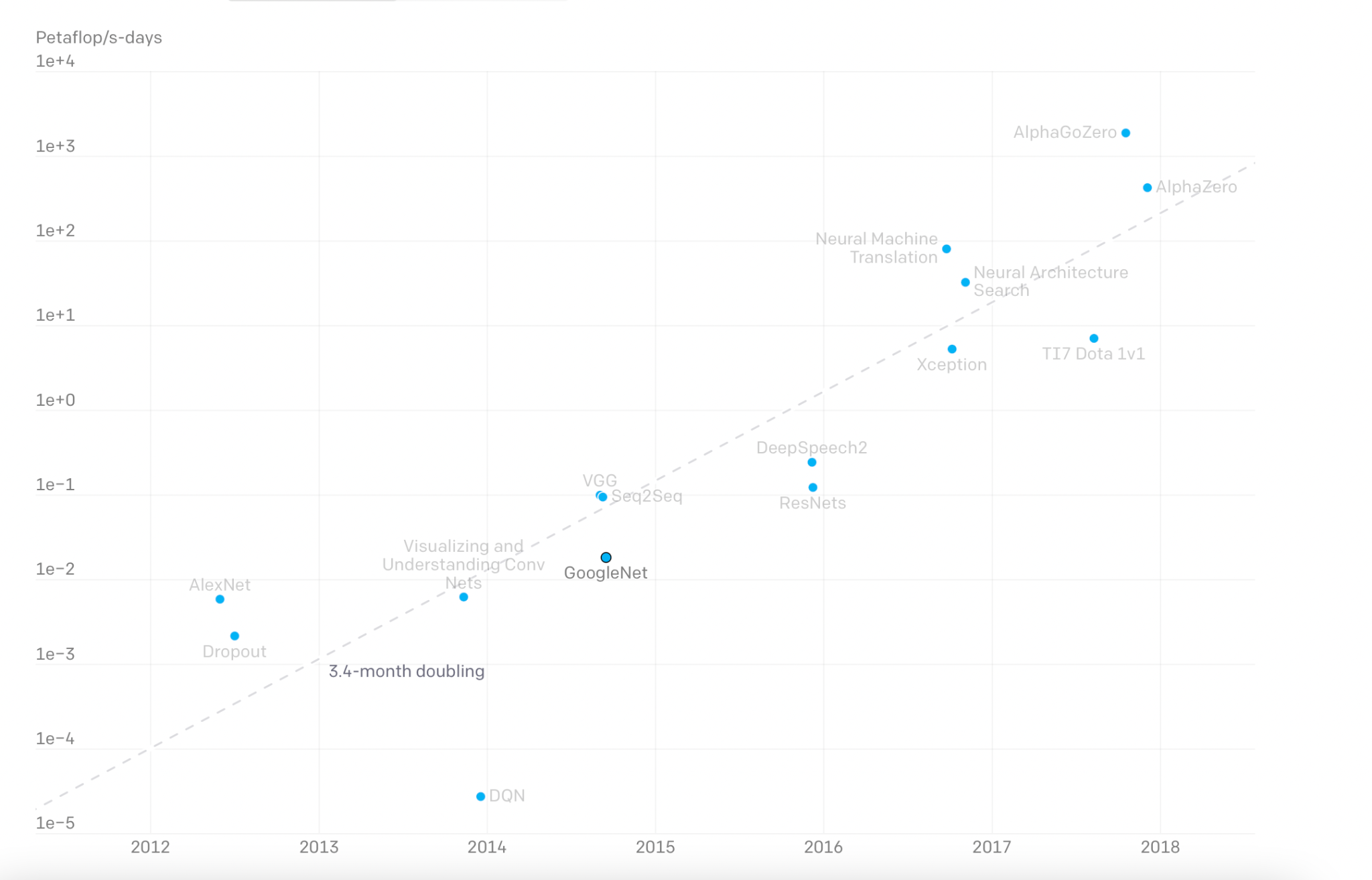
Sam Gulinello

EC 601

Project 1

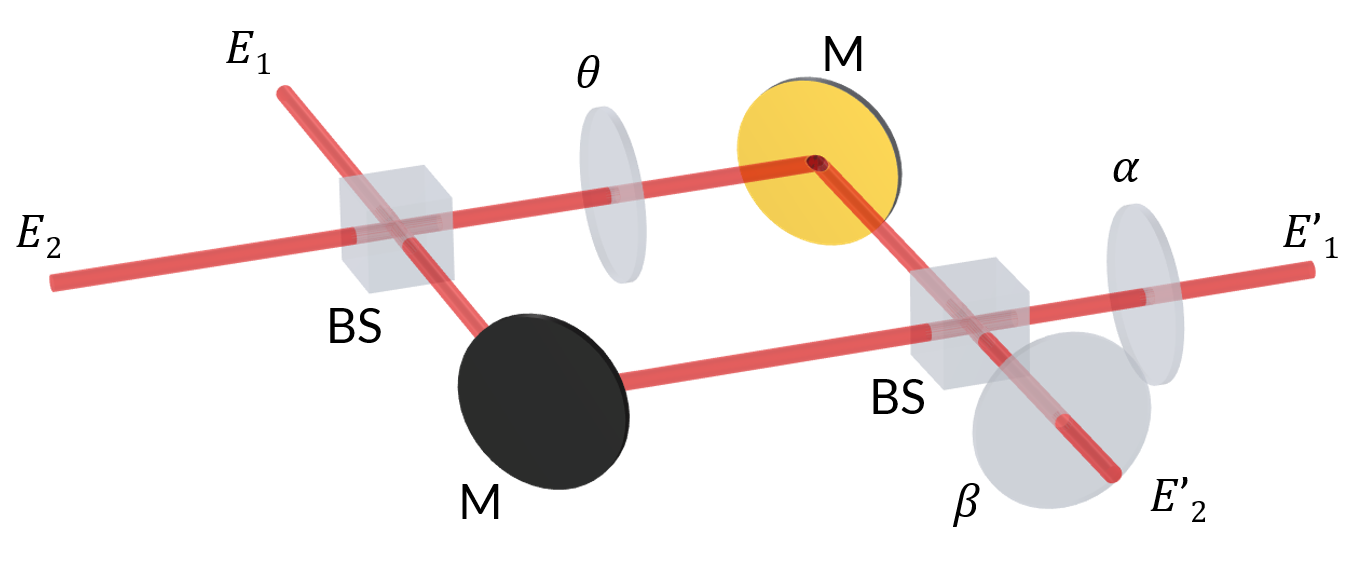
Photonics for Machine Learning

In recent years, the applications of Machine Learning (ML) have grown at a very rapid pace and it is easy to admit that it is here to stay. As the popularity and uses of ML increase the demand for powerful computing systems rises as well. This rise in computational demand has outpaced Moores Law, forcing industries to adapt and create new algorithms and hardware that accelerate the training of complex computing systems.

The right graph shows just how demanding the most recent ML applications have become. Shown is a consistent increase in PetaFlops/Days over the course of about 6 years. The graph is slightly outdated but the upward trend shown has continued and is expected to keep rising.

Historically companies turned to GPU’s to train their ML models because of their processing power. This approach worked for a while but eventually needed some help. In 2013 Google introduced a product called the *“Tensor Processing Unit” (TPU).* This device was designed to accelerate the process of training large ML models by handling the demanding linear calculations associated with these networks. It is designed to solely perform matrix multiplications at a very fast rate by using specialized processing pipelines. It is specifically the matrix multiplications that demand the large quantities of computing power. The general purpose design of the GPU won’t cut it and AI developers needed racks of TPU’s to train their models.

