Project Acronym: **Fun-COMP**

Project Title: **Functionally scaled computing technology:** From novel devices to non-von Neumann architectures and algorithms for a connected intelligent world

WP4

Memcomputing with N-vN Devices and Networks

(Leader IBM)

Deliverable D4.4: (Mixed Precision) **Memcomputing: A new approach for fast energy-efficient computation**

Deliverable ID: D4.4

Deliverable title: (Mixed Precision) Memcomputing: A new approach for fast energy-efficient computation

Revision level: Final

Partner(s) responsible: IBM

Contributors: IBM (Syed Ghazi Sarwat, Manuel Le Gallo, Abu Sebastian), OXFORD (Harish Bhaskaran & team), MUENSTER (Wolfram Pernice & team), UNEXE (David Wright & team)

Dissemination level: PU[1]

[1] CO: Confidential, only for members of the Fun-COMP consortium (including the Commission Services); PU: Public.

Summary

The world is generating exponentially increasing amounts of digital information- each day 2.5 quintillion bytes of data are created, captured, and consumed, and this number is only projected to grow with the proliferating internet-of-things, and the emerging artificial intelligence. The data needs to be processed fast and efficiently, and has necessitated new thinking about how our computers must fundamentally work. In FunCOMP, we are working towards overcoming the fundamental limitations of our conventional sequential electrical computing by designing a parallelized photonic in-memory computing (or memcomputing) architecture using phase-change memory arrays and coherent on-chip frequency combs. We harness the distinct property of light: wavelength division multiplexing, to execute matrix-vector multiplication operations in parallel and experimentally demonstrate a photonic tensor core capable of operating at Tera-Multiply-Accumulate (TMAC)/s speeds. In this report, we discuss our photonic memcomputing architecture, and we showcase a demonstrator example of a highly energy-efficient and fast convolutional processing engine for image detections in demanding AI applications.

Contents

1. Introduction and background

In the early 1900s, Alan Turing conceptualized with adequate mathematical proof a hypothetical system he called the universal computing machine. What he was alluding to at that time is the present-day computer: a machine by definition capable of storing, communicating, and computing data. In their earliest implementations, computers executed fixed algorithms, performing simple calculations, and lacked the ability to be restructured for dynamically changing and diverse tasks. As the demand for computation and reprogramming grew, came the concept of a von Neumann architecture, proposed by John von Neumann, a decade after Turing's proposal. The architecture physically separates the computational, or the central processing unit (CPU), from the memory unit, so enabling re-programmable logical and computational operations. The very requirement to shuttle data back and forth between the CPU and memory incurs tremendous costs on energy, and latency (delays), which in the recent past have turned into troublesome computing bottlenecks. The 'memory wall' bottleneck, for example, is a result of the finite bandwidth at which data can be moved across the interconnects and be accessed at the memory. This means even with the fastest imaginable processors, the computation throughput would decisively be governed by the data transfer speeds, resulting in the processor spending a lot of time being idle. This disparity between the data demand and supply at the CPU, defined by bits per second, is only widening as processors are getting faster, and while many computer performance metrics have exponentially improved in the 40 years of the semiconductor technology, the latency that defines the time it takes to access and compute data from the memory has improved far less dramatically. Similarly, the energy expenditure on accessing and shuttling data has become strikingly large compared to energy spent on the computation itself.

Motivated to overcome these growing limitations, in the Fun-COMP project, we have set out to create novel computer architectures that can radically redefine computing from being compute (processor) centric (the von-Neumann architecture) to data (memory) centric. Our approach also substitutes electricity with light, such that information gets represented, transferred, and computed as optical signals, in a framework that we refer to as *photonics memcomputing*. Such a scheme provides extremely low latencies and a very large computational throughput, the latter a result of the ability to use wavelength division multiplexing (WDM) to carry out parallel computations. Our research importantly leverages the recent advances in integrated photonics, including hybrid integration of soliton microcombs at microwave line rates, ultra low-loss silicon nitride waveguides, and high-speed on-chip detectors and modulators. To that end, we have demonstrated a highly parallelized, fast, and scalable integrated photonics accelerator – a form of photonics co-processor (or photonics tensor processor core – TPU). Such first-of-its-kind hardware can be considered as the optical analog of an application-specific integrated circuit, that is capable of operating at speeds of trillions of multiply-accumulate (MAC) operations per second (10^{12} MAC operations per second or tera-MACs per second). We will briefly describe this technology in this article.

