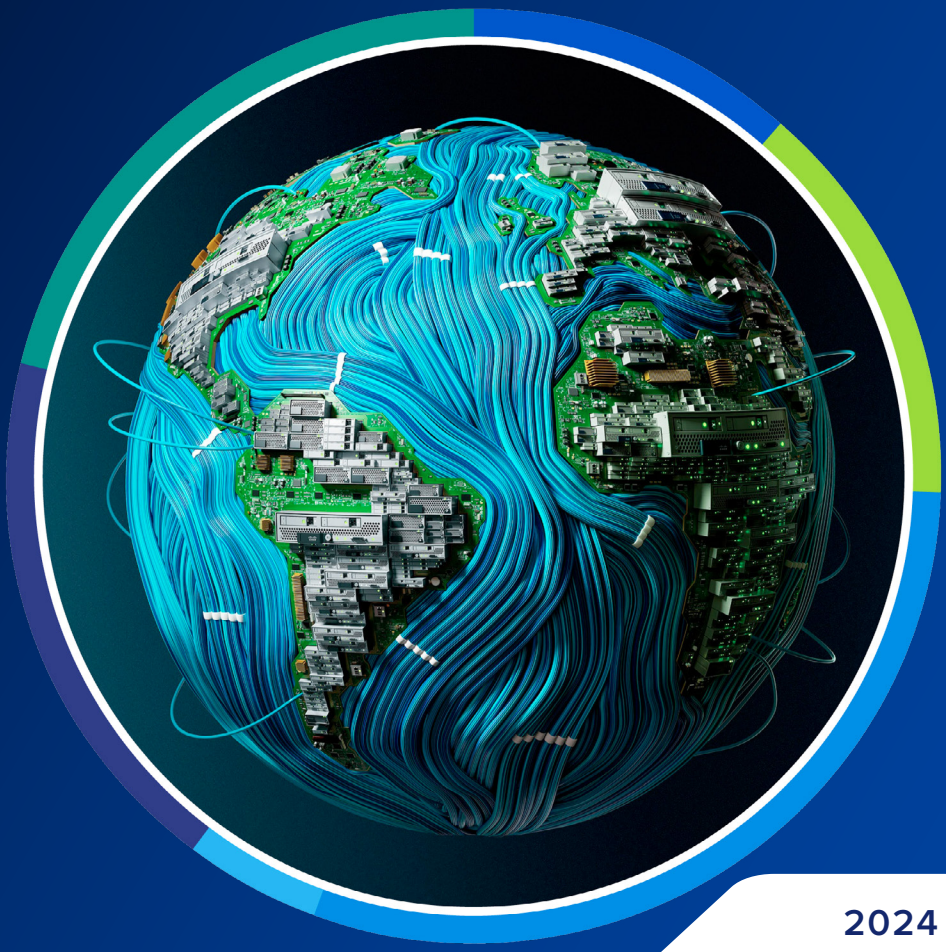
EPRI



2024 White Paper

Powering Intelligence

Analyzing Artificial Intelligence and Data Center Energy Consumption

EXECUTIVE SUMMARY

Key Messages

- In the United States, powering data centers, providing clean energy for manufacturing, supporting industrial onshoring, and electrifying transportation are driving renewed electric load growth. Clusters of new, large point loads are testing the ability of electric companies to keep pace.
- Data centers are one of the fastest growing industries worldwide. Between 2017 and 2021, electricity used by Meta, Amazon, Microsoft, and Google—the main providers of commercially available cloud computing and digital services—more than doubled.
- A fundamental uncertainty in projecting data center load growth comes from the broad emergence of artificial intelligence (AI) technologies in business and daily life—punctuated by the explosion into public consciousness of generative AI models, such as OpenAI's ChatGPT, released in November 2022. While AI applications are estimated to use only 10%–20% of data center electricity today, that percentage is growing rapidly.
- AI models are typically much more energy-intensive than the data retrieval, streaming, and communications applications that drove data center growth over the past two decades. At 2.9 watt-hours per ChatGPT request, AI queries are estimated to require 10x the electricity of traditional Google queries, which use about 0.3 watt-hours each; and emerging, computation-intensive capabilities such as image, audio, and video generation have no precedent.
- To provide an early assessment of potential data center load growth at the national level, EPRI has developed low, moderate, high, and higher growth scenarios for data center loads from 2023 to 2030. Data centers grow to consume 4.6% to 9.1% of U.S. electricity generation annually by 2030 versus an estimated 4% today.
- While the national-level growth estimates are significant, it is even more striking to consider the geographic concentration of the industry and the local challenges this growth can create. Today, fifteen states account for 80% of the national data center load, with data centers estimated to comprise a quarter of Virginia's electric load in 2023. Concentration of demand is also evident globally, with data centers projected to make up almost one-third of Ireland's total electricity demand by 2026.
- With the shift to cloud computing and AI, new data centers are growing in size. It is not unusual to see new centers being built with capacities from 100 to 1000 megawatts—roughly equivalent to the load from 80,000 to 800,000 homes. Connection lead times of one to two years, demands for highly reliable power, and requests for power from new, non-emitting generation sources can create local and regional electric supply challenges.
- EPRI highlights three essential strategies to support rapid data center expansion:
    1. Data center ***efficiency improvements and increased flexibility***.
    2. ***Close coordination between data center developers and electric companies*** regarding data center power needs, timing, and flexibility, as well as electric supplies and delivery constraints.
    3. ***Better modeling tools*** to plan the 5–10+ year grid investments needed to anticipate and accommodate data center growth without negatively impacting other customers and to identify strategies for maintaining grid reliability with these large, novel demands.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 2

Key Messages 2

Potential Impacts of Artificial Intelligence on Data Center Load Growth 4

EPRI U.S. Data Center Load Projections 4

Data Center Power Demands Are Concentrated in a Few Regions 5

A Roadmap to Support Rapid Data Center Expansion 6

Introduction 7

Research Questions 7

Data Centers in the United States 7

Data Centers' Primary Electricity-Consuming Hardware and Equipment 9

AI and Data Center Power Consumption Insights 10

Immense Volumes of Data are Being Processed Daily 10

History of Energy Efficiency in the Data Center Industry 11

Uneven Geographic Distribution Creates Imbalance in Data Center Load 12

AI Implications for Power Consumption 14

Chat GPT and Other Large Language Models (LLMs) 15

Forecasting Data Center Load Growth to 2030 17

Four Scenarios Based on Historical Data, Expert Insights, and Current Trends 17

Energy Efficiency, Load Management and Clean Electricity Supply 18

Energy-Efficient Training Algorithms 18

Energy-Efficient Hardware 19

Energy-Efficient Cooling Technologies 19

Scalable Clean Energy Use 20

Monitoring and Analytics 20

Reducing Data Centers' Environmental Footprint 21

Actions to Support Rapid Data Center Expansion 21

Improve Data Center Operational Efficiency and Flexibility 22

Increase Collaboration through a Shared Energy Economy Model for Sustainable Data Centers 22

Better Anticipate Future Point Load Growth through Improved Forecasting and Modeling 23

Appendix A: State-Specific Scenarios 24

Projected Data Center Load Scenarios for Top 15 States 24

Regional Differences in Data Center Capacities by Metropolitan Area 27

Projections of Potential Power Consumption for 44 States 28

Appendix B: Insights Into the Energy Use of AI Models 29

References 31

Potential Impacts of Artificial Intelligence on Data Center Load Growth

Data center operation is one of the fastest growing industries worldwide. The International Energy Agency recently projected that global data center electricity demand will more than double by 2026. In the United States, the national outlook could resemble the global outlook, but is highly uncertain.

One key uncertainty that could change the trajectory of data center load growth is the use of generative AI models. Both public and corporate imaginations were triggered by the release of OpenAI's ChatGPT on November 30, 2022. Evidence about how widely these tools will be used and how much they will change computational needs is just starting to emerge. These early applications were estimated to require about ten times the electricity—from 0.3 watt-hours for a traditional Google search to 2.9 watt-hours for a ChatGPT query—to respond to user queries. Creation of original music, photos, and videos based upon user prompts and other emerging AI applications could require much more power. With 5.3 billion global internet users, widespread adoption of these tools could potentially lead to a step change in power requirements. On the other hand, history has shown that demand for increased processing has largely been offset by data center efficiency gains.

EPRI U.S. Data Center Load Projections

Drawing on public information about existing data centers, public estimates of industry growth, and recent electricity demand forecasts by industry experts, EPRI prepared four scenarios of potential electricity consumption in U.S. data centers during the period from 2023 to 2030 (**Figure ES-1**). The blue line in the figure, running from 2000 to 2020, traces historical data center electricity consumption estimates. From 2000 to 2010, data center load grew as centers expanded across the country to support the emerging internet. From 2010 to 2017, despite continued growth in computing demands and data storage this load growth flattened due to efficiency gains and the replacement of small, relatively inefficient corporate data centers with large, cloud computing facilities. In recent years, load growth has likely accelerated, driven by emerging AI applications and COVID-era increases in demand for services like streaming and video conferencing. The light blue area highlights uncertainty in a range of data center electricity consumption estimates for 2021 to 2023. Colored bands show the four projections, which combine estimates of increased data processing needs with assumptions about efficiency gains. The widths of these bands carry forward the uncertainty about the 2023 starting load level:

- ***Low growth****—3.7% annual load growth based on a Statista projection of data center financial growth issued prior to the release of ChatGPT.*

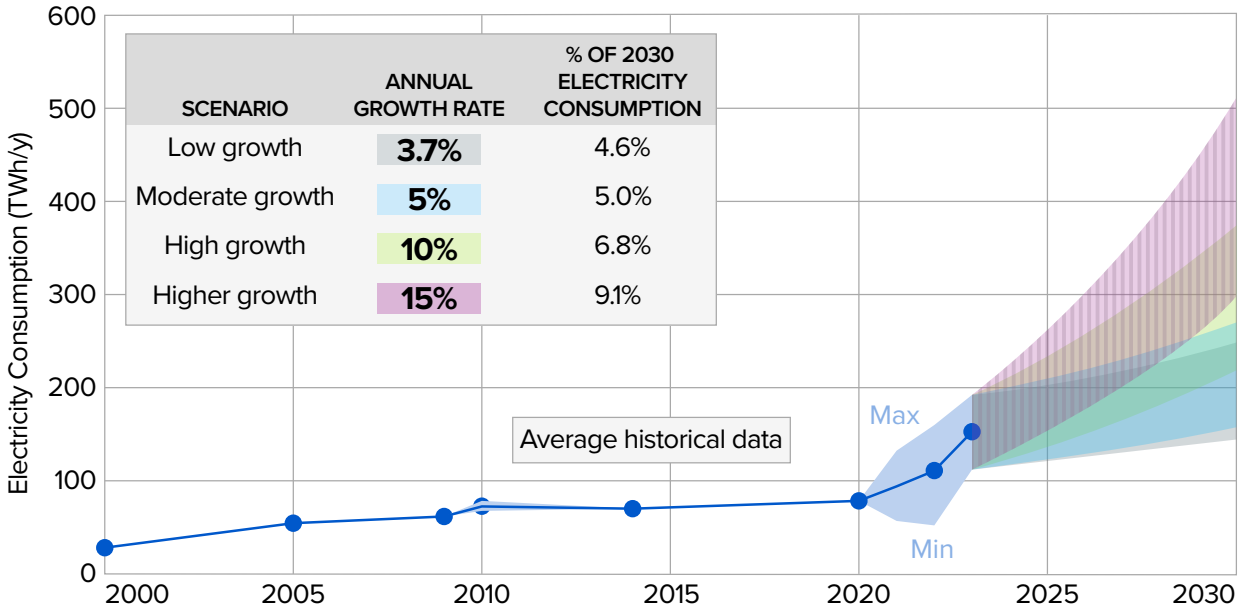


Figure ES-1. *Projections of potential electricity consumption by U.S. data centers: 2023–2030 . % of 2030 electricity consumption projections assume that all other (non-data center) load increases at 1% annually.*

- ***Moderate growth**—5% annual load growth based on an expert assessment commissioned by EPRI.*
- ***High growth**—10% annual load growth consistent with both a McKinsey estimate and another expert assessment commissioned by EPRI in summer 2023.*
- ***Higher growth**—15% annual growth based upon a commissioned expert assessment consistent with rapid expansion of AI applications and limited efficiency gains.*

The estimates of data centers' share of total U.S. electricity consumption in 2030—9.1%, 6.8%, 5.0%, and 4.6%—assume that all other loads increase at 1% per year. Data centers accounted for about 4% of the total load in 2023 (average estimate).

Data Center Power Demands Are Concentrated in a Few Regions

Fifteen states accounted for an estimated 80% of the national data center load in 2023. Ranked from highest to lowest, they are Virginia, Texas, California, Illinois, Oregon, Arizona, Iowa, Georgia, Washington, Pennsylvania, New York, New Jersey, Nebraska, North Dakota, and Nevada. Concentration of demand is also evident globally, with the International Energy Agency recently projecting that data centers in Ireland could account for almost one-third of Ireland's total electricity demand by 2026.

The map in **Figure ES-2** shows the effect in 2030 of applying the annual U.S. data center growth rates (averaged across the four scenarios) to project state-level loads against a backdrop of 1% annual growth in other loads. With evenly spread growth, the data center share of load in Virginia increases to almost 50% in the higher growth scenario and to 36% when averaged across the four scenarios. The shares in other states vary widely with five other states projected to approach 20% or more of electricity demand under these simplified assumptions. In reality, load growth is unlikely to be spread evenly. Data centers favor sites where internet connections are strong; where electricity prices, land costs, and disruptive events are low; where skilled labor is available; near population centers and users; and where the centers can develop backup power to ensure power supply (usually natural gas or diesel generators). The additional

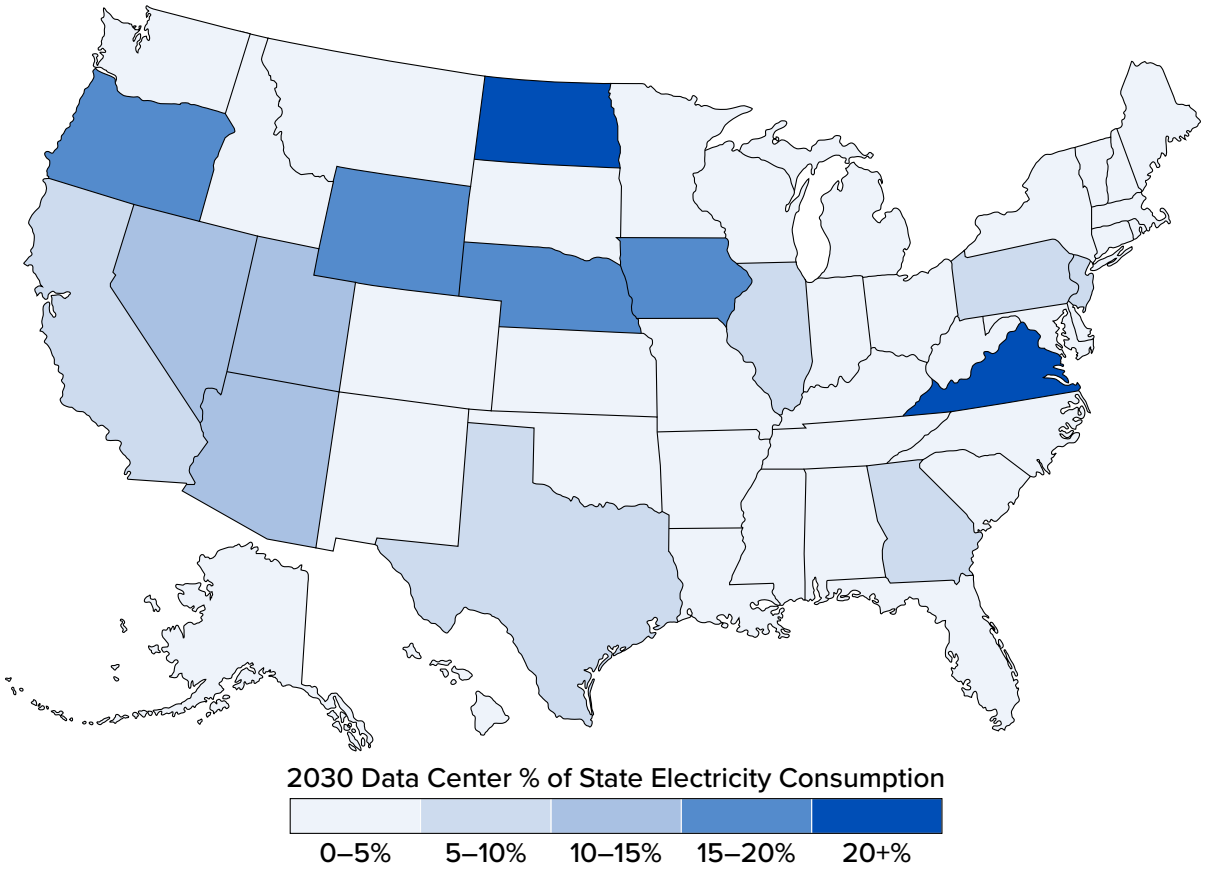


***Figure ES-2.** 2030 projected data center share of electricity consumption (assumes average of the four growth scenarios and that non-data center loads grow at 1% annually) [4, 8, 9]*

requirement of some developers for new, clean electricity generation sources adds to the challenge of developing and delivering this new generation.

