Data Centers to Pay at Least 85% of Power Costs." *Data Center Dynamics*. July 11. www.datacenterdynamics.com/en/news/ohio-regulators-approve-settlement-requiring-data-centers-to-pay-at-least-85-of-power-costs/.

Solstice Power Technologies. 2024. "Solstice and Google Partner to Advance Inclusive Clean Energy Access." PR Newswire. July 23. www.prnewswire.com/news-releases/solstice-and-google-partner-to-advance-inclusive-clean-energy-access-302203203.html#:~:text=This%20collaboration%20between%20Solstice%20and%20Google.org%20represents,stewardship%20and%20equitable%20access%20to%20clean%20energy.

St. John, J. 2025a. "Utilities Are Flying Blind on Data Center Demand. That's a Big Problem." *Canary Media*. February 25. www.canarymedia.com/articles/utilities/utilities-are-flying-blind-on-data-center-demand-thats-a-big-problem.

_____. 2025b. "Google's New Plan to Keep Its Data Centers from Stressing the Grid." *Canary Media*. August 28. www.canarymedia.com/articles/utilities/google-ai-data-center-flexibility-help-grid.

Terrell, M. 2025. "How We're Making Power Centers More Flexible to Benefit Power Grids." *Google The Keyword*. August 4. www.blog.google/inside-google/infrastructure/how-were-making-data-centers-more-flexible-to-benefit-power-grids/.

Tilton, J. 2025. "Big Tech Tests Data Center Flexibility." *IEEE Spectrum.* June 12. www.spectrum.ieee.org/dcflex-data-center-flexibility.

Trane. 2024. "Go with the Flow: Is It Time Yet for Liquid Cooling? www.trane.com/commercial/north-america/us/en/about-us/newsroom/blogs/go-with-the-flow-is-it-time-for-liquid-cooling.html.

Trueman, C. 2024. "Nvidia: PUE in an Ineffective Efficiency Metric for AI Workloads and Needs Replacing." Data Center Dynamics. May 14. www.datacenterdynamics.com/en/news/nvidia-pue-is-an-ineffective-efficiency-metric-for-ai-workloads-and-needs-replacing/.

Uptime Institute. 2025. *Uptime Institute Global Data Center Survey*. https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2025.

Vertiv. 2025. "Liquid Cooling Options for Data Centers." www.vertiv.com/en-us/solutions/learn-about/liquid-cooling-options-for-data-centers/.

Vincent, M. 2025. "8 Trends That Will Shape the Data Center Industry in 2025." Data Center Frontier. www.datacenterfrontier.com/cloud/article/55253151/8-trends-that-will-shape-the-data-center-industry-in-2025.

Wade, C., M. Blackhurst, J. DeCarolis, A. de Queiroz, J. Johnson, and P. Jaramillo. 2025. "Electricity Grid Impacts of Rising Demand from Data Centers and Cryptocurrency Mining Operations." Carnegie-Mellon University. www.energy.cmu.edu/_files/documents/electricity-grid-impacts-of-rising-demand-from-data-centers-and-cryptocurrency-mining-operations.pdf.

Walton, R. 2025. "There Aren't Enough AI Chips to Support Data Center Projections, Report Says." *Utility Dive*. July 9. www.utilitydive.com/news/not-enough-ai-chips-to-support-data-center-projections-london-economics/752371/.

Wang, N. and C. Assadi. 2025. *Future-Proof AI Data Centers, Grid Reliability and Affordable Energy: Recommendations for States*. Washington, DC: ACEEE: https://www.aceee.org/white-paper/2025/04/future-proof-ai-data-centers-grid-reliability-and-affordable-energy.

Wang, P., S. Kowalski, Z. Gao, J. Sun, C.-M. Yang, D. Grant, P. Boudreaux, S. Huff, and K. Nawaz. 2024. "District Heating Utilizing Waste Heat of a Data Center: High-Temperature Heat Pumps." *Energy and Buildings* 315. www.sciencedirect.com/science/article/abs/pii/S0378778824004432#:~:text=Information%20technology%20equipment%20consumes%20about,and%20insufficient%20infrastructure%20%5B6%5D.

Wyant, C., M. Verma, and W. Kanj. 2025. "Homegrown Energy: How Household Upgrades Can Meet 100 Percent of Data Center Demand Growth." Rewiring America. www.rewiringamerica.org/research/homegrown-energy-report-ai-data-center-demand.

Zeitlin, M. 2025. "The Country's Biggest Grid Has a Plan to Manage Data Centers' Power Use. Everyone Hates It." *Heatmap Daily*. September 4.

Zhang, M. 2024. "PUE (Power Usage Effectiveness): Optimizing Data Centers." *Dgtl Infra*. March 1. www.dgtlinfra.com/pue-power-usage-effectiveness/.