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 D4.1: Correlations in Big Data Extracted Using Computing In-Memory

Deliverable ID: D4.1

Deliverable title: Correlations in Big Data Extracted Using Computing In-Memory

Revision level: FINAL

Partner(s) responsible: IBM

Contributors: IBM (Syed Ghazi Sarawt, Manuel Le Gallo, Abu Sebastian)

Dissemination level: PU[1]

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

Summary

As computations are increasingly becoming more and more data-centric, in Fun-COMP we are developing novel computational architectures, namely computational memory, where the physics of phase-change photonic memory devices is used to perform certain computational tasks. In our previous reports and discussions, we illustrated how such photonic memcomputing architectures can enable the accelerations of convolutional neural networks for demanding AI applications. Here we discuss yet another computational paradigm, one that harnesses wavelength division multiplexing (a distinct property of light), and the accumulative property unique to phase-change materials (i.e. their progressive crystallization dynamics). We discuss that such a computational memory can be utilized for accelerating demanding statistical problems, such as correlation detections, for processing real-world data sets in big data and artificial intelligence.

Contents

1. Introduction and background

There is often a particular technology associated with a particular era and which helps to define the workings of our economy and society in that era. While today such technology can probably be identified as digital computing, it is widely suggested that in the coming years artificial intelligence (AI) will be the dominating technology theme. Artificial intelligence is a technology developed to smartly handle (make sense-out-of) and process the increasing amounts of structured and unstructured data, which our gadgets and scientific tools incessantly generate. Because such computations are extremely data-centric, it is becoming more and more clear that we must start to rethink how our future computers must work. In today's computing systems based on the conventional von Neumann architecture (see Figure 1A), there are distinct memory and processing units. The processing unit comprises the arithmetic and logic unit, a control unit, and a limited amount of cache memory. The memory unit typically comprises dynamic random-access memory, where information is stored in the charge state of a capacitor. Performing an operation (such as an arithmetic or logic operation), f, over a set of data stored in the memory, A, to obtain the result, $f(A)$, requires a sequence of steps in which the data must be obtained from the memory, transferred to the processing unit, processed, and stored back to the memory. This results in a significant amount of data being moved back and forth between the physically separated memory and processing units, and thus costs time and energy, which introduces a bottleneck in the achievable performance (the so-called von Neumann bottleneck). Thus, even with the fastest imaginable processors, the computational throughput would decisively be governed by the data transfer speeds, resulting in the processor spending a lot of time essentially 'doing nothing'.

To overcome the von Nuemann bottleneck, a promising prospect is of transitioning to a hybrid architecture where certain computational operations can be performed at the same physical location as where the data is stored (see Figure 1B and C). Such a memory unit that facilitates co-located computation is referred to as *computational memory*. The essential idea is not to treat memory as a passive storage entity, but to exploit the physical attributes of memory devices to realize computation exactly at the place where the data is stored. In Fun-COMP, we are taking a step further, and pursuing such non-von-Neumann or in-memory computer architectures, in the optics domain. Our approach is to substitute electricity with light, such that information gets represented, transferred, and computed as optical signals, and compute information with the same devices that store them. Such a scheme provides extremely low latencies (since data is shuttled at the speed of light), a very large computational throughput (resulting from the use of wavelength division multiplexing and in-memory computing), and near-zero power dissipation (since photons, are bosons).

More specifically, Fun-COMP research aims at creating a photonic engine that is targeted at accelerating statistical methods, such as temporal correlation detection, and clustering of correlated and uncorrelated random processes in continuous data streams. Correlation detection is at the core of computational methods that combine different multiple signals, and it is widely exploited across many applications, including the Internet of Things (IoT), life sciences, networking, large scientific experiments. One can view correlation detection as a key constituent of unsupervised learning, where one of the objectives is to

find correlated clusters in data streams. Our research leverages the recent advances in integrated photonics, including ultra low-loss silicon nitride waveguides, and high-speed on-chip detectors and modulators. We briefly summarize this technology in this article. Specifically, we discuss the use of the crystallization physics of non-volatile phase-change memory technologies for real-world correlation detection between data streams, such as those produced in social media platforms.

