S&P Global

AI and energy: The big picture

AI and cloud services growth is driving up US electricity needs, creating a supply-demand imbalance that will require innovations in grid technology, a balance between datacenter expansion and power infrastructure, and cross-sector coordination to mitigate environmental impacts.

Ben Levitt
Research and Analysis Associate Director, Gas, Power, and Energy Futures, S&P Global Commodity Insights
ben.levitt@spglobal.com

Converging trends are challenging US power infrastructure

Rapid expansion of datacenters to meet demand for cloud and AI services is one of several converging trends that will strain US power sector infrastructure in the coming years. After a decade and a half of stagnation, US electricity demand is set for significant near-term growth. This surge will be driven not only by datacenters but also by the development of new manufacturing and industrial facilities; the activities of cryptocurrency mining operations; the electrification of vehicles, buildings and industrial processes; and the increased need for heating and cooling due to extreme temperatures. This growth in electricity demand coincides with other challenges that threaten the adequacy of US power sector infrastructure:

- An increasingly challenging development environment for new power generation and transmission infrastructure
- A swift transition toward intermittent and weather-dependent resources: Over the next 10 years, the share of US generation from wind and solar is expected to rise from 15% to 40%
- Proposed regulations that could further constrain development and use of coal, oil and gas-fired resources
- A growing focus on reshoring clean energy supply chains and implementation of trade barriers that will likely increase the costs of new power supply resources

Highlights

Rapid expansion of datacenters to meet growing demand for cloud and AI services is one of several converging trends that will strain the US power sector's infrastructure in the coming years.

Much attention to date has centered solely on datacenter electricity demand. However, this misses the bigger picture. As datacenter capacity underpins a larger share of economic activity, it will contribute to reshaping historical patterns of electricity consumption throughout the broader economy. Early evidence indicates these trends may offset some of the rise in direct electricity consumption from datacenters.

In the US, near-term outlooks for gas- and coal-fired electricity generation are higher as demand growth driven by AI and cloud services outpaces the development of new power supply and transmission infrastructure. New AI tools for the power sector may help, but they will take time to implement and may not fully resolve issues of power grid adequacy and rising emissions from AI workloads.

Efforts to expand power sector infrastructure development will likely remain a focus of US federal government action as maintaining and growing the country's global lead in AI depends on the adequacy of power infrastructure.

AI-related datacenter development timelines are misaligned with the power system

In the US, electricity demand growth driven by AI and cloud services is outpacing the development of new power supply and transmission infrastructure. While it typically takes two to three years to design, permit and build a datacenter, development timelines for power generation resources often extend to five years or more.

