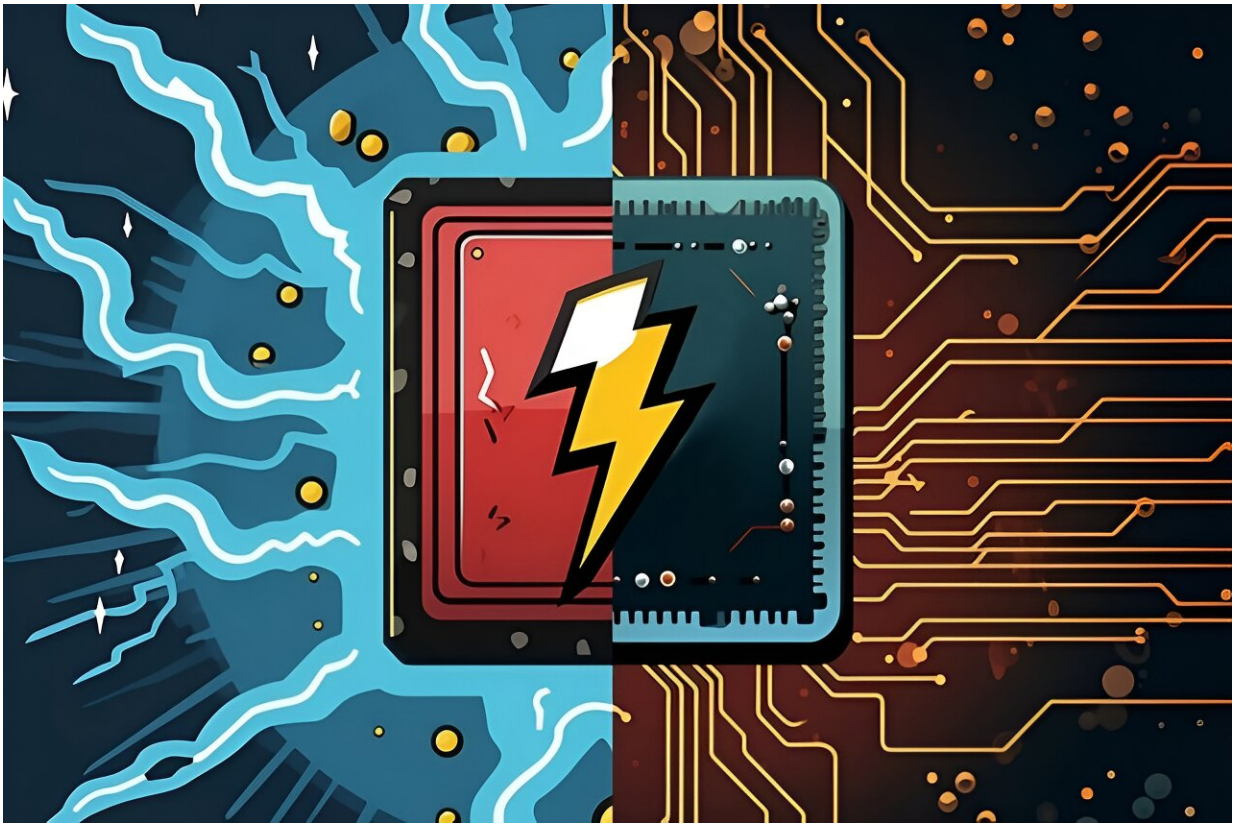
# System combines light and electrons to unlock faster, greener computing

September 11 2023, by Alex Shipps



MIT researchers introduce Lightning, a reconfigurable photonic-electronic smartNIC that serves real-time deep neural network inference requests at 100 Gbps. Credit:Alex Shipps/MIT CSAIL via Midjourney

Computing is at an inflection point. Moore's Law, which predicts that

the number of transistors on an electronic chip will double about every two years, is slowing down due to the physical limits of fitting more transistors on affordable microchips. Increases in computer power are slowing down as the demand grows for high-performance computers that can support increasingly complex artificial intelligence models.

This inconvenience has led engineers to explore new methods for expanding the computational capabilities of their machines, but a solution remains unclear.

Photonic computing is one potential remedy for the growing computational demands of machine-learning models. Instead of using transistors and wires, these systems utilize photons (microscopic light particles) to perform computation operations in the analog domain.

Lasers produce these small bundles of energy, which move at the speed of light like a spaceship flying at warp speed in a science fiction movie. When photonic computing cores are added to programmable accelerators like a network interface card (NIC, and its augmented counterpart, SmartNICs), the resulting hardware can be plugged in to turbocharge a standard computer.