2. Discussion

Photonic Correlation Detection:

Fun-COMP's all-optical information processing benefits from all-optical memory solutions. We utilize phase-change materials (PCMs) for photonic memory elements. PCMs provide strong optical contrast in their refractive index when reversibly switched between their amorphous and crystalline phase states, which are also used to represent 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 (programmed) by controllably switching fractions of amorphous and crystalline parts in the PCM cell (i.e. via fractional, or partial, crystallization). 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 makes the integration of PCM devices into the silicon photonics platform both feasible and highly attractive.

For in-memory correlations, we exploit yet another unique dynamic behavior of phase-change memory, which is 'accumulation'- the dynamic evolution of the transmission levels in photonic devices upon application of a stream of optical signals. Depending on the operation to be performed, a suitable optical signal is applied to the photonic memory devices. The transmission of the devices evolves in accordance with the optical input, and the result of the computation is imprinted in the memory array. Note that this approach is different from schemes such as logical and matrix-vector multiplications, where photonic devices are generally programmed once before being used for processing application (e.g. inference), and do not dynamically evolve (in transmission) during use.

Because of the broadband operation window of PCM devices, 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 memory cells can be both readout and programmed in parallel on multiple wavelengths. Thus, at any time instance, many devices are dynamically changed, and the devices with similar changes can encode correlated events. In our Fun-COMP demonstrations, we focus on a specific aspect phase-change materials - their crystallization dynamics - which allows for the accumulative behaviour we require for correlation detection. These dynamics capture the progressive reduction (upon multiple excitations) in the size of the amorphous region due to the phase transition from amorphous to crystalline. When an optical pulse (typically referred to as a SET pulse) is applied to a photonic memory cell in the amorphous state, such that the temperature reached in the cell via photo-thermal effects is sufficiently high, but below the melting

temperature, a part of the amorphous region crystallizes. At the micro and nanometer scale, which is the case in Fun-COMP devices, the crystallization mechanism can be dominated by crystal growth, due to the large amorphous–crystalline interface areas. Thus, from a circuit-theoretic representation, a photonic memory cell device can be viewed as a generic memristor, with the size of the amorphous state serving as an internal state variable that is tunable.

Figure 1D-E sketches the correlation detection scheme using our Fun-COMP photonic computational memory. Here a process (denoted as P_n) defines the type of dataset, which, for example, maybe finance, weather station, or social media related, and events (denoted as E_n) represent the event-based binary (0 or 1) data streams, within each process. Take for example the problem of detecting correlations in rainfall across the various Cantons of Switzerland. Here, rainfall detection is a process, and events are the weather stations, which provide data-streams of either 1 or 0 for time stamps when there is rain and no-rain, respectively. Each event is assigned to a single photonic phase-change memory unit cell and some unique wavelength λ_n. Whenever the event takes the value 1, a SET pulse is applied to the corresponding cell, such that it partially crystallizes, and the optical