Note also that because of the impact and technological significance our Fun-COMP photonic memcomputing technology for fast-emerging AI applications, we brought together a round-table of leading experts to critically discuss these advances, and to draft a research road-map. What this naturally resulted in is a cross-disciplinary engagement to communicate the ideas and discoveries to the wider public, and research audience, in the format of a comprehensively detailed book. In this book, which is due for publication in late-2021, we take a ground-up approach in elaborating the concepts, from device fundamentals to systems design. Details of the book will be posted in the Fun-COMP website (www.fun-comp.org) post publication

2. Discussion

Photonic Memcomputing:

Our all-optical information processing benefits from all-optical memory solutions that do not require detours through electronic circuitry, thus eliminating energy-inefficient electro-optical conversion operations. To that end, we utilize phase-change materials (PCMs) for photonic memory elements. PCMs provide strong optical contrast in the refractive index when reversibly switched between the amorphous and crystalline phase states, representing the logical states. In the crystalline PCM state, most of the incoming light is absorbed, representing for example a "0". In the amorphous state, most of the light is transmitted, thus representing a "1". Intermediate transmission states can be chosen by controllably switching fractions of amorphous and crystalline material in the PCM cell. The switching process can be induced with optical laser pulses on a picosecond timescale and thus allows for ultrafast operation of PCM-photonic devices. The change in refractive index is a broadband optical property of PCMs, and therefore such memory elements can be addressed in a wide wavelength range. In particular, this includes the 1500-1600 nm IR range and so makes the integration into the silicon photonics platform as a building block feasible and highly attractive. This enables the PCM-based photonic memories to be directly operated with other required active elements (lasers, optical amplifiers, modulators, detectors) in the form of a scalable photonic integrated circuit (a PIC).

Such PCM-memories allow for implementing cumulative data storage and thus are highly attractive for realizing in-memory computing applications. Particularly attractive about this concept is the possibility to transfer computationally expensive operations from the electronic domain to the photonics domain, where ultrafast modulation speeds and reduced latency are readily available. This concerns especially essential computing operations in artificial neural network implementations and which limit the system performance in terms of speed and energy efficiency - such as matrix-vector multiplications (MVMs). Additionally, because of the broadband operation window, such PCM memory cells can be combined with wavelength division multiplexing (WDM) strategies. This feature allows parallel access in the frequency domain as a route for upscaling both computational capacity as well as memory access. This way the same memory cell can be read out in parallel at multiple wavelengths, and these parallel signals can act as inputs to photonic neuromorphic hardware. This is especially attractive for realizing ultrahigh bandwidth MVM systems in a fully integrated format. Such systems can be employed for range of applications that fall under the categority of mixed-precision computing. One such application is solving systems of linear equations $Ax = b$. Here an iterative refinement algorithm is used, whereby an initial solution is chosen as the starting point, and is iteratively updated with a low-precision error-correction term, z. The error-correction term is computed by solving $Az = r$ inexactly, using the residual $r = b\text{-}Ax$, calculated with high precision. Another arguably more impactful application domain is deep learning inference where the MVM systems are used to implement (albeit with reduced precision) the MVM operations associated with each synaptic layer in a deep neural network (DNN). The other operations such as nonlinear activations are implemented in high precision in a digital processor.

A particularly attractive DNN for mixed-precision deep learning is convolutional neural networks (CNNs) (*Nature* 589, 52–58 (2021)). CNNs are highly effective for applications such as image classification, autonomous navigation, and audio analysis in the frequency domain. In state-of-the-art CNNs, many convolutional "hidden layers" are applied to an input signal before feeding the processed data to fully connected layers for classification (i.e. for ouput generation). Each of the convolution layers takes in an input image, performs convolutional operations (with a filter or 'kernel') to extract features, and generates an output image. To perform each convolution operation, a filter is passed over the input image inspecting a small window of pixels at a time. These filters are relatively small

even for state-of-the-art CNNs, and hence an MVM photonic processor based on photonic PCM-cells was developed to implement these convolution layers in a CNN. Importantly, through WDM capability, we are able to efficiently parallelize several convolution operations within the confines of the same Fun-COMP photonic processor, thus achieving significant savings on computational time and space complexities (see **Figure 1**).