MIT researchers have now harnessed the potential of photonics to accelerate modern computing by demonstrating its capabilities in machine learning.

Dubbed "Lightning," their photonic-electronic reconfigurable SmartNIC helps deep neural networks—machine-learning models that imitate how brains process information—to complete inference tasks like image recognition and language generation in chatbots such as ChatGPT. The prototype's novel design enables impressive speeds, creating the first photonic computing system to serve real-time machine-learning inference requests.

The group will present their findings at the Association for Computing Machinery's Special Interest Group on Data Communication (SIGCOMM) this month. The abstract has been published in the *Proceedings of the ACM SIGCOMM 2023 Conference*.

Despite its potential, a major challenge in implementing photonic computing devices is that they are passive, meaning they lack the memory or instructions to control dataflows, unlike their electronic counterparts. Previous photonic computing systems faced this bottleneck, but Lightning removes this obstacle to ensure data movement between electronic and photonic components runs smoothly.

"Photonic computing has shown significant advantages in accelerating bulky linear computation tasks like matrix multiplication, while it needs electronics to take care of the rest: memory access, nonlinear computations, and conditional logics. This creates a significant amount of data to be exchanged between photonics and electronics to complete real-world computing tasks, like a machine learning inference request," says Zhizhen Zhong, a postdoc in the group of MIT Associate Professor Manya Ghobadi at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

"Controlling this dataflow between photonics and electronics was the Achilles' heel of past state-of-the-art photonic computing works. Even if you have a super-fast photonic computer, you need enough data to power it without stalls. Otherwise, you've got a supercomputer just running idle without making any reasonable computation."

Ghobadi, an associate professor at MIT's Department of Electrical Engineering and Computer Science (EECS) and a CSAIL member, and her group colleagues are the first to identify and solve this issue. To accomplish this feat, they combined the speed of photonics and the dataflow control capabilities of electronic computers.

Before Lightning, photonic and electronic computing schemes operated independently, speaking different languages. The team's hybrid system tracks the required computation operations on the datapath using a reconfigurable count-action abstraction, which connects photonics to the electronic components of a computer.

This programming abstraction functions as a unified language between the two, controlling access to the dataflows passing through. Information carried by electrons is translated into light in the form of photons, which work at light speed to assist with completing an inference task. Then, the photons are converted back to electrons to relay the information to the computer.

By seamlessly connecting photonics to electronics, the novel count-action abstraction makes Lightning's rapid real-time computing frequency possible. Previous attempts used a stop-and-go approach, meaning data would be impeded by a much slower control software that made all the decisions about its movements.

"Building a photonic computing system without a count-action programming abstraction is like trying to steer a Lamborghini without knowing how to drive," says Ghobadi, who is a senior author of the paper.

"What would you do? You probably have a driving manual in one hand, then press the clutch, then check the manual, then let go of the brake, then check the manual, and so on. This is a stop-and-go operation because, for every decision, you have to consult some higher-level entity to tell you what to do."

"But that's not how we drive; we learn how to drive and then use muscle memory without checking the manual or driving rules behind the wheel. Our count-action programming abstraction acts as the muscle memory in

Lightning. It seamlessly drives the electrons and photons in the system at runtime."

## An environmentally friendly solution

Machine-learning services completing inference-based tasks, like ChatGPT and BERT, currently require heavy computing resources. Not only are they expensive—some estimates show that ChatGPT requires $3 million per month to run—but they're also environmentally detrimental, potentially emitting more than double the average person's carbon dioxide. Lightning uses photons that move faster than electrons do in wires, while generating less heat, enabling it to compute at a faster frequency while being more energy-efficient.

To measure this, the Ghobadi group compared their device to standard graphics processing units, data processing units, SmartNICs, and other accelerators by synthesizing a Lightning chip. The team observed that Lightning was more energy-efficient when completing inference requests.

"Our synthesis and simulation studies show that Lightning reduces machine learning inference power consumption by orders of magnitude compared to state-of-the-art accelerators," says Mingran Yang, a graduate student in Ghobadi's lab and a co-author of the paper. By being a more cost-effective, speedier option, Lightning presents a potential upgrade for data centers to reduce their machine learning model's carbon footprint while accelerating the inference response time for users.

**More information:** Zhizhen Zhong et al, Lightning: A Reconfigurable Photonic-Electronic SmartNIC for Fast and Energy-Efficient Inference, *Proceedings of the ACM SIGCOMM 2023 Conference* (2023). DOI: 10.1145/3603269.3604821

*This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.*

Provided by Massachusetts Institute of Technology