Brain-Inspired Computing: Adventure from Beyond CMOS Technologies to Beyond von Neumann Architectures

ICCAD Special Session Paper

Hussam Amrouch*, Jian-Jia Chen†, Kaushik Roy‡, Yuan Xie§,
Indranil Chakraborty‡, Wenqin Huangfu§, Ling Liang§, Fengbin Tu§, Cheng Wang‡, Mikail Yayla†
*University of Stuttgart, †Technical University of Dortmund, ‡Purdue University,
§University of California, Santa Barbara
E-mail: amrouch@iti.uni-stuttgart.de

***Abstract*—The goal of this special session paper is to introduce and discuss different breakthrough technologies as well as novel architectures and how they together may reshape the future of Artificial Intelligent. Our aim is to provide a comprehensive overview on the latest advances in brain-inspired computing and how the latter can be realized when emerging technologies, using beyond-CMOS devices, are coupled with novel computing paradigms that go beyond von Neumann architectures. Different emerging technologies like Ferroelectric Field-Effect Transistor (FeFET), Phase Change Memory (PCM), and Resistive RAM (ReRAM) are discussed, demonstrating their promising capability in building neuromorphic computing architectures that are inspired by nature. In addition, this special session paper discusses various novel concepts such as Logic-in-Memory (LIM), Processing-in-Memory (PIM), and Spiking Neural Networks (SNNs) towards exploring the far-reaching consequences of beyond von Neumann computing on accelerating deep learning. Finally, the latest trends in brain-inspired computing are summarized into algorithm, technology, and application-driven innovations towards comparing different PIM architectures.**

***Index Terms*—FeFET, PCM, ReRAM, photonic, neuromorphic, DNN, SNN, Processing-in-Memory, emerging technology**

I. Introduction

The unprecedented shift towards data-centric computing, driven by the massive amount of data that deep neural networks (DNNs) demand, makes specialized brain-inspired hardware accelerators inevitable. In order to overcome the memory bottleneck, architectures that go beyond von Neumann principles are key because they offer processing capability for the data where it resides. Hence, the continuous need for moving the data back and forth between processing elements and memory blocks is eliminated. Towards realizing brain-inspired computing, emerging non-volatile memory technologies can play a substantial role. Several technologies are gaining a remarkable attraction due to their promising capability to build efficient neuromorphic hardware. In this special session paper, we discuss three main technologies; Ferroelectric Field-Effect Transistor (FeFET), Phase Change Memory (PCM), and Resistive RAM (ReRAM) as well as how they can be employed to accelerate deep learning and build efficient neuromorphic hardware.

Ultra-Efficient Deep Learning using FeFET: Due to its CMOS compatibility, FeFET technology is gaining more and more attention from the semiconductor industry. For example, GlobalFoundries demonstrated the fabrication of FeFETs using their commercial state-of-the-art 28nm HKMG CMOS through a dual-mask patterning [1]. They have also showed that their 10 MiB chips feature 1ns read latency and a very good yield. Furthermore, Intel has recently demonstrated the fabrication of FeFETs with an endurance that reaches up to 10^{12} cycles [2]. FeFET technology can be used not only to build area-efficient ultra-low power non-volatile memories (NVM), but it also holds a large promise for neuromorphic applications.

In Section II, we explain how Binarized Neural Networks (BNNs) can be trained in the presence of errors that may stem from underlying FeFET-based NVM devices when they are used to store the model's data (i.e., parameters, inputs, activations). We show how hardware/software co-design is a key to obtain accurate inference based on unreliable FeFET devices. We also discuss how FeFET can be employed to build PIM-based XNOR which accelerates further the BNN's inference because weights are stored inside the XNOR logic eliminating the need for memory communications. Finally, we briefly discuss how error-aware BNN training help in implementing robust SNNs providing large energy savings.

