THE GYPSUM ADVOCATE

AI raises energy demand of data centers

By Paige Gross
States Newsroom

The next time you're on a Zoom meeting or asking ChatGPT a question, picture this: The information zips instantaneously through a room of hot, humming servers, traveling hundreds, possibly thousands of miles, before it makes its way back to you in just a second or two.

It can be hard to wrap your mind around, said Vijay Gadepally, a senior scientist at Massachusetts Institute of Technology's Lincoln Laboratory, but large data centers are where nearly all artificial intelligence systems and computing happens today.

"Each one of these AI models has to sit on a server somewhere, and they tend to be very, very big," he said. "So, if your millions or billions of users are talking to the system simultaneously, the computing systems have to really grow and grow and grow."

As the United States works to be a global AI superpower, it's become a home to hundreds of data centers — buildings that store and maintain the physical equipment needed to compute information.

For users of the new and increasingly popular AI tools, it might seem like the changes have been all online, without a physical footprint. But the rise of AI has tangible effects — data centers and the physical infrastructure needed to run them use large amounts of energy, water and other resources, experts say.

"We definitely try to think about the climate side of it with a critical eye," said Jennifer Brandon, a science and sustainability consultant. "All of a sudden, it's adding so much strain on the grid to some of these places."

'Pretty intense'

As society traded large, desktop computers for sleek laptops, and internet infrastructure began supporting AI models and other software tools, the U.S. has built the physical infrastructure to support growing computing power.

Large language models (LLMs) and machine learning (ML) technologies — the foundation of most modern AI tools — have been used by technologists for decades, but only in the last five to seven years have they become commercialized and used by the general public, said David Acosta, cofounder and chief artificial intelligence officer of ARBOai.

To train and process information, these fast-learning AI models require graphic processing units (GPUs), servers, storage, cabling and other networking equipment, all housed in data centers across the country. Computers have been storing and processing data off-site in dedicated centers for decades, but the dot-com bubble in the early 2000s and the move to cloud storage demanded much more storage capacity over the last decade.

As more things moved online, and computing hardware and chip technology supported faster processing, AI models became attainable to the general public, Acosta said. Current AI models use thousands of GPUs to operate, and training a single chatbot like ChatGPT uses about the same amount of energy as 100 homes over the course of a year. "And then you multiply that times the thousands of models that are being trained," Acosta said. "It's pretty intense."

The United States is currently home to more than 3,600 data centers, and about 80% of them are concentrated in 15 states, Data Center Map shows. The market has doubled since 2020, Forbes reported, with 21% year over year growth. For many years, nearly all of the country's data centers were housed in Virginia, and the state is considered a global hub with nearly 70% of the world's internet traffic flowing through its nearly 600 centers. Texas and California follow Virginia, with 336 and 307 centers, respectively.

Tech companies that require large amounts of computing power, the private equity firms and banks that invest in them and other real estate or specialized firms are the primary funders of data centers. In September, BlackRock, Global Infrastructure Partners, Microsoft and AI investment fund MGX invested $30 billion into new and expanded data centers primarily in the U.S, and said they will seek $100 billion in total investment, including debt financing.

Investment in American data center infrastructure is encouraging considering the global "AI arms race," we're in, Acosta said. "If you own the data, you have the power," Acosta said. "I just think we just make sure we do it ethically and as preemptive as possible."

Like a city

Current estimates say data centers are responsible for about 2% of the U.S.' energy demand, but Anthony DeOrsey, a research manager at sustainable energy research firm Cleantech group, projects data centers will be about 10% of demand by 2027.

As data centers are developed in new communities across the country, residents and their state legislators see a mix of financial benefits with energy and environmental challenges.

The development of data centers brings some infrastructure jobs to an area, and in busy data center communities, like Virginia's Loudoun and Prince William counties, centers can generate millions in tax revenue, the Virginia Mercury reported.

Local governments can be eager to strike deals with the tech companies or private equity firms seeking to build, but the availability and cost of power is a primary concern. New large data centers require the electricity equivalent of about 750,000 homes, a February report from sustainability consultancy firm BSI and real estate services firm CBRE.

Under many state's utilities structures, local residents can be subjected to electric price increases to meet big electric needs of data centers. Some legislators, like Georgia State Sen. Chuck Hufstetler, have sought to protect residential and commercial customers from getting hit with higher utility bills.