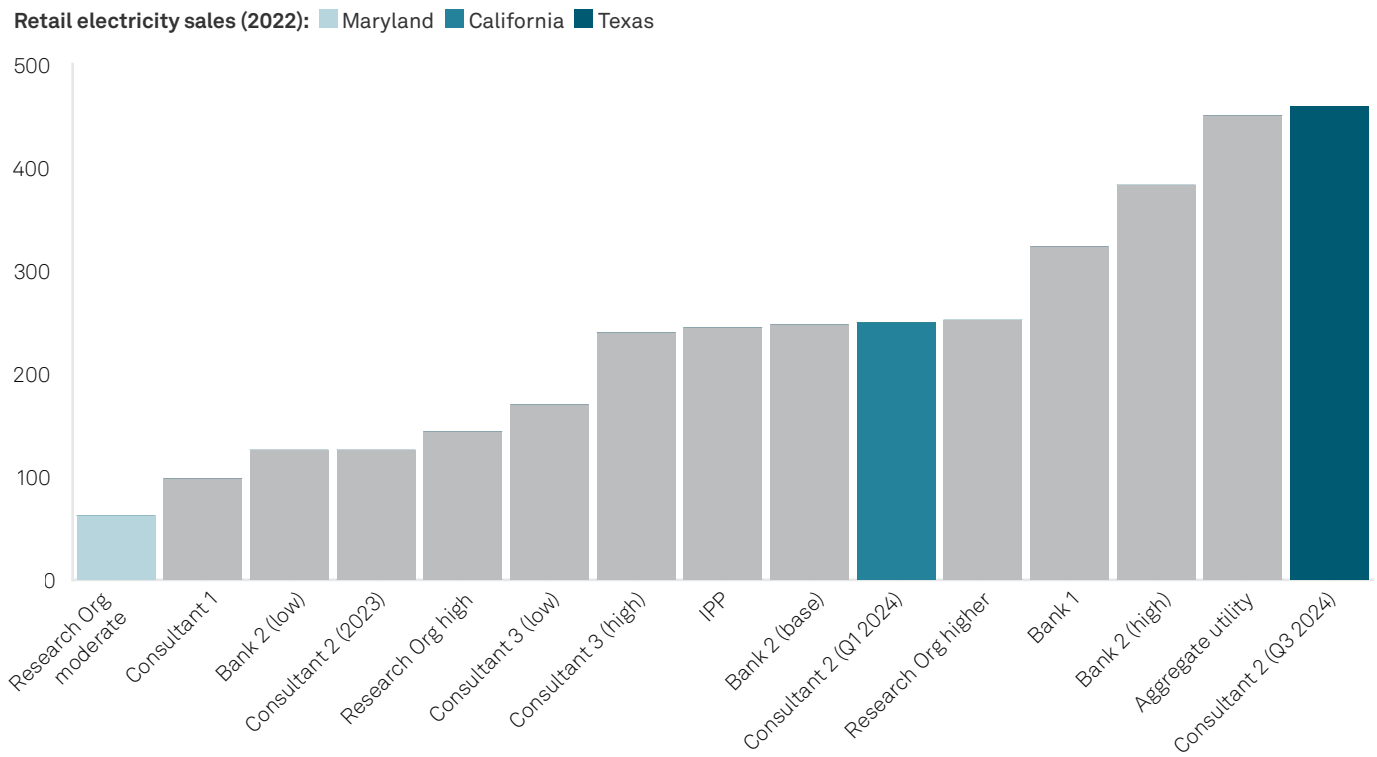
Moreover, utilities have had minimal time to plan for the expected surge in demand. For instance, OpenAI released ChatGPT 3.5 to the public in November 2022, but utilities only began broadly assessing generative AI's impacts on grid demand **toward the end of 2023**. Without adequate lead time to obtain project approvals, new utility infrastructure will reflect a regulatory lag.

As a result, supporting rapid, near-term growth in electricity demand will require squeezing more from the existing power generation fleet and delivery infrastructure. Regions with greater load growth will scramble for capacity in the near term and rely on delayed fossil plant retirement. The remaining fossil generation fleet will run harder. And utilities will look to "grid enhancing technologies" such as dynamic line ratings, which often use machine learning, to increase utilization of the existing grid. Utilities in Texas, New York, Ohio and Indiana have already begun using these technologies.

Uncertainty clouds datacenter load forecasts and implications for electricity consumption elsewhere in the economy

While new datacenters present significant growth opportunities for utilities, uncertainty over the scale, timing and location of new datacenters brings new risks. Opinions vary on how far and how fast datacenter power consumption will grow. A survey of industry stakeholders suggests that datacenter-driven demand growth through 2030, expressed in terms of equivalent state-level electricity demand, could range from "Maryland" to "Texas."

Estimates for new US datacenter demand for 2023–2030 (TWh)



Data compiled June 20, 2024.
IPP = independent power producer.
Estimates are drawn from public documents released by companies in the information, consulting and financial services sectors.
Sources: S&P Global Commodity Insights; public documents.

On the one hand, there is no question that datacenters will drive electricity demand growth. The US has 5-6 GW of datacenter capacity under construction, which will expand the existing fleet by roughly a quarter. Meanwhile, vacancy rates, especially in primary markets, have fallen over the past five years.