Photonics Neuromorphic Computing using PCM: Enabling energy-efficient hardware emulation of key functionalities of the brain is critical for realizing brain-inspired computing. In particular, synapses and neurons efficiently implemented at the device/circuit level not only provide building blocks for executing the event-driven bio-plausible spiking neural networks (SNNs), but also open up promising new avenues for crossbar-based analog PIM. However, hardware implementations of various spiking neuron models such as (Hodgkin-Huxley and Leaky-Integrate-Fire) and synapses on CMOS platforms are inefficient in terms of latency, energy, and area. Emerging device technologies, especially various types of non-volatile memories (NVMs) such as Resistive RAM (ReRAM) [3] and Phase Change Memory (PCM) [4], have been extensively investigated for developing such neuro-

978-1-6654-4507-8/21/$31.00 ©2021 IEEE

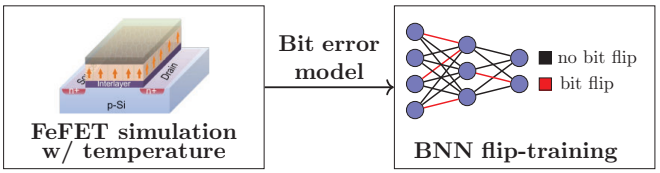

Fig. 1: Errors due to temperature, stemming from underlying FeFET devices, are modeled and then injected during the BNN training. As a result, robust BNNs against run-time temperature errors are acquired [7].

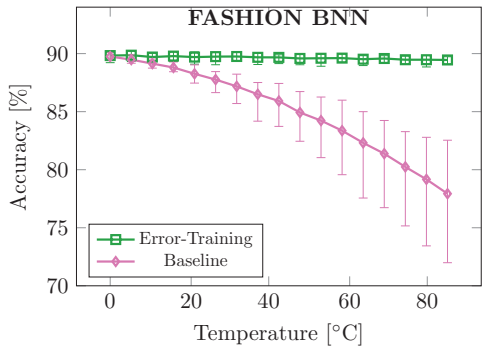

Fig. 2: Accuracy of convolutional BNNs (trained on the FashionMNIST) plotted over temperature in error-prone FeFET-based on-chip memory [7]. In the rose plot, the accuracy is shown when no countermeasures against the temperature-induced errors are employed. In the green plot, the accuracy is shown for a BNN acquired with error-tolerance training.

mimetic devices, although the majority of effort so far has been made within the electrical domain, relying on the modulation of device resistances. In Section III, we discuss the recent demonstrations of ultra-fast photonic computing devices based on PCMs that can pave the way for fast neuromorphic computing beyond the electrical domain [5]. Instead of relying on the changes in resistivity in conventional memristive technologies, phase-change photonic neuro-mimetic devices exploit optical characteristics (such as transmission and reflection) in response to the modulation of the complex refraction index associated with amorphous-crystallized phase changes. We demonstrate an all-photonic SNN inferencing engine for image classification tasks. The proposed photonic neuromorphic systems can potentially overcome limitations of electrically driven NVM-based neuromorphic systems such as high write latency, sneak paths and IR drops [6].

Algorithm, Technology, and Application-Driven Innovations: In Section IV, we present the latest trends in brain-inspired computing, and summarize these studies into algorithm, technology, and application-driven innovations. In the algorithm level, we present two mainstream brain-inspired algorithms, deep neural networks and spiking neural networks. We also talk about the hardware design driven by these two types of neural networks. In the technology level, we discuss Processing-in-Memory (PIM), a promising architecture inspired by the in-memory computing nature of our brain. We compare PIM technologies based on DRAM/SRAM/Non-Volatile Memory, by analyzing their different targeting problems, advantages, and challenges. In the application level, we demonstrate how brain-inspired techniques motivate system designs for new applications, with a focus on bio-informatics. We believe more cross-layer innovations will emerge in the field of brain-inspired computing and reshape the future AI.

II. Brain-Inspired Computing with Ferroelectric FETs: Opportunities and Challenges for BNNs

We first introduce BNNs and their advantages to other models in Sec. II-A. Then, we introduce FeFET, one of the most promising emerging NVMs in II-B, and discuss its trade-offs. In Sec. II-C, we show that, despite reliability issues, FeFET memory can be used as on-chip memory for BNN accelerators in traditional von Neumann systems. Finally, in Sec. II-D, we present our visions of highly efficient FeFET-based beyond von Neumann systems for BNN execution.