A Roadmap to Support Rapid Data Center Expansion

The most serious challenges to data center expansion are local and regional and result from the scale of the centers themselves and mismatches in infrastructure timing. A typical new data center of 100 to 1000 megawatts represents a load equal to that of a new neighborhood of 80,000 to 800,000 average homes. While neighborhoods require many years to plan and build, data centers can be developed and connected to the internet in one to two years. New transmission, in contrast, takes four or more years to plan, permit, and construct. And developing and connecting new generation can also take years.

EPRI highlights three essential strategies to support rapid data center expansion. These strategies emphasize increased collaboration between data center developers and electric companies.

1. ***Improve data center operational efficiency and flexibility***. Although gains in data center operational efficiency have plateaued in recent years, there are clear opportunities for further improvement, including more efficient IT hardware; lower electricity use for cooling, lighting, and security; and more efficient AI development and deployment strategies. Efforts to increase both temporal and spatial (i.e., spreading compute geographically) flexibility are critical to helping accommodate these new loads.
2. ***Increase collaboration between data center developers and electric companies***. Developing a deeper understanding of data center power needs, timing, and potential flexibilities—while assessing how they match available electric supplies and delivery constraints—can create workable solutions for all. Enabled by technology and supporting policies, data center backup generators, powered by clean fuels, could support a more reliable grid while reducing the cost of data center operation. Shifting the data center-grid relationship from the current "passive load" model to a collaborative "shared energy economy"—with grid resources powering data centers and data center backup resources contributing to grid reliability and flexibility—could not only help electric companies contend with the explosive growth of AI but also contribute to affordability and reliability for all electricity users.
3. ***Improve point load forecasting to better anticipate future point load growth and modeling of transient system behavior to maintain reliability***. Forecasts need to make better projections describing new point load locations, magnitudes, and timing alongside better techniques for making decisions—to build or not build long lead-time infrastructure—while facing the economic, regulatory, and political uncertainty associated with siting these large point loads. Also, real-time modeling of data center operational characteristics in an increasingly inverter-based grid is needed to maintain reliability.

INTRODUCTION

Research Questions

As the number and size of data centers expand to support continued growth in data processing, internet traffic, and rapid expansion in artificial intelligence (AI) applications, some critical questions emerge:

- How rapidly can we expect data centers to expand, and how does the rapid growth in AI change their power demands?
- What is the impact of these developments on electric load and resource adequacy?
- What implications do these trends have for future electricity infrastructure planning?
- How can the data center and electric utility industries work together to support rapid data center expansion?

Data Centers in the United States

As of March 2024, there were approximately 10,655 data centers globally; half of them, 5,381, were in the United States. Just over three years ago, in January 2021, there were approximately 8,000 data centers, with about one-third of them in the United States [1].

The construction of new data centers is accelerating at a rapid pace, largely driven by demand for AI-powered tasks such as speech recognition, tailored diagnostics, logistics, internet of things (IoT), and generative AI. The expansion of interest in generative AI is particularly notable due to the overnight popularity of ChatGPT, released on November 30, 2022, marking the public-facing start of a technology race.

Data centers vary significantly in design and purpose and are generally grouped into two categories, small or large scale. *Small-scale data centers*, representing about 10% of U.S. data center load [2], typically cater to localized operations and service small businesses, government facilities, or specific departmental needs within larger corporations. They include server rooms/closets embedded in buildings and "edge data centers," which are strategically located on the outer edges of networks to bring computing capabilities closer to users who are geographically distant from large cloud data centers [3]. Though the electricity demands of each installation are relatively modest—500 kilowatts (kW) to 2 megawatts (MW)—they account for roughly half of all servers [3]. Market research analysts have projected the global edge data center market to grow at a compound annual growth rate (CAGR) of 22.1% to 2030 [4], highlighting the rising importance of small-scale and edge data centers in digital infrastructure.

Large-scale commercial data centers are designed to serve extensive operations and often serve multiple businesses or even entire industries. These data centers seek proximity to customers and a skilled workforce and can benefit from lower land costs, property taxes, labor rates, energy prices, and risk of severe weather or seismic activity [5]. **Figures 1–3** show maps of various large-scale facility types, which include:

- Enterprise data centers, which are owned and operated by single companies for their exclusive computing and networking use. These account for about 20–30% of total load [2, 6].
- Co-location centers, where several businesses may rent space to house their servers and other hardware with shared energy and cooling infrastructure.
- Hyperscale data centers, which are capable of rapidly scaling up their operations to meet the vast computing needs of cloud giants like Amazon AWS, Google Cloud, and Microsoft Azure. Given their large scale and recent emergence, they are often at the forefront of electricity consumption and efficiency innovations. Hyperscale and co-location centers together account for the lion's share of U.S. data center load—about 60%–70% [7].

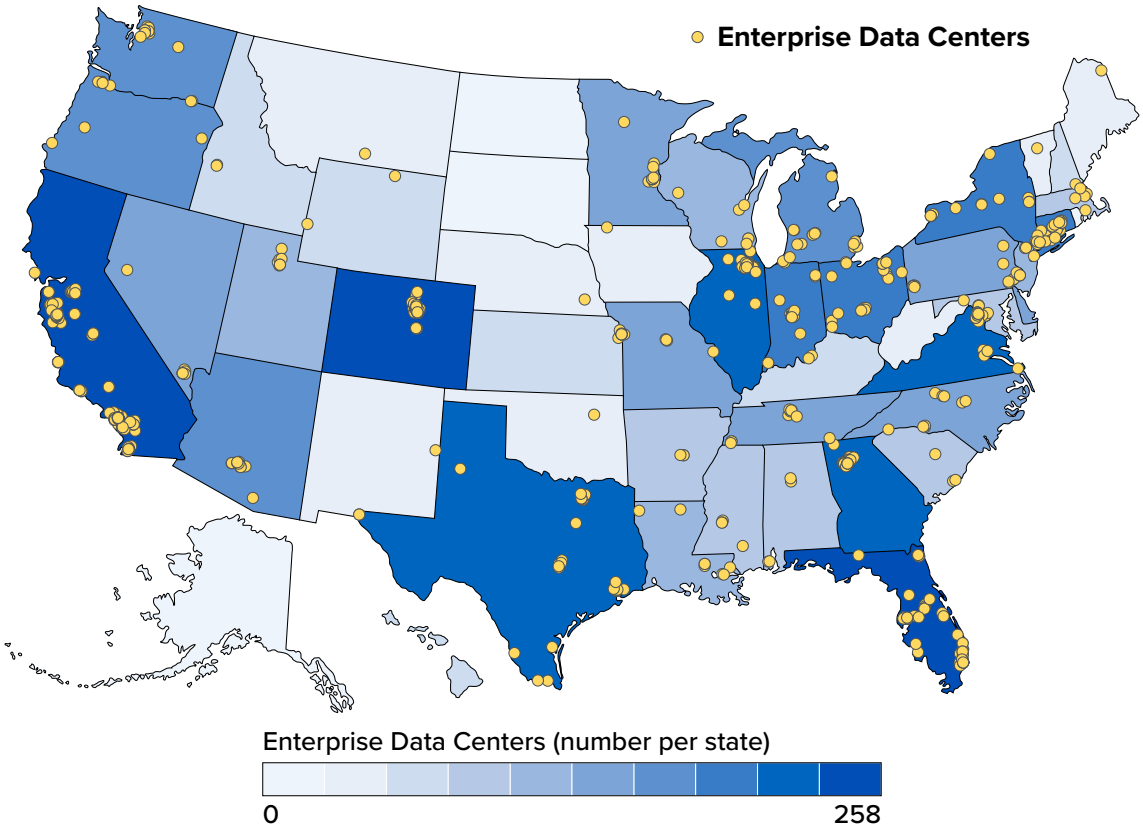


Figure 1. U.S. enterprise data center distribution as of 2022 [4, 8, 9]

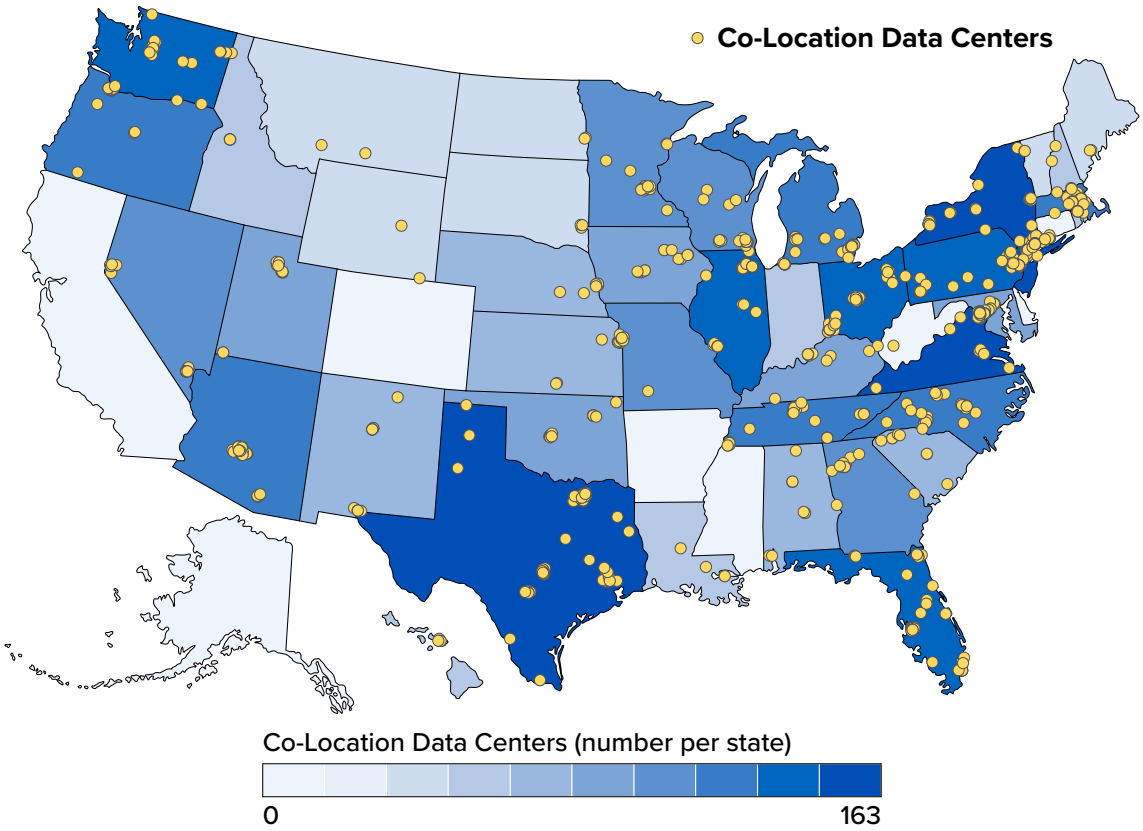


Figure 2. U.S. co-location data center distribution as of 2022 [4, 8, 9]

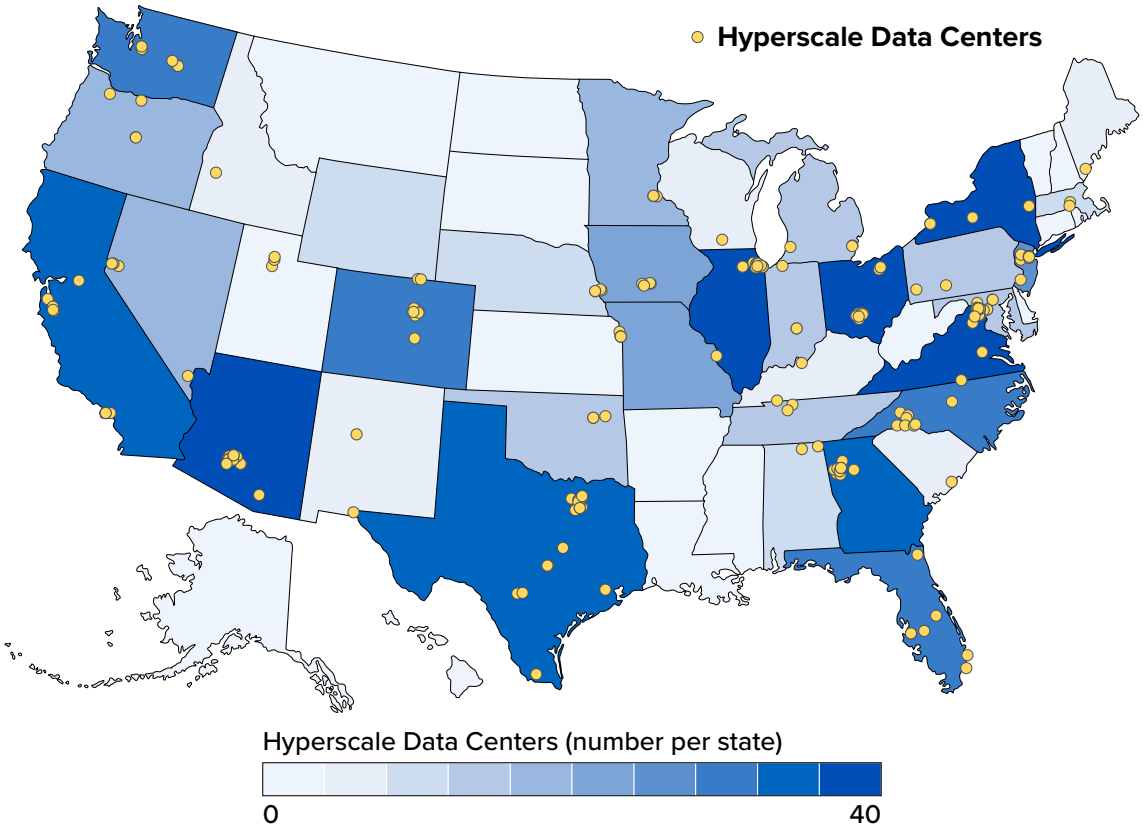


Figure 3. U.S. hyperscale data center distribution as of 2022 [4, 8, 9]

Data Centers' Primary Electricity-Consuming Hardware and Equipment

The electricity needs of data centers are determined primarily by the three constituent hardware categories. Each category's proportion of energy consumption can vary depending on the data center's age, configuration, type, and function [10, 11, 12, 13]. The three main categories and their energy consumption [2, 13, 14] are:

- IT equipment, typically composing 40%–50% of data center energy consumption, encompasses the following foundational hardware units:
  - Servers, which are the workhorses, responsible for data processing and computational tasks
  - Storage systems, which include both traditional hard disk drives (HDDs) and the faster, more energy-efficient solid-state drives (SSDs), crucial for data retention
  - Network infrastructure, which comprises switches, routers, and other components, ensuring seamless data transfer and connectivity
- Cooling systems, typically composing 30%–40% of data center energy consumption, are critical to maintaining an optimal temperature within data centers to prevent hardware malfunction and ensure longevity. While data centers historically used traditional HVAC, advanced cooling technologies in data centers have transitioned towards systems that are specialized for data center use. Please refer to the section **Energy Efficiency and Load Management** below for more details.
- Auxiliary components, typically composing 10%–30% of data center energy consumption, are used for various operational needs and include uninterruptible power supplies, security systems, and lighting.

