Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

Updated November 16, 2026

Data Centers and Cloud Computing: Information Technology Infrastructure for Artificial Intelligence

Artificial intelligence (AI) systems, including their development, deployment, and operation, rely on an information technology (IT) infrastructure with components of hardware, software, networks, data, and hosting facilities (these components are collectively known as *AI compute*). Data centers, physical facilities that house and run large computer systems, are the primary means to store much of this AI compute. Cloud computing, an internet-based IT service model, enables AI developers and users to remotely access AI models and tools as well as computing resources hosted in data centers. Some AI researchers have suggested that the performance of certain AI models (e.g., large language models) improves with the scaling of three factors: the number of model parameters, the size of the training datasets, and the amount of computing resources used for training. Collectively known as AI scaling laws, these patterns drive the current paradigm of AI development: More capable AI systems require larger models, more data, and more computing resources. Researchers assert that these scaling requirements have led to "unprecedented" capital investment in AI infrastructure, the magnitude and speed of which "have surpassed previous investment booms" in U.S. history. The rush to build AI data centers has raised several public concerns, including the energy and water consumption needed for data center operations; the associated potential impacts on power grids, utility bills, and environment effects; and a lack of disclosure and public involvement in the decisionmaking process. Several cloud service providers (CSPs) leverage loans to support AI infrastructure build-out, and some experts assert that this debt is contributing to increasing financial risks in stock and lending markets.

This In Focus introduces concepts of data centers and cloud computing and policy considerations for Congress. It focuses on the roles of data centers and cloud computing as AI infrastructure. For data center-related energy and water issues, see CRS Report R48646, *Data Centers and Their Energy Consumption: Frequently Asked Questions*; CRS Report R49326, *Data Centers and the Electricity Grid: Frequently Asked Questions*; and CRS Report R49057, *Data Centers and Water: Frequently Asked Questions*.

From General IT Data Centers to AI Data Centers

The term *data center* has been defined in federal laws. For instance, in the Energy Independence and Security Act of 2007, Congress defined it as any physical facility that "primarily contains [IT] equipment used to process, store, and transmit digital information" (P.L. 110-140 §453).

Since the early 2000s, data centers have served as an important IT resource for the operation of organizations, for example, in telecommunications, financial services, health care, and government sectors. An increasing number of private and public entities use data centers to support their expanding IT needs, particularly as data volumes continue to grow. The United States reportedly has over 5,000 data centers in operation. Many of them serve general IT functions. General IT or traditional data centers connect consumers with access to online services and content such as websites, emails, file sharing, and video streaming. Many CSPs build and operate geographically distributed but interconnected general IT data centers to allow their users to remotely access a shared pool of computing resources and services, such as data processing chips, data storage devices, databases, software, and networks, as well as web hosting, data analytics, and application deployment.

AI data centers are different from general IT data centers in workload type, primary hardware, and power demand. General IT data centers performing routine, sequential computing tasks such as hosting and managing enterprise and consumer data typically use general-purpose central processing units (CPUs). Since the current AI development cycle started in late 2022, the demand for high-performance, specialized computing chips, such as graphics processing units (GPUs) and high-bandwidth memory (HBM), has grown rapidly in order to handle large-scale AI workloads. These chips, along with datasets, algorithms, and software required for AI, are largely housed in data centers often managed by CSPs. A data center primarily used for AI model training and inference may require tens of thousands of GPUs or other specialized chips designed for complex computational tasks.

Leading AI developers and some industry analysts have argued that existing data center capacities cannot meet the AI demand due to intensive computational tasks in model training and surging inference requests (i.e., running a trained model) as a result of the increasing AI adoption, use, and proliferation of agentic AI. Billions of dollars in private investment have been reportedly announced to fund the construction or expansion of over 1,000 data centers throughout the country. This data center surge is likely driven by two factors: AI scaling laws and the associated expectation of a lasting computing shortage and the view that controlling AI infrastructure is a strategic competitive advantage and key to market dominance.

Defining AI Data Centers

An advanced GPU designed for AI data centers can draw up to 1,000 watts of electricity when operating at its full capacity. The electrical energy entering the chip is converted into a matching amount of thermal energy (i.e., heat). A data center's cooling system (either air or liquid based) must use additional power to dissipate the heat in order to keep the housed chips from overheating. By comparison, a typical data-center CPU draws about one-

fourth of that power (and therefore requires less cooling). The sheer number of GPUs and supporting power distribution and cooling infrastructure require more physical space in an AI data center than in a traditional one.

Because chips in an AI data center require massive amounts of power to run, the facility's maximum computing capacity is directly bound by its available power supply. In turn, how much electric power the data center can draw at its peak workload—measured in megawatts (MW; 1 MW = 1 million watts) or gigawatts (GW; 1 GW = 1,000 MW)—has become the primary metric when referring to its scale. For example, according to some industry analysts, to be considered a "hyperscale data center," the facility should be able to draw over 100 MW of power at any moment, house at least 5,000 computer servers, and occupy at least 10,000 square feet. A data center running at a continuous demand of 100 MW for 24 hours would consume 2,400 MW-hours (MWh) of energy. That single-day electricity consumption is roughly equal to the daily use of 80,000 average U.S. households.