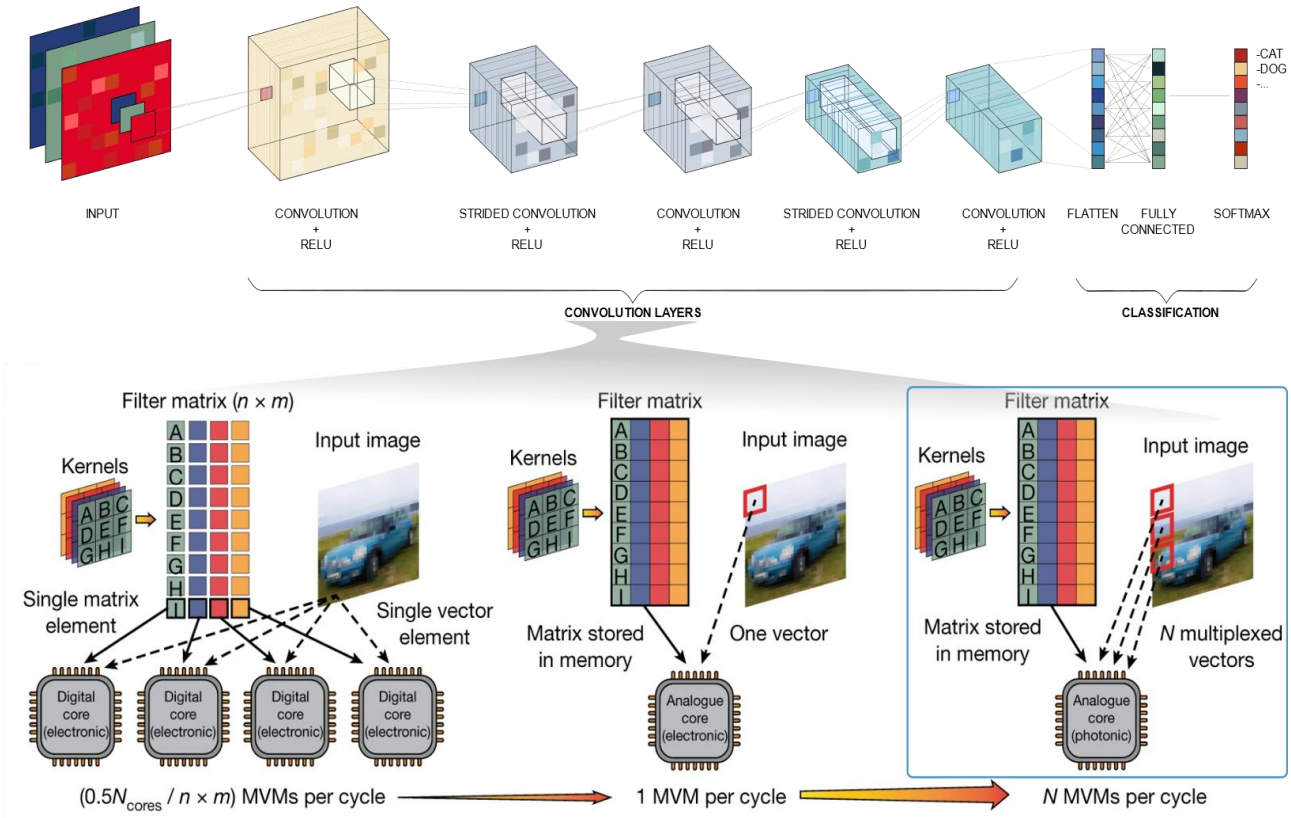

Figure 1. An Illustration of a convolution neural network, with multiple densely connected convolution layers. Each of the convolution layers takes in an input image, performs convolutional operations (which are matrix-vector or MAC operations) with pre-defined filters to extract features, and generates an output image to fully connected layers for classification. The bottom panel is a comparison of digital and analog electronic architectures with our photonic tensor core architecture, for purposes of MAC operation. Digital electronics (left) require many sequential processing steps distributed across multiple cores to compute convolutional operations on an image, whereas an entire MVM can be performed in one step using analog electronic in-memory computing (center). Photonic in-memory computing (right), using wavelength multiplexing as an additional degree of freedom, enables multiple MVM operations in a single time step, alongside the data-transfers at the speed of light.

