Four reasons to be optimistic about AI’s energy usage

technologyreview.com/2025/05/20/1116337/ai-energy-use-optimism

Will Douglas Heaven May 20, 2025

The day after his inauguration in January, President Donald Trump announced Stargate, a $500 billion initiative to build out AI infrastructure, backed by some of the biggest companies in tech. Stargate aims to accelerate the construction of massive data centers and electricity networks across the US to ensure it keeps its edge over China.

This story is a part of* MIT Technology Review*’s series “Power Hungry: AI and our energy future,” on the energy demands and carbon costs of the artificial-intelligence revolution.

The whatever-it-takes approach to the race for worldwide AI dominance was the talk of Davos, says Raquel Urtasun, founder and CEO of the Canadian robotruck startup Waabi, referring to the World Economic Forum’s annual January meeting in Switzerland, which was held the same week as Trump’s announcement. “I’m pretty worried about where the industry is going,” Urtasun says.

She’s not alone. “Dollars are being invested, GPUs are being burned, water is being evaporated—it’s just absolutely the wrong direction,” says Ali Farhadi, CEO of the Seattle-based nonprofit Allen Institute for AI.

But sift through the talk of rocketing costs—and climate impact—and you’ll find reasons to be hopeful. There are innovations underway that could improve the efficiency of the software behind AI models, the computer chips those models run on, and the data centers where those chips hum around the clock.

Here’s what you need to know about how energy use, and therefore carbon emissions, could be cut across all three of those domains, plus an added argument for cautious optimism: There are reasons to believe that the underlying business realities will ultimately bend toward more energy-efficient AI.

1/ More efficient models

The most obvious place to start is with the models themselves—the way they’re created and the way they’re run.

AI models are built by training neural networks on lots and lots of data. Large language models are trained on vast amounts of text, self-driving models are trained on vast amounts of driving data, and so on.

But the way such data is collected is often indiscriminate. Large language models are trained on data sets that include text scraped from most of the internet and huge libraries of scanned books. The practice has been to grab everything that's not nailed down, throw it into the mix, and see what comes out. This approach has certainly worked, but training a model on a massive data set over and over so it can extract relevant patterns by itself is a waste of time and energy.

There might be a more efficient way. Children aren't expected to learn just by reading everything that's ever been written; they are given a focused curriculum. Urtasun thinks we should do something similar with AI, training models with more curated data tailored to specific tasks. (Waabi trains its robotrucks inside a superrealistic simulation that allows fine-grained control of the virtual data its models are presented with.)

It's not just Waabi. Writer, an AI startup that builds large language models for enterprise customers, claims that its models are cheaper to train and run in part because it trains them using synthetic data. Feeding its models bespoke data sets rather than larger but less curated ones makes the training process quicker (and therefore less expensive). For example, instead of simply downloading Wikipedia, the team at Writer takes individual Wikipedia pages and rewrites their contents in different formats—as a Q&A instead of a block of text, and so on—so that its models can learn more from less.

Training is just the start of a model's life cycle. As models have become bigger, they have become more expensive to run. So-called reasoning models that work through a query step by step before producing a response are especially power-hungry because they compute a series of intermediate subresponses for each response. The price tag of these new capabilities is eye-watering: OpenAI's o3 reasoning model has been estimated to cost up to $30,000 per task to run.

But this technology is only a few months old and still experimental. Farhadi expects that these costs will soon come down. For example, engineers will figure out how to stop reasoning models from going too far down a dead-end path before they determine it's not viable. "The first time you do something it's way more expensive, and then you figure out how to make it smaller and more efficient," says Farhadi. "It's a fairly consistent trend in technology."

One way to get performance gains without big jumps in energy consumption is to run inference steps (the computations a model makes to come up with its response) in parallel, he says. Parallel computing underpins much of today's software, especially large language models (GPUs are parallel by design). Even so, the basic technique could be applied to a wider range of problems. By splitting up a task and running different parts of it at the same time, parallel computing can generate results more quickly. It can also save energy by making more efficient use of available hardware. But it requires clever new algorithms to coordinate the multiple subtasks and pull them together into a single result at the end.