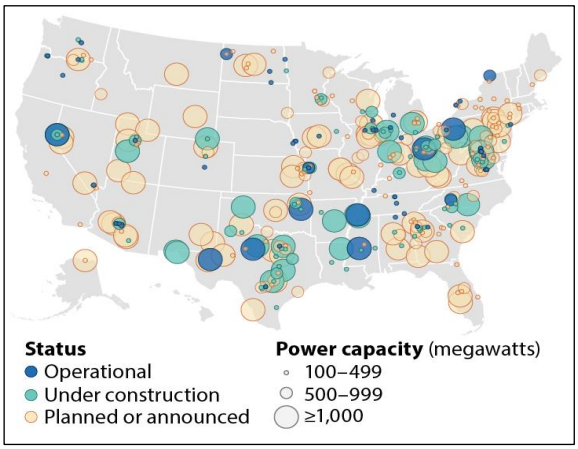
Lawmakers have introduced several bills addressing various public concerns associated with AI data centers, using power demand thresholds to define covered facilities. These power thresholds vary, suggesting that it may be a legislative challenge to differentiate between traditional and AI data centers. Some bills introduced in the 119th Congress provided definitions of AI data centers using power demand thresholds between 20 MW and 100 MW along with other characteristics, such as the proportion of computing resources used for AI purposes. For example, the Artificial Intelligence Data Center Moratorium Act (H.R. 9442/S. 4214) set the threshold as "a maximum rated power capacity or total peak power load in excess of 20 [MW]." In the Artificial Intelligence Environmental Impacts Act of 2026 (H.R. 9183/S. 4727), a covered data center "has a maximum rated power capacity or total peak power load in excess of 50 [MW]." The RECLAIM Act (S. 5689) set the threshold as "a contracted peak power demand of not less than 100 [MW]."

Policy Considerations for Congress

Several dynamics define the current landscape of AI data centers and associated cloud computing services. First, the investment and build-out, if fully realized, could happen at a massive scale. **Figure 1** illustrates data centers over 100 MW—either operational, under construction, or planned—across the country. Some data center projects have surpassed 1 GW, and the five largest CSPs were expected to commit approximately $4.6 trillion in AI capital expenditure between 2022 and 2030.

Second, data centers and cloud computing are becoming strategically important infrastructure for AI. They were recently at the center of the global AI race debate following the 2026 Nvidia controversy over alleged export-control workarounds. After data centers in third-party countries acquired large numbers of advanced GPUs, AI developers located in countries subject to U.S. export controls were reportedly able to access those computing resources remotely through CSPs located there. This widely reported loophole raises important questions related to national and economic security.

Figure 1. U.S. Locations of Data Center Projects Over 100 Megawatts



Source: CRS, based on data downloaded from Jillian Bommarito and Michael James Bommarito, "Data Center Policy Across the 50 States: A Survey of Incentives, Land Use, and Energy Regulation," June 18, 2026, Server Country.

Notes: Server Country uses primary sources such as company announcements, regulatory filings, and news reports to identify data center projects. This information should be considered illustrative rather than comprehensive. The operational category includes existing data centers with planned expansions.

Lastly, there is public opposition to AI data centers. AI infrastructure development and operation may have various impacts on natural resources and the economies of local communities near data centers. The land use, permits, and construction of data centers are largely subject to state and local regulation, though some federal requirements may apply. Federal and state authorities and local communities all potentially influence AI data center development. Leading AI developers and CSPs may also align their data center projects with public interests and policy goals.

AI infrastructure, including data centers and cloud computing services, is vital for AI development and operation. The criticality of such infrastructure raises policy issues such as infrastructure access and security, energy reliability and efficiency, and costs of building such infrastructure, all of which implicate national security interests. These issues and the current policy landscape present some legislative considerations. For example, Congress may consider how to facilitate enhancing AI compute to support innovation and maintain U.S. leadership in AI. Congress may consider legislation to ensure that data center construction, expansion, or operation aligns with public interests, such as community benefits, disclosure of natural resource consumption and environmental impacts, and regional and national economic development. Congress may also consider leveraging data centers to pace AI development, such as through legislation that pauses AI data center construction until other laws are enacted to regulate the release of AI products, address economic and workforce impacts of AI, and ensure AI data centers meet specified criteria (as proposed in H.R. 9442/S. 4214).

Ling Zhu, Specialist in Science and Technology Policy

IF12899

Disclaimer

This document was prepared by the Congressional Research Service (CRS). CRS serves as nonpartisan shared staff to congressional committees and Members of Congress. It operates solely at the behest of and under the direction of Congress. Information in a CRS Report should not be relied upon for purposes other than public understanding of information that has been provided by CRS to Members of Congress in connection with CRS's institutional role. CRS Reports, as a work of the United States Government, are not subject to copyright protection in the United States. Any CRS Report may be reproduced and distributed in its entirety without permission from CRS. However, as a CRS Report may include copyrighted images or material from a third party, you may need to obtain the permission of the copyright holder if you wish to copy or otherwise use copyrighted material.