Key Results:

In our memcomputing based all-optical neural network, the synaptic matrix weights are implemented in the PCM-memory cells, based on $Ge_2Sb_2Te_5$ chalcogenide phase-change material. The memory cells are operated in a transmission modulation mode, where the optical output power is regulated in a non-volatile manner depending on the PCM structural state (amorphous-crystalline volume fraction), where the PCM cells are optically programmed with high precision for differing transmission states. By arranging multiple PCM memory cells in an interconnected matrix form, the photonic analog of electronic crossbar array, or matrix, is realized on-chip (see **Figure 2**a). The matrix is designed as a waveguide crossbar array with directional couplers that equally distribute the input power to all PCM-cells. By using different wavelengths, interference inside the waveguides can be avoided and the summation of the individual products (of the matrix-vector multiplications) can be performed by adding

the light to the output waveguides, also by using directional couplers. This way the optical power at the output comprises the input optical power weighted by the programmed states (levels) of the PCM-memory cells, thus directly encoding the computationally expensive multiply-accumulate (MAC) operation (that lies at the heart of convolution processing) in the optical domain. We verified the applicability of such a concept using a suite of simulations and experimental trials, before arriving at an optimal design. The photonic circuits were fabricated using a three-step electron-beam lithography process on a silicon nitride on silicon oxide on silicon wafer. The complete circuit was designed using FDTD method and GDShelpers, a design framework for integrated photonic circuitry. The key chip regions are magnified in the optical micrographs shown in Figure 2b. The coupling of light into the optical chip is achieved using broadband total internal reflection (TIR) couplers. The TIR couplers provide access to a wide wavelength spectrum and thus allow the coupling of multiple wavelengths into the chip. The PCM-cells (of area 3×3 μm^2) acting as the matrix elements are deposited on top of waveguide crossings. Each matrix cell can be optically switched for programming each matrix element (in this case the light is coupled to the chip using Bragg-grating couplers because operation at a single wavelength (1550 nm) is sufficient).

In a conventional CNN, to a convolution layer input images of dimension $n \times n$ and d_{in} channels are fed, where the channel may represent the color tones (red, green, and blue). If there are d_{out} convolution kernels of size $k \times k$, each convolutional filter is of dimension $k \times k \times d_{in}$ and the resulting output image is of dimension $(n-k+1) \times (n-k+1)$ with d_{out} channels. To perform each convolution operation, a filter is passed over the input image inspecting a small window of pixels at a time. A pixel-wise MAC operation between the filter and the current filter window is carried out to calculate a single pixel of the output image. This corresponds to $(n-k+1)^2 \times k^2 \times d_{in} \times d_{out}$ MAC operations per convolution layer, leading to a significant computational bottleneck. In order to build efficient hardware to perform the convolution operations, one approach is to combine all the convolutional filters into a large filter matrix. The filter matrix will be of dimension $(k^2 \times d_{in}) \times d_{out}$. It is constructed by stacking the kernel matrices into the columns of the final filter matrix. In the same way, the pixels of the input image are rearranged by stacking the pixels of the filter volume, $(k \times k \times d_{in})$, into the rows of the input matrix. Hence a single convolution operation involves $(n-k+1)^2$ such MVM operations between the filter matrix and the input vectors of $k^2 \times d_{in}$ dimension. In the electronic domain, these MVM operations are typically multiplexed in time with parallelization afforded only by physically replicating the filter matrix (i.e. physically replicating the hardware). In this work, we exploit the optical WDM to overcome this fundamental limitation by encoding multiple input vectors of dimension $k^2 \times d_{in}$ onto multiple lines of a coherent frequency comb. These optical input vectors can then be applied to a single $(k^2 \times d_{in}) \times d_{out}$ filter matrix simultaneously, thus eliminating duplicated physical hardware and sequential operations.

