Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

February 5, 2025

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

The advancement of artificial intelligence (AI), a "critical and emerging technology," presents policy considerations for U.S. leadership, with implications for economic competitiveness and national security. AI systems, including their development, model training, deployment, operation, applications, and services, rely on an information technology (IT) infrastructure with components of hardware, software, networks, data, and facilities. Data centers are the primary means to house much of this IT backbone. Internet-based remote computing services (i.e., *cloud computing*) enable AI developers and users to access computing resources hosted in geographically distributed data centers. Not only may AI innovation and competition hinge on the availability of and access to advanced, secure, and sustainable computing resources, but such IT infrastructure may also be deemed "a strategic national asset." Related issues have attracted congressional attention in recent years (e.g., a Senate committee hearing on advanced computing research and a House committee hearing on "powering AI").

The U.S. government has increasingly focused on policy directions to support building a robust, domestic AI infrastructure. For example, President Biden issued Executive Order (E.O.) 14141, "Advancing United States Leadership in Artificial Intelligence Infrastructure," on January 14, 2025, providing a federal plan to build AI infrastructure in the United States. On January 21, 2025, President Trump announced a private joint venture with potential investment of up to $500 billion to fund AI infrastructure, including plans to build up to 20 data centers in the country. The National Telecommunications and Information Administration (NTIA), a Department of Commerce agency serving as the President's principal advisor on telecommunication and information policies, issued a request for comments (RFC) to inform policymaking for "sustainable, resilient and secure" growth of data centers to power critical and emerging technologies, including AI. The RFC received 58 comments, ranging from the importance of allocating an estimated $175 billion of potential funds to U.S.-backed global AI infrastructure projects to challenges faced by domestic data centers, such as regulatory obstacles for siting new facilities and access to energy, supply chains, and operational workforce.

This In Focus introduces the use of data center and cloud computing infrastructure for AI development and AI-enabled services and selected policy issues for congressional consideration. For more on AI technologies and policies, see CRS Report R47644, *Artificial Intelligence: Overview, Recent Advances, and Considerations for the 118th Congress*; for more on energy issues related to selected data centers, see CRS Report R45863, *Bitcoin, Blockchain, and the Energy Sector*.

Overview of Data Centers

In its simplest form, a data center is a facility that houses and powers a large computer system. Data centers have evolved to house multiple enterprise-level, interconnected computer servers (e.g., a cluster of servers called a *server farm*). Personal computers and smart devices connect users to these servers to access online services (e.g., websites, emails, and file sharing). Many data centers have expanded to support cloud computing services, allowing users to remotely access computing resources such as data processing chips, software, data storage, networks, and applications and services hosted by these centers.

The term *data center* has been defined in federal laws in the context of energy efficiency and federal use of data centers. For instance, the Energy Independence and Security Act of 2007 (P.L. 110-140, §453(a)(1)) defines a data center as a facility that "contains electronic equipment used to process, store, and transmit digital information." In its guidance (M-25-03) for federal agencies to implement the Federal Data Center Enhancement Act of 2023 (P.L. 118-31, §5302), the Office of Management and Budget specified that a data center (1) is composed of permanent structures and operates in a fixed location; (2) houses IT equipment, including servers and other high-performance computing devices, or data storage devices; and (3) hosts information and information systems accessed by other systems or by users on other devices.

Data Centers for AI

The ever-increasing demand for data storage and processing capacities, especially for intensive computational tasks such as AI training, has led to construction and operation of *hyperscale data centers*. According to industry analysts, to be considered a hyperscale data center, a facility must contain at least 5,000 servers and occupy at least 10,000 square feet of physical space, with a power demand exceeding 100 megawatts (MW). If a data center had a continuous power demand of 100 MW for 24 hours, it would consume 2,400 MW-hours (MWh) of energy.

These large facilities provide powerful computing resources (such as memories, central processing units [CPUs], and graphics processing units [GPUs]) to handle vast amounts of data and large-scale workloads. (CPU and GPU are two major types of computing chips found in most personal computers and servers.) General-purpose workloads typically require only a CPU, while a GPU is generally considered better than a CPU to handle AI computational tasks. According to the chip manufacturer Nvidia, an AI-

ready data-center server can support one to eight high-performance GPUs, each of which consumes hundreds of watts of electricity. Cutting-edge chip technologies also support high-speed (at the gigabit-per-second level) GPU-to-GPU data communication among hundreds of GPUs across multiple servers, enabling the creation of a massive AI server farm in a data center.