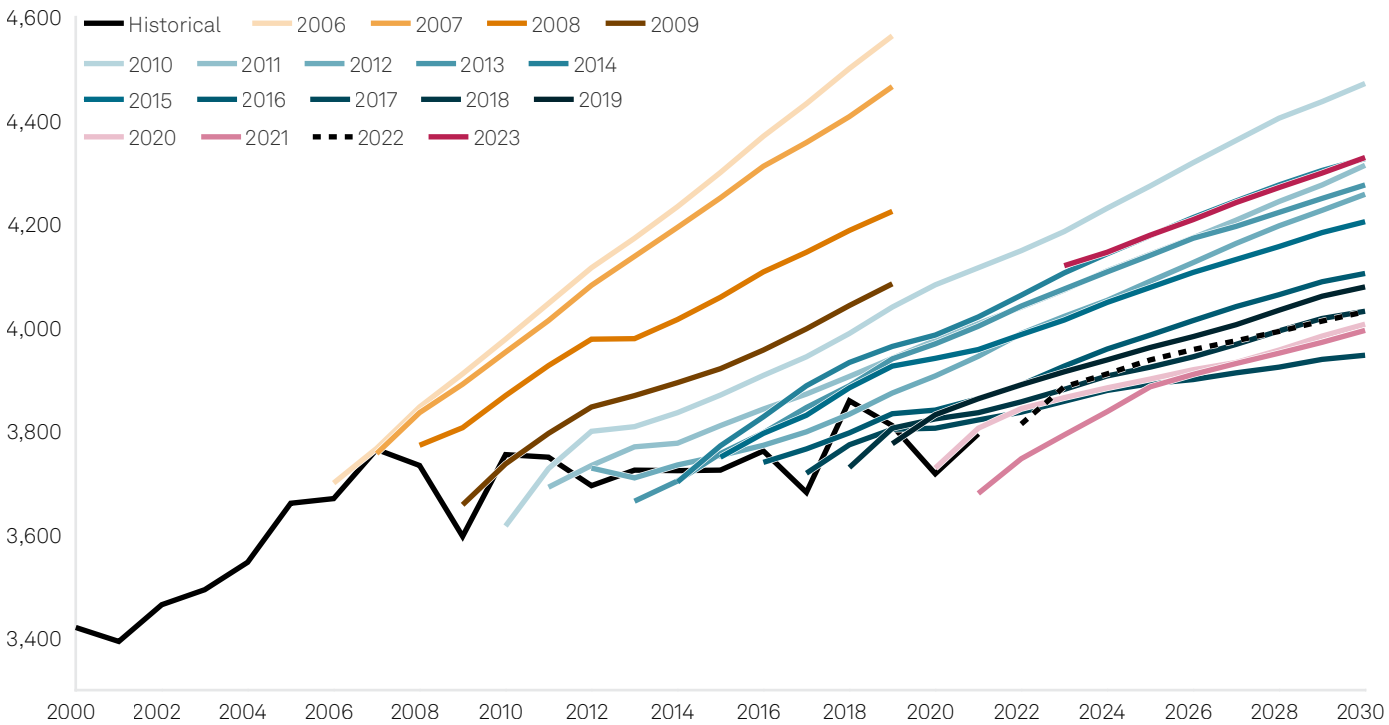
On the other hand, the timing and scale of new datacenter-driven electricity demand will depend on several factors and emerging trends:

- Evolution of AI demand and commercial prospects for AI technologies
- Efforts to improve grid connection bottlenecks
- Sufficiency of skilled labor in regions where development is planned
- Evolution of hardware and AI computational and algorithmic energy efficiency
- Adequacy of grid infrastructure and datacenter hardware supply chains
- Datacenters' use of behind-the-meter generation (electricity generated, stored and consumed directly by datacenters)
- Incrementality of new AI workloads — i.e., adding new workloads versus displacing others
- Capability of devices to perform AI tasks without data transmission to and from a datacenter

Skepticism over electricity demand forecasts is warranted. For decades, utilities and grid operators have over-forecast load as the industry has underestimated the impacts of energy efficiency improvements and shifts in industrial activity that pushed power demand below expectations.

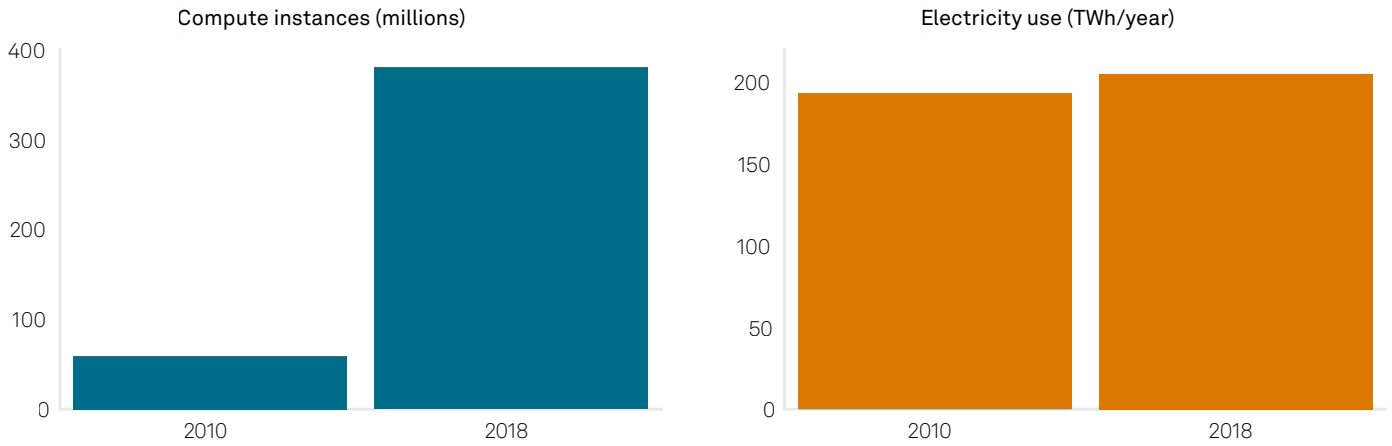
Similar concerns about the IT sector's appetite for electricity emerged during the rise of internet services. In 2018, researchers found that demand for internet services rose much faster than associated electricity consumption.

