POWERING DOWN

How Optics Technologies Can Reduce the Energy Costs of Artificial Intelligence and Boost Data center Efficiency

By **Rachel Berkowitz**

AS YOU DRIVE EAST ALONG THE COLUMBIA RIVER from Portland, Oregon, coastal woodlands complete their transition to inland desert as the city of The Dalles comes into view. Here, a dam along the river has long provided for humanity's increasing energy demands—including, since 2006, the energy needs of some of Google's largest data centers. But even sites like The Dalles, with its abundant hydropower, are reaching a crossroads as the energy requirements of servers, storage, and networks used in generative artificial intelligence (AI) quickly surpass those of basic cloud computing. With new energy-efficient approaches to data processing and storage, the photonics industry can help.

ChatGPT brought generative AI into the household vernacular. The term refers to computer programs that can compose emails, answer questions, write software, or otherwise create content. These tools can, for example, improve customer experiences in the form of chatbots, or help with repetitive tasks such as data entry. Many AI programs operate by learning the patterns in vast amounts of human-created content, from which they synthesize and generate new text, images, or videos.

But these programs are expensive because of the vast amount of information that they must process to learn the patterns that drive them. "Running servers already requires a lot of energy, carbon, and water," says Shaolei Ren, an electrical and computer engineer at the University of California, Riverside, who focuses on AI sustainability. Generative AI programs such as ChatGPT consume much more power than traditional search tools.

The International Energy Agency (IEA) reported in January 2024 that a single Google search consumes 0.3 watt-hours of electricity, but one query to ChatGPT requires 2.9 watt-hours of electricity—roughly the equivalent of turning on a 60-watt lightbulb for a few minutes to generate a page of text. A recent *Wall Street Journal* article noted that AI data centers could consume between 20 to 25 percent of US

The energy requirements of servers like the ones shown here are growing rapidly as the demands of generative artificial intelligence (AI) quickly surpass those of basic cloud computing

power requirements by the end of the decade. That's an increase from 4 percent today. In addition to the energy demands, AI consumes a lot of water for cooling, most of which evaporates and is considered permanently lost. Based on the IEA projection, Ren calculates the data center on-site water consumption required by global AI in 2026 would reach between 15- to 75-billion L, equivalent to the residential water consumption of between 2 to 10 million people in Europe.

"It's not about computing faster or storing better, the main point is consuming less power during computation," says Zeev Zalevsky, a professor of optoelectronics at Bar-Ilan University in Israel. He and his colleagues, along with many other research groups around the world, are developing technologies for both faster data processing and larger data storage. They believe it's possible to meet this growing demand for computing power by using photons instead of electrons.

A photonic processor uses light, rather than electrical current, to perform digital computations. Photons pass through waveguides and other components to process information. "Photons have a larger bandwidth to carry information, and they are less sensitive to noise when transmitting high-frequency information than electronic information carriers," says Zalevsky.

Photonic processors work well for linear spatial transformations—such as reflecting or rotating a plane—and were within easy reach of early optical computing systems that pass light through an interferometer. But combining photonic processors with each other to solve more generic computational problems, in which multiple signals must interact, has remained a challenge.

The invention of neural networks as computational models has changed that. The basic processing module in a personal laptop is built from Boolean logic gates, whose nonlinear operations are hard to replicate with light. In contrast, neural network models use a processer where adding complexity is merely a structural add-on. These basic computing units, called neurons, can easily be realized using light—for example, photons passing through a collection of silica cores. A neural network is made up of multiple neurons that are organized into layers and connected to each other via weights that represent the strength of one neuron's influence on another. By connecting each pair of optical signals with proper weights, they can be trained to solve different types of problems.

For Zalevsky, the clear direction for optical computing is to build a standalone photonic processor that can be integrated with an optical communications system and does not depend on today's silicon chip infrastructure. "I prefer it not be integrated with silicon logic gate-based components, but rather to be integrated within fibers with all the infrastructure available in optical communications," he says. "That industry has proven technology to inject electronic-world information onto photonic carriers and transmit it via fibers."

Zalevsky's approach starts with a multicore optical fiber. Unlike communications fibers, where photonic information channels are guided through different cores that must be isolated from each other to avoid disturbances, his technology

requires that the propagating light gets exchanged between cores. This exchange provides the main functionalities of the neural network, where each layer's function is to take inputs, do a weighted summation, and perform inferences based on these summations. "The photons in the channels interact with each other, so light is exchanged between the cores, and you get the weighted summation of the input photons," he says. The activation function of the neural network's basic processing cell can also be realized easily using the nonlinear effects of light in optical fibers—for example, the fact that doubling the optical input intensity does not simply double the output intensity.