In E.O. 14141, the term *AI data center* means a data center used primarily to develop or operate AI, and the term *frontier AI data center* means an AI data center capable of being used to develop an AI model that matches or surpasses the state-of-the-art AI model with regard to its performance or computational resources used in its development. These frontier AI models and related services typically require more power than traditional IT operations supported by data centers (e.g., data storage, retrieval, and transmission). For example, ChatGPT was estimated to use 2.9 watt-hours to respond to a user query, while a traditional Google search query uses about 0.3 watt-hours.

Multiple industry reports indicate that data processing demands of AI and related cloud computing services have spurred new construction and upgrades of data centers. The computing resources hosted by these centers would in turn lead to increased power demand. For example, one report estimated that the computing capacity (measured by the power demand) of data centers under construction in North America in the first half of 2024 reached a record-high 3,872 MW, up by 69% from a year earlier. Nearly 80% of this capacity has been pre-leased, with AI and cloud service providers contributing to significant portions of such demand. A report commissioned by the Department of Energy estimated that data centers accounted for about 4.4% of total U.S. electricity consumption (or about 176 million MWh) in 2023, "equivalent to the average annual consumption of 14 million households." According to another report, new hyperscale data centers have been built with capacities from 100 to 1,000 MW, "roughly equivalent to the load from 80,000 to 800,000 homes."

Investments in AI Infrastructure

Many experts believe that AI development and operation require significant capital investment. According to the annual *AI Index Report 2024*, costs of training large foundation models alone could "run into millions of dollars and are rising." Given rental prices charged by major cloud service providers for accessing computing hardware, the report estimated that training costs in 2023 for OpenAI's GPT-4 and Google's Gemini Ultra were around $78 million and $191 million, respectively (excluding other costs, such as data acquisition and labor). In a January 24, 2025, Facebook post, Meta's CEO revealed that the company was building a 2,000-MW data center and planned to spend $60-$65 billion in AI capital expenditures and acquire over 1.3 million GPUs by the end of 2025. Three days earlier at the White House, a group of companies announced a joint venture called Stargate, planning to invest up to $500 billion over the next four years to build AI infrastructure in the United States. Stargate's first data center, reportedly under construction, would be used by OpenAI and operated by Oracle (one of the largest cloud-based database vendors).

The tech industry and financial market have raised questions about large investments in AI infrastructure, due to the emergence of high-performing AI models developed by the less-known Chinese company DeepSeek. The company claimed in a non-peer-reviewed technical report that it has developed a large language model (LLM) called DeepSeek-V3 using efficient and cost-effective approaches. The model required 2.8 million GPU hours (equal to about 57 days) for its full training on a cluster of 2,048 Nvidia H800 GPUs. Assuming the cloud rental price at $2 per GPU hour, DeepSeek reported its total training cost for the model was $5.6 million—a fraction of what is currently spent by leading U.S. AI companies (e.g., $78 million for GPT-4). Nvidia reportedly developed the H800 chip in March 2023 as a modified version of its more powerful H100 GPU to comply with U.S. export control regulations. It is no longer permissible to export the H800 chip since the Department of Commerce's Bureau of Industry and Security updated its rules on advanced computing chips in October 2023 (see Category 3A090 in Supplement No. 1 to 15 C.F.R. Part 774).

According to its report, DeepSeek evaluated its V3 model against several leading LLMs, including those developed by Alibaba, Meta, OpenAI, and Anthropic. In six benchmark tests in language understanding, graduate-level science and mathematics questions, high school-level math competition questions, coding, and memory-chip security, the company claimed that DeepSeek-V3 outperformed the competitors in three and ranked second in the other three. The widely reported relatively low-cost advancement in AI model training has raised questions about the necessity of large AI investments, the efficiency of AI development, and U.S. leadership in the field.

Selected Policy Issues for Congress

AI infrastructure, including data centers, cloud computing services, and energy resources, is vital for AI development and operation. The criticality of such infrastructure raises policy issues such as data security, infrastructure access and security, energy reliability and efficiency, and costs of building such infrastructure, all of which implicate national security interests. Investment, construction, and operation of AI infrastructure may impact job and workforce development opportunities, and the economies and natural resources of local communities near data centers.

Previous Congresses saw bills to assess, understand, and address potential impacts of growth in AI, data centers, and needs for advanced computing resources. For example, the Department of Energy AI Act (S. 4664, 118th Congress) would have required the Secretary of Energy to report to Congress on data centers for advanced computing, their hardware and software needs for AI, and national security risks. The Remote Access Security Act (H.R. 8152, 118th Congress) would have expanded export controls to include *remote access* (e.g., through a cloud computing service) to a commodity, software, or technology for purposes such as training AI models.

Ling Zhu, Analyst in Telecommunications Policy

IF12899

Disclaimer