US total electricity sales projections (TWh)



Data compiled April 2024.
Data from the US Energy Information Administration's Annual Energy Outlook Retrospective Review.
Sources: IHS Markit; US Energy Information Administration.

Global datacenter statistics

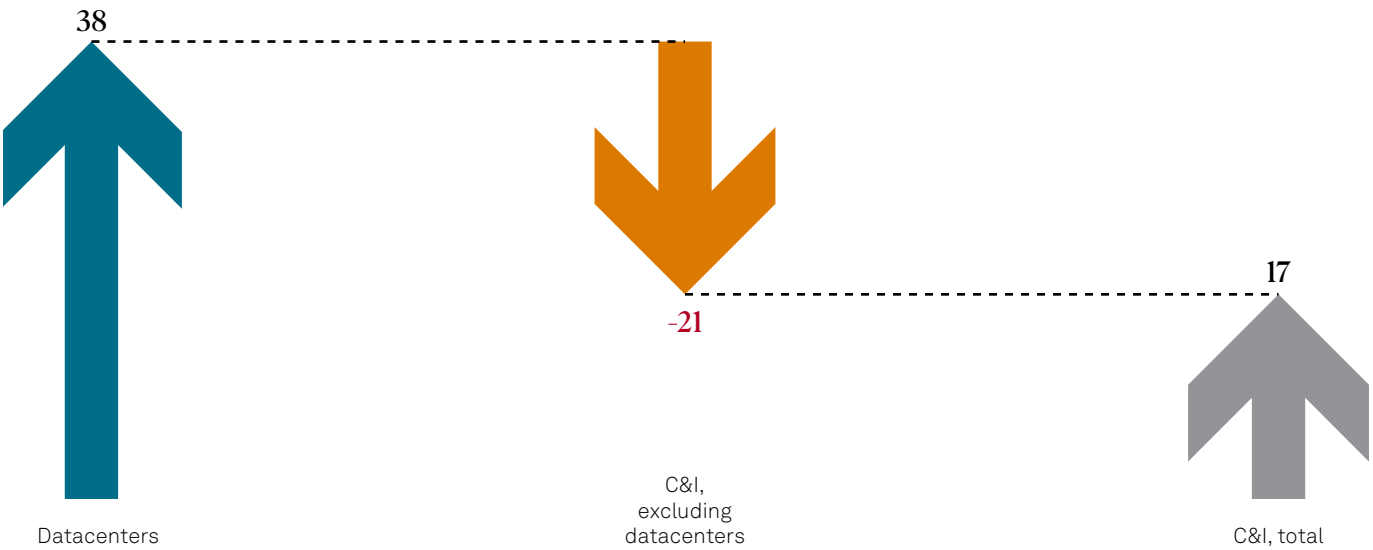


Data compiled May 22, 2024.
A compute instance is defined as a virtual machine with its own set of resources (CPU, RAM, storage) running on physical hardware.
Datacenter electricity demand includes electricity consumed by traditional, hyperscale and cloud datacenters.
Sources: S&P Global Commodity Insights; Eric Masanet et al., Recalibrating global data center energy-use estimates, Science, February 28, 2020.

While US technology executives expect significant growth in datacenter-driven electricity demand, they also signal that AI's energy demands may be overestimated. Executives from Google and NVIDIA have suggested that efficiency gains in AI will surpass expectations as energy-efficient hardware and optimized models continue to develop. (Of course, such statements should be taken with a grain of salt, as the big tech players continue to release increasingly dense and power-hungry computing infrastructure.) Some have also indicated that certain forecasts double-count projects, while others point to the likelihood that US datacenter capacity will be overbuilt, potentially leading to overestimation of the associated power demand.