Assessing data center energy efficiency is crucial to gauging how effectively they use electricity. These assessments help to identify trends, drive improvements, and set benchmarks for electricity usage; and play a key role in operational strategy [15, 16].

AI AND DATA CENTER POWER CONSUMPTION INSIGHTS

Immense Volumes of Data are Being Processed Daily

Data centers' worldwide electricity use in 2022 totaled 300 million megawatt-hours (MMWh), or 1.2% of all load, a 45% increase from 2015 [17]. In the United States in 2023, data centers accounted for about 4% of total electricity consumption or 150 MMWh, equivalent to the average annual consumption of 14 million households [9, 18].

Since 2017, *annual* data volumes have soared, tripling to around 4,750 exabytes (an exabyte being a billion gigabytes) by 2022, showcasing the immense volume of information being processed and transmitted globally every day [19]. In 2022, the *daily* generation of data—including captured, copied, or consumed—reached approximately 13 exabytes, a surge partly attributable to the burgeoning impact of AI models [17]. Concurrently, in 2022, global data transmission network energy use was reported to be around 260–360 MMWh, roughly equal to data center power use [17, 20]. **Figure 4** illustrates the dramatic rise in global consumer IP traffic.

Data centers are facing a significant challenge with internet traffic growing nearly 12-fold in the past decade, a trend paralleled by increasing AI-related workload demands [19]. The historical precedent is showcased in **Figure 5**, which contrasts the U.S. data storage supply versus estimated demand from 2009 to 2020, underscoring a growing deficit and the need to address these trends [22].

Despite the immense growth in network traffic and data generation, load growth has been much slower due to efficiency gains and consolidation.

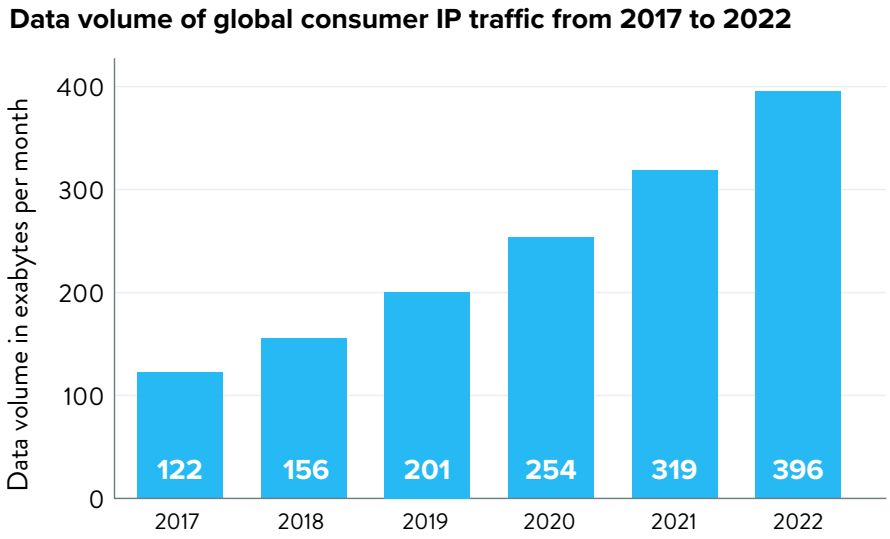


Figure 4. Trends in global consumer IP traffic, 2017–2022 [21]

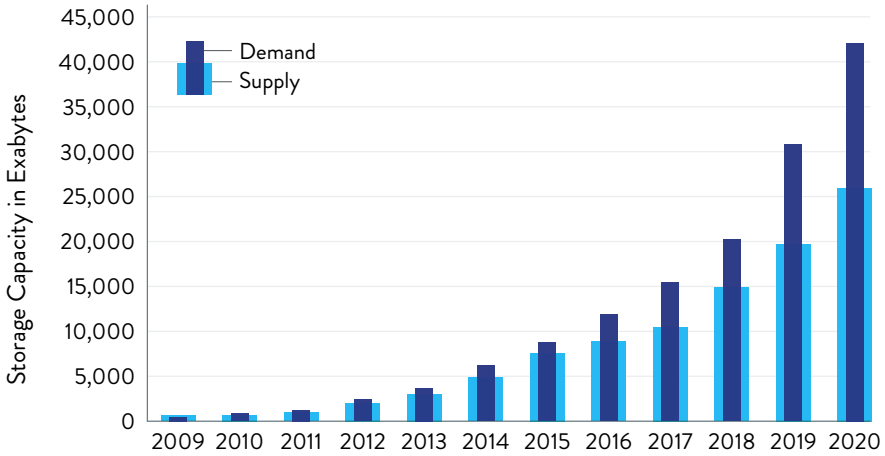


Figure 5. U.S. data storage supply vs. demand, 2009–2020 [22, 23]

History of Energy Efficiency in the Data Center Industry

Over the last 25 years, U.S. data center load growth, as shown in **Figure 6**, has experienced three phases:

1. Energy consumption grew in the early 2000s driven by the rapid expansion of internet infrastructure and the dot-com boom [24].
2. From 2010–2020, electricity consumption stabilized as data center expansion was offset by equally rapid improvements in energy efficiency achieved both through improvements at individual facilities and through the transition from small data centers to more efficient cloud facilities [25, 26].
3. Recent load growth in data centers is driven mainly by the expanding demand for cloud services, big data analytics, and AI technologies—which require significant computational resources—and a slowing of efficiency gains [27].

Efficiency gains in individual data centers have been led by advancements in server efficiency, which have been significant, leading to reduced power consumption per unit of computing power [28]. Power and cooling equipment, required to operate the IT components, has also improved its efficiency. Power Usage Effectiveness (PUE) and Data Center Infrastructure Efficiency (DCIE), key efficiency metrics in the data center industry, are defined in the box on the next page. **Figure 7** shows the U.S. PUE declined from 2007 to 2022, illustrating notable efficiency gains in

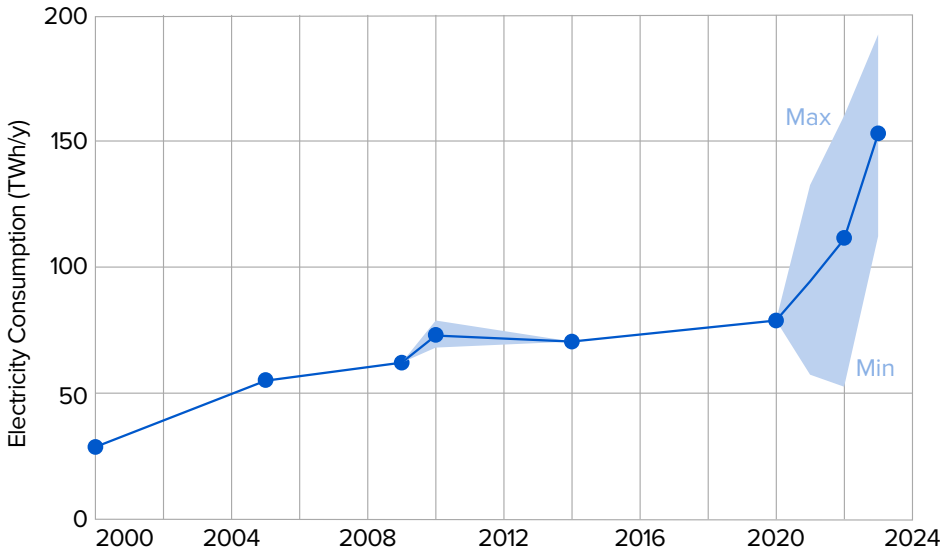


Figure 6. U.S. data center load growth from 2000 to 2023. The graph's light blue area indicates the uncertainty range based on two datasets estimating recent-to-current data center loads [4, 8, 9]

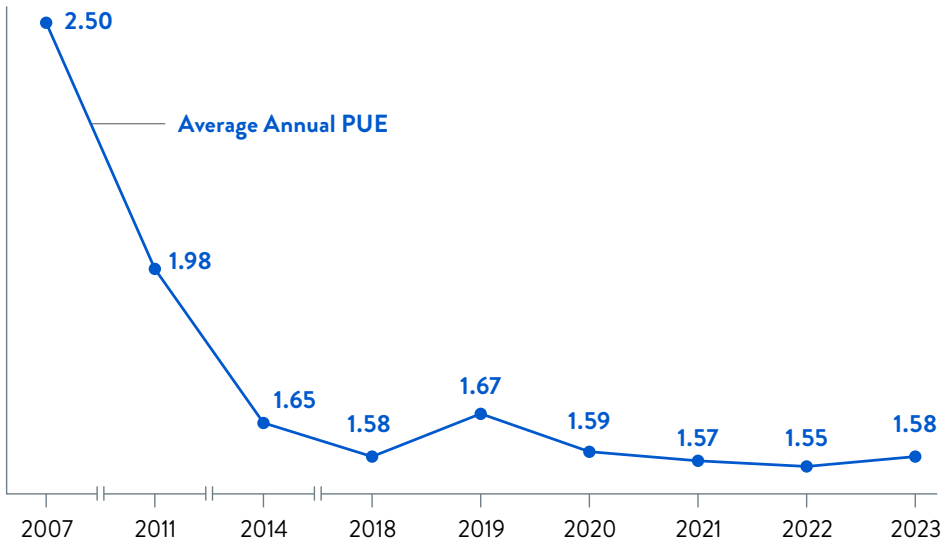


Figure 7. Average annual PUE in data centers, 2007–2023 [26, 29]

non-compute power demands [26, 29]. The recent stabilization at an average PUE of 1.6 suggests slowed progress in energy-saving strategies amidst the recent, rapid buildout [29]. With increased cooling needs of GPUs, advanced technologies (as discussed later in the report) could restart the downward drive in cooling and ancillary equipment power needs, with projected PUEs in advanced facilities potentially approaching 1.2.

DEFINITIONS OF KEY EFFICIENCY METRICS FOR DATA CENTERS

Power Usage Effectiveness (PUE)—A metric that quantifies a data center's energy efficiency by dividing the total energy usage by the energy consumed by the IT equipment alone. A lower PUE indicates higher efficiency, with 1.0 implying that all energy use is for computation.

Data Center Infrastructure Efficiency (DCIE)—A measure of a data center's energy efficiency calculated as the percentage of energy used directly by IT equipment out of the total energy consumption. Higher DCIE values signify greater efficiency in non-computational functions.

Uneven Geographic Distribution Creates Imbalance in Data Center Load

The geographic distribution of data centers is notably uneven, creating economic opportunity but also localized grid stress. For example, in 2022, data centers accounted for 1.2% of worldwide electricity, but 20% of electricity consumption in Ireland [30]. Similarly, the United States shows uneven growth in data center investments, reflecting a diverse landscape of regional opportunities and challenges. Data centers consume more power in Virginia, for instance, than in any other state. [9, 17]

Fifteen states accounted in 2023 for 80% of the national data center load. Ranked from highest to lowest in estimated load, each presents both opportunities and challenges as shown in **Table 1**.

Table 1. Opportunities and challenges for states ranked in the Top 15 for data center growth [29, 31, 32, 33, 34, 35, 36, 37]

STATE	OPPORTUNITIES	CHALLENGES
Virginia	Unparalleled network infrastructure; proximity to federal government agencies	Community pushback; regulatory scrutiny, particularly concerning environmental impact; transmission
Texas	Business-friendly; ample land availability	Electric grid reliability and pace of expansion
California	Robust technological ecosystem	High real estate and power costs; stringent environmental regulations
Illinois	Strategic location; significant tax incentives; nuclear generation and increasing renewable energy investments to address sustainability	Transmission constraints and rapidity of development
Oregon	Low electricity rates, low carbon emissions, moderate climate, tax incentives, and skilled workforce	Complex environmental regulations; demand for green energy solutions, and pace of growth
Arizona	Solar electricity, low risk of natural disasters; recent market growth	Water scarcity; need for sustainable cooling solutions
Iowa	Low electric rates; renewable energy availability	Geographic limitation (distant from major U.S. data hubs)
Georgia	Availability of land and power; friendly business environment	Balancing rapid expansion with local resource impacts
Washington	Abundant renewable energy resources	High energy costs; strict regulatory measures
Pennsylvania	Strategic location near major East Coast markets; relatively low energy costs	Aging infrastructure
New York	Hub for financial services; high connectivity demand	Space limitations; high energy costs
New Jersey	Close to major metropolitan areas; robust fiber-optic infrastructure ; transmission capacity from recent build out	High property and energy costs
Nebraska	Low energy costs; generous tax incentives	Remote location might limit connectivity options
North Dakota	Significant tax incentives; low cost of operations	Limited connectivity; need for more robust infrastructure
Nevada	Tax abatements; low electricity prices	Water scarcity; need for sustainable cooling solutions

Table 2 presents estimates of data center consumption in 2023, 2030, and the projected consumption as a percentage of state electricity demand (%EC) for the 15 states. For detailed graphs of each state's projections as well as a table showing 44 states that are pertinent to the U.S. data center market, see **Appendix A**.

Table 2. Current and projected load growth in Top 15 states [4, 8, 9]

FORECASTED SCENARIOS: PROJECTIONS OF DATA CENTER ELECTRICITY CONSUMPTION IN TOP 15 STATES (2023—2030)										
	2023 Load		Low-growth scenario (3.71%)		Moderate-growth scenario (5%)		High-growth scenario (10%)		Higher-growth scenario (15%)	
STATE	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)
Virginia	33,851,122	25.59%	43,683,508	29.28%	47,631,928	31.10%	65,966,260	38.47%	89,880,357	46.00%
Texas	21,813,159	4.59%	28,149,002	5.47%	30,693,306	5.94%	42,507,676	8.04%	57,917,564	10.64%
California	9,331,619	3.70%	12,042,078	4.43%	13,130,525	4.81%	18,184,686	6.54%	24,777,000	8.70%
Illinois	7,450,176	5.48%	9,614,151	6.53%	10,483,145	7.08%	14,518,285	9.54%	19,781,455	12.56%
Oregon	6,413,663	11.39%	8,276,574	13.39%	9,024,668	14.43%	12,498,415	18.93%	17,029,342	24.14%
Arizona	6,253,268	7.43%	8,069,590	8.81%	8,798,975	9.53%	12,185,850	12.73%	16,603,465	16.58%
Iowa	6,193,320	11.43%	7,992,230	13.44%	8,714,623	14.48%	12,069,029	18.99%	16,444,294	24.21%
Georgia	6,175,391	4.26%	7,969,093	5.08%	8,689,396	5.51%	12,034,090	7.48%	16,396,690	9.92%
Washington	5,171,612	5.69%	6,673,757	6.77%	7,276,977	7.34%	10,078,009	9.88%	13,731,490	13.00%
Pennsylvania	4,590,240	3.16%	5,923,520	3.78%	6,458,929	4.11%	8,945,079	5.61%	12,187,850	7.49%
New York	4,067,385	2.84%	5,248,796	3.40%	5,723,219	3.69%	7,926,182	5.05%	10,799,583	6.75%
New Jersey	4,038,360	5.42%	5,211,341	6.46%	5,682,378	7.00%	7,869,621	9.44%	10,722,517	12.44%
Nebraska	3,959,520	11.70%	5,109,601	13.75%	5,571,442	14.81%	7,715,984	19.41%	10,513,184	24.71%
North Dakota	3,915,720	15.42%	5,053,079	18.00%	5,509,811	19.31%	7,630,631	24.89%	10,396,888	31.11%
Nevada	3,416,707	8.69%	4,409,122	10.28%	4,807,649	11.10%	6,658,195	14.75%	9,071,924	19.07%

The four load growth projection scenarios reflect national-level estimates of data center growth applied to state-level estimates of current demand. This analytical approach is explained in more detail in the section **Forecasting Data Center Load Growth to 2030.*