A. Binarized Neural Networks (BNNs)

BNNs are one of the most resource-efficient and hardware-friendly NN models to date. In BNNs, the weights and activations are binarized. Due to this, instead of integer or floating-point-based multiply-accumulate (MAC) operations, simple XNOR and bitcount can be employed for computations. This leads to significant latency and energy reductions. With binarized weights and activations, the operations of layers can be computed with

$$popcount(XNOR(W_i^\ell, X^{\ell-1})) > s,$$

where ℓ describes the layer index, W_i^ℓ the weights of neuron i, $X^{\ell-1}$ the inputs to layer ℓ, *popcount* accumulates the number of '1s', and s is a learnable threshold parameter (the comparison returns a binary activation value [8], [9]). In addition to the reduction of the required memory size from the binarization of the weights/activations and the simplification of operations, the on-and off-chip communication overhead is also significantly decreased.

One additional advantage of BNNs is the error-tolerance. In the case of integer or floating-point-based NNs, the position of bit errors matters. For example, in floating-point NNs, one bit error can cause predictions to become useless, if a bit error occurs in the exponent field [10], while in integer values, the flip of the most significant bit causes a change with large magnitude as well. In contrast to this, in BNNs, a flip of one bit in a binary weight or activation causes a change of computation results by merely 1. Furthermore, due to the binary activation function, values with large magnitude get saturated. Since BNNs have simplified logic in computations, it brings to the fore the memory technology used. As on-chip memory, SRAM is typically used. However, high leakage power and large area pose difficult challenges for efficient system designs. Using a non-volatile memory, for example based on FeFET, considerably reduces the overall inference

cost. Therefore, in efficient BNN inference systems, the inefficient SRAM memories should be replaced with efficient non-volatile FeFET memories.

B. Ferroelectric Field-effect Transistor (FeFET) Memory

FeFET is considered to be one of the most promising memory technologies. The reason for the ability of FeFET to store logic '0' and logic '1' lies in the available dipoles inside the FE. The directions of these dipoles can switch, if a sufficiently strong electric field is applied. This state is non-volatile, because the dipoles retain their direction when the field is turned off. The logic '0' and logic '1' can be read out from the FeFET based on the intensity of the current returned (e.g. high or low), which can be converted into the digital domain with sensing circuits.

The three main advantages of FeFET over other NVMs are as follows: (1) FeFET is fully CMOS-compatible, which means that it can be fabricated using current manufacturing processes. This has been demonstrated by GlobalFoundries [1]. (2) FeFET-based memories can perform read and write operations within 1ns latency. This reduces the differences compared to traditional SRAM technology, while the energy usage of FeFET is significantly lower [1]. (3) FeFET memory has the potential to enable extremely low-density memory, since a cell consists of merely one transistor.

One of the major disadvantages of FeFET is the susceptibility to errors. Manufacturing variability (during production) and temperature influences (at run-time) can cause variations in the FE properties. This shrinks available noise margins and may cause errors. To use FeFET despite the errors, for example as on-chip memory for BNN inference systems, it is necessary to extract the error models for the stored bits. With the error model, the impact of the temperature-induced bit errors on the inference accuracy of BNNs can be evaluated.

C. BNNs with FeFET-based Memory in von Neumann Systems

In Fig. 1, we show the steps for extracting the temperature-dependent error model of FeFET transistors. The entire FeFET device is implemented and modeled in the Technology CAD (TCAD) framework (Synopsys Sentaurus). We consider variation in the underlying transistor and the added ferroelectric layer. After incorporating the temperature and variation effects in our calibrated TCAD models, we perform Monte-Carlo simulations for the entire FeFET device. Then, for a certain read voltage, we extract the probability of error, i.e. we calculate the probability that logic '0' is read as logic '1' and a logic '1' is read as logic '0'. Details on device physics modeling and reliability analysis for FeFET under the effects of temperature variability (run-time) and manufacturing (design-time) variability can be found in [11] and [12], respectively.

With the acquired bit error model, we then evaluate the resiliency of BNNs against temperature-induced bit errors, assuming a system that uses FeFET-based on-chip memory. The system architecture is a von Neumann system, i.e. memory and processing elements are separated. The system uses traditional off-chip memory (e.g., reliable DRAM) and unreliable emerging on-chip FeFET memory.