To illustrate the convolutions in the optical domain experimentally, we show in Figure 2c examples of processing four input vectors in parallel, for edge detection. In this case, four pixels of the input image (in the top panel, the Waterloo underground station logo, and in the bottom panel, a Zebra) are obtained per image kernel simultaneously, therefore shortening the processing time by a factor of four. The kernel size used for this experiment is 2×2 and the input dimension of the image, $d_{in} = 1$, leading to a 4×4 filter matrix. The convolutions highlight the different edges (orientations (horizontal/vertical)) which can be seen, for example in the representation of the bricks in the Waterloo Logo. We also experimentally performed the task of MNIST (a commonly utilized benchmarking database) digit recognition with a CNN, as illustrated in Figure 2d. The CNN employed in our experiments consists of the input layer taking the pixel data (28x28 pixels, single-channel) that is then passed to a convolution layer consisting of four 2x2 kernels plus subsequent Rectified Linear Unit (ReLU) activation, resulting in an output of dimension 27x27x4 (valid padding). The output from the

convolution step is flattened and fed to a fully-connected layer with ten neurons. The probabilities for every digit are obtained from the final classification using the softmax function. The network was trained via software and the weights of the filter kernels were programmed to the states of the PCM-cells in the on-chip matrix. The experimental implementation of the CNN reached an accuracy of 95.3% showing good agreement with the calculated prediction accuracy of 96.1%. To estimate the ultimate performance capabilities of the system, we explored the scaling capabilities in terms of matrix size, modulation speed, and the number of parallel vectors.

