Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

September 18, 2023

Semiconductors and Artificial Intelligence

The increasing popularity of artificial intelligence (AI) has drawn congressional attention, and many Members are considering proposals to regulate the quickly evolving landscape. Technical progress in AI has been enabled in large part by advances in the underlying computational hardware—also known as semiconductors, integrated circuits, microelectronics, or simply *chips*—that offer increased processing power to improve the development of AI systems. This In Focus describes the types of semiconductors used in AI, concerns related to their supply chains, and challenges for the regulation of semiconductors to promote U.S. competitiveness in AI.

Artificial Intelligence Models

AI refers broadly to computational systems that can learn from data and make decisions such as predictions, recommendations, or classifications. AI systems can be implemented for diverse applications, including speech/visual recognition, autonomous driving, robotic process automation, and as virtual assistants.

A popular class of AI systems is deep neural networks, which use algorithms, or models, to mimic neurons in the brain to identify complex patterns. This AI model typically involves two stages: training and inference. In the training phase, the model is fed data that can be labeled (e.g., thousands of pictures of dogs to learn all the variations of dogs) or unlabeled to identify patterns. In the inference phase, the trained model is used to enable predictions and guide decisions, such as autonomous driving systems recognizing dogs as obstacles to avoid. The training phase typically requires the most computational power.

Generally, the accuracy of AI models increases with training on larger amounts of data, which in turn requires more computational power. Popular large language models, such as GPT-3, are trained on billions or trillions of text data to process and generate text. Given the large data sets associated with AI models, some of the largest AI models can take weeks or months to train, using thousands of chips and costing millions of dollars. These high costs are due in large part to the electricity required to operate the hardware.

Semiconductor Use in AI Models

Semiconductors are tiny electronic devices designed to enable functions such as processing, storing, sensing, and moving data or signals. AI models employ different types of chips, including memory chips to store large amounts of data and logic chips to process the data. According to forecasts from Gartner, revenues from semiconductors used in AI may increase rapidly from around $44 billion in 2022 to $120 billion in 2027.

Early AI models used commercial, off-the-shelf logic chips called *central processing units* (CPUs) for training and inference. Although CPUs are still sufficient for inference, leading AI models now primarily train using *graphics processor units* (GPUs) originally designed for video rendering. GPUs enable parallel processing of information; by contrast, CPUs process information serially. Parallel processing allows the AI model to train faster using large amounts of data by dividing tasks and executing them simultaneously. Additionally, many chip design firms are increasingly offering custom logic chips designed for particular applications, including AI training, called *application-specific integrated circuits* (ASIC) or *accelerators*. Logic chips used for AI applications are also referred to generally as *AI chips*.

To train the largest AI models, many logic chips are connected together in large clusters with other semiconductor hardware (e.g., memory and networking chips) in data centers or supercomputing facilities. For example, Meta is building a supercomputer for AI research that is anticipated to contain 16,000 GPUs, and a startup called Inflection AI is building a cluster of 22,000 GPUs for its AI model. Some supercomputers built by private firms such as Meta, Tesla, and NVIDIA are larger than many nationally owned supercomputers around the world.

Companies that train AI models may purchase and maintain their own chip hardware infrastructure or may train their models remotely using the cloud by paying fees to access the hardware they need. According to the Federal Trade Commission, “cloud services can be expensive and are currently provided by only a handful of firms, raising the risk of anticompetitive practices.” Top cloud service providers in the United States for AI applications include Amazon Web AI Services, Microsoft Azure AI, and Google Cloud AI.

AI training typically benefits from improving two technical parameters for AI chips: higher processing power and faster chip-to-chip transfer speeds. A common metric used to market the processing power of different AI chips is a measurement of the number of mathematical operations a chip can do in one second, calculated in trillions of operations per second (TOPS). Chip-to-chip transfer speeds are generally reported by measuring how fast a chip can send information, or bytes, into and out of the chip in gigabytes per second (1 gigabyte is 1 billion bytes).

Many large AI models, such as GPT models from OpenAI, and leading AI research papers do not explicitly report the amount of computational power used to train the AI model. Additionally, there are no standard methods or tools to measure the amount of computational power used to train

AI models, as TOPS may be calculated differently by different companies and may not be the most optimal metric to evaluate and compare AI models. Transparency in computing usage for AI training and standard methods for measuring computing power globally may support regulatory efforts for AI.