In Fig. 2, we show that the impact of the temperature bit errors can be substantial if no bit error training is used and when no attention is paid to the asymmetry of the bit error rates (rose curve). We find accuracy degradation of over 25% for the FASHION dataset at the highest operating temperature $85°$. When applying methods to increase the error tolerance of BNNs, e.g with bit flip injection during training (green curve), we achieve bit error tolerance for the entire range of operating temperature. More details about the system model, methods, experiments, and BNN architectures can be found in [7].

D. Beyond von Neumann: FeFET-based XNOR Logic-In-Memory for BNNs

One of the most fundamental challenges in existing von Neumann-based architectures is the memory wall. Compared to the latency of processing elements, the data movements cause latencies that are orders of magnitudes higher. To conquer this challenge, the Logic-In-Memory (LIM) design paradigm has been proposed, in which computations are performed inside the memory. In the last few years, several studies have explored LIM-based architectures. For example, for conventional SRAM memories [13] and emerging NVMs [14], boolean logic functions (e.g., XNOR, NAND, etc.) have been successfully integrated inside the memory.

Here, we focus on LIM designs for BNNs. The LIM architecture is in a stark contrast to a traditional von Neumann architecture, where logic and memory are separated. In the LIM architecture, the binary weights are stored in a pair of complementary FeFET gates, which also implement the logic function XNOR. This means, the weights are already stored in the memory, and no additional data movement is required for the weights. The FeFET-based XNOR gates are connected in a row to perform binary multiplication, while for the popcount and activation, analog or digital circuits can be employed [14].

However, as shown above, FeFETs are inherently prone to errors, and error models for more advanced system setups were not explored in the literature yet. Furthermore, the error tolerance of BNNs is not fully exploited yet in BNN circuit design, although several recent studies have proposed methods to increase the error tolerance significantly with minimal accuracy cost [7], [15]. Still, the core design in [14] has served as a template to build more advanced hardware, such as the neuron circuits in Spiking Neural Networks (SNNs), as demonstrated in [16], [17]. Recently, the impact of executing BNNs in SNN hardware was explored in [16]. In that study, a neuron circuit is composed out of multiple FeFET-based XNOR gates in a row, where popcount is performed by Kirchhoff's circuit law, and a membrane capacitor (serving as a sum-of-product accumulator) is connected to a comparator (enabling the conversion of the time to first spike to a discrete representation). The membrane capacitor size, and therefore, energy, latency, and area, is optimized by exploiting the error tolerance of BNNs. We believe that there is further potential

for exploiting the error tolerance of BNNs to build even more efficient FeFET-and BNN-based SNN circuits.

III. Brain-inspired computing with Phase-change Photonic Devices: Opportunities and Challenges

In this section, we focus on how to build large-scale SNN systems with PCM-based photonic devices. First, we show how the contrasting optical properties of the PCM $Ge_2Sb_2Te_5$ (GST) can be leveraged to realize basic neuromorphic elements such as neurons and synapses. We further demonstrate an in-memory photonic dot product engine, and an all-photonic SNN inferencing engine for image-classification tasks. We conclude the section with a discussion highlighting future opportunities and key challenges for photonic neuromorphic.

A. PCM-based photonic spiking neuron

Basic functional blocks of an SNN consist of spiking neurons and weighted synaptic connections. The bio-plausible integrate-and-fire (IF) spiking neuron model and its variants have been extensively used in large scale SNNs and demonstrated satisfactory performance on various AI tasks such as image classfications [18]. We demonstrate photonic IF neuron based on a GST-embedded ring resonator, leveraging the distinctive optical characteristics in the crystalline and amorphous states of GST materials [19]. Conceptually, the writing of neuron's membrane potential is realized by exploiting the phase change dynamics of GST under the heating of incident EM waves, while the reading operations rely on the ring resonator's transmission characteristics [20].

