Hewlett Packard Enterprise

Hewlett Packard Labs

# Coherent Optical Computing: On a New Memristor and Machine Learning-assisted Photonics

Author: Ray Beausoleil

It’s coherence that makes optics cool. By harnessing this process, there’s a different way for us to compute in parallel. Matrix-vector multiplication we can do much faster using optical coherence than using electronics. This is what I want to talk about in greater detail. Coherence becomes interference, and interference provides us with a much better way of computing for certain types of especially hard problems.

First, something about who I am.

I am a Hewlett Packard Enterprise Senior Fellow, and I lead Hewlett Packard Labs’ Large-Scale Integrated Photonics Lab. Put those two together, and you have someone open-minded and passionate about learning, even if that means his ideas occasionally need reconfiguring. But I’m also someone who has, I hope, a professional life that has led to equally passionate ideas about how the industry and discipline I work in could and should be changed. It is my hope that in this series of explorations, I might share those passions—and those of other leading technologists in HPE—with you.

We’ve had optical communication for some time now, but there remains so much more waiting in the wings. Imagine a photonic integrated circuit that transmits and receives but also learns and remembers. Think of photons moving down waveguides and across chips, constructively and destructively interfering with each other along the way. Imagine the kind of issues that would experience a phase change from insurmountable to soluble.

I’m asking you to imagine such things because those of us who are devoted to coherent optical computing already are. This series will give me an opportunity to investigate these issues and express them in the hope that it will start a bigger conversation. Part of the way I intend to do that is by bringing in the work and opinions of my colleagues.

## Every problem is an optimization problem

The Von Neumann computational models we've had for a very long time have gone as far as they can with silicon CMOS-based technology because Moore’s Law—an empirical observation made long ago by Gordon Moore that the number of transistors on a chip doubles every 18 months—no longer applies. At this juncture, thinking about computation differently gives us a chance to tackle problems that CMOS electronics doesn't do a particularly good job with and provides us with opportunities to engage in post-Moore’s Law thinking.

There's a huge opportunity for machine learning to have an impact on science, to help human scientists solve incredibly difficult problems that outstrip our brain's ability to do pattern matching.

You might be training neural networks; you might be doing a singular value decomposition for data reduction; you might be doing system identification, very often through techniques of linear algebra that have been around for a very long time. But at the core of every machine learning problem that these techniques are used to solve is an optimization problem. So the key to making machine learning go faster and to enable even more complex problems to be tackled is to figure out how to do optimization better. It's that simple.

## You do not need to do anything exotic

If what you want is a minimum size, weight, and power in computing, you do not need to do anything exotic in the quantum mechanical sense. Your first stop should be memristors, non-linear electrical components linking electric charge and magnetic flux. Memristors exist at electronic size scales, are relatively cheap, and integrate very nicely with existing large-scale integrated electronic technologies.

Cat Graves is the principal research scientist and team lead for emerging accelerators at our AI Research Lab. Her team works on the memristor side of things, using different kinds of circuits plus memristors to target different computations, particularly in this exploding machine learning area.

One of those was the Dot-Product Engine, which utilizes a crossbar of these devices in order to accelerate one of those key linear algebra operations that come up all the time in machine learning: vector-matrix multiplication. This is an element shared by coherent optical computing. Cat's team has been playing with an analog version of a content-addressable memory circuit, which effectively accelerates lookup tables. The analog flavor also accelerates traversing decision trees, which form the basis of machine learning models like XGBoost or Random Forests.

This could enable real-time, low-latency ML processing at a relatively low cost, as well as operating on natively-analog signals from sensors that could eliminate expensive digital-to-analog conversion. In heavy-edge environments where you need to process lots of data quickly (and throw most of it away), this could actually keep up with those rates.

## Low energy

The way to do really low-energy computing is to go right up to what I call the "stochastic limit," where you begin to see quantum fluctuations. Then you harness those quantum fluctuations in such a way that you can ensure the device will still give you a yes or no answer the way that you expect a computer to operate, or, if we're doing analog computation, gives us a reliable, real number result that is, in fact, converging towards what I would call the classical value.