Zalevsky and his colleague Eyal Cohen, who together co-founded CogniFiber, have used this new technology to demonstrate a basic computational function. In their prototype, 90 silica cores and 13 laser-light information channels comprise a neural network that "learns" the output power corresponding to a specific input pattern. It then uses the output to decipher a variety of given laser inputs. "Our world is wired with millions of sensors and floating information that cannot go straight to the final user," Zalevsky says. For example, the devices that monitor a smart home's environment constantly deliver data to a user's phone. A photonic processor could weed out nonrelevant information and point out abnormalities that may be of interest—changes in patterns due to attacks or device malfunction.

Compared to a small electronic processor of similar complexity, Zalevsky and Cohen's photonic processor could perform the same computation orders-of-magnitude faster using orders-of-magnitude lower power. Making a more powerful or complex version would simply involve adding or connecting more fiber cores. "If I connect more fibers, I can easily increase the complexity. That's not true for electronic chips," says Zalevsky. The challenge is how best to interface the fiber-based processor with a real-world system. That's a software question that Zalevsky's team hopes to resolve in a few years.

Optical communication networks hint at other ways to reduce power consumption in information-processing networks. Since the late 1990s, engineers have explored light-based networking based on optical switches developed by the telecommunications industry to connect subscribers. Here, arrays of micromirrors known as optical MEMS (micro-electro-mechanical system) switches are angled to change the direction of light beams so they "switch" paths.

For most of the 20th Century, mechanical switches dominated the telecommunications world. "The switchboards even look like the rows of servers at data centers," says Niels Quack, an associate professor in the School of Aerospace and Mechanical Engineering at the University of Sydney. Unlike mechanical switchboards, however, scaling up modern data processing networks requires that information be swiftly transferred from one chip to another, then from one server to another, and, eventually, from one data center to another.

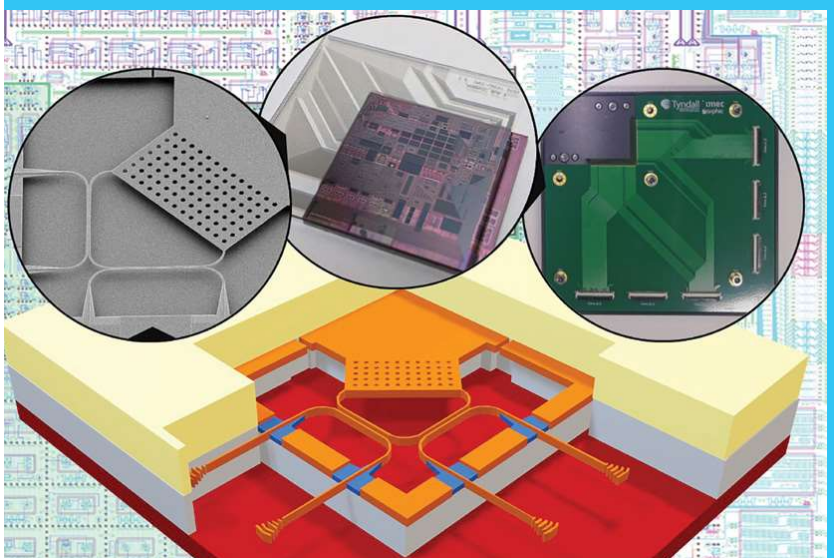
In 2023, Google reported that optical circuit switches are beginning to replace electronic switches for data center networking. The connections that send information from one server to another will use an optical switching system unit named Palomar that consists of 136 individually controllable micromirrors. When light enters Palomar via an input fiber, the mirrors can be realigned rapidly, so that each light signal gets redirected to the correct output fiber. The unit uses just 108 W of electrical power to hold the mirrors in place. In contrast, each standard 136-port electronic switch in data centers consumes around 3,000 W of electricity.

Google's in-house optical switching is state-of-the-art, says Quack. The fun part, for him, is finding out how to make photonic switches even better. That means combining optics and MEMS technology into photonic integrated circuits (PICs) that can be fabricated in the same high-volume semiconductor manufacturing processes that make silicon chips for everyday microprocessors. Instead of copper interconnects that route electronic information on a conventional electronic microchip, a photonic circuit with integrated MEMS routes light by moving a nanoscale mechanical waveguide—paving the way for efficient scaling of integrated optical circuit switches for data center servers.

For today's silicon PICs, most switch designs consume energy to maintain light-based connections. Electrical current heats the entire silicon-based waveguide producing an optical change on the material, thereby changing the path of the light. Not only does this process consume energy, but the change is temporary; once the heating stops, the connection—and flow of information—is lost.

Arka Majumdar, a professor of electrical and computer engineering and physics at the University of Washington, and Roger Fang, a photonics engineer at LightIC Technologies, developed a photonic switch that maintains information flow by triggering a change in the waveguide that persisted after the heating source was removed. They introduced a graphene layer that served as an atomically thin heater instead of heating the entire waveguide. The result was a 70-fold reduction in switching energy compared to state-of-the-art electrically actuated phase-change switches. "We activate the transition, using a tiny amount of energy in just one spot," explains Fang. A subsequent voltage pulse of light switches the waveguide back to the original state.