Further, as datacenters underpin an increasing share of global economic activity, they are reshaping electricity consumption in other economic subsectors. Datacenters are the key infrastructure supporting a large and growing portion of the economy (i.e., the "digital economy"). The US digital economy has seen robust growth for decades and is now one of the largest economic subsectors. The digital economy is reshaping how we shop, work, and spend our time and money, which affects patterns of electricity consumption. From 2018 to 2023, overall retail sales of electricity to US commercial and industrial customers, including datacenters, only rose by about 1 kWh for every 2 kWh of increased datacenter sales, indicating reductions in electricity consumption for other commercial and industrial categories.

Change in US retail electricity sales, 2018–2023, by component (TWh)



Data compiled May 2024.
C&I = commercial and industrial.
Sources: S&P Global Market Intelligence 451 Research; US Energy Information Administration.

Power constraints are encouraging innovation

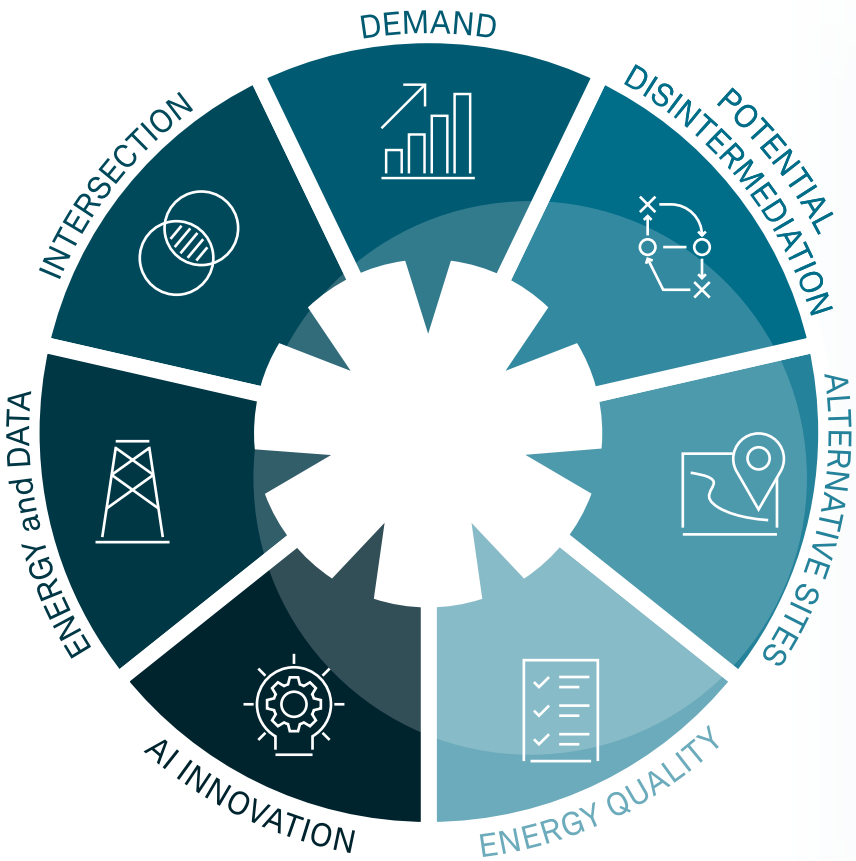
Timelines for new datacenter-driven demand may be longer than anticipated. Already, the need for incremental grid infrastructure has delayed datacenter development timelines in key markets such as Northern Virginia and Silicon Valley. As spare grid capacity dwindles, further delays are anticipated, making regional grids a key bottleneck for datacenter development. Queues for large loads have emerged in Texas and Ohio. American Electric Power Co. (AEP), one of the largest utilities in the country, with customers across 11 states, recently informed investors that no spare capacity is available to support new datacenter requests in key subregions such as central Ohio.

In pursuit of utility power, adaptation is also manifesting in datacenter operators' site selections. For example, xAI plans to build a large AI datacenter in Memphis, while Amazon Web Services is planning two hyperscale AI facilities in Mississippi — markets that have yet to see any datacenter development at scale.

Some are seeking to minimize their reliance on the grid through demand response and flexibility, while others are looking to avoid reliance on the grid altogether by colocating with an existing power supply resource. Owners of nuclear plants in competitive markets could strike deals with datacenters like the one made between Talen and Amazon Web Services in early 2024. Others have envisioned more dramatic solutions such as the development of new, private utilities and grids that can develop faster to match technology-sector demand.