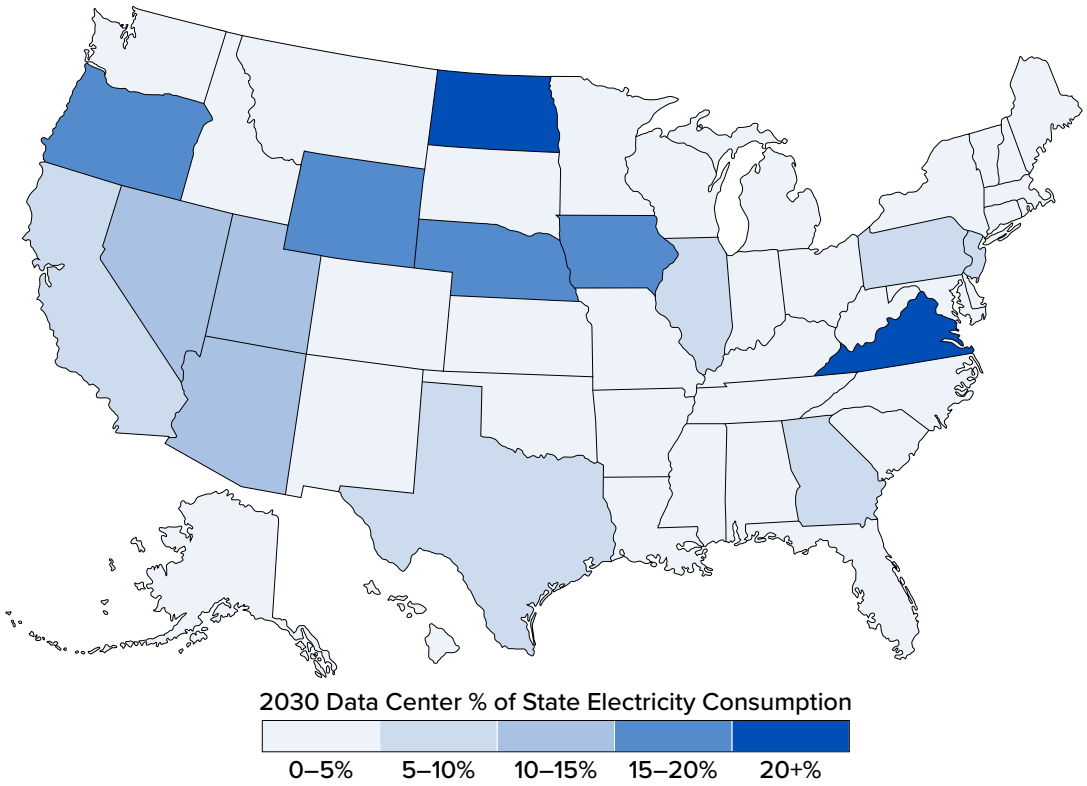


Figure 8. 2030 projected data center share of electricity consumption (assumes average of the four growth scenarios and that non-data center loads grow at 1% annually) [4, 8, 9]

The map in **Figure 8** depicts the projected data center share of state electricity demand in 2030, calculated by applying the annual U.S. data center growth rates (averaged across the four scenarios) to project state-level data center loads and assuming other loads grow at 1% annually. (The scenarios are explained in the section **Forecasting Data Center Load Growth to 2030**.) The potential for a rapidly rising share of data center power demand in many states accentuates the need for customized energy strategies that align with the specific demands and infrastructure capabilities of each state's grid. State-level projections also underscore the critical need for innovation in energy management and the optimization of localized infrastructure to accommodate the rising energy demands associated with expanding data center workloads.

AI Implications for Power Consumption

In the latter half of the 20th century, AI applications typically involved rule-based strategies and small machine-learning models that used very little electricity. However, as the 21st century unfolded, AI systems witnessed exponential growth in their complexity and computational requirements [38, 39]. On a global level, the United States has been leading in the development of prominent AI systems, with the creation of 16 such systems since 2022, compared to the United Kingdom's eight and China's three [39].

Key AI-related technological drivers contributing to escalating data center electricity demands include:[1]

- The exponential growth of data generation: The dramatic rise in global consumer IP traffic represents a reflection of the "big data" wave, part of which has resulted from feeding AI models with diverse and large datasets [9, 10, 40, 41]. The surge in data availability not only fuels the sophistication and accuracy of AI algorithms but also underscores the symbiotic relationship between increasing internet usage and AI advancement. Of course, this has required expanded storage, increased processing capabilities, and escalating electricity demands [21].
- The increasing complexity of AI models: Initially constituted as rule-based entities functioning through coded

1 While cryptocurrency mining, with its distinct computational processes and energy patterns for blockchain transaction verification and cryptocurrency generation, also impacts energy loads, it is excluded from this study to maintain focus on traditional data center operations and AI-driven computations. In 2022, global crypto mining was estimated to have consumed around 110 million MWh, accounting for 0.4% of annual global electricity demand, around one-third the usage of traditional data centers [17]. A separate assessment is warranted to understand the potential power needs and flexibility of cryptocurrency power demands.

instructions, AI models have undergone a monumental transformation, becoming increasingly complex and capable over time [42], in turn increasing their computational demands. As an illustration of the staggering increase in computation demand, note that in 1957, the Perceptron Mark I, the first real-world implementation of a one-layer neural network that could classify images, utilized 695,000 floating-point operations per second (FLOPS)—an assessment of AI complexity and computation intensity. In 2020, however, GPT-3 required a staggering $3.14x10^{23}$ FLOPS, an increase of 18 orders of magnitude, and at present each subsequent AI model is requiring even greater amounts.

- The continuous operational demands of a digital ecosystem: In the modern era, data centers function ceaselessly to uphold the demands of a globalized society that thrives on connectivity. Data centers facilitate uninterrupted services, ensuring 24/7 availability in various sectors including business, e-commerce, and entertainment. Maintaining constant uptime requires robust backup power solutions.

Energy contributions of AI annual workloads are categorized into three major areas [7, 18, 39, 43, 44]:

- Model development (10% of the energy footprint): Models are developed and fine-tuned before training.
- Model training (30% of the energy footprint): Algorithms learn by processing a vast array of data to make predictions or decisions without exact input-response relations preprogrammed, which requires substantial computation efforts and high energy expenditure for extended periods.
- Use/inference (60% of the energy footprint): Includes the deployment and utilization of developed AI models in real-world applications and requires computational resources for interpreting new data and generating outcomes or predictions based on pre-trained models.

For detailed information on AI model types, specific models and their descriptions, and the electricity consumption of each, see **Appendix B**.

Chat GPT and Other Large Language Models (LLMs)

Over the last year, the surge in popularity of generative AI sparked by the public release of Open AI's ChatGPT has created new concerns about AI's potential impact on future computing energy needs. **Figure 9** shows the increase in web traffic—starting from zero—for prominent generative AI platforms including ChatGPT, which is illustrated by the dark blue line [45].

ChatGPT garnered 100 million global users in only two months, which was rapidly followed by tech giants like Microsoft, Alphabet, Meta, and Bing launching their own large-language model chatbots. From a power usage perspective, these LLMs create a new frontier with ultimate impact to be determined, in part, by how widely the 5.3 billion internet users adopt the new features being rolled out. [46, 47].

For example, Google plans to implement LLMs to boost its search engine's ability to recognize and respond to user queries in a more conversational and natural style [48]. At

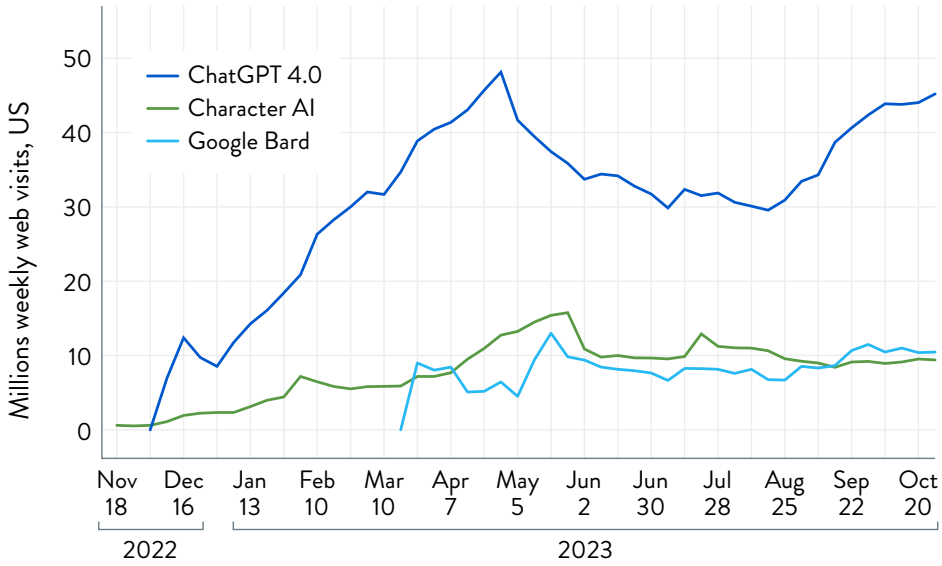


Figure 9. U.S. web traffic trends to AI platforms, 2022–2023 [45]

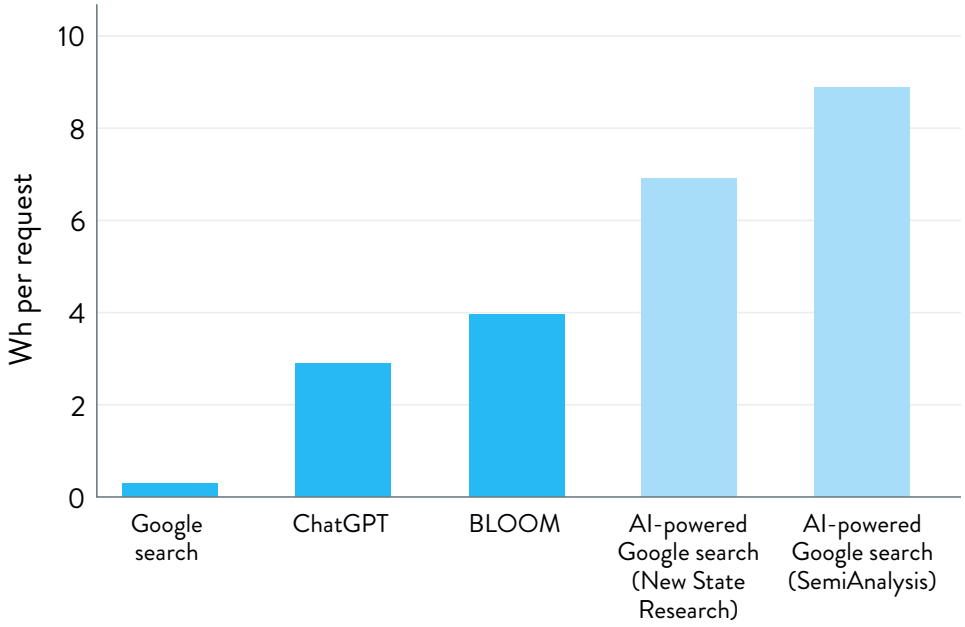


Figure 10. Electricity consumption per request [47]

2.9 watt-hours (Wh) per ChatGPT request, AI queries are estimated to require 10 times the electricity of traditional Google queries, which use about 0.3 Wh each [47]. Implementing LLMs in every Google search could necessitate 80 gigawatt-hours (GWh) daily or 29.2 terawatt-hours (TWh) yearly electricity consumption, according to SemiAnalysis [34]. New Street Research's similar analysis suggests the need for around 400,000 servers, consuming 62.4 GWh daily or 22.8 TWh yearly [47]. As shown in **Figure 10**, the BLOOM model's electricity usage averages 3.96 Wh per request, while ChatGPT's is slightly lower at 2.9 Wh per request; however, if Google integrated similar AI into its searches, the electricity per search could increase to between 6.9–8.9 Wh [47].

The explosive growth in investments aimed at building and deploying new AI capabilities are raising concerns over the overall electricity consumption and environmental impact of AI and data centers and the ability of the United States to maintain its leadership position.

FORECASTING DATA CENTER LOAD GROWTH TO 2030

Four Scenarios Based on Historical Data, Expert Insights, and Current Trends

Drawing on public information about existing data centers, public estimates of industry growth, and recent electricity demand forecasts by industry experts, EPRI prepared four projections—using low (3.7%), moderate (5%), high (10%), and higher (15%) growth scenarios described in **Table 3** below—of potential electricity consumption in U.S. data centers from 2023 to 2030. See **Figure 11** for a graph of the projections. These projections are based on a bounding analysis of various data sources surveyed as of November 2023 [1, 2, 4, 8, 14]. The analysis reflects historical trends for the AI industry, internet traffic, demand for storage, coupled with the computational intensity and prevalence of AI models. All of these factors are uncertain, including the development of business models and updates for LLMs, rate of increase in mature applications, and efficiency gains in computational and non-computational aspects of data centers.

The graph's blue line depicts average historical data center electricity consumption. The light blue area indicates the uncertainty in recent historical projections of data center power use, and the colored swaths show the four projection scenarios [4, 8].

Under the 15% higher growth scenario, EPRI's projections show data center electricity usage rising to an average of 403.9 TWh/year. Under the 10% high growth scenario, data center energy usage rises to a mid-range of 296.4 TWh/yr. Using the moderate growth 5% scenario, the projection predicts a mid-range of 214.0 TWh/yr. Under the 3.7% low growth scenario, the graph shows the projection at a mid-range of 196.3 TWh/yr. The mid-range estimates of data centers' share of total U.S. electricity consumption in 2030—9.1%, 6.8%, 5.0%, and 4.6%—assume that other loads grow at 1% annually. An examination of regional variations is found in **Appendix A**.

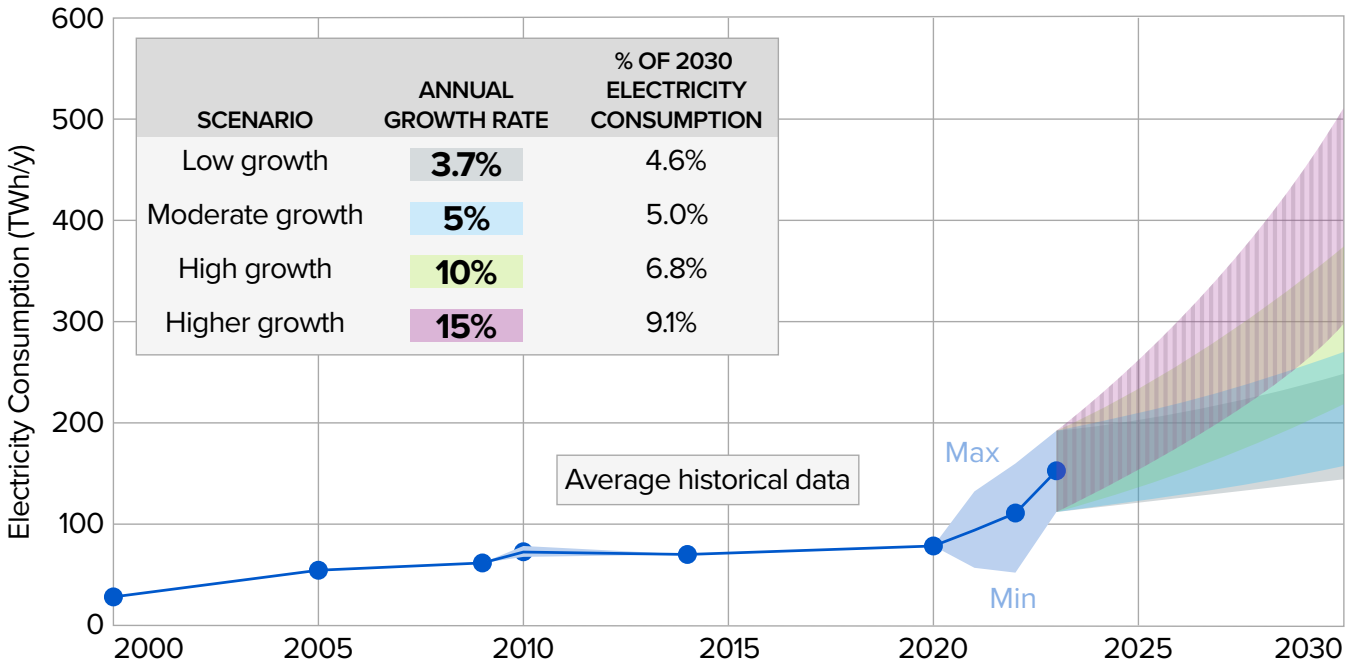


Figure 11. Projections of potential power consumption in U.S. data centers scenarios, 2023–2030 [1, 2, 4, 8, 14]

Table 3. Forecasted load projections: Parameters of power consumption in each of the four U.S. data center scenarios, 2022-2030 [4, 8, 9]

COMPOSITION OF GROWTH SCENARIOS (2023—2030)				
GROWTH SCENARIO	CAGR (%)	AVERAGE 2023 DATA CENTER LOAD (MWH)	AVERAGE PROJECTED LOAD, 2030 (MWH)	CHANGE IN GROWTH (Δ)
Higher Growth	15%	152,120,846	403,906,136	166%
High Growth	10%	152,120,846	296,440,493	95%
Moderate Growth	5%	152,120,846	214,049,306	41%
Low Growth	3.7%	152,120,846	196,305,818	29%

Each growth scenario's characteristics are described in **Table 3**. The load projections combine estimates of today's data center power usage with assessments of potential future technological advances and computational demands. It is essential to understand that while these scenarios are based upon the latest available data and subject-matter expert (SME) insights, the factors affecting them—such as consumer demand, technological advancements, operational efficiencies, and evolving industry standards—are changing almost daily.