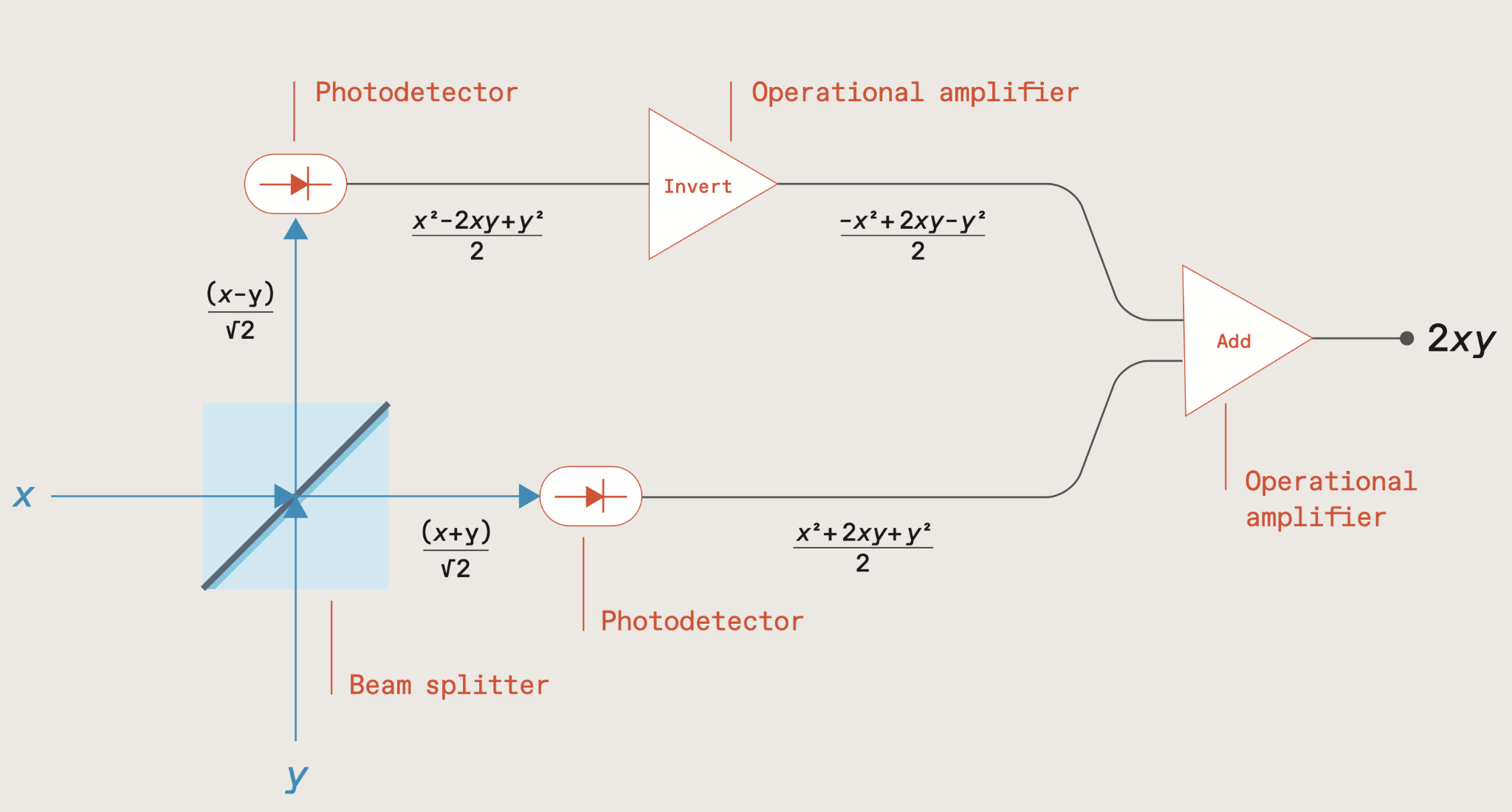
While TPU’s can successfully process data much faster than other technologies it comes at the cost of energy consumption. The amount of energy consumed has both financial and environmental impacts that need to be addressed. Smart algorithms and system design can only solve the problem of power consumption so much. The issue lies with the fundamental building block of these devices, the transistor, which relies on electrical current to function. Generating this necessary current to operate the transistors not only requires large amounts of power but also creates large amounts of heat in the process. In an attempt to solve this power issue, scientists have turned to the physics of photonics to create more efficient and powerful computing devices to accelerate the training of large ML networks.

The optical component commonly used to create these processors is called the Mach Zehnder Interferometer (MZI). The MZI is made up of two beam splitters as well two mirrors and three phase shifters (pictured below).



The MZI works by feeding two beams of light (E₁, E₂) into the first beam splitter. This splitter has a 50:50 ratio of splitting the beams creating an equal output on either end. These beams go through a series of phase shifters () and meetup at the second 50:50 splitter to again produce two beams of light (E’₁, E’₂). Through some complex linear algebra a relationship forms between the value of the phase shifting and the values of E₁, E₂ and E’₁, E’₂. This relationship can then be mapped to a multiplication result. Because it is done with photons, this calculation occurs essentially at the speed of light. A processing chip can be constructed of many MZIs to make up the ML network. The values of θ, ⍶, and ꞵ can be controlled by an electrical voltage across the phase shifter. Signals representing matrices can be sent to the chip and the output can be recorded and converted to a digital signal.