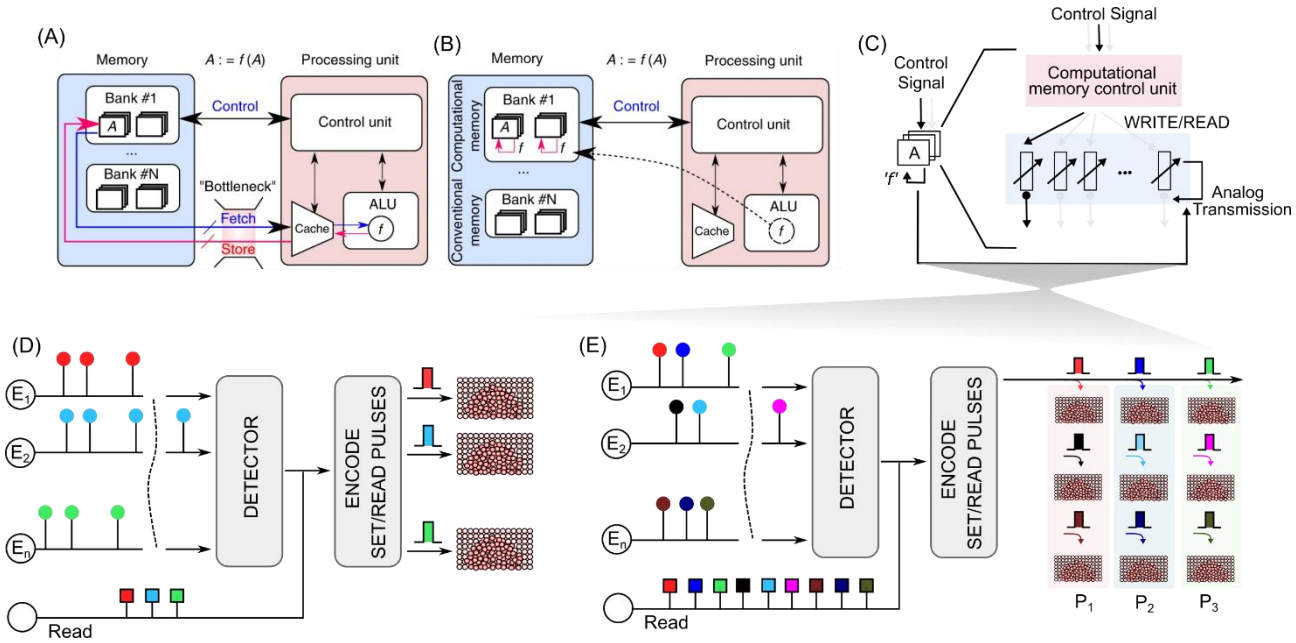

Figure 1. The concept of computational memory. (A) A sketch of the von Neumann computer architecture, where the memory and computing units are physically separated. *A* denotes information stored in a memory location. To perform a computational operation, *f*(*A*), and to store the result in the same memory location, data is shuttled back and forth between the memory and the processing unit. (B) An alternative architecture where *f*(*A*) is performed in place in the same memory location. (C) One way to realize computational memory is by relying on the state dynamics of a large collection of photonic non-volatile devices. Depending on the operation to be performed, a suitable optical signal is applied to the memory devices. The optical transmission through the devices evolves in accordance with the optical input, and the result of the computation can be retrieved by reading the transmission later. (D) Temporal correlation detection using wavelength division multiplexing of events. Multiple events in data streams can be programmed simultaneously into the transmission states of multiple devices. (E) Correlation detection across many data processes (sources) can be also performed by wavelength division multiplexing of these processes. Notably, the photonic implementation enables multiple correlation operations in a single time step, alongside data-transfers at the speed of light.

transmission through it decreases. The amplitude or the width of the SET pulse dictates the crystalline volume created. The correlated events are identified and grouped at some time instance by measuring the transmission across all devices. Unique to photonics, the provision for WDM enables parallelism in both the read and write operations. This enables manifold improvement in the computational latency (O(1) time complexity against $O(n^2)$, where n is the number of devices), and significant energy savings since information is transmitted as light. Figure 1E illustrates a further improvement of this concept. Here, WDM is utilized in availing parallelism in the write and read operations of not only multiple events, but equally multiple processes. Each process is assigned a number of PCM photonic cells (accumulators), and each event for that process has a designated wavelength. Such a novel approach provides further improvements in the latency.