In fact, we did a study as part of a DARPA program that showed that we could use coherent feedback control in order to reduce the number of photons we had propagating around this chip down to a point where you'd actually see quantum effects. Then we had to control those quantum effects so that we'd still end up with a reliable classical computation.

Think of it this way:

1. It takes energy to create a photon.
2. The fewer photons we make, the less energy we burn.
3. Currently, our best photodetectors need about 3,000 photons to represent a classical bit.
4. Could we reduce this number to 30? Yes, but at this low number, quantum mechanics gets in the way.
5. Sometimes when we count these photons, we'll get 30. Sometimes we'll get 31, or 32, or even 19. This is "quantum fluctuations" in action.
6. If we say that more than 20 photons are a "1", and we count 19, then we incorrectly call it a "0."
7. We can use coherent quantum feedback control to dramatically reduce the probability that we'll make that mistake.

Years ago, we were trying to understand how to manifest and use coherence in photonic circuits that do computation. However, nano-photonics are still not as remotely nano as electronics, at least not in the kinds of form factors we use now. But nature gives us four ways to process information.

The Heisenberg uncertainty principle and Schrodinger's equation, taken together, are enough to do both silicon electronics and memristors in post-Moore's Law computation. In addition, we have coherence and entanglement. If you go all the way to the end of that list, you're talking about quantum computing. But we've never really spent much time exploring the utility of coherence, even classical coherence, to say nothing of quantum coherence, to see whether or not it offers us computational advantages.

Thomas Van Vaerenbergh, one of our photonics research engineers, has made the point that coherence gives you more degrees of freedom. That, in turn, produces a computational benefit. With less hardware, you are still able to process more information.

You are now beginning to see people utilizing this approach. Companies like Lightmatter are harnessing the parallelism of coherent optics to do linear algebra.

It will be difficult to outperform systems like analog electronics. But in the long term, Thomas sees an advantage.

If you can represent information with a coherent system, you only need one wavelength for a big calculation, and then you can imagine that if you have multiple wavelengths, multiple colors going through your chip in parallel at the same time, so you can augment the amount of information that is processed on a single chip surface simultaneously.

These tools add to the interconnectable kit we can customize to help deal with the enduring problems we face at the end of Moore's Law. These problems have to be recognized as real, not just hypothetical. When a customer buys an HPC installation, they don't buy the machine as a collector's item. They buy it because they have problems to solve. So the vision for the future of coherent optical computing has to be the creation of practical architecture.

## Solve for x

Frequently updated linear algebra computations enable machine learning problems to be solved more efficiently. But a key step after each layer of linear algebra is a non-linear "rectification" to force the state of a neuron to take on a value in a particular range. If we encode our state in photons and use coherent optical waveguides on a chip to do matrix multiplication, we still have to perform this last non-linear step. In current photonic neural networks, the optical information is converted back to electronics in order to do some of these non-linear things. So photonics is being used for coherence and linear algebra but not for the non-linear parts of the computation. This approach is what's used for a huge fiber-optic implementation of an optimizer called an Ising machine.

Some years back, we engaged Defense Advanced Research Projects Agency (DARPA) on a project to investigate putting all of that on just one chip. Our thinking is that if you can successfully get both the linear coherence and the non-linear computations on one chip, you have an optimization engine you can use to solve machine learning problems much more rapidly than you can using current CMOS-based computers.

## What will the coherent optical computing architecture look like?

The idea is to be able to use essentially a unified HPC application programming interface to specify what your problem is and then try it out on several different accelerators with different solvers and optimizers to see which one for your problem is the most efficient in time and energy.

It might be more expensive, but what you really care about is time-to-solution.

What we want to give you is a portfolio of choices. What you have right now are no choices. Your problem either can be solved on a supercomputer with a particular interconnect architecture, or it cannot. What has to happen as optical scientists is that we must focus on solving particular classes of problems and do so at extremely low energy. This is the origin of accelerators, and accelerators are the beginning of the future.

# Learn more at

hpe.com/us/en/hewlett-packard-labs



a00128626enw

Hewlett Packard Enterprise