The largest, most powerful models won't be used all the time, either. There is a lot of talk about small models, versions of large language models that have been distilled into pocket-size packages. In many cases, these more efficient models perform as well as larger ones, especially for specific use cases.

As businesses figure out how large language models fit their needs (or not), this trend toward more efficient bespoke models is taking off. You don't need an all-purpose LLM to manage inventory or to respond to niche customer queries. "There's going to be a really, really large number of specialized models, not one God-given model that solves everything," says Farhadi.

Christina Shim, chief sustainability officer at IBM, is seeing this trend play out in the way her clients adopt the technology. She works with businesses to make sure they choose the smallest and least power-hungry models possible. "It's not just the biggest model that will give you a big bang for your buck," she says. A smaller model that does exactly what you need is a better investment than a larger one that does the same thing: "Let's not use a sledgehammer to hit a nail."

2/ More efficient computer chips

As the software becomes more streamlined, the hardware it runs on will become more efficient too. There's a tension at play here: In the short term, chipmakers like Nvidia are racing to develop increasingly powerful chips to meet demand from companies wanting to run increasingly powerful models. But in the long term, this race isn't sustainable.

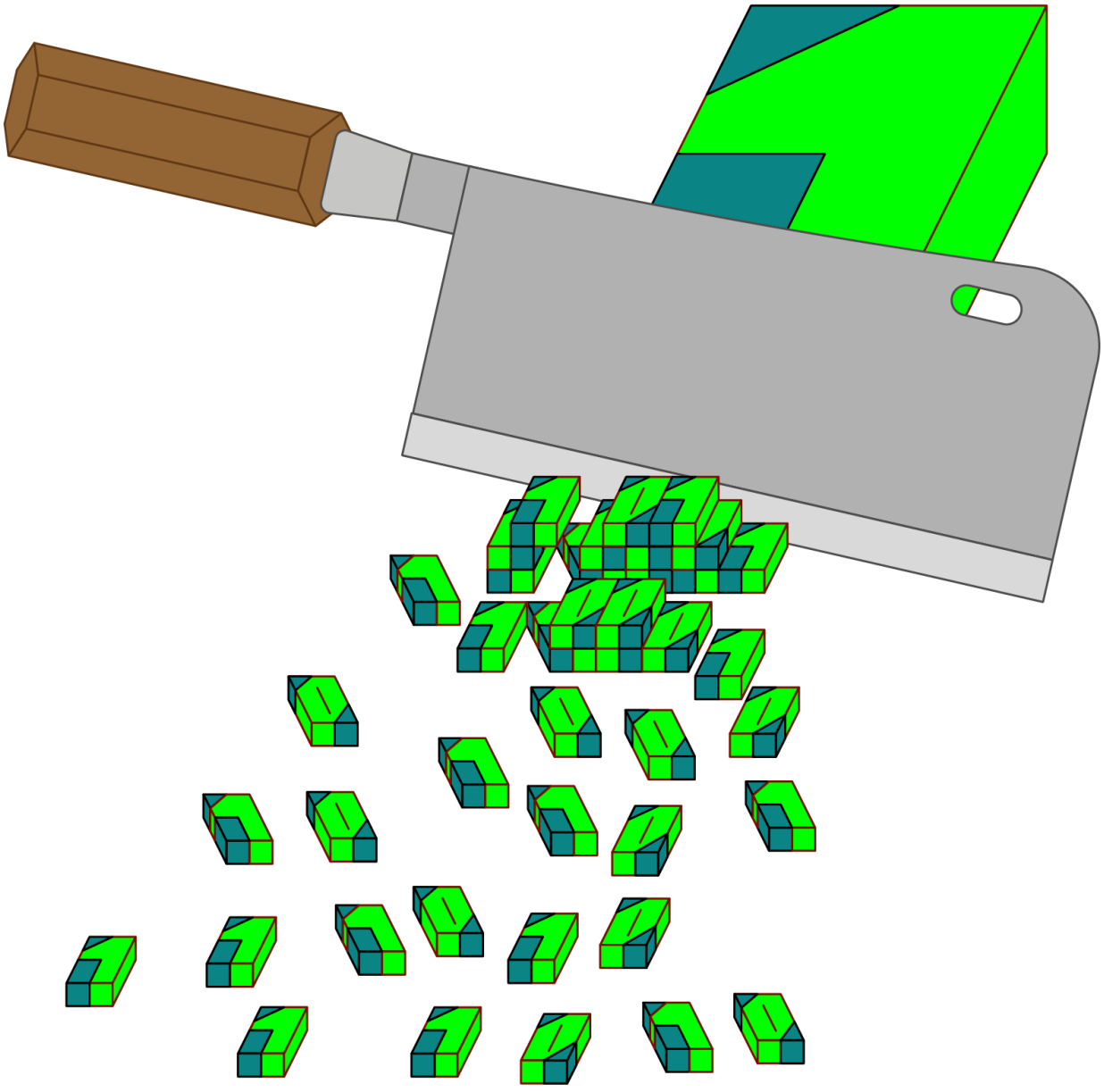
"The models have gotten so big, even running the inference step now starts to become a big challenge," says Naveen Verma, cofounder and CEO of the upstart microchip maker EnCharge AI.

