The Dawn of Energy Efficient Computing: Optically Accelerating the Fast Fourier Transform Core

Iman Kundu, Edward Cottle, Florent Michel, Joseph Wilson, Nick New*
Optalysys Ltd., 8 Flemming Court, Wakefield WF10 5HW, United Kingdom
**E-mail address: nick.new@optalysys.com*

Abstract: We present a novel approach in optical processing by accelerating Fourier transform through integration of silicon photonics and free-space optics. We introduce Fourier transform based multiply and accumulate operations, that require less number of operations compared to the traditional multiply and accumulate operations used in computer architecture.

1. Introduction

Over the last decade, the wide adoption of artificial intelligence (AI) and machine learning algorithms has led to an increase in the demand for data processing. Optical computing has introduced a paradigm shift in the architecture of computers, where computation is performed by modifying the properties of light, rather than electrons [1,2]. Recent advances in silicon photonic (SiPh) manufacturing techniques permit integration of disparate photonic devices and transistors in the same die [3]. Most optical computing architectures are designed to perform multiplication and accumulation (MAC) operations that enable acceleration of AI tasks [1]. At Optalysys, we adopt a fundamentally different approach – instead of accelerating MAC operations, we exploit the instantaneous speeds and ultrahigh bandwidths offered by free space optics to accelerate optical Fourier transform (FT). Our FT technology naturally accelerates convolution operations, as well as correlation and poly-/multinomial products, which are reduced from quadratic to linear time complexity using spectral methods [4,5]. Thus, our technology is a hardware accelerator for many applications where MAC reduction can be achieved by manipulating data in the frequency domain.

2. Optalysys Etile

The Optalysys Etile is a SiPh based modular FT accelerator core with integrated mixed-mode driver circuits and a digital backend. In an Etile (Fig. 1a), coherent light from a single-mode laser is split into mulitple channels, each of which are amplitude modulated to encode a 4-bit unsigned number. The channels carrying the modulated light are then arranged in a 11x11 pixel grid and coupled into a free space optic module. Inside the free space optic module, the modulated light propagates through a lens which transforms the incident light and changes both amplitude and phase of the light. The FT light is then coupled back into the SiPh die, where the amplitude and phase information of the FT light is detected using coherent receivers. In a single clock cycle, the Etile processes a 11x11 matrix of 4-bit unsigned number and generates a 10-bit complex number (4-bit real, 4-bit imaginary, 2-bit sign).

A digital backend supports digital data communication between the Etile and a host. The modularity of the Etile design allows integration of four Etiles with our Echip Logic Core to form a multi-chip-module (MCM) solution, where the FT cores are connected to the logic core using high datarate die-to-die interconnects. In this MCM, a mathematical complex number is decomposed into four terms (signed real and imaginary numbers) and each decomposed term is sent to each Etile simultaneously. The FT of the complex number is processed digitally in the Echip to compute large FTs of any dimension and precision.

3. Results

The expected FT performance of the Etile is compared against two CPUs [6] and state-of-the-art NVIDIA GPUs [7]. Assuming a convention that a Fourier transform of size N corresponds to $5\,N\log_2 N$ operations, Figure 1b shows the number of operations per Joule of energy consumed. The Etile is about 50 times more efficient than the NVIDIA A100 GPU. Saliently, despite using a newer architecture, the NVIDIA A100 performs only marginally better than the V100. This illustrates the fact that progress on FT on electronic hardware is significantly slower than the evolution in raw matrix-multiply operations per second.

Functions like the convolution have linear time complexity in the spectral domain, as opposed to quadratic scaling in the spatial or temporal domain. In this way, with an appropriate spectral representation of an input and kernel function, a convolution of the two functions can be obtained via a simple point-wise complex multiplication followed by an inverse FT. Such an operation is what we describe as a Fourier multiplication. Many layers of multichannel convolutions are used to build a CNN. Typically, a 3-dimensional tensor is an input to a single multi-channel convolutional layer. A single channel of a 3-dimensional output tensor is formed by convolving each

channel of input with a kernel function and summing across the input channel dimension. The FT of the input can be taken at inference time, with FT kernels held in memory. Point wise complex multiplications take place for each input pixel with the corresponding kernel pixel, followed by a summation across the depth dimension, which we describe as the Fourier MAC operation. To achieve this efficiently in electronics, a typical MAC hardware can be used to build up the complex multiplications and summations. Finally, the inverse FT of each channel of the output tensor is taken yielding the output of the multichannel convolution layer.

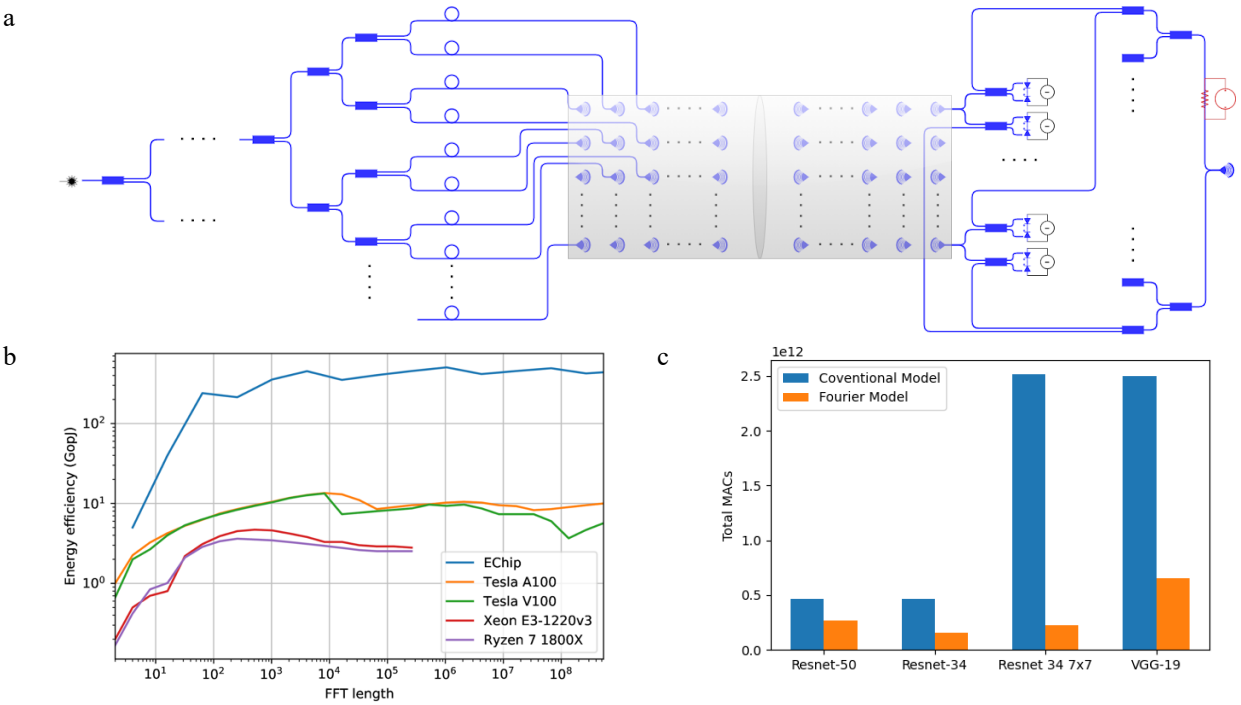

Fig. 1 (a) Schematic illustration of the SiPh architecture used in the Optalysys ETile. (b) Comparison of the predicted energy efficiency of the ETile against existing CPUs and GPUs. (c) Predicted reduction in MAC operations (with a batch size of 128) using the ETile across specific computer vision models.

Using our Fourier MAC formalism, fewer operations need to be performed in order to produce an output (Figure 1c). In the case of the typically used 3x3 kernels in deep learning, we observe a factor 4.5 reduction in total MAC operations that make up the layer. This is due to complex multiplication requiring 4 real number multiplications to attain, with a reduction factor of 2 of the number of elements requiring multiplication due to symmetry in the FT of real input matrices. FT of the 3-dimensional input/output tensors are taken at inference time, crucially the amount of data IO required to achieve this is modest compared to the total number of MACs required for the 4-dimensional multichannel convolution operation.

Conclusions

The Optalysys Etile accelerates one of the most demanding computing processes – the Fourier transform operation by integrating SiPh with free space optics. Furthermore, by introducing FT based MACs, we significantly reduce the number of operations, making the Etile a universal accelerator for any AI/big data hardware.

3. References

1. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441–446 (2017).
2. D. A. B. Miller, "Meshing optics with applications," Nat. Photonics **11**, 403–404 (2017).
3. C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, L. Alloatti, M. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin, B. R. Moss, R. Kumar, F. Pavanello, A. H. Atabaki, H. M. Cook, A. J. Ou, J. C. Leu, Y.-H. Chen, K. Asanović, R. J. Ram, M. A. Popović, and V. M. Stojanović, "Single-chip microprocessor that communicates directly using light," Nature **528**, 534–538 (2015).
4. George B. Arkfen and Hans J. Weber, *Mathematical Methods for Physicists*, Fourth Edition (Academic Press Inc., 1995).
5. Ronald Bracewell, *The Fourier Transform and Its Applications*, Second Revised Edition (Mcgraw-Hill College, 1986).
6. M. Frigo and S. G. Johnson, "FFTW: an adaptive software architecture for the FFT," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)* (1998), Vol. 3, pp. 1381–1384 vol.3.
7. "Fast Fourier Transforms for NVIDIA GPUs," Accessed: Apr. 26, 2021. [Online]. Available: https://developer.nvidia.com/cufft.