Key Results:

In our all-optical neural network (correlation detector), the dynamic 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. A sketch of the photonic engine for the correlation detection task is shown in Figure 2A. Lights of different wavelengths, encoding for example the various events, are multiplexed into a common photonic waveguide, that represents the multiplexer (MUX) unit. These wavelengths are then de-multiplexed by a cascade of optimally designed ring resonators that serve as the drop filters, and direct these wavelengths to corresponding N-vN (non-von Neumann) PCM unit cells (represented by D_n), where they can either crystallize the phase change material or be used only for the reading of the transmission states. Briefly, the output comprises multiple wavelengths modulated in their amplitudes by the transmission states of the PCM cells. Each output wavelength meets some resonance condition of the ring resonator $\lambda_n=2\pi r_n n_g/n$, where r_n is the ring radius, n is an integer, and n_g is the effective index of the photonic structure. The outputs from every unit cell are read through a photodetector, signals of which are acquired and analyzed with digital circuitry. Suppose an example of correlation detection between four events at some time instance. The corresponding devices to these events are programmed to state that is defined by the SET pulses, i.e. higher amplitude optical pulses create large crystalline volume and hence smaller optical transmission. Correlation between these four events is established by determining the transmission state of each device. Crucially, the devices with lower optical transmissions are more correlated than those with intermediate and high transmissions.

We verified the applicability of such a concept using 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 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 grating couplers. The 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 to 10 μm²) acting as the functional elements are deposited on top of the waveguide. The ring resonators have different radii (increasing linearly from 40 μm) and have a high optical quality factor (around 10,000). We constructed a custom-built setup for experimentally validating this novel concept (see Figure 2C). The setup is operated using tunable continuous wave (CW) laser sources and electro-optical modulators (EOMs). Individual laser sources for pump and probe pulses are spectrally aligned to the resonance wavelengths of the ring resonators. Suitable pulse sequences, encoding some data stream in which correlations are to be detected are applied to the devices. The transmitted signals are then recorded with high-speed photodetectors and subsequently amplified. These signals are then post-processed using a data acquisition unit and analyzed on a standard computer. There are additional provisions in the setup such as the EDFA (erbium-doped fiber amplifier) for amorphization pulses and polarization controllers. While both programming and read operations are performed with the same optical route, a provision is also provided to decouple these operations. This is realized using a different optical route, one that directly connects an additional CW laser to all the unit cells, via optical circulators.