While Google, Amazon and Oracle have all made recent headlines with their plans to purchase electricity supplied by new small modular reactors (SMRs) and enhanced geothermal systems, construction will take years and involve higher levels of financial and technological risks.

AI demand growth, datacenters, energy and real estate: Key considerations



Demand
AI demand will increase datacenters' role as ratepayers to energy providers. Considering the magnitude of demand, such expansion will require a delicate balance of changing supply/demand dynamics, regulatory considerations and balanced load operations to ensure rate equity and load accessibility across an expanding customer base.

Potential disintermediation
Increasing "behind-the-grid" energy sources such as small modular reactors (SMRs) will likely curb some energy shortages but may increase regulatory complexity over time.

Alternative sites
Datacenters may replace aging alternative sites with access to high-power energy sources (such as retired smelting facilities).

Energy quality
Alternative energy sources such as renewables are generally inadequate to meet the energy demands of AI datacenters, at least in the near term, likely increasing fossil fuel demand in the absence of nuclear and other power options.

AI innovation
Increased algorithm improvement, agentic AI, decoupled models training, inferencing and execution standards may reduce power consumption somewhat.

Energy and data
Datacenter real estate sites will increasingly prioritize proximity to power generation sources, distribution lines and increased data-throughput resources.

Intersection
Demand for AI will have knock-on effects on datacenters, technology, real estate, power generation, power distribution and other sectors.

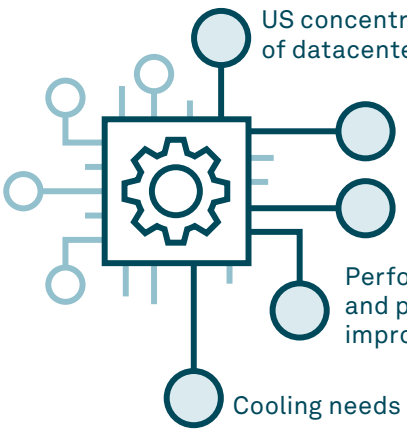
Data compiled November 2024.

Power availability is one of many factors influencing datacenter distribution

Datacenter development will remain "hyperlocalized" in regional clusters in Northern Virginia, Phoenix, Atlanta, Dallas and Columbus, Ohio. However, even in these regions, grid constraints or local pushback may counteract ambitions. Regional distribution will be shaped not only by power availability and lead times but also by connectivity and fiber, latency needs and customer proximity, state and local development policies and tax incentives, water resource availability, skilled labor availability, cost of land and electricity, and natural disaster risk.

State policies, in particular, present a shifting landscape. In Maryland, Aligned Data Centers canceled a major development project in 2023 after the state's Public Service Commission rejected its request for an exemption to install 168 diesel-fueled backup generators. In mid-2024, the state passed the Critical Infrastructure Streamlining Act to ease future approval of such projects. In Georgia, a bill that would have suspended tax abatements for new datacenters passed through the legislature, but the governor vetoed the measure in May 2024 to continue promoting datacenter development. In the same month, Tennessee passed legislation expanding datacenter tax breaks.

5 key factors influencing US power demand resulting from AI



US concentration of datacenters
Determining the portion of global datacenters that will reside in the US will be critical to understand the inflection between the current power grid's capacity and AI's potential to surge demand.

Parasitic loads
Power requirements for parasitic loads (e.g., when a generator powers devices that do not contribute to the net electric yield).