Companies like Microsoft and OpenAI are losing money running their models inside data centers to meet the demand from millions of people. Smaller models will help. Another option is to move the computing out of the data centers and into people's own machines.

That's something that Microsoft tried with its Copilot+ PC initiative, in which it marketed a supercharged PC that would let you run an AI model (and cover the energy bills) yourself. It hasn't taken off, but Verma thinks the push will continue because companies will want to offload as much of the costs of running a model as they can.

But getting AI models (even small ones) to run reliably on people's personal devices will require a step change in the chips that typically power those devices. These chips need to be made even more energy efficient because they need to be able to work with just a battery, says Verma.

That's where EnCharge comes in. Its solution is a new kind of chip that ditches digital computation in favor of something called analog in-memory computing. Instead of representing information with binary *0*s and *1*s, like the electronics inside conventional, digital computer chips, the electronics inside analog chips can represent information along a range of values in between *0* and *1*. In theory, this lets you do more with the same amount of power.



SHIWEN SVEN WANG

EnCharge was spun out from Verma's research lab at Princeton in 2022. "We've known for decades that analog compute can be much more efficient—orders of magnitude more efficient—than digital," says Verma. But analog computers never worked well in practice

because they made lots of errors. Verma and his colleagues have discovered a way to do analog computing that's precise.

EnCharge is focusing just on the core computation required by AI today. With support from semiconductor giants like TSMC, the startup is developing hardware that performs high-dimensional matrix multiplication (the basic math behind all deep-learning models) in an analog chip and then passes the result back out to the surrounding digital computer.

EnCharge's hardware is just one of a number of experimental new chip designs on the horizon. IBM and others have been exploring something called neuromorphic computing for years. The idea is to design computers that mimic the brain's super-efficient processing powers. Another path involves optical chips, which swap out the electrons in a traditional chip for light, again cutting the energy required for computation. None of these designs yet come close to competing with the electronic digital chips made by the likes of Nvidia. But as the demand for efficiency grows, such alternatives will be waiting in the wings.

It is also not just chips that can be made more efficient. A lot of the energy inside computers is spent passing data back and forth. IBM says that it has developed a new kind of optical switch, a device that controls digital traffic, that is 80% more efficient than previous switches.