Despite not consuming any power to maintain its setting, the biggest problem with phase-change material switches like Fang's is reliability. The graphene heater can switch the waveguide 1,000 times before it ceases to perform reliably but billions of cycles are needed to meet the demands of modern computing. MEMS technologies, on the other hand,



Combining optics with micro-electro mechanical systems (MEMS) could enable energy-efficient microchips that can be produced using existing infrastructure

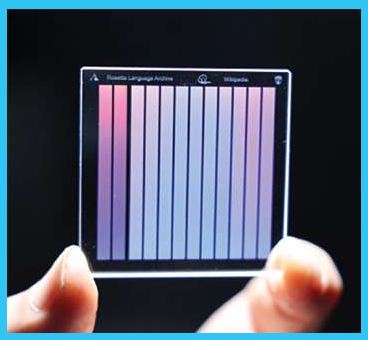
New laser-writing technology allows the storage capacity within the area of a DVD-sized disk to reach up to petabyte level ▶

have been proven reliable over billions of cycles in consumer electronics.

For photonic switches, the optimal power consumption is a trade-off between optical loss and physical size when scaling to large systems. Because of the materials involved, Quack believes that MEMS can optimize this trade-off. A single chip could host millions of MEMS components while keeping the on-chip optical loss very low and consuming very little power. "That's the real sales pitch for MEMS," he says. His team developed silicon photonic MEMS, using electrostatic actuators to manipulate the motion of a freestanding, half-micron-wide waveguide. Now, he's working on mechanical and materials aspects of the design, such as piezoelectric actuation mechanisms that could lead to even higher efficiency.

Quack says that silicon photonic MEMS have the potential to reach the industry-standard goal for optical loss (of less than 3dB, meaning that the output power after the switch retains half the power of the input) in the next five years at a price point of $1 to $10 per-optical-port. "These chips could help to replace the MEMS mirror-based technology from the late 1990s by novel, more efficient photonic-integrated circuits needed for AI within the next decade," he says.

Optical technologies offer one more boon for mitigating AI energy demands: long-term data storage. What started as a global need to archive and process several terabytes (TB) of data is burgeoning into petabytes (PB) and exabytes (EB) of data. Today, data is stored on hard drives and magnetic tape—both of which degrade over time. Routinely migrating it onto new drives is an energy- and time-consuming process. "Optical data storage emerges as a beacon of hope for cost-effective and long-term data management," says Jing Wen, a photonics researcher at the University of Shanghai for Science and Technology (USST).



Introducing nanostructures into fused quartz using a femtosecond laser leads to a 5D memory crystal that remains stable for 1,000 years. The sample shown here archives 13,000 pages written in 1,500 human languages (2 columns, 4 GB) and Wikipedia text from 2020 (10 columns, 24 GB) ◀

In 2014, Peter Kazansky, a professor of optoelectronics at the University of Southampton, and his colleagues demonstrated a new optical-based technology for long-term data storage. "We physically modify silica glass with light to create very small nanostructures," he says. These structures last indefinitely, like a CD or DVD. The difference is that it's possible to add dimensions to the nanostructure, and to include layers of nanostructures in a CD-sized disk—potentially thousands.

To create the nanostructure, a laser beam targets a tiny spot on the silica-glass storage material. The resulting volumetric pixel, or voxel, is a half-micron thick and several microns deep. But voxels are not simple voids. They can be nanogratings or elongated nanovoids whose orientation, or shape and size, can be controlled by the polarization and intensity of the light—even adding fourth and fifth dimensions to data storage, based on properties of how the light travels through the material.

In a similar vein, along with Min Gu, dean of the Institute of Photonic Chips at USST, and colleagues—has developed a 3D version with PB capacity. Their technology combines dual data-writing lasers and a newly developed light-sensitive material for high-density storage. The researchers demonstrated how to write and read data on ultrathin disks, in 100 layers, stacked 1 mm apart. The result was a storage capacity of up to 1.6 PB for a CD-sized disk area. "The significance of this research for AI data centers cannot be overstated," says Wen.

In an era where AI is advancing by leaps and bounds and infiltrating every corner of everyday life, the insatiable thirst for data has propelled data centers into an unprecedented energy crisis. As algorithms evolve and datasets balloon in size, conventional storage and processing solutions strain under their weight. The surge in AI-driven data center energy demands presents a pressing challenge in the quest for sustainable digital infrastructure—and the optics and photonics community are leading the way with solutions.

RACHEL BERKOWITZ *is a freelance science writer with a PhD in geophysics from the University of Cambridge. Her work has appeared in* New Scientist, Physics Magazine, Physics Today, Science News, Scientific American, *and the newsrooms of several US national laboratories.*

Photo credits: University of Sydney (left); University of Shanghai for Science and Technology (top right); Arch Mission Foundation/SPhotonix/University of Southampton (middle right).