Another way to perform multiplication using light is with OpAmps, and Photodetectors instead of Phase Shifters. With this approach a beam splitter is still used initially but the two beams of light are measured for their electric fields. One beam's electric field value is inverted using an OpAmp and added to the other beam's electric field value. The sum of those values is proportional to the product of the intensities of the original beams of light. Similar to the approach of using MZI's, once the optical circuit can produce products at the speed of light, the necessary matrix operations can be done much faster.



Two companies in the Boston area are devoted to developing this technology. Lightmatter and Lightelligence are both start-ups that have successfully used the Mach Zehnder Interferometer technology to process ML data with light. Both companies are active in publishing their experiments promoting a great collaborative environment in this field. Along with their research they also touch on some of the limitations of using photonics and the challenges both companies face. The first and rather prominent draw back to a photonic chip is that it only really excels at performing linear calculations. It is designed to do matrix multiplication and doesn’t do non-linear functions well. This has resulted in the chips being more of a “hybrid” design where photonics chips are being integrated with digital circuits and more traditional processors. Another issue with the photonic devices is accuracy. Due to the quantum nature of the light particles it is not always very accurate when converting the light signal back to digital. Companies are working with the light properties to make them more detectable and less prone to measurement errors. The last pressing limitation of photonic chips is their size. Compared to the bottom limits of a transistors size the photonic elements are very large. This has resulted in a balance between chip size and performance. Google’s TPU is able to pack a large amount of processing power in a small device. A photonic device of the same size would be far less powerful. Companies will need to find a way to scale the technology while maintaining their power consumption advantages.

Along with solving the problems mentioned before, engineers will also need to create tools for developers to interact with these new technologies. Lightmatter CEO, Nicholas Harris, claims that their devices will be compatible with the most popular Machine Learning software libraries like PyTorch and TensorFlow but special compilers and firmware will need to be developed. Software will need to be what brings these photonic processors to life and integrates them with an already developed AI industry.

Photonic chips seem to have a very large social impact. Being able to accelerate ML training in a more sustainable way would not only advance the field of AI but also help in the fight against climate change. The promise of these photonic chips would have a great benefit to climate initiatives focused on energy conservation as well as appeal to large businesses trying to lower the cost of powering their supercomputers. This increase in efficiency will also aid in the development of society altering ML applications. The dreams of using ML to improve medical imaging and diagnostics, self-driving technology, language processing, image classification, etc. can all become a reality with the promise of an efficient way to accelerate the training of those large, demanding networks.

The field of photonics still has a lot of challenges to face before it can really disrupt the processor industry. The theory of using light to compute data efficiently has been proven but the problem that engineers face now is how to use that theory to create a truly superior type of processor. The promises of power and efficiency are what make the topic worth exploring. It seems the Computer Engineering industry is just at the beginning of realizing this new technology and the next decade should be interesting to see where it ends up.

Resources

<https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu>

[https://lightmatter.co](https://lightmatter.co/)

[https://www.lightelligence.ai](https://www.lightelligence.ai/)

<https://openai.com/blog/ai-and-compute/>

<https://news.engineering.pitt.edu/bringing-light-into-computers-to-accelerate-ai-and-machine-learning/>

<https://interestingengineering.com/innovation/transistors-moores-law>

<https://spectrum.ieee.org/the-future-of-deep-learning-is-photonic>

<https://medium.com/lightmatter>

<https://ieeexplore.ieee.org/document/1446917>

<https://iopscience.iop.org/article/10.1088/0031-9120/35/1/308/meta>

<https://www.comsol.com/blogs/analyzing-an-optical-computation-device-with-simulation/>

<https://www.nature.com/articles/nphoton.2017.93>

https://cloud.google.com/tpu/docs/intro-to-tpu#:~:text=intermediate%20calculation%20results.-,How%20a%20TPU%20works,neural%20networks%20at%20fast%20speeds.