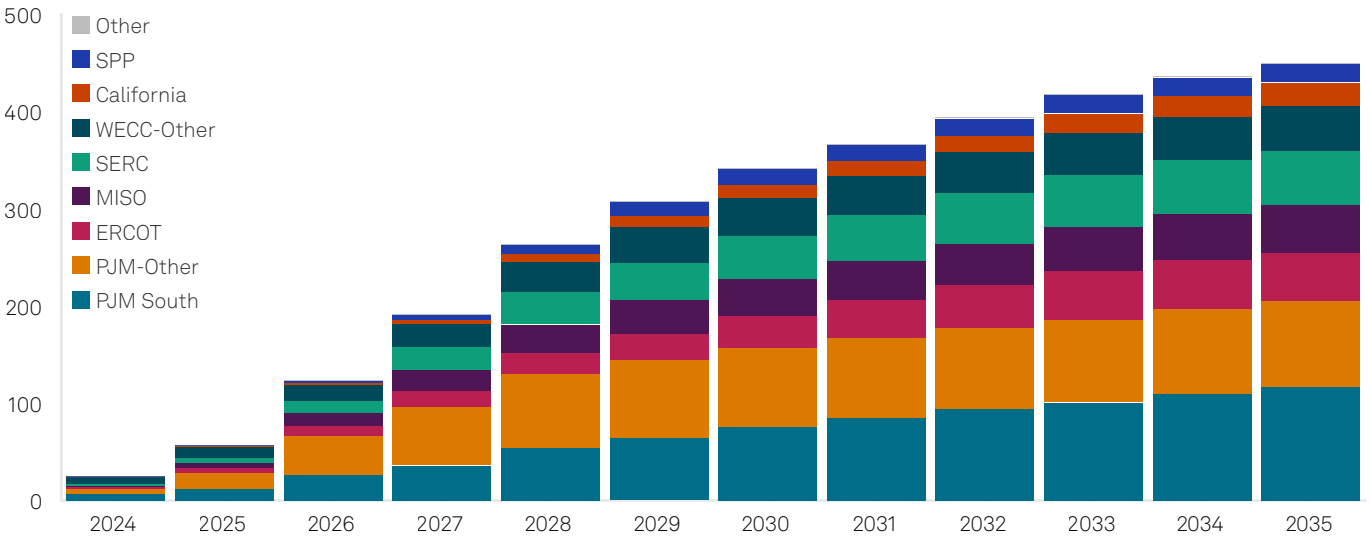
Number of queries per year
The number of training and inference workloads and their annual growth.

Performance and process improvements
General/organic performance and power improvements through activities such as operating improvements, process efficiencies and other innovation; specifically, improvements in compute speed and power densities.

Cooling needs
Power requirements for cooling as a function of total power demand will be significant considering AI's current power draw; innovations such as inferencing solutions and agent-based approaches may mitigate this concern but remain some way off.

Data compiled November 2024.

Utility outook for cumulative new electricity demand from datacenters by power market (TWh)



As of August 2024.
Data compiled Aug. 6, 2024.
PJM-Other includes PJM West and PJM Mid-Atlantic. Other includes NYISO, ISO-NE and FRCC. WECC-Other includes Desert Southwest, Northwest Power Pool and Rockies.
Source: S&P Global Commodity Insights.

Datacenters increase emissions in the near term, but longer-term implications are more uncertain

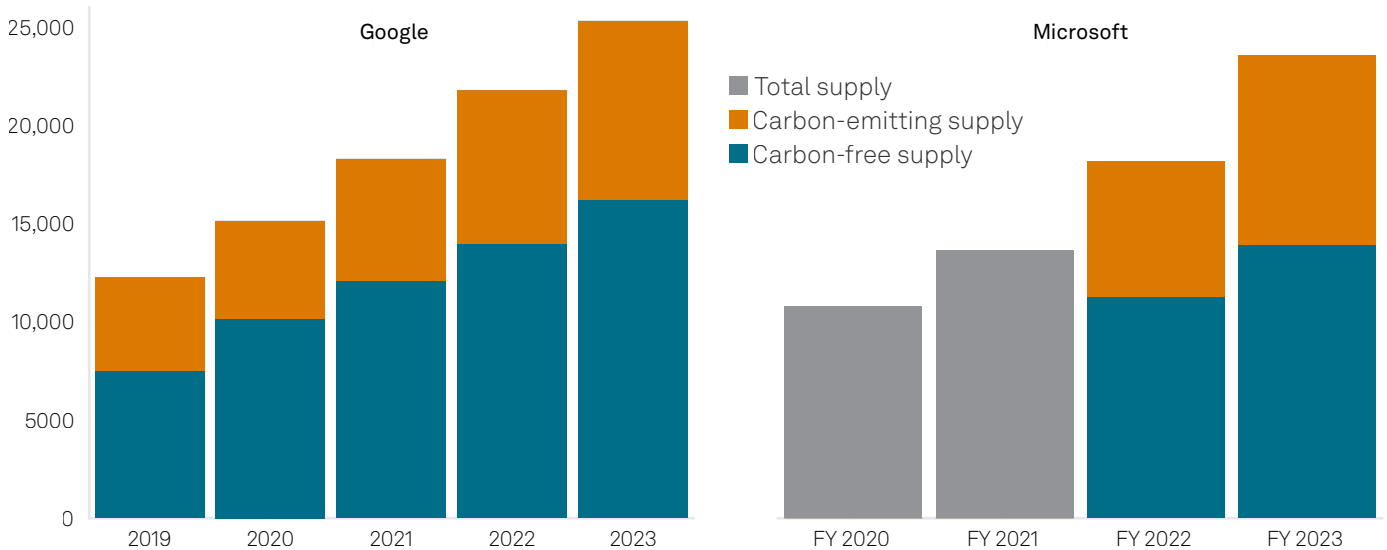
S&P Global Commodity Insights continues to expect that new datacenter-driven demand will increase the near-term outlook for gas- and coal-fired electricity generation, as technology companies appear to be prioritizing competitive pressures over environmental goals. With demand growing faster than new supply, the power sector must lean more heavily on existing fossil generation resources to supply incremental demand. Data from Google and Microsoft reveals that rising datacenter demand has already increased power supply from carbon-emitting resources.