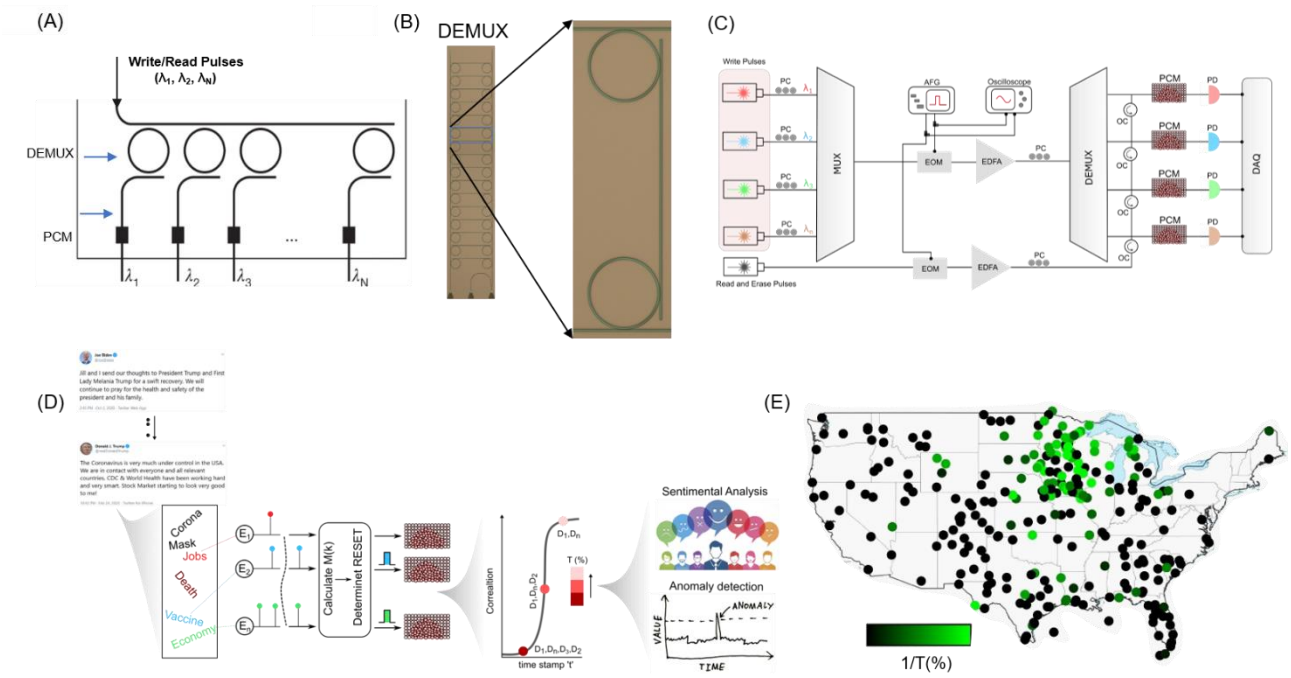

Figure 2. An Illustration of a photonic correlation accelerator. The three important components are the multiplexer (a waveguide) that combines the inputs of different wavelengths, a demultiplexer (ring resonators), which separate-out the inputs selectively, and the phase change photonic memory (PCM), which stores the information. (B) An optical micrograph of one proposed architecture based on (A). Inset highlights a device for better clarity. The top ring-resonator is used as a de-multiplexer and the bottom as a coupler. (C) A schematic of the experimental setup for correlation detection measurements. All devices are on-chip, while components such as lasers, photodetectors, modulators, circulators, and amplifiers are off-chip. (D) An illustration of correlation detection (to be performed experimentally) on a social media platform tweeter. The photonic engine finds the correlations between tweets and uses these correlations for statistical sentimental analysis and anomaly detection. (E) Illustration of correlations detection (to be performed experimentally) between rainfalls using data from weather stations across the USA. The map of the devices' optical transmission levels shows that the correlated group have lower transmission value.

We are performing experiments for real-time processing of event-based data streams, for a few interesting and relevant applications, such as social media, and weather forecasting. Figure 2D illustrates the use case of our photonic accelerator for correlations of live tweets (in social media). The objective is to smartly and efficiently process tweets to find any commonalities (correlation). The correlations can then be utilized for secondary statistical analysis, such as detecting sentiments, market analysis, and more. In Figure 2E, we also illustrate the example of weather forecasting. Correlations between rainfall events based on the geographical proximity between the corresponding weather stations can be efficiently captured by our photonic engine. Devices with similar transmission states (color) are correlated. The algorithm for these measurements includes defining the SET pulse amplitudes using some variable. This variable is a function of a collective momentum term (M(k)) that sums all the binary 1's across all the events at any time instance. The degree of correlation between the events is established with an additional multiplication factor, which scales with the collective momentum.

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 approaches. Fun-COMP technologies utilize one of two important properties of phase change materials: (i) the ability to perform matrix–vector multiplications, and (ii) the rich crystallization dynamics. In this article, we discussed the latter approach, using which we aim to demonstrate a high-level computational primitive or machine-learning hardware using computational memory. Crucially, we are developing a photonics approach to finding correlations between data streams smartly and efficiently. In this approach, the transmission of photonic PCM devices receiving correlated inputs evolve to a low value, and by monitoring these transmission values we detect temporal correlations. To this end, we have performed simulations, fabricated devices, and have built a measurement setup. We are now experimentally performing the tasks of identifying real-time correlations in data streams on the social media platform Twitter, and weather forecasting. We believe that our demonstrations will provide an exemplary showcase of Fun-COMP photonic technologies, that show promise to not only remove the computing bottleneck in modern machine learning hardware, but also enable new functionalities. There are important investigations planned for the future. These include on-chip integration of active components, such as the laser sources (as soliton-combs) and modulators, a road-map for a system-level architecture to execute end-to-end AI workloads, and the benchmarking of the figures-of-merits. We will also research newer applications, tailored for such photonic in-memory correlation detection. An immediate goal will be to construct an anomaly detector using the correlation principles across the high-traffic computing nodes in data centers.