As is shown in Fig.3 (a), a ring resonator comprises a pair of rectangular waveguides optically coupled to a ring waveguide. The transmissions of the THROUGH and DROP ports reach a peak or dip when resonant conditions of the ring is satisfied. By incorporating a GST element on top of a fraction of the ring waveguide, light propagation through the waveguide is modulated due to the tunable evanescent coupling between the GST element and the adjacent ring [20]. Specifically, amorphization of GST is triggered when the local device temperature is elevated above the melting point of GST due to the considerable heating from the incident electromagnetic (EM) wave under "WRITE" pulses. When the crystallographic states of GST evolve between 100% amorphous (a-GST) and 100% crystalline (c-GST), the imaginary component of the refractive index varies by over 10x [21], leading to significant change of optical attenuation of the PCM and thus continuous modulations of the port transmission. During "READ" operation with an incident EM wave at the resonant wavelength, crystalline (amorphous) GST induces high (low) transmission in THROUGH Port and (low) high transmission in DROP Port.

Moreover, incoming spikes of opposite polarities are considered by connecting two ring resonators with an interferometer. As is illustrated in Fig.3 (a), the DROP port of the positive ring resonator and the THROUGH port of the negative ring resonator are connected to the interferometer, forming the integration unit of IF neuron. The output magnitude of the interferometer can reflect the combined effects of positive and

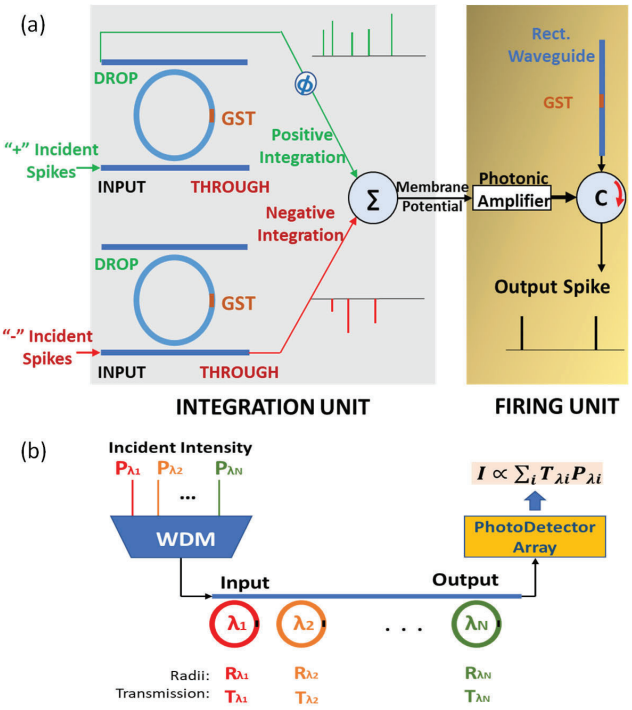

Fig. 3: Photonic neuromorphic building blocks: (a) IF Neuron. (b) Synapse and dot-product engine.

negative inputs, serving the role of membrane potential of an IF neuron. Note that the change of PCM states during "WRITE" operations are retained when write pulses are gone, enabling non-volatility for "READ" operations. Thus at every time-step, the membrane potential integration is proportional to the amplitude of the resultant incident spike to the neuron.

Once the GST reaches full amorphization, the membrane potential exceeds its threshold, resulting in the 'firing' action of a spike which is implemented by an additional firing unit. The firing unit is made of a photonic amplifier, a circulator and a rectangular waveguide with an embedded GST element initially in the crystalline state. For a rectangular waveuguide with GST, the transmission is low (high) in crystalline (amorphous) state. The device is designed so that 'read' and 'write' phases for the 'integration unit' and the 'firing unit' alternate in successive cycles. When the output from integration unit is strong enough to amorphize the GST in the rectangular waveguide, a large transmission in the rectangular waveguide will generate an output spike, followed by a 'RESET' pulse that resets the GST to the initial crystalline state which corresponds to the resting potential of neurons.