AI Chip Design and Manufacturing

U.S.-headquartered companies, both established firms and start-ups, lead globally in the design of specialized logic chips for AI applications. However, the large majority of these chip-design firms rely wholly on contract manufacturing services to produce and package their designs. As the highest performance AI chips require the most advanced manufacturing processes in the world, the majority of AI chip designers rely on the two logic chip manufacturing firms currently capable of producing their designs: Taiwan Semiconductor Manufacturing Company (TSMC) and Samsung.

The top AI chip designer by revenue and usage in AI research is U.S.-headquartered NVIDIA, one of the first companies to mass market GPUs in the early 2000s. Leading GPU products from NVIDIA used in AI applications, in order of increasing computational power, are marketed by the names V100 (2017), A100 (2020), and H100 (2022). Each successor chip can transfer information into and out of the chip faster than its predecessor, enabling higher-speed communications between large clusters of chips and faster AI training. These higher performance metrics may enable an AI model to train faster than it would using other commercially available GPUs and, in turn, may lead to relatively lower costs.

Top U.S.-headquartered AI chip design start-ups include, by company valuation, SambaNova, Cerebras, and Graphcore. Smaller entities such as start-ups often face challenges to prototyping and producing their designs due to the high cost of and limited access to contract manufacturing services from companies such as TSMC. As competitiveness in AI benefits from advancements in chip hardware, promoting access to prototyping and manufacturing services for U.S.-based firms may boost long-term innovation.

Export Controls on AI Chips

In October 2022, the Department of Commerce implemented controls that require licenses for exports to China and Macau of certain advanced logic and other chips that can be used for applications such as AI training and for building supercomputers. Controls apply to those logic chips with chip-to-chip transfer speeds of 600 gigabytes per second or more and computational power over 600 TOPS.

Under this definition, exports of leading AI chips, including NVIDIA's A100 and H100, to China and Macau are restricted. In recent years, China accounted for about a quarter of total annual revenues for NVIDIA. In November 2022, NVIDIA began marketing an A800 chip, which had a lower chip-to-chip transfer speed of 400 gigabytes per second (compared with 600 gigabytes per second in the A100), to "provide alternative products not subject to the new license requirements" to customers in China, according to its annual report. Similarly, in March 2023, NVIDIA marketed an H800 chip that does not require a license as an alternative to the newest H100 products, which fall under the controls.

Additionally, the October 2022 export controls restrict chip manufacturing facilities globally from manufacturing certain advanced chips for Chinese-headquartered chip design firms without a license if the manufacturer uses U.S.-origin technology or software (i.e., Advanced Computing Foreign Direct Product Rule). As the United States is a global leader in the production of semiconductor manufacturing equipment, this rule would apply to most foreign chip manufacturing firms, including TSMC, which previously produced advanced chips for Chinese AI chip design companies such as Biren. The rules require licenses to export certain advanced manufacturing equipment to chip manufacturers in China and Macau to impede the manufacturers' ability to produce advanced chips.

The export controls are designed to limit the ability of China and Macau to buy or produce certain advanced chips that can be used for AI applications. However, there are no controls on Chinese AI firms to use cloud service providers inside or outside of the country to train AI models.

Selected Federal Actions and Considerations for Congress

In January 2021, Congress enacted the National Artificial Intelligence Initiative Act of 2020 (Division E of P.L. 116-283), which seeks to advance U.S. leadership in AI research and development. Part of the act seeks to establish a roadmap for a National AI Research Resource, a shared research infrastructure for AI researchers and students. Directed by Executive Order 13859, the National Institute of Standards and Technology conducted a study that recommended the federal government "commit to deeper, consistent, long-term engagement in AI standards development activities," including the development of "metrics to quantifiably measure and characterize AI technologies, including but not limited to aspects of hardware and its performance." As Congress considers legislation to regulate the AI landscape, standard methods and tools to measure, for example, how much computational hardware AI models used for training may help govern these technologies.

In August 2022, President Biden signed P.L. 117-167, known as the CHIPS and Science Act. The act appropriated $39 billion to expand domestic semiconductor manufacturing capacity and $11 billion for research and development of next-generation semiconductor technologies. Congress may exercise its oversight authority with respect to the effectiveness of expanding domestic manufacturing capacity for advanced logic chips and improving manufacturing accessibility for smaller entities.

Additionally, as many AI models are trained using cloud services, Congress may consider export control reforms that enable the Department of Commerce to exercise regulatory authority over providers of cloud services that sell access to large amounts of computational power.

IF12497

Manpreet Singh, Analyst in Industrial Organization and Business

Disclaimer