Over time, in addition to uncertainty regarding datacenter electricity demand, power sector fossil generation demand will depend on the pace of adding renewables and other clean energy resources. Most companies driving the current datacenter build-out are directing major investments toward clean energy, accounting for about one-half of all US corporate renewables procurement. At the same time, challenges with building transmission, interconnection backlogs, local opposition to development and permitting delays continue to threaten the pace of renewable additions. Residual demand not met by renewables will primarily be supplied by gas-fired generation. It is not surprising, then, that technology companies have recently expanded investments into clean energy technologies such as advanced geothermal and SMRs to avoid longer-term reliance on fossil generation.

Overall, S&P Global Commodity Insights does not expect that datacenters will produce a significant and sustained increase in total US natural gas demand. By 2040, if all 450 TWh of new datacenter-driven electricity demand (high-end forecast) were incremental and supplied entirely by gas-fired generation, the outlook for total demand for US natural gas would increase by 9 Bcf/d, or 8%, according to data from S&P Global Commodity Insights. However, S&P Global Commodity Insights expects clean energy to supply a large share of incremental datacenter demand over the longer term. If one-quarter of the projected datacenter load were supplied by gas-fired generation, this would translate to a 2% increase in total US gas demand in 2040.

New AI tools for the power sector may help but will not fully resolve issues of power grid adequacy and rising emissions from AI workloads. Initial applications of AI for the power sector do not provide the step-change improvements needed to meet the scale of new power demands. For example, applications such as accelerating security-constrained unit commitment calculations, improving forecasts for electric vehicle charging and nondispatchable power supply technologies, and enhancing vegetation management practices are useful but only offer incremental improvements. Moreover, the electricity required to train and run such models would partially offset the efficiency gains. Ambitions to use AI to "operate" the grid may hold greater potential but will require more time to develop and may introduce new security risks. For instance, using AI in power system operations introduces an additional attack vector for breaches into critical systems, such as through training data "poisoning" that could cause models to learn incorrect behaviors.

Global electricity consumption by supply type (GWh)



Data compiled June 18, 2024.
About 70% of Google's total electricity consumption occurs in the US.
Source: S&P Global Commodity Insights.

Looking forward: Tricky balancing act to support US technological superiority without harming consumers

US leadership in AI is bolstered by one of the largest fleets of datacenters globally. According to data from S&P Global Market Intelligence 451 Research's Datacenter KnowledgeBase, the US hosts approximately 38% of the world's operational datacenter capacity, providing the computational resources necessary to manage complex AI algorithms and large datasets. Though the US is poised to maintain its lead in AI in the coming years, this depends on meeting datacenters' increasing power demands.

In support of this goal, efforts to expand power infrastructure will likely remain a focus of federal action. The US government has announced plans — and in some cases has already acted — to bolster grid development through Federal Energy Regulatory Commission orders targeting transmission planning, cost allocation and interconnection queue backlogs, in addition to directives and financial support from the US Energy Department. However, overbuilding to meet new AI-driven electricity demand could impose financial burdens on utilities and consumers. Striking a balance between meeting demand and avoiding excess capacity will be a crucial challenge for the US power sector in the coming years.

This article was authored by a cross-section of representatives from S&P Global and, in certain circumstances, external guest authors. The views expressed are those of the authors and do not necessarily reflect the views or positions of any entities they represent and are not necessarily reflected in the products and services those entities offer. This research is a publication of S&P Global and does not comment on current or future credit ratings or credit rating methodologies.

Contributors

Dan Thompson
S&P Global Market Intelligence
Principal Research Analyst, 451 Research

Matt Tompkins
S&P Global
Senior Editor

Aneesh Prabhu
Senior Director, Ratings Analytical

Cat VanVliet
S&P Global
Associate Director, Data Visualization