B. PCM-based photonic synapse and dot product engine

The integrated micro-ring resonator with embedded PCM can also implement synaptic devices. Leveraging similar

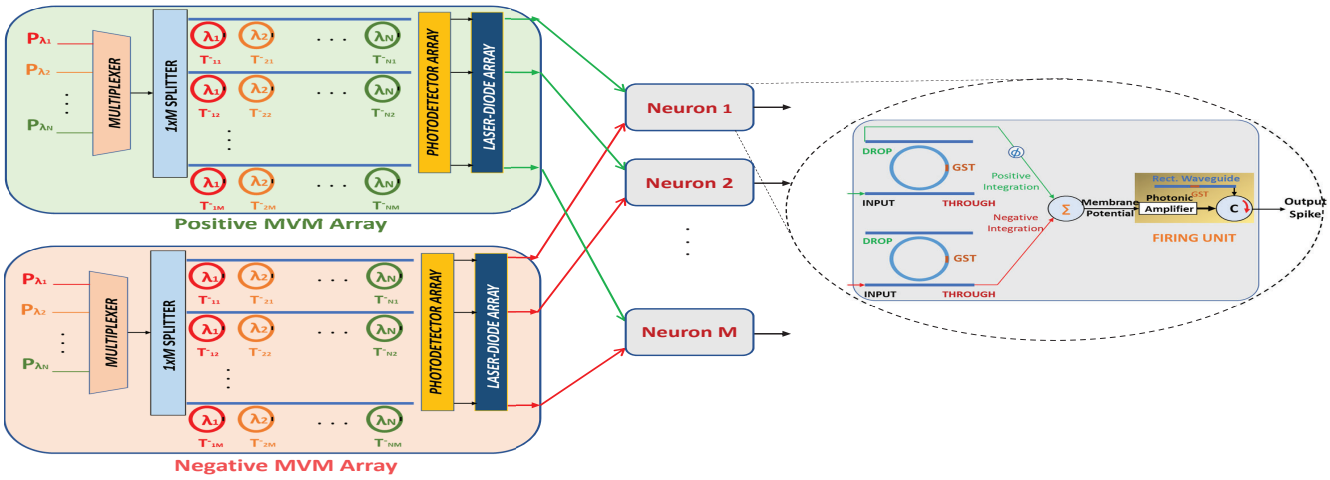

Fig. 4: All-photonic MVM engine with both synapse and neuron arrays.

mechanism as used for the aforementioned photonic IF neurons, synaptic connection with varying magnitudes can be realized based on the impact of the embedded GST on the transmission of waveguides [22]. A single-bus ring resonator can achieve the desirable synaptic functionality. Due to the contrasting imaginary refractive index of a-GST and c-GST, modulation of transmission can be obtained following the PCM amorphization dynamics. Specifically, a-GST will have the minimal transmission in the rectangular bus waveguide (weakest synaptic connection), while multi-level synaptic weights can be represented by the partially crystallized GST leading to intermediate levels of transmission. The synaptic weights are retained following the non-volatility of GST states.

The PCM-based photonic synapse paves a pathway towards implementing weighted sum operations, which are ubiquitous in both SNNs and regular deep artificial neural networks. It has been demonstrated that the proposed ring resonator devices with GST components can be linked by a sharing rectangular bus to execute a parallel summation of input weighted by a vector of transmission [23]. By leveraging the wavelength division multiplex (WDM) technology [24], input spikes can represent a vector by the magnitude of incident EM waves at multiple wavelength channels $P_{\lambda i}$. Therefore, as is illustrated in Fig. 3 (b), an in-memory dot product engine can be constructed if the selective wavelengths are matched with the resonance wavelengths of the synaptic resonators. At the output port of bus waveguide, we obtain a multi-wavelength spike with weighted amplitudes. This spike is then fed to a photodiode (PD) array, which produces a current array with the magnitude governed by $I \propto \sum T_{\lambda i} \cdot P_{\lambda i}$. Note that each synaptic resonator needs to represent its synaptic weight by the transmission at a distinctive resonant wavelength, requiring a designed differentiation method (such as varying ring diameters among the connected resonators). For accurate dot-product operation, it is necessary to achieve significant isolation between the wavelength channels to minimize channel-to-channel crosstalk. To this effect, the constraint on the input vector size of the proposed photonic dot product engine is determined by the ratio of the free spectral range (FSR) and the full-width at half maximum (FWHM) of the individual ring resonator. Based on the proposed single-bus ring resonