High-Performance Neuromorphic Computing Based on Photonic Technologies

P. Stark[1], J. Weiss[1], R. Dangel[1], F. Horst[1], J. Geler-Kremer[1,2] and B.J. Offrein[1]
[1]IBM Research – Europe, 8803 Rüschlikon, Switzerland
[2]EMPA, 3602 Thun, Switzerland
Author e-mail address: ofb@zurich.ibm.com

Abstract: Dedicated technology platforms gain interest for enhancing the performance and efficiency of neuromorphic computing. We demonstrate integrated optic devices for convolutional signal processing and neural network training.

1. Introduction

Recently, new computing paradigms gained tremendous interest. This trend originates from several developments, especially the slow-down in traditional computing hardware scaling and the increasing importance of processing unstructured data. While the human brain is capable of interpreting complex information, such as in image or speech recognition, at reasonable speeds (100's ms) and ultra-low power (20W), large scale and power-hungry computing systems are required to tackle such tasks. Brain-inspired architectural concepts regained attention, holding the promise to improve the power efficiency and performance of computing systems. A wide range of neural networking concepts has been explored, differing in architecture, signal transmission scheme, synaptic control, activation function and complexity. The performance of such networks has in many applications reached levels beyond human capabilities. However, these novel compute schemes are still executed on legacy hardware systems based on the Von Neumann architecture. Consequently, the signal processing and architectural properties known from the biological brain, such as dispersed compute kernels (neurons) linked through weighted connections (synapses) to a complex network, is only implemented at software level. Despite the existence of a wide range of neural networks, they all build on the same basic operations to infer the output result: the calculation of the weighted interconnects and the nonlinear activation function. Computing the interconnections in the network scales with the square of the number of neurons and constitutes the bulk of the compute effort. Accelerators such as graphic processing units (GPUs) and tensor processing units (TPUs) are designed to parallelize and optimize the underlying multiply & accumulate (MAC) operation.

2. Neuromorphic computing hardware

It is our goal to analyze the power drivers and performance bottlenecks in today's computing systems to overcome those, rather than realizing a fully new disruptive hardware platform, by extending the capabilities of today's silicon CMOS technology in a 'more than Moore' approach. In today's computing systems, the central processing unit (CPU) is connected to memory through a bus. Data is fetched from memory, processed in the CPU, and written back to memory. In general, the energy required to transfer the data exceeds that for processing. In addition, the data transfer itself poses a serious performance and latency bottleneck, making signal processing a serial operation. Furthermore, the handling of digital signals represents an overhead as multiple operations are required for processing all bits, representing the values involved.

We overcome these bottlenecks by employing calculation concepts incorporating 'in-memory computing', parallel operation and analog signal processing. Neural network accelerators based on these concepts will be integrated in legacy digital computing systems as dedicated high-performance engines with a focus on performing MAC and especially vector matrix multiplications, representing the synaptic operation in a neural network. Such engines were evaluated to provide performance and efficiency improvements up to several orders of magnitude compared to today's systems [1].

3. The prospects of photonic systems

Various technology platforms are under development for realizing the analog signal processing accelerators described above. Memristive structures integrated in the back-end-of-line of a CMOS process are an exciting candidate, providing high density and tight integration with all functions of the neural network [2]. A strong focus is on establishing the memristor technology with the desired specifications and stability, several concepts are under evaluation [3]. Here we focus on opportunities for applying photonic technologies [4-6].

Photonic signal processing structures exhibit some inherent properties, in line with the requirements described above. The multiply operation can be performed as a transmission or gain factor, a reflection or diffraction efficiency or a change of the phase of the optical signal. The accumulation function is obtained by collecting multiple optical beams in an output waveguide or on a detector. When coherent light is used, interference phenomena must be considered, which offer the additional advantage of complex valued signal processing. A wide range of systems have been implemented, ranging from 3D bulk optical systems based on spatial light modulators [7], fibers [8] and integrated optic structures [9]. The inherent parallel nature of optical systems, the ability to process ultra-high bandwidth signals and real-time operation, make photonic systems a unique platform for computing. Furthermore, non-linear optical materials enable ultra-fast and accurate control of the device architecture, for example adapting the synaptic weights of an accelerator engine in real time with the incoming data stream.

Whether these prospects make optics the candidate of choice depends strongly on the application. It is critical to target problems leveraging the described properties. A detailed system-level performance analysis is required, comparing the available technological options for making the final judgement. In the following, we will discuss some potential applications in convolutional neural networks and for neural network training.

3. Integrated optic convolutional signal processor

Integrated optical finite impulse response (FIR) filters, implemented as cascaded Mach-Zehnder interferometer structures, are well-known in integrated optics for a range of applications, such as tunable dispersion equalizers [10], add-drop filters or dynamic gain equalizers [11]. Here we demonstrate their application for convolutional signal processing exploiting the fact that an FIR filter represents a 1D convolution. Our convolutional optical processor is fabricated in silicon photonics technology. The tunable couplers are realized as balanced Mach-Zehnder interferometers and can be controlled by the voltage applied across the thermo-optic phase shifter in one of the arms. Figure 1a shows the device in a subassembly with the electrical connections for applying the control voltages. In Figure 1b, the operation of a two-stage device for edge detection is demonstrated.

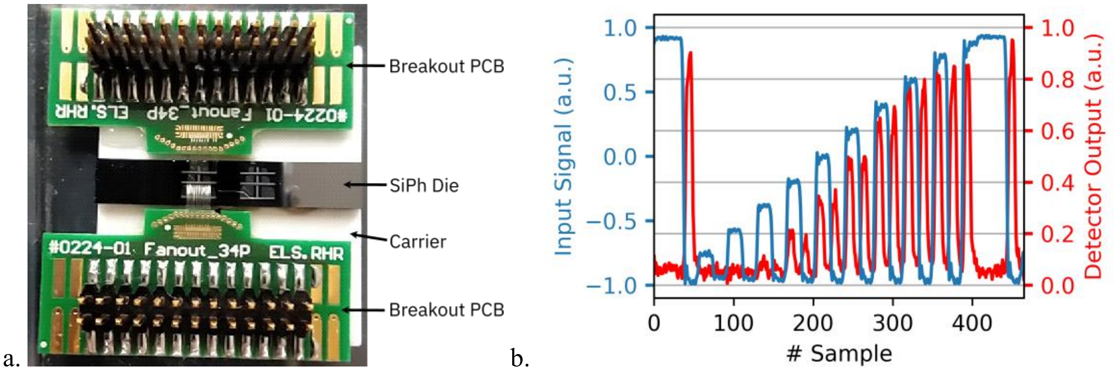

Fig. 1. Silicon photonics convolutional optical processor (a) and edge detection measurement (b).

This device is well suited as a vehicle to analyze the performance efficiency (Operations per Watt). We consider the power dissipation of the device itself as well as the power consumption of the peripheric infrastructure required to operate the device. These peripheric elements, such as the laser, the electronics to control the phase shifters, as well as the detector and transimpedance amplifier dominate the power metric. The total power dissipation is below 1W at an operating efficiency of 1.5 TOps/W. Here we assume power efficient bariumtitanate phase shifters as the Mach-Zehnder tuning elements [12], also a non-volatile optical phase shifting element based on this technology could be applied [13]. The operating efficiency is in the same range as today's top performing digital systems such as the GPU or TPU. By moving to more efficient integrated elements, there is room for further improvement. This example shows that photonic devices might be able to outperform complex digital systems for specific tasks. However, the photonic devices will have a challenge in addressing general purpose operations as the device size and non-negligible propagation losses limit the functionality to be integrated on a chip.

4. Synaptic optical crossbars for neural network inference and training

Crossbar structures are an efficient means for interconnecting two activation layers in a neural network [1]. Photonic crossbar arrangements were for example implemented as discrete waveguide arrays interconnected by a transmission tuning element [14]. The transmission values are mostly set or adapted through an external signal. Such structures are well suited for neural network inference.

By exploiting the parallelism of optics, other types of crossbar structures based on diffraction can be realized [4]. A diffractive structure is established in a photorefractive material through the interference of optical beams. The interference pattern generates free carriers in the photorefractive material, in the high intensity regions, that diffuse and get trapped in regions with low intensity. The charge separation in the photorefractive layer is turned into a refractive index corrugation through the Pockels effect [15] and remains available over the carrier decay time. During this period, the applied diffraction pattern can be read out by an optical beam. Multiple of such corrugation patterns can be written in the layer, enabling vector matrix operations with multiple input and output beams. As the performance of such a crossbar system scales quadratically with the number of inputs, large performance enhancements can be obtained. A distinct advantage of using a photorefractive material as the weight storage medium is the ability to accurately control the stored values by applying the appropriate optical interference pattern and energy. This is a critical feature for neural network training, making such structures suited for inference and training of neuromorphic architectures.

5. Conclusions

Integrated optic devices offer opportunities for executing functions in neuromorphic computing. The ability to process ultra-high-speed signals, low latency and parallelism make photonics of interest for dedicated applications, as accelerators in larger networks. A holistic system-level analysis, taking the performance, power and interface requirements into account, is essential for obtaining insight in the areas where optics can make a difference.

6. Acknowledgements

The authors gratefully acknowledge the IBM BRNC operations team for their valuable support. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 828841 (ChipAI). This study was partially funded through the Swiss National Science Foundation grant no. 175801 Novel Architectures for Photonic Reservoir Computing.

7. References

[1] T. Gokmen and Y. Vlasov, “Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations,” Front. Neurosci., vol. 10, Jul. 2016, doi: 10.3389/fnins.2016.00333.

[2] G. W. Burr et al., “Recent Progress in Phase-Change Memory Technology,” IEEE J. Emerg. Sel. Top. Circuits Syst., vol. 6, no. 2, pp. 146–162, Jun. 2016, doi: 10.1109/JETCAS.2016.2547718.

[3] D. Ielmini and H.-S. P. Wong, “In-memory computing with resistive switching devices,” Nat. Electron., vol. 1, no. 6, pp. 333–343, Jun. 2018, doi: 10.1038/s41928-018-0092-2.

[4] P. Stark, et al., “Opportunities for integrated photonic neural networks,” Nanophotonics, vol. 9, no. 13, 2020, pp. 4221-4232. https://doi.org/10.1515/nanoph-2020-0297

[5] Shastri, B.J., Tait, A.N., Ferreira de Lima, T. et al., “Photonics for artificial intelligence and neuromorphic computing”, Nat. Photonics 15, 102–114 (2021). https://doi.org/10.1038/s41566-020-00754-y

[6] A. R. Totović, et al., “Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 5, pp. 1-15, Sept.-Oct. 2020, Art no. 8800115, doi: 10.1109/JSTQE.2020.2975579.

[7] H. J. Caulfield and S. Dolev, “Why future supercomputing requires optics,” Nat. Photonics, vol. 4, no. 5, pp. 261–263, May 2010, doi: 10.1038/nphoton.2010.94.

[8] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, “Parallel photonic information processing at gigabyte per second data rates using transient states,” Nat. Commun., vol. 4, pp. 1364–1367, 2013, doi: 10.1038/ncomms2368.

[9] A. N. Tait et al., “Neuromorphic photonic networks using silicon photonic weight banks,” Sci. Rep., vol. 7, no. 1, p. 7430, Dec. 2017, doi: 10.1038/s41598-017-07754-z.

[10] F. Horst, R. Germann, U. Bapst, D. Wiesmann, B. J. Offrein and G. L. Bona, "Compact tunable FIR dispersion compensator in SiON technology," IEEE Photonics Technology Letters, vol. 15, no. 11, pp. 1570-1572, Nov. 2003, doi: 10.1109/LPT.2003.818671.

[11] B. J. Offrein, F. Horst, G. L. Bona, R. Germann, H. W. M. Salemink and R. Beyeler, "Adaptive gain equalizer in high-index-contrast SiON technology," IEEE Photonics Technology Letters, vol. 12, no. 5, pp. 504-506, May 2000, doi: 10.1109/68.841267.

[12] S. Abel et al., “Large Pockels effect in micro- and nanostructured barium titanate integrated on silicon,” Nat. Mater., vol. 18, no. January, 2019, doi: 10.1038/s41563-018-0208-0

[13] S. Abel, et al., "Multi-Level Optical Weights in Integrated Circuits," 2017 IEEE International Conference on Rebooting Computing (ICRC), Washington, DC, 2017, pp. 1-3, doi: 10.1109/ICRC.2017.8123672.

[14] Feldmann, et al. “Parallel convolutional processing using an integrated photonic tensor core”, Nature 589, 52–58 (2021). https://doi.org/10.1038/s41586-020-03070-1

[15] C. Denz, Optical Neural Networks. 1998.