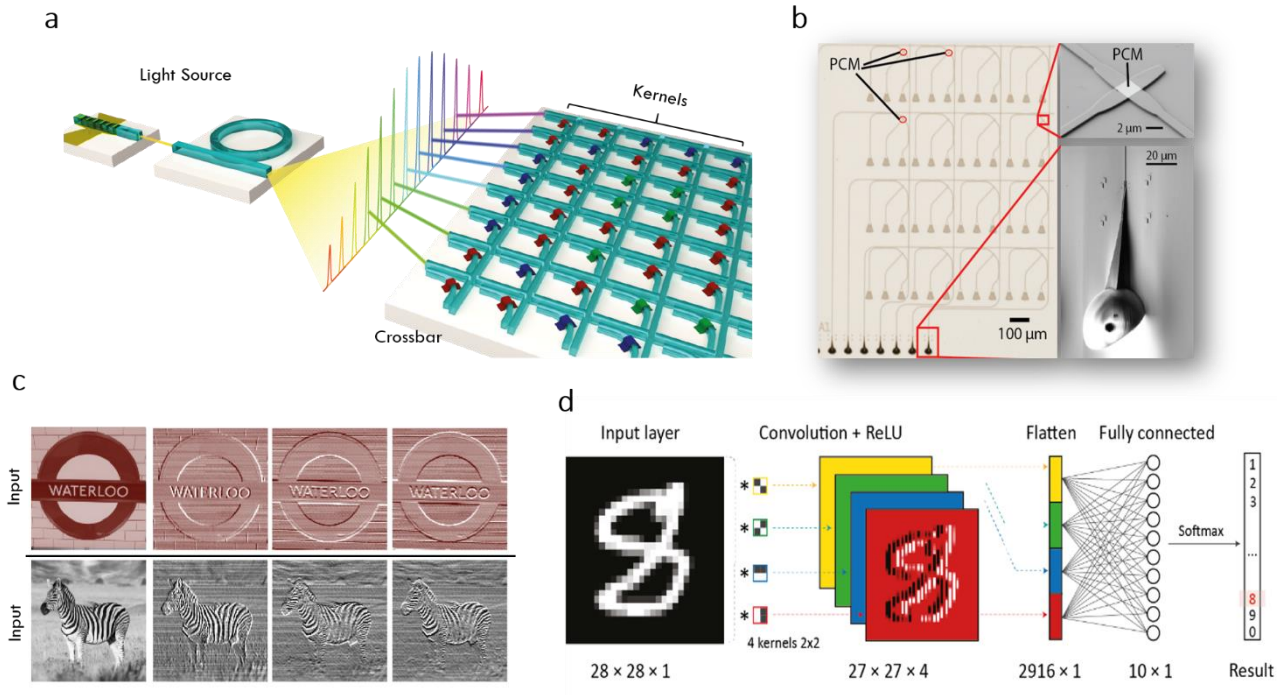

Figure 2. (a) Illustration of a FunCOMP photonic architecture to compute convolution operations. A laser generates a broadband frequency comb. Individual comb teeth that form the input vectors are modulated at high speeds, multiplied with a matrix (crossbar) of non-volatile phase-change memory cells, and summed along each column on a photodetector. Every column of the crossbar encodes a convolution filter or kernel. (b) Optical micrograph of a fabricated 4×4 matrix with 3D printed input and output couplers to enable broadband operation. The close-up SEM images on the right show the 3D printed couplers (bottom) and the waveguide crossings with the PCM (top) in more detail. (c) Convolution using parallel MVM operations. The original input images are shown on the left and to the right the output images using four different edge detection image kernels. The size of the four image kernels is 2×2 corresponding to a 4×4 filter matrix. In each time step, four input vectors are processed simultaneously via wavelength division multiplexing. (d) The layer structure of a convolutional neural network used to test the photonic tensor core with the MNIST database for digit recognition.

Benchmarking projections show that our Fun-COMP memcomputing hardware has a MAC rate that is (i) around ten to twenty times faster than modern electronic CPUs, GPUs, etc. (including ~30 times faster than Google's work-horse TPU), (ii) many orders of magnitude faster than electronic neuromorphic implementations (e.g. 1000 times faster than HICANN) and (iii) as fast as other photonic neuromorphic systems currently in development. Moreover, the energy consumed per MAC using the Fun-COMP approach is already 'best in class' (e.g. at 0.04 pJ some 500 times less than state-of-art GPUs and 5000 times lower than HICANN) and the computational density is (in spite of the relatively large size of integrated photonic components) also projected to be better than all other relevant technologies. These are some remarkable improvements, given this is only the first instance of our demonstration of this exciting new technology. More importantly, such an approach more broadly

suggests that integrated photonics are coming of age and in some cases can begin to match and even challenge electronic computation.

3. Conclusions and next steps

The memcomputing technologies developed in the Fun-COMP project have the potential to perform data processing with orders of magnitude higher speed than any other state-of-the-art techniques, thus achieving (even exceeding) the goals set on the roadmap set for AI hardware. Our demonstrations of convolutional processing on a photonics memcomputing engine are an exemplary showcase of such technologies, that show promise to remove the computing bottleneck in modern machine learning applications. There are, as with any promising technology however, important challenges to discuss and tackle. For example, neural network based accelerators commonly have hundreds of millions of parameters, thus for photonic technologies to be commensurate, the scaling challenge must be overcome (i.e. large-scale PCM-based PICs need to be designed and fabricated). Similarly, R&D directions must be laid in solving input-output challenges, which will require efficient on-chip electro-optical converters, temperature controllers, and power supplies. This will most likely be enabled by the co-integration of PICs with CMOS. Newer optical components and materials will be needed too, both for more efficient processing, as well as for more specific data-processing units, such as non-linear thresholding for fully-connected layers. Among other things, these are some very interesting aspects that we aim to explore and discuss in the remainder of the Fun-COMP project. We will also work toward conceiving system-level architecture based on Fun-COMP MVM photonic processors to execute end-to-end AI workloads. We will more rigorously benchmark the figures-of-merits, and henceforth optimize the various variables in achieving the optimal performance in data processing speed, bandwidth, and computational energy, first through a suite of simulations and then experimentally. We will also thoroughly assess the fabrication costs, scalability, and challenges for system deployment of optical memcomputing technologies in an industrial environment. Such an assessment will be done at the end of the Fun-COMP project. While we add to our demonstration of an optical convolutional neural network-based, we will also research newer concepts, tailored critically for such photonic memcomputing processors. An immediate research goal is to project a few important brain-inspired algorithms onto our hardware, for the task of sequential learning. We will aim to show how, much like the human brain, the photonic hardware could detect patterns and find correlations in live video processing, for important applications such as autonomous driving. Whilst engaged in achieving this research goal, we will also put together comprehensive reports (via journal and conference presenations, as well as Fun-COMP public deliverables) of the fundamentals, advances, and opportunities relating to our technology so that more and more research groups and industrial firms indulge. We hope that a concomitant slew of interest would result in research advances and discussions, that would help make this technology a commercial reality.