ENERGY EFFICIENCY, LOAD MANAGEMENT AND CLEAN ELECTRICITY SUPPLY

With the escalating demands of AI and data center operations, there is a critical need for new, innovative strategies that leverage advances in hardware, system monitoring, computational algorithms, clean electricity procurement, and operational flexibility [49]. Key considerations include:

- Energy efficiency: Adopt advanced cooling solutions, power management systems, and leverage efficiency advances in computational and supporting hardware to reduce overall electricity consumption.
- Scalability: Implement modular designs and virtualization techniques to ensure that infrastructure can handle future demands without disproportionate increases in energy use.
- Carbon-free energy (CFE) use: Transition to carbon-free electricity sources for data center operations and low-carbon technologies for backup power to support data center expansion without increasing carbon emissions and make energy costs more certain.
- Monitoring and analytics: Utilize real-time monitoring tools to track electricity consumption, detect inefficiencies, and optimize operations.
- Research and development (R&D): Invest in innovations that drive both performance and sustainability, such as green energy sources or AI-driven optimization of work load to meet latency, grid, environmental, and other objectives.

The remainder of this section discusses in more detail strategies that cover energy-efficient algorithms, hardware, and cooling technologies; scalability and renewable energy use; and monitoring, analytics, and R&D.

Energy-Efficient Training Algorithms

The initial phases of AI algorithm development have been heavily focused on enhancing accuracy and augmenting performance capabilities. However, as the performance of the algorithms has increased and recognizing the exponential growth in computational demands, the paradigm is beginning to shift to also value the efficiency of model development. Recent studies document applications where a slight compromise on model accuracy has yielded substantial reductions in electricity consumption [7, 18, 43, 50]. The techniques utilized include:

- Pruning: This technique aims to reduce or eliminate unnecessary elements in neural networks, thereby maintaining robust performance while reducing computational complexity [43, 51].
- Quantization: This method reduces the numerical precision of computations, effectively conserving electricity without compromising significantly on accuracy [14, 51].
- Knowledge distillation: This approach involves developing a smaller, more manageable model that mirrors the functionalities of a larger, more intricate structure, reducing computational requirements [14, 51].

Energy-Efficient Hardware

Computational hardware is becoming more efficient, venturing beyond general-purpose central processing units (CPUs) to embrace an array of specialized hardware. These hardware variants are customized for specific tasks, streamlining power usage, and enhancing overall efficiency. This specialized hardware includes:

- Tensor processing units (TPUs): Specifically designed to expedite machine learning (ML) tasks, these units provide pronounced performance and energy efficiency enhancements [11, 18]. For example, Google's Cloud TPUv4 showed not only a 10-times leap forward in ML system performance over TPUv3, but it also boosted energy efficiency by 2–3% compared to contemporary ML data structures and algorithms [52].
- Field-programmable gate arrays (FPGAs): Recognized for their versatility as non-hard etched processors, FPGAs can be reprogrammed for specific tasks, providing improved performance and lower per-unit energy consumption [11, 28]. Though savings are task-dependent, FPGAs have shown reductions in memory and bandwidth usage as much as 75% when compared to traditional CPUs and graphics processing units (GPUs) [28, 53, 54].
- Power capping: Some processing chips, such as GPUs, can operate at reduced power levels. For example they can reduce direct power consumption by 10% while also reducing cooling needs.

Energy-Efficient Cooling Technologies

Heat is a byproduct of computation, and traditional cooling methods are energy-intensive, composing around 35% of data center electricity use. However, innovative solutions are emerging, some of which include:

- Liquid cooling: Utilizing liquids to absorb and dissipate heat can use less electricity than traditional air-cooling systems [18, 50]. A recent study, which examined the shift from 100% air cooling to a combination of 25% air cooling and 75% liquid cooling, highlighted the efficiency gains—leading to a notable decrease in PUE—from transitioning to hybrid cooling systems in data centers. The study observed a 27% reduction in facility power consumption and a 15.5% decrease in overall energy usage across the data center site [55]. **Table 4** shows an overview of various innovative cooling ttechnologies currently being adopted or considered in data centers, highlighting vendor-reported technology-readiness level (TRL) and energy-saving estimates [56, 57, 58, 59, 60, 61].
- Economizer use: An economizer can evaluate outside temperature and humidity, and use exterior air to help cool data center infrastructure when appropriate, minimizing reliance on mechanical cooling methods and leading to significant electricity savings [18, 62]. A 2015 study found that air-side economizers yielded cooling coil load savings of 76–99% in comparison to conventional cooling systems in data centers; and the total cooling energy savings of the economizers ranged from 47.5%–67.2% [62].

Table 4. Emerging cooling technologies with vendor-reported TRLs and energy savings [56, 57, 58, 59, 60, 61]

EMERGING COOLING TECHNOLOGIES		
TECHNOLOGY	**TECHNOLOGY-READINESS LEVEL (TRL)**	**CLAIMED EFFICIENCY DIFFERENTIAL (%)**
Air-Assisted Liquid Cooling	9	This technology offers up to a 50% reduction in energy usage compared to traditional air cooling, with the potential to reach a PUE of less than 1.1 [56].
Immersion Cooling	8	Immersion cooling promises substantial energy savings from 50–95% compared to traditional air-cooling methods [57, 58].
Microconvective Liquid Cooling	6	This emerging technology proposes an 18% energy saving and a PUE of 1.02, alongside a 90% reduction in water usage compared to other liquid systems, indicating its potential for more sustainable operation [59].
Radiative Cooling	6	This solution offers 50–70% energy savings, with the benefits of zero water use and low maintenance [60].
Two-Phase Liquid Immersion Cooling	7	This technology claims a 41% energy saving compared to air cooling, noting its water conservation and space-saving benefits [61].

- Heat reuse: Heat generated by computation can be used for various applications such as heating adjacent buildings, particularly in cold climates, thereby reducing overall energy usage [18, 42, 63]. Since 2016, Amazon's 1.1 million-square-foot Doppler building has been estimated to recover 3200 MWh of excess heat from a nearby data center; this is projected to continue over the next 25 years. This heat, which would otherwise have been wasted and would have required cooling equipment, is redirected through the district's energy system, demonstrating an energy-efficient approach to energy reutilization [25].

Scalable Clean Energy Use

As digital services proliferate and demand for computational power intensifies, scalable clean energy supplies are important to avoid increases in greenhouse gas emissions [64]. Corporate commitments to acquire carbon-free electricity on an annual or hourly-matched basis are emerging and can play a significant role in reducing data center emissions impacts. These include:

- Clean electricity procurement from the grid and clean onsite generation: Data center owners have been instrumental in driving the corporate shift towards contracting for renewable energy to provide their power needs. In 2021, Apple, Google, Meta, and Microsoft matched their operational electricity consumption, predominantly from data centers, on an annual basis with the purchase or generation of renewable electricity—2800 MWh, 18,300 MWh, 9400 MWh, and 13,000 MWh respectively [55, 56, 57, 58]. Meanwhile, Amazon's operations consumed 30,900 MWh, 85% of which was matched on an annual basis by generation from renewable sources, and the company aims to reach 100% renewable energy by 2025 [25, 16, 41]. Moreover, a growing number of organizations are working towards 24/7 CFE, which entails matching their electricity demand with carbon-free sources in the same region on an hourly basis. This hourly matching will require flexible technologies such as batteries that can shift solar or wind output to times when they are needed as well as firm clean capacity such as nuclear, fossil plants with carbon capture and storage, or geothermal, that typically operate around the clock. Spurring deployment of flexible and clean firm assets can help speed the path to a net-zero power sector [42, 44].
- Cleaner onsite backup power systems: Backup power systems at most existing data centers typically operate for less than 100 hours annually when the grid or primary power supply are unavailable. Accordingly, they constitute only a small portion of a data center's environmental footprint. Shifting from the most common backup technology, diesel generators, to lower-emitting alternatives, like battery energy storage systems (BESS) or cleaner fuels—such as renewable natural gas, biodiesel, or clean hydrogen or ammonia, especially when the latter are integrated with fuel cells—can reduce backup GHG emissions and, in some instances, allow more frequent operation of these resources, creating the potential for them to serve as a grid resource when/if needed [9, 42].
- Clean onsite or nearby technologies such as nuclear generation or renewable generation coupled with long-duration energy storage that can match the growing size of data centers: With currently proposed data centers reaching 1 GW or more at a single site, the scale of power demand is escalating rapidly. In the near term, uprating, relicensing, or restarting existing nuclear plants near data centers could provide one solution. Amazon's purchase of a data center in Pennsylvania co-located with a Talen nuclear power plant provides one example of utilizing existing nuclear. Looking forward, small modular reactors (SMRs) offer a scalable power solution that can grow with the demands of a data center. Companies such as NuScale are exploring scalable capacities of 250–600MW for SMRs [9, 42]. Standard Power has chosen NuScale's SMR technology to power two facilities it plans to develop, one in Ohio and the other in Pennsylvania [69].

Monitoring and Analytics

Advances in monitoring and analytics of power consumption play a crucial role in realizing operational savings in data centers. These processes enable precise tracking of energy usage, identification of inefficiencies, and implementation of advanced technologies, thus driving cost reduction and enhancing overall efficiency:

- Efficient server management: Traditionally, data centers have grappled with up to 30% server underutilization, where servers consume energy but don't fully utilize their computational capabilities. However, with the adoption of innovations like advanced scheduling and dynamic resource allocation, some companies are aiming to reduce underutilization rates to below 10% within the next five years [18, 40, 63, 70]. In addition, the implementation of virtualization and containeriza-

tion can enhance server efficiency significantly, potentially increasing server capacity utilization by 45% by having a single physical server handle more workloads through virtual or containerized environments. If successful, this is estimated to reduce the number of physical servers needed, leading to about 20% less energy consumption per unit of computation over the next decade [15, 33, 42].

- Flexible computation strategies: Optimizing data center computation and geographic location to respond to electricity supply conditions, electricity carbon intensity, and other factors in addition to minimizing latency enables data centers to actively adjust their electricity consumption [71]. For example, some could achieve significant cost savings—as much as 15%—by optimizing computation to capitalize on lower electric rates during off-peak hours, reducing strain on the grid during high-demand periods [38, 72]. With technological and regulatory advances, these strategies could evolve to incorporate real-time energy market dynamics enabling data centers to not only adjust their operations based on grid demands but also actively participate in energy markets to optimize their benefits and support grid stability.

Reducing Data Centers' Environmental Footprint

The previous sections focus on actions that data center owners and operators are actively pursuing to diminish their carbon footprint, focusing primarily on onsite direct emissions such as from onsite generation (Scope 1 emissions) and emissions associated with the purchase of electricity (Scope 2 emissions) [64]. These strategies involve reducing their electricity needs through the adoption of advanced computation, cooling, and operational technologies, shifting toward cleaner onsite backup power, and moving towards various strategies for matching their hourly loads with carbon-free electricity [73]. Several of the hyperscale companies have fully matched their annual power purchases with carbon-free electricity on an annual basis and are moving forward on hourly matching. Progress is slower on shifting to cleaner backup power (although this, as noted earlier, represents only a small fraction of their environmental footprint).

In recent years, some companies have taken the additional step of quantifying and setting reduction targets for their (Scope 3) indirect emissions, which include emissions associated with supply chains and end-user services [7, 44, 66]. Key actions include sourcing materials from environmentally responsible vendors, minimizing the carbon footprint associated with transportation and logistics, and ensuring that the lifecycle of data center components is managed sustainably, from manufacturing to end-of-life disposal and recycling [42, 74, 75].

ACTIONS TO SUPPORT RAPID DATA CENTER EXPANSION

Data centers are one of the fastest growing industries worldwide. These facilities—and advanced cloud computing and AI technologies that are proliferating and driving further growth—represent large point loads and are at the leading edge of an anticipated global rise in electricity demand driven by efficient electrification and production of low-carbon fuels.

In the United States, data center power demand growth, coupled with increasing electricity demands from EVs, heat pumps, electrification in industry, and the onshoring of manufacturing incentivized by the CHIPS Act, Inflation Reduction Act (IRA), and Infrastructure Investment and Jobs Act (IIJA), is placing both immediate and sustained pressure on the electric grid to accommodate new loads.

Clusters of new, large point loads create several challenges. Data centers' speed from breaking ground to operation—often within two or three years—requirements for highly reliable power, and requests for power generated by new, non-emitting generation sources can create local and regional electric supply challenges and test the ability of electric companies to keep pace. The most serious challenges to data center expansion are local and result from the scale of the centers themselves and mismatches in infrastructure timing.

EPRI highlights three essential strategies to support rapid data center expansion. These strategies, each of which is explained below, emphasize increased collaboration between data center developers and electric companies and are:

- Improve data center operational efficiency and flexibility
- Increase collaboration through a shared energy economy model for sustainable data centers
- Better anticipate future point load growth through improved forecasting and modeling

Improve Data Center Operational Efficiency and Flexibility

Over the past decade, economy-wide electricity demand in the United States has remained relatively flat in large part due to enhanced energy efficiency, which has offset potential increases driven by economic expansion and population growth. Specific to data centers, power demands from rapid expansion in computation, communication, and data storage were largely offset by efficiency gains for over a decade. This is largely due to technological advancements in computation, improved cooling systems, sophisticated energy management strategies, and the replacement of many small data centers with more efficient cloud data centers. However, since around 2018, efficiency gains have slowed, data center expansion accelerated (in part due to lifestyle changes caused by the pandemic), and AI has proliferated, leading to an increase in data center power consumption.

Meeting the increasing electricity demands of AI and data centers while limiting the growth of CO_2 emissions necessitates a comprehensive strategy that intertwines technological advancements that improve efficiency with power purchase and production strategies that favor low-carbon resources and that increase both temporal and spatial flexibility to link intense operation periods to the availability of low-cost, low-carbon generation.

Computational efficiency gains require investing in the next generation of energy-efficient processors and server architectures and enhancing AI training algorithms for greater computational efficiency. From an architectural viewpoint, virtualization stands out, with its capability to run multiple virtual machines on one physical server, potentially cutting hardware needs by 30–40% with consequent electricity savings [9, 75]. Implementations like software-defined infrastructure (SDI) offer dynamic resource allocation in real time, potentially increasing allocation efficiency by 30%, potentially increasing spatial flexibility in computation loads. Hybrid cloud solutions provide a balance between on-premises infrastructure and shared cloud services, potentially providing locational flexibility by reducing onsite requirements by 25% during peak periods.

In addition, continued gains in data center infrastructure efficiency can be achieved through more effective cooling technologies, adopting energy management systems that leverage AI for optimized power usage, and setting stringent industry targets for energy consumption. Continuous monitoring and analytics can help data centers better anticipate and react to dynamic energy needs, ensuring optimal operational efficiency and rapid adaptability [42, 50, 78]. Embedding real-time monitoring tools within AI and data center ecosystems can facilitate immediate insights into fluctuations in electricity usage. Pilot projects to explore and validate novel energy conservation methods, which document and disseminate findings broadly, can accelerate adoption of proven sustainable strategies [10, 70].

Increase Collaboration through a Shared Energy Economy Model for Sustainable Data Centers

Electric companies are challenged as they must meet the increasing and uncertain load from data centers while also ensuring reliability, affordability, and sustainability for all customers. Developing a deeper understanding of data center power needs, timing, and potential flexibilities—while assessing how they match available electric supplies and delivery constraints—can create workable solutions for all.

EPRI, in collaboration with major data center builders/operators/owners and the electric companies that power these facilities, is exploring sustainable approaches to powering the growing wave of AI data centers. Enabled by technology and supporting policies, data center backup generators, powered by clean fuels, could support a more reliable grid while reducing the cost of data center operation. Shifting the data center-grid relationship from the current "passive load" model to a collaborative "shared energy economy"—with grid resources powering data centers and data center backup resources contributing to grid reliability and flexibility—could not only help electric companies contend with the explosive growth of AI but also contribute to affordability and reliability for all electricity users.

This new paradigm of collaboration between data centers and electric companies, which transforms data centers from passive consumers to active participants in maintaining the grid, is crucial for ensuring electric companies are prepared for the explosive growth of AI. Under this model, data centers move from being a burden on the grid—acting as passive loads demanding specific power levels within defined timeframes and at affordable rates—to becoming partners in a sustainable future, serving as a grid reliability resource. The goal is the complete integration of grid and data center power resources. Clean power generators co-located with data centers act as both grid and data center power sources. During grid outages, these resources can seamlessly form a microgrid to provide uninterrupted power to data

centers, eliminating the need and cost of standard diesel backup generators.

More research is needed into how data centers and electric grids can collaborate in a shared energy economy model, as well as the benefits and challenges of doing so. Focusing on U.S. AI training data centers using backup generators powered by clean fuels, EPRI suggests a study of the economic, environmental, social, and technological implications of this shared energy economy model compared to other, more traditional models. The results of this study could provide suggestions and guidelines for data centers and electric grids to adopt and implement the shared energy economy model, or parts of it, in their operations and planning.

Better Anticipate Future Point Load Growth through Improved Forecasting and Modeling

The lead time for constructing and bringing a large data center online is around two to three years, while adding new electric infrastructure (generation, transmission, substations) can take four or many more years. This highlights the need for better forecasting and decision tools to anticipate where and when data center connection requests may appear and characterizing the operational characteristics of that load, especially as the size of interconnection requests grow from hundreds of MW to thousands of MW.

In the current environment, electricity companies are often receiving multiple requests for the same project from the owner and from developers trying to support the owner. Also, a single data center project may seek interconnection information in multiple locations. And the ramp up to full power demand and operational characteristics on the data centers can vary widely, depending upon their function (e.g., cloud, AI training, AI inference). Therefore, new approaches are needed not only to project where load will grow, but also its operational characteristics and opportunities for flexible operation.

EPRI's Load Forecasting Initiative (https://msites.epri.com/lfi), initiated in late 2023, has research activities underway to help address some of these key uncertainties.

APPENDIX A: STATE-SPECIFIC SCENARIOS

Projected Data Center Load Scenarios for Top 15 States

Figures A1 through A15 apply the projected U.S. load growth rates under EPRI's higher-, high-, moderate-, and low-growth scenarios to 2023 estimated state-level data center loads. The figures show projections for the 15 states with the highest data center demands in 2023, comprising around 80% of U.S. data center load in that year. As noted above, the projections utilize the projected national growth rate and do not reflect the deferential regional growth rates implied by Integrated Resource Plan analyses that have emerged recently.