Granville Martin, an Eastern Shore, Connecticut-based lawyer with expertise in finance and environmental regulation, said the same problem has come up in his own community. "The argument was, the locals didn't want this data center coming in there and sucking up a bunch of the available power because their view — rightly or wrongly, and I think rightly — was well, that's just going to raise our rates," Martin said.

Some states are exploring alternative energy sources. In Pennsylvania, Constellation Energy made a deal to restart this nuclear power plant at Three Mile Island to provide carbon-free electricity to offset Microsoft's power usage at its nearby data centers.

But climate experts have concerns about data centers outside of their power demand.

"The general public is largely unaware that cooling industrial facilities, whatever they might be, is actually a really, really important aspect of their function," Martin said.

The equipment in data centers, many of which run 24/7, generate a lot of heat. To regulate temperature, most pump water through tubing surrounding the IT equipment, and use air conditioning systems to keep those structures cool. About 40% of data center's energy consumption is used for cooling, the Cleantech group found.

Some have a closed-loop system, recycling grey water through the same system, but many use fresh drinking water. The amount of water and energy used in cooling is enormous, Brandon, the sustainability consultant, said.

"The current amount of AI data centers we have takes six times the amount of water as the country of Denmark," she said. "And then we are using the same amount of energy as Japan, which is the fifth largest energy user in the world, for data centers right now."

Is it worthwhile?

Energy is now a material issue to running an AI company, DeOrsey said, and unrestrained, quickly evolving AI models are very expensive to train and operate. DeOrsey pointed to Chinese AI company DeepSeek, which released its attempt at a cost-conscious, energy efficient large language model, R1, in January.

The company claims it trained the model on 2,000 chips, much fewer than competitors like Open AI, ChatGPT's parent company, and Google, which use about 16,000 chips. It's not yet clear if the model lives up to its claims of energy efficiency in use, but it's a sign that companies are feeling the pressure to be more efficient, DeOrsey said.

"I think companies like DeepSeek are an example of companies doing constrained optimization," he said. "They're assuming they won't just get all the power they need, they won't be able to get all of the chips they need, and just make do with what they have."

For Gadepally, who is also chief tech officer of AI company Radium Cloud, this selective optimization is a tool he hopes more companies begin using. His recent work at MIT's Lincoln Laboratory Supercomputing Center focused on the lab's own data center consumption. When they realized how hot their equipment was getting, they did an audit.

Gadepally said simple switches like using cheaper, less-robust AI models cut down on their energy use. Using AI models at off-peak times saved money, as did "power capping" or limiting the amount of power feeding their computer processors. The difference was nominal — you may wait a second or two more to get an answer back from a chatbot, for example.

With Northeastern University, MIT built software called Clover that watches carbon intensity for peak periods and makes adjustments, like automatically using a lower-quality AI model with less computing power when energy demand is high.

"We've been kind of pushing back on people for a long time saying, is it really worth it?" Gadepally said. "You might get a better, you know, knock-knock joke from this chatbot. But that's now using 10 times the power than it was doing before. Is that worth it?"

Gadepally and Acosta both spoke about localizing AI tools as another energy and cost saving strategy for companies and data centers. In practice, that means building tools to do exactly what you need them to do, and nothing more, and hosting them on local servers that don't need to send their computing out potentially hundreds of miles away to the nearest data center.

Health care and agricultural settings are a great example, Acosta said, where tools can be built to serve these specialized settings rather than processing their data at "bloated, over-fluffed" large data centers.

Neither AI developer sees any slowdown in the demand for AI and processing capabilities of data centers. But Gadepally said environmental and energy concerns will come to a head for tech companies when they realize they could save money by saving energy, too. Whether DeepSeek finds the same success as some of its American competitors is yet to be seen, Gadepally said, but it will probably make them question their practices.

"It will at least make people question before someone says, 'I need a billion dollars to buy new infrastructure,' or 'I need to spend a billion dollars on computing next month,'" Gadepally said. "Now they may say, 'did you try to optimize it?'"

Radium Cloud's newest data center in Raleigh, N.C. Such facilities might soon consume up to 10 percent of U.S. electricity. Photo courtesy of Vijay Gadepally

Gypsum High alumni dinner

The annual Gypsum High School alumni basket dinner will be held May 24 at City Auditorium, 621 Maple. Registration will be from 5 to 6 p.m. The dinner starts at 6. Drinks will be provided, but you should bring your own table service.