3/ More efficient cooling in data centers

Another huge source of energy demand is the need to manage the waste heat produced by the high-end hardware on which AI models run. Tom Earp, engineering director at the design firm Page, has been building data centers since 2006, including a six-year stint doing so for Meta. Earp looks for efficiencies in everything from the structure of the building to the electrical supply, the cooling systems, and the way data is transferred in and out.

For a decade or more, as Moore's Law tailed off, data-center designs were pretty stable, says Earp. And then everything changed. With the shift to processors like GPUs, and with even newer chip designs on the horizon, it is hard to predict what kind of hardware a new data center will need to house—and thus what energy demands it will have to support—in a few years' time. But in the short term the safe bet is that chips will continue getting faster and hotter: "What I see is that the people who have to make these choices are planning for a lot of upside in how much power we're going to need," says Earp.

One thing is clear: The chips that run AI models, such as GPUs, require more power per unit of space than previous types of computer chips. And that has big knock-on implications for the cooling infrastructure inside a data center. "When power goes up, heat goes up," says Earp.

With so many high-powered chips squashed together, air cooling (big fans, in other words) is no longer sufficient. Water has become the go-to coolant because it is better than air at whisking heat away. That's not great news for local water sources around data centers. But there are ways to make water cooling more efficient.

One option is to use water to send the waste heat from a data center to places where it can be used. In Denmark water from data centers has been used to heat homes. In Paris, during the Olympics, it was used to heat swimming pools.

Water can also serve as a type of battery. Energy generated from renewable sources, such as wind turbines or solar panels, can be used to chill water that is stored until it is needed to cool computers later, which reduces the power usage at peak times.

But as data centers get hotter, water cooling alone doesn't cut it, says Tony Atti, CEO of Phononic, a startup that supplies specialist cooling chips. Chipmakers are creating chips that move data around faster and faster. He points to Nvidia, which is about to release a chip that processes 1.6 terabytes a second: "At that data rate, all hell breaks loose and the demand for cooling goes up exponentially," he says.

According to Atti, the chips inside servers suck up around 45% of the power in a data center. But cooling those chips now takes almost as much power, around 40%. "For the first time, thermal management is becoming the gate to the expansion of this AI infrastructure," he says.

Phononic's cooling chips are small thermoelectric devices that can be placed on or near the hardware that needs cooling. Power an LED chip and it emits photons; power a thermoelectric chip and it emits phonons (which are to vibrational energy—a.k.a. temperature—as photons are to light). In short, phononic chips push heat from one surface to another.

Squeezed into tight spaces inside and around servers, such chips can detect minute increases in heat and switch on and off to maintain a stable temperature. When they're on, they push excess heat into a water pipe to be whisked away. Atti says they can also be used to increase the efficiency of existing cooling systems. The faster you can cool water in a data center, the less of it you need.

4/ Cutting costs goes hand in hand with cutting energy use

Despite the explosion in AI's energy use, there's reason to be optimistic. Sustainability is often an afterthought or a nice-to-have. But with AI, the best way to reduce overall costs is to cut your energy bill. That's good news, as it should incentivize companies to increase efficiency. "I think we've got an alignment between climate sustainability and cost sustainability," says Verma. "I think ultimately that will become the big driver that will push the industry to be more energy efficient."

Shim agrees: “It’s just good business, you know?”

Companies will be forced to think hard about how and when they use AI, choosing smaller, bespoke options whenever they can, she says: “Just look at the world right now. Spending on technology, like everything else, is going to be even more critical going forward.”

Shim thinks the concerns around AI’s energy use are valid. But she points to the rise of the internet and the personal computer boom 25 years ago. As the technology behind those revolutions improved, the energy costs stayed more or less stable even though the number of users skyrocketed, she says.

It’s a general rule Shim thinks will apply this time around as well: When tech matures, it gets more efficient. “I think that’s where we are right now with AI,” she says.

AI is fast becoming a commodity, which means that market competition will drive prices down. To stay in the game, companies will be looking to cut energy use for the sake of their bottom line if nothing else.

In the end, capitalism may save us after all.