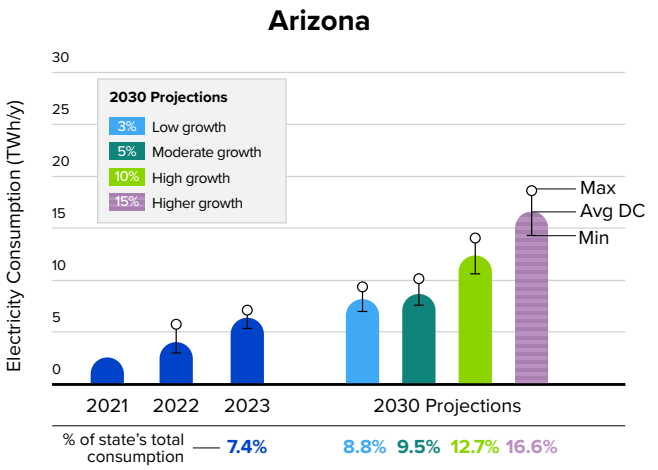


Figure A1.
Projected electricity consumption in Arizona data centers

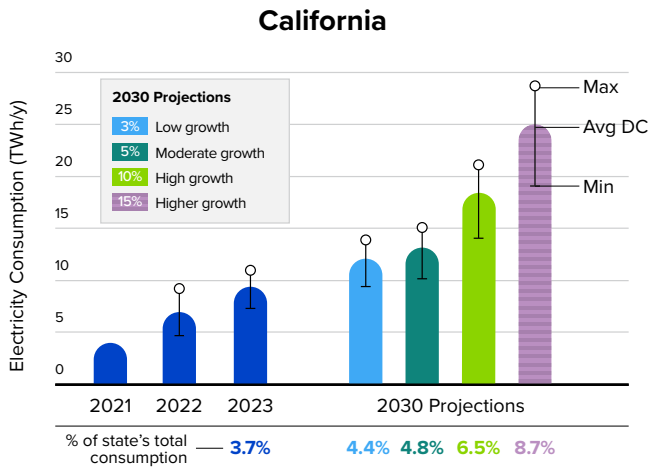


Figure A2.
Projected electricity consumption in California data centers

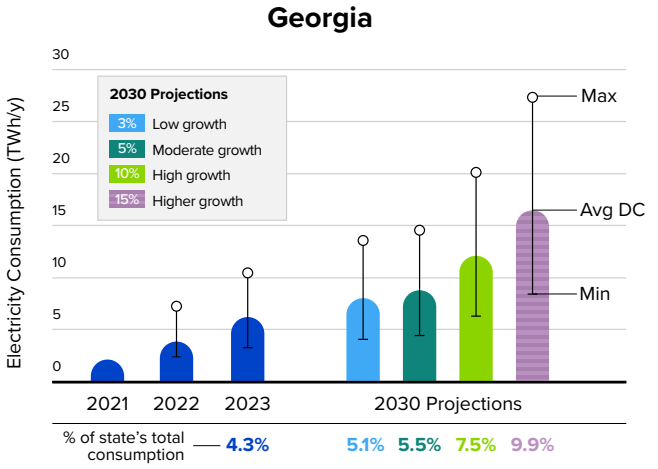


Figure A3.
Projected electricity consumption in Georgia data centers

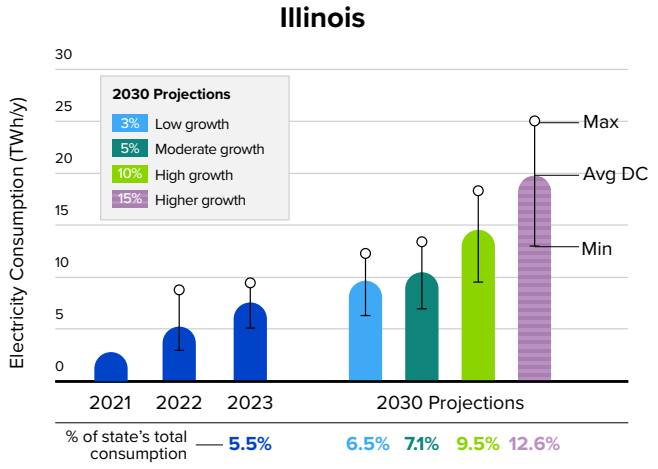


Figure A4.
Projected electricity consumption in Illinois data centers

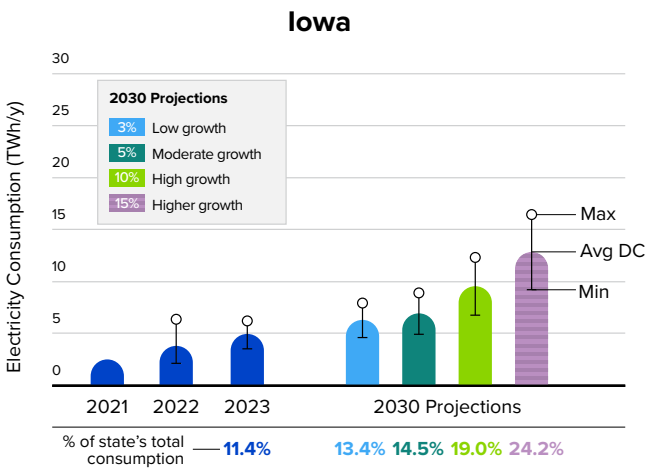


Figure A5.
Projected electricity consumption in Iowa data centers

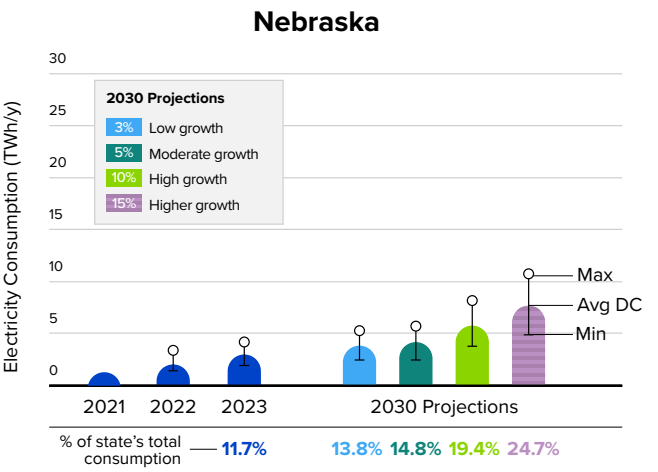


Figure A6.
Projected electricity consumption in Nebraska data centers

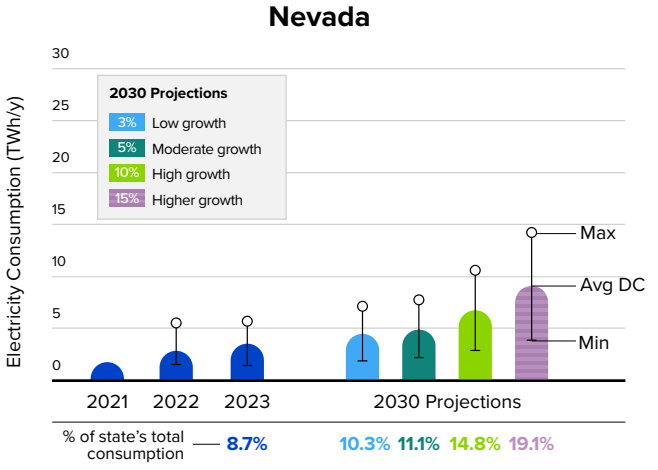


Figure A7.
Projected electricity consumption in Nevada data centers

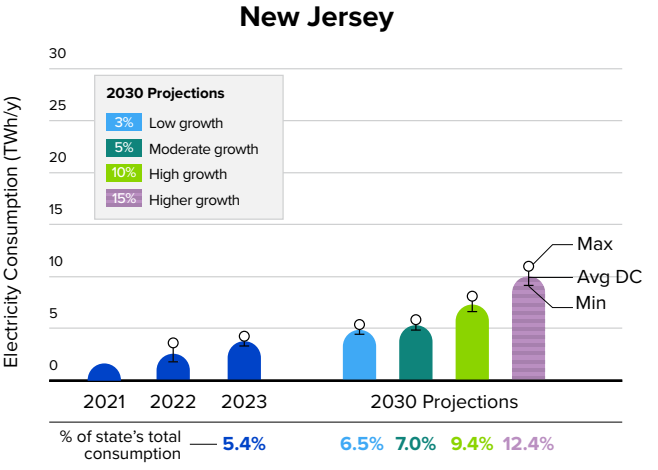


Figure A8.
Projected electricity consumption in New Jersey data centers

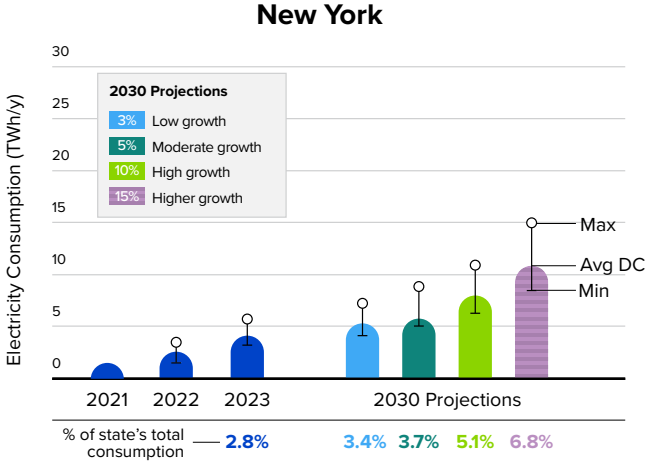


Figure A9.
Projected electricity consumption in New York data centers

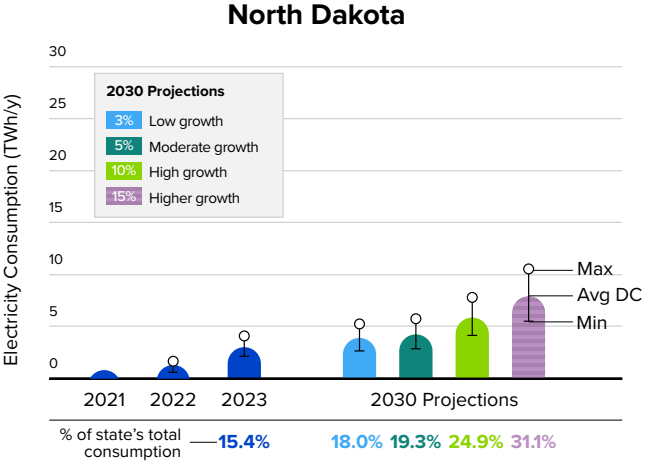


Figure A10.
Projected electricity consumption in North Dakota data centers

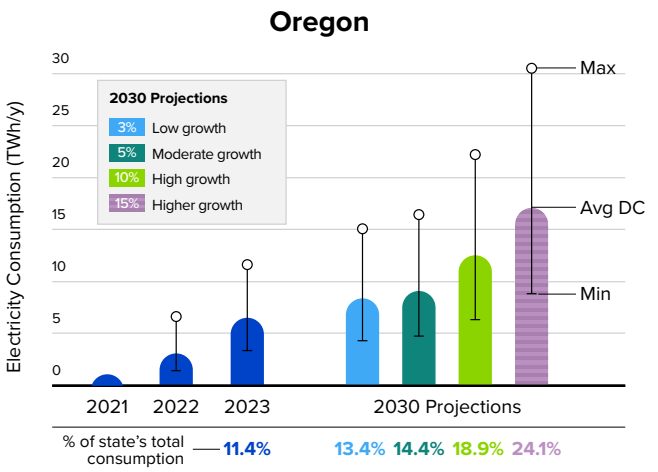


Figure A11.
Projected electricity consumption in Oregon data centers

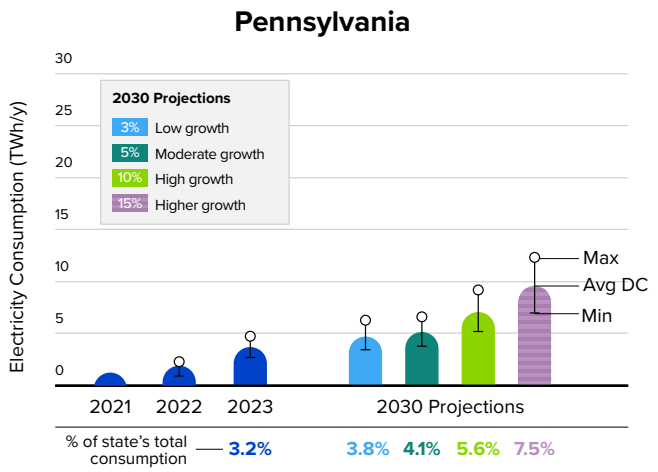


Figure A12.
Projected electricity consumption in Pennsylvania data centers

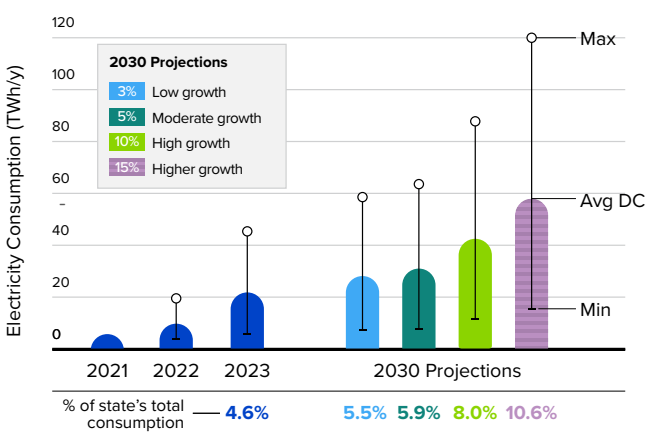


Figure A13.
Projected electricity consumption in Texas data centers

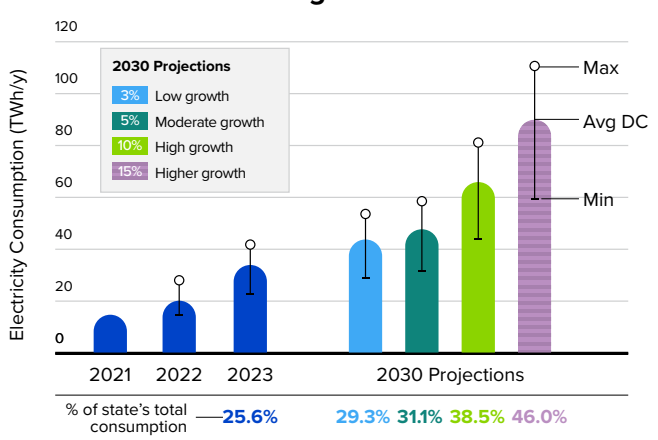


Figure A14.
Projected electricity consumption in Virginia data centers

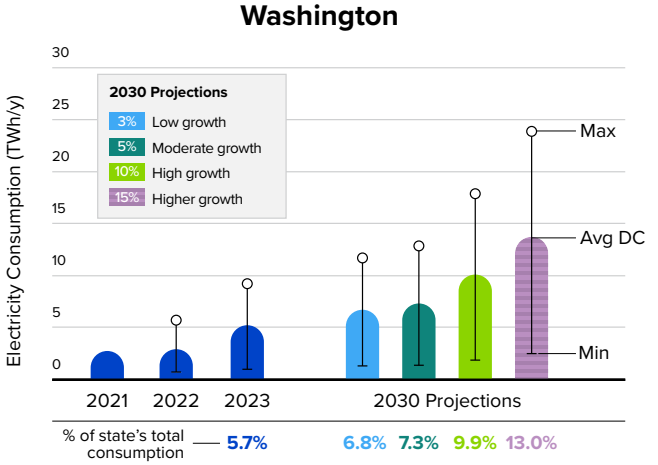


Figure A15.
Projected electricity consumption in Washington data centers

Regional Differences in Data Center Capacities by Metropolitan Area

Data center development is heavily clustered in a few counties/cities across the country rather than evenly spread within states, exacerbating power delivery challenges. **Figure A16** provides a snapshot for leading metropolitan areas of current data center capacity (measured in MW); additional capacity under development; absorption rates, reflecting the percentage of capacity leased by customers over a specific period of time; and vacancy rates, indicating unutilized space within these data centers.

Northern Virginia is the clear leader in terms of current capacity and current construction. Other regions, such as Dallas-Ft. Worth, Silicon Valley, Chicago, New York Tri-State, and Atlanta, highlight current construction activity that is projected to lead to a 50% or more increase in power demands. [29].

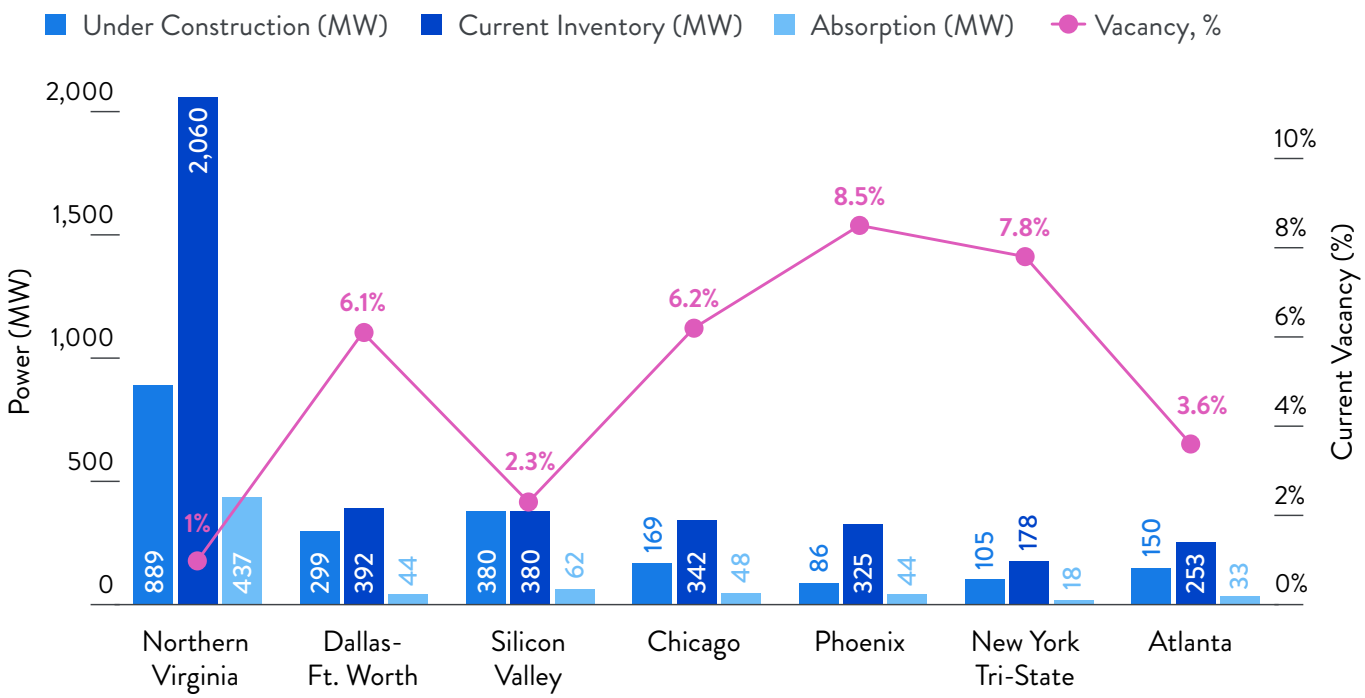


Figure A16. Data center development: Key U.S. regions (2022)

Projections of Potential Power Consumption for 44 States

Table A1 presents a detailed view of the energy consumption from data centers in each of the 44 states that had significant data center load in 2023 and contrasts it with projections for 2030. These projections are categorized into three scenarios: low growth, moderate growth, high growth, and higher growth [1, 2, 4, 8, 14].

Table A1. Projections to 2030 of potential power consumption for states with significant data center load in 2023 [4, 8, 9]

FORECASTED SCENARIOS: PROJECTIONS OF POTENTIAL POWER CONSUMPTION BY STATE (2023–2030)										
	2023 Load		Low-growth Scenario (3.71%)		Moderate-growth Scenario (5%)		High-growth Scenario (10%)		Higher-growth Scenario (15%)	
STATE	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)
Alabama	1,489,200	1.71%	1,921,753	2.05%	2,095,454	2.23%	2,902,030	3.07%	3,954,074	4.13%
Arizona	6,253,268	7.43%	8,069,590	8.81%	8,798,975	9.53%	12,185,850	12.73%	16,603,465	16.58%
California	9,331,619	3.70%	12,042,078	4.43%	13,130,525	4.81%	18,184,686	6.54%	24,777,000	8.70%
Colorado	1,509,640	2.66%	1,948,130	3.18%	2,124,215	3.46%	2,941,861	4.73%	4,008,345	6.34%
Connecticut	262,800	0.95%	339,133	1.14%	369,786	1.24%	512,123	1.71%	697,778	2.31%
Florida	1,384,080	0.56%	1,786,099	0.67%	1,947,540	0.73%	2,697,180	1.01%	3,674,963	1.37%
Georgia	6,175,391	4.26%	7,969,093	5.08%	8,689,396	5.51%	12,034,090	7.48%	16,396,690	9.92%
Hawaii	8,760	0.10%	11,304	0.12%	12,326	0.13%	17,071	0.18%	23,259	0.24%
Idaho	148,920	0.57%	192,175	0.68%	209,545	0.74%	290,203	1.03%	395,407	1.40%
Illinois	7,450,176	5.48%	9,614,151	6.53%	10,483,145	7.08%	14,518,285	9.54%	19,781,455	12.56%
Indiana	192,720	0.19%	248,697	0.23%	271,176	0.25%	375,557	0.35%	511,704	0.48%
Iowa	6,193,320	11.43%	7,992,230	13.44%	8,714,623	14.48%	12,069,029	18.99%	16,444,294	24.21%
Kansas	8,760	0.02%	11,304	0.03%	12,326	0.03%	17,071	0.04%	23,259	0.05%
Kentucky	1,620,600	2.15%	2,091,319	2.58%	2,280,347	2.80%	3,158,091	3.84%	4,302,962	5.16%
Louisiana	78,840	0.08%	101,740	0.10%	110,936	0.11%	153,637	0.15%	209,333	0.20%
Maine	26,280	0.22%	33,913	0.27%	36,979	0.29%	51,212	0.40%	69,778	0.55%
Maryland	96,360	0.16%	124,349	0.19%	135,588	0.21%	187,778	0.29%	255,852	0.40%
Massachusetts	1,062,369	2.08%	1,370,944	2.50%	1,494,860	2.72%	2,070,257	3.72%	2,820,766	5.01%
Michigan	525,600	0.52%	678,266	0.63%	739,572	0.68%	1,024,246	0.95%	1,395,555	1.28%
Minnesota	824,316	1.24%	1,063,747	1.49%	1,159,895	1.62%	1,606,359	2.23%	2,188,696	3.01%
Missouri	972,360	1.21%	1,254,791	1.45%	1,368,208	1.58%	1,894,855	2.18%	2,581,777	2.95%
Montana	578,160	3.71%	746,092	4.43%	813,529	4.81%	1,126,670	6.54%	1,535,111	8.71%
Nebraska	3,959,520	11.70%	5,109,601	13.75%	5,571,442	14.81%	7,715,984	19.41%	10,513,184	24.71%
Nevada	3,416,707	8.69%	4,409,122	10.28%	4,807,649	11.10%	6,658,195	14.75%	9,071,924	19.07%
New Hampshire	17,520	0.16%	22,609	0.19%	24,652	0.21%	34,142	0.29%	46,519	0.40%
New Jersey	4,038,360	5.42%	5,211,341	6.46%	5,682,378	7.00%	7,869,621	9.44%	10,722,517	12.44%
New Mexico	402,960	1.48%	520,004	1.78%	567,005	1.94%	785,255	2.66%	1,069,926	3.60%
New York	4,067,385	2.84%	5,248,796	3.40%	5,723,219	3.69%	7,926,182	5.05%	10,799,583	6.75%
North Carolina	2,672,676	1.92%	3,448,981	2.30%	3,760,724	2.50%	5,208,289	3.44%	7,096,399	4.62%
North Dakota	3,915,720	15.42%	5,053,079	18.00%	5,509,811	19.31%	7,630,631	24.89%	10,396,888	31.11%
Ohio	2,363,886	1.58%	3,050,500	1.90%	3,326,225	2.07%	4,606,545	2.84%	6,276,510	3.83%
Oklahoma	1,226,400	1.76%	1,582,620	2.12%	1,725,668	2.30%	2,389,907	3.16%	3,256,296	4.26%

FORECASTED SCENARIOS: PROJECTIONS OF POTENTIAL POWER CONSUMPTION BY STATE (2023–2030)										
	2023 Load		Low-growth Scenario (3.71%)		Moderate-growth Scenario (5%)		High-growth Scenario (10%)		Higher-growth Scenario (15%)	
STATE	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)	MWh/y	% of Total State Electricity Consumed (%EC)
Oregon	6,413,663	11.39%	8,276,574	13.39%	9,024,668	14.43%	12,498,415	18.93%	17,029,342	24.14%
Pennsylvania	4,590,240	3.16%	5,923,520	3.78%	6,458,929	4.11%	8,945,079	5.61%	12,187,850	7.49%
Rhode Island	17,520	0.23%	22,609	0.28%	24,652	0.30%	34,142	0.42%	46,519	0.57%
South Carolina	2,023,560	2.45%	2,611,323	2.93%	2,847,352	3.18%	3,943,346	4.36%	5,372,888	5.84%
South Dakota	70,080	0.52%	90,435	0.63%	98,610	0.68%	136,566	0.94%	186,074	1.28%
Tennessee	1,327,140	1.30%	1,712,621	1.56%	1,867,419	1.70%	2,586,220	2.34%	3,523,777	3.16%
Texas	21,813,159	4.59%	28,149,002	5.47%	30,693,306	5.94%	42,507,676	8.04%	57,917,564	10.64%
Utah	2,562,037	7.68%	3,306,206	9.10%	3,605,044	9.84%	4,992,686	13.13%	6,802,635	17.08%
Virginia	33,851,122	25.59%	43,683,508	29.28%	47,631,928	31.10%	65,966,260	38.47%	89,880,357	46.00%
Washington	5,171,612	5.69%	6,673,757	6.77%	7,276,977	7.34%	10,078,009	9.88%	13,731,490	13.00%
Wisconsin	148,920	0.21%	192,175	0.26%	209,545	0.28%	290,203	0.39%	395,407	0.53%
Wyoming	1,857,120	11.26%	2,396,538	13.24%	2,613,154	14.27%	3,619,002	18.73%	4,930,962	23.90%

APPENDIX B: INSIGHTS INTO THE ENERGY USE OF AI MODELS

To better appreciate how AI uses such enormous amounts of electricity, it can be useful to understand more about AI models and how they work. AI models are typically divided into three types:

- Process automation and optimization (PAO), which focuses on streamlining and enhancing operations
- Predictive analytics (PA), which deals with forecasting trends and patterns
- Natural language processing (NLP), which interprets and generates human language

Moreover, machine learning (ML), a subset of AI, employs statistical methods to enable machines to improve at tasks with experience. Deep learning (DL), a further subset of ML, involves neural networks with multiple layers that autonomously learn from vast amounts of data. ML and DL have evolved significantly, with industrial applications overtaking academic contributions in recent years. Industry's edge stems from its vast data access, advanced computing capacities, and robust financial backing, positioning it above academia and nonprofit sectors in this subset of the AI domain. ML's broad capabilities enable advancements in PAO, PA, and NLP models, while DL's complex neural networks further refine these applications [81, 82].

Each AI model type has distinct energy implications due to its unique computational requirements. Understanding these distinctions is essential for assessing the broader energy impact of AI's spread. **Table B1** provides a comparative analysis of various AI models' energy consumption and key characteristics, offering a view of their energy footprints and their computational complexity [44, 51, 57, 63, 84, 85, 86, 87, 88].

Table B1. Comparative analysis of AI model load consumption and characteristics [44, 51, 57, 63, 84, 85, 86, 87, 88]

LOAD CONSUMPTION: BY SPECIFIC AI MODEL					
MODEL NAME	AI TYPE	YEAR	TRAINING (DAYS)	CONSUMPTION (MWH)	MODEL DESCRIPTION
T5	PA	2019	~20	85.7	A versatile model trained to convert text inputs into text outputs, suitable for various tasks like translation and summarization.
Meena	NLP	2019	~30	232	A chatbot model developed by Google designed to engage in conversations and understand context more naturally.
Evolved Transformer	PAO	2019	~7	7.5	A machine learning model designed using neural architecture search for improved performance on tasks.
Switch Transformer	PAO	2020	~27	179	A variant of the Transformer model designed to handle a large number of parameters more efficiently by dynamically routing activations to a subset of experts.
GShard-600B	PAO	2020	~3	24.1	Google's model optimized for large-scale multitask training, aiding in handling vast amounts of parameters.
ChatGPT-3	NLP	2021	~34	1,287	A state-of-the-art language model by OpenAI known for generating coherent and contextually relevant sentences over long passages.
BERT	PA	2021	~6	2.8	A model that understands the context of words in a sentence by analyzing them in both directions (left-to-right and right-to-left), widely used in sentiment analysis and other prediction tasks.
Gopher	NLP	2022	~23	1,066	Large language models on many tasks, particularly answering questions about specialized subjects like science and the humanities, such as logical reasoning and mathematics.
BLOOM	PA	2022	~117	433	Multilingual and open source, the Bloom model, which has emerged from the BigScience participatory project, aims to help advance research work on large language models
ChatGPT-4	NLP	2023	~100	62,318.8	An advanced version of OpenAI's ChatGPT series, designed for more nuanced and context-aware language generation.
OPT-175B	NLP	2023	~33	324	A state-of-the-art language model by Meta known for generating coherent and contextually relevant sentences over long passages.

REFERENCES

1. Statista. (2023). Data Center Market in the United States - Statistics & Facts. Statista. [Link]

2. PreScouter. (2023). Data Center Subject Matter Expert Interviews for EPRI Research. PreScouter Interviews, 2023(1).

3. U.S. Department of Energy. (2017). Small Data Centers, Big Energy Savings: An Introduction for Owners and Operators - Final Report. U.S. Department of Energy: Better Buildings, AR(17).

4. CBRE. (2023). Global Data Center Trends 2023. CBRE Research: Intelligent Investment Report, 2023(1). [Link]

5. Koomey, J., Brill, K., Turner, P., Stanley, J., & Taylor, B. (2007). White Paper: A Simple Model for Determining True Total Cost of Ownership for Data Centers. *Uptime Institute, Inc. White Papers*, 1.

6. Hoosain, M.S., et al. (2023). Tools Towards the Sustainability and Circularity of Data Centers. *Circular Economy and Sustainability*, 2023(3), 173-197.

7. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.

8. Cushman & Wakefield. (2023). Global Data Center Market Comparison. Cushman & Wakefield's Data Center Advisory Group, 2023 (1). [Link]

9. U.S. Energy Information Administration. (2023). Energy Information Administration, Electric Power Annual 2023. [Link]

10. Andrae, A. S. (2020). New perspectives on internet electricity use in 2030. *Engineering and Applied Science Letters*, 3(2), 19-31.

11. Koot, M., & Wijnhoven, F. (2021). Usage impact on data center electricity needs: A system dynamic forecasting model. *Applied Energy*, 291(116798).

12. Rahmani, R., Moser, I., & Seyedmahmoudian, M. (2017). A Complete Model for Modular Simulation of Data Centre Power Load. *Journal of IEEE Transactions on Automation Science and Engineering*, Vol. 14(8).

13. Shehabi, A., Smith, S. J., Sartor, D., Brown, R. E., Herrlin, M., Koomey, J., ... & Lintner, W. (2016). United States data center energy usage report. *Lawrence Berkeley National Laboratory*, Berkeley, CA.

14. Mysore, M., Woetzel, J., & Gupta, S. (2022). Investing in the Rising Data Center Economy. McKinsey & Company. [Link]

15. Koomey, J. (2008). Worldwide electricity used in data centers. *Environmental Research Letters*, 3(034008).

16. PJM Resource Adequacy Planning Department. (2022). 2023 Load Forecast Supplement. PJM Resource Adequacy Planning Department. [Link]

17. International Energy Agency (IEA). (2022). World Energy Outlook: 2022. IEA: Annual World Energy Outlook Reports, 2022(1). [Link]

18. Shehabi, A., Smith, S. J., Masanet, E., & Koomey, J. (2018). Data center growth in the United States: decoupling the demand for services from electricity use. *Environmental Research Letters*, 13(124030).

19. TeleGeography. (2023). The State of the Network: 2023 Edition. TeleGeography Annual Reports, 2023(1). [Link]

20. Malmodin, J., et al. (2023). ICT sector electricity consumption and greenhouse gas emissions: 2020 outcome. SSRN, 2023(1).

21. CISCO. (2022). 2022 Global Hybrid Cloud Trends Report. CISCO: Annual Global Hybrid Cloud Reports, 2022(1). [Link]

22. Khanboubi, Y.E., & Hanoune, M. (2019). Exploiting Blockchains to improve Data Upload and Storage in the Cloud. *International Journal of Communication Networks and Information Security (IJCNIS)*, 11(3), 357-364.

23. Daigle, B. (2021). Data Centers Around the World: A Quick Look. *United States International Trade Commission: Executive Briefings on Trade*, 2(1).

24. Andrae, A.S., & Edler, T. (2015). On global electricity usage of communication technology: trends to 2030. *Challenges*, 6(1), 117-157.

25. Amazon. (2022). Annual Corporate Sustainability Report. Amazon Sustainability Reports, 2022(1). [Link]

26. Avelar, V., et al. (2012). PUE: A Comprehensive Examination of the Metric. The Green Grid, White Paper 49(1).

27. Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systemic

review of Green AI. WIREs Data Mining Knowledge Discovery, 13(1507).

28. Tibaldi, M., & Pilato, C. (2023). A Survey of FPGA Optimization Methods for Data Center Energy Efficiency. EEE Transactions on Sustainable Computing, 8(3), 343-362.

29. Davis, J., et al. (2022). Uptime Institute: Global Data Center Survey 2023. Uptime Institute: Planning & Strategy UI Intelligence Report, 78(1).

30. EIRGRID & SONI. (2022). Ireland Capacity Outlook: 2022-2031. Electric Power Transmission Operator in Ireland & System Operator for Northern Ireland, AR(22)

31. Miller, R. (2023). Virginia State Legislators Target Data Center Development with New Bills. Data Center Frontier: Special Reports. [Link]

32. CBRE. (2023). Dallas-Fort Worth Records Unprecedented Data Center leasing Activity in First Half of 2023. CBRE Group, Inc. [Link]

33. Sodhi, R. (2023). How California's New Emissions Disclosure Law Will Affect Data Centers. Security Bloggers Network. [Link]

34. Patel, D., & Ahmad, A. (2023). The Inference Cost of Search Disruption – Large Language Model Cost Analysis. SemiAnalysis Reports, 2(1). [Link]

35. DCF Staff. (2022). Why Phoenix is an Increasingly Hot Data Center Market. Data Center Frontier: Special Reports. [Link]

36. JetCool Technologies, Inc. (2023). Drive Faster Computer Sustainability with Microconvective Cooling: How Microconvective Cooling Technology Provides Future-Ready Flexibility Meeting Data Centers Where They Are Today. JetCool Technologies, Inc. White Papers, 1. Doesn't have an integrated link in the original white paper, but here is a link to where you can find the content online. [Link]

37. Miller, R. (2023). Atlanta Prepares for Data Center Building Boom Amid Growing Interest from Hyperscale Users. Data Center Frontier: Special Reports. [Link]

38. Desislavov, R., Martinez-Plumed, F., & Hernandez-Orallo, J. (2023). Trends in AI Interface Energy Consumption: Beyond the Performance-vs-Parameter Laws of Deep Learning. Sustainable Computing: Informatics and Systems, 38(100857).

39. Stanford University: Human-Centered Artificial Intelligence. (2023). Artificial Intelligence Index Report 2023. Stanford University: HAI. [Link]

40. Costenaro, D., & Duer, A. (2012). The Megawatts behind Your Megabytes: Going from Data-Center to Desktop. American Council for an Energy-Efficient Economy.

41. Imperva. (2023). 2023 Imperva Bad Bot Report. Imperva Bot Reports, 2023(1). [Link]

42. National Renewable Energy Laboratory. (2022). 2022 Standard Scenarios Report: A U.S. Electricity Sector Outlook. National Renewable Energy Laboratory. [Link]

43. Mittal, S. (2019). A Survey of Techniques for Improving Energy Efficiency in Embedded Machine Learning Systems. arXiv Preprint, arXiv:1904.10462.

44. Patterson, D., et al. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Google & University of California-Berkeley. [Link]

45. SimilarWeb. (2023). ChatGPT and Competitors: Weekly Visits Desktop & Mobile Web, U.S. *SimilarWeb Data Comparison*, 2(1).

46. OECD. (2022). Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint. *OECD Digital Economy Papers: OECD Publishing*, 2022(341).

47. Vries, A.D. (2023). The Growing Energy Footprint of Artificial Intelligence. *Joule*, 7(1), 2191-2194.

48. Reid, E, et al. (2023). Supercharging Search with Generative AI. Google Blog Products. [Link]

49. U.S. Environmental Protection Agency. (2022). 16 More Ways to Cut Energy Waste in the Data Center. ENERGY STAR. [Link]

50. National Renewable Energy Laboratory. (2017). The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. National Renewable Energy Laboratory. [Link]

51. Xu, W., et al. (2021). Accelerating Federated Learning for IoT in Big Data Analytics with Pruning, Quantization and Selective Updating. IEEE, 9(1), 38457-38766.

52. Jouppi, N., & Patterson, D. (2023). Google's Cloud TPU v4 provides exaFLOPS-scale ML with industry-leading

efficiency. Google Cloud Products. [Link]

53. Nguyen, T., et al. (2020). The Performance and Energy Efficiency Potential of FPGAs in Scientific Computing. *Lawrence Berkeley National Laboratory (LBNL)*.

54. Qasaimeh, M., et al. (2019). Comparing Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernals. *IEEE*, 978(1), 7281-7289.

55. Heydari, A., et al. (2022). Power Usage Effectiveness Analysis of a High-Density Air-Liquid Hybrid Cooled Data Center. *Proceedings of the ASME 2022 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems*, 25(27).

56. JetCool Technologies, Inc. (2023). Drive Faster Computer Sustainability with Microconvective Cooling: How Microconvective Cooling Technology Provides Future-Ready Flexibility Meeting Data Centers Where They Are Today. *JetCool Technologies, Inc. White Papers*, 1.

57. Meta. (2022). OCP Summit 2022: Grand Teton. Facebook Engineering. [Link]

58. Page Southerland Page, Inc. (2022). White Paper: Two-Phase Liquid Immersion Cooling Case Study. Page Southerland Page White Papers, 1.

59. SkyCool Systems. (2019). Harnessing the Cold of the Sky and Space to Enable Electricity-free Cooling. SkyCool Systems: Radiative Cooling Technology Case Study. [Link]

60. Submer & Wyoming Hyperscale. (2022). Immersion Cooling for Hyperscalers: Powering Farming of the Future. Submer White Papers & Wyoming Hyperscale White Box, 1. [Link]

61. Texas Advanced Computing Center & Green Revolution Cooling. (2023). Advanced Cooling Advances Science Case Study. *TACC Case Studies*, 1.

62. Ham, S.W., Kim, M.H., Choi, B.N., & Jeong, J.W. (2015). Energy saving potential of various air-side economizers in a modular data center. *Applied Energy*, 135(15), 258-275.

63. Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. *Computer*, 40(12).

64. U.S. Environmental Protection Agency. (2022). Annual Emissions Report by Industry Sector. U.S. Environmental Protection Agency. [Link]

65. Apple. 2022 Environmental Progress Report. Apple Sustainability Reports, 2022(1). [Link]

66. Google. (2022). Environmental Report. Google Sustainability Reports, 2022(1). [Link]

67. Meta. 2022 Sustainability Report: For a Better Reality. Meta Sustainability Reports, 2022(1). [Link]

68. Microsoft. 2022 Environmental Sustainability Report. Microsoft Sustainability Reports, 2022(1). [Link]

69. Larson, A. (2023). NuScale Gets a Win with SMRs for Data Centers in Ohio and Pennsylvania. PowerMag Publishing. [Link]

70. Chhabra, S., & Singh, A. (2023). Dynamic Resource Allocation Method for Load Balance Scheduling over Data Center Networks. *Journal of Web Engineering*, 2211(02352).

71. Ghatikar, G., et al. (2012). Demand Response Opportunities and Enabling Technologies for Data Centers: Findings from Field Studies. *Lawrence Berkeley National Laboratory (LBNL)*.

72. Mehra, V., & Hasegawa, R. (2023). Supporting power grids with demand response at Google Data Centers. Google Cloud Products. [Link]

73. Abhyankar, N., et al. (2021). 2030 Report: Powering America's Clean Economy. *University of California-Berkeley: Goldman School of Public Policy*, 2021 (1).

74. Horner, N., Shehabi, A., & Azevedo, I. (2016). Known unknowns: Indirect energy effects of information and communication technology. *Environmental Research Letters*, 11(10).

75. Raizada, A., & Singh, K. (2020). Worldwide energy consumption of hyperscale data centers: a Survey. *International Research Journal on Advanced Science Hub*, 02(11).

76. U.S. Environmental Protection Agency. (2018). Quantifying the Multiple Benefits of Energy Efficiency and Renewable Energy. U.S. Environmental Protection Agency. [Link]

77. Howard, A. (2022). Data Center Building Report – 2022. OMDIA: Annual Reports, 2(1). [Link]

78. Data Center Frontier. (2023). The Power Problem:

Transmission Issues Slow Data Center Growth. Data Center Frontier: Special Reports. [Link]

79. Masanet, E., et al. (2020). Recalibrating global data center energy-use estimates. *Science (AAAS)*, 367(6481), 984-986.

80. Mytton, D., & Ashtine, M. (2022). Sources of data center energy estimates: A comprehensive review. *Joule*, 6(1), 2032–2056.

81. Movva, R., Lei, J., Longpre, S., Gupta, A., & DuBois, C. (2022). Combining Compressions for Multiplicative Size Scaling on Natural Language Tasks. *Proceedings of the 29th International Conference on Computational Linguistics*.

82. Sevilla, J., et al. (2022). Compute Trends Across Three Eras of Machine Learning. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1(8).

83. Bastian, M. (2023). GPT-4 Has More Than a Trillion Parameters – Report. The Decoder Reports, 1. [Link]

84. BigScience. (2022). Introducing the World's Largest Open-source Multilingual Language Model: BLOOM. BigScience Analysis, 1. [Link]

85. DeepChecks. (2023). LLM Models Comparison: GPT-4, Bard, LLaMA, Flan-UL2, BLOOM. DeepChecks Analysis, 1. [Link]

86. Hoffman, J., et al. (2023). Training Compute-Optimal Large Language Models. arXiv Preprint, arXiv:2203.15556.

87. Rae, J., Irving, G., & Weidinger, L. (2021). Language Modelling at Scale: Gopher, Ethical Considerations, and Retrieval. Google DeepMind Research, 1.

88. Zhang, S., et al. (2023). OPT: Open Pre-trained Transformer Language Models. arXiv Preprint, arXiv:2205.01068

About EPRI

Founded in 1972, EPRI is the world's preeminent independent, non-profit energy research and development organization, with offices around the world. EPRI's trusted experts collaborate with more than 450 companies in 45 countries, driving innovation to ensure the public has clean, safe, reliable, affordable, and equitable access to electricity across the globe. Together, we are shaping the future of energy.

PRINCIPAL INVESTIGATORS

JORDAN ALJBOUR, *Strategic Insight Engineer*
jaljbour@epri.com

TOM WILSON, *Executive Technical Leader*
twilson@epri.com

PRINCIPAL INVESTIGATOR AND EPRI CONTACT

POORVI PATEL, *Manager Strategic Insight*
704.232.4551, ppatel@epri.com

For more information, contact:

EPRI Customer Assistance Center
800.313.3774 • askepri@epri.com

3002028905 May 2024

EPRI
3420 Hillview Avenue, Palo Alto, California 94304-1338 USA • 650.855.2121 • www.epri.com

© 2024 Electric Power Research Institute (EPRI), Inc. All rights reserved. Electric Power Research Institute, EPRI, and TOGETHER…SHAPING THE FUTURE OF ENERGY are registered marks of the Electric Power Research Institute, Inc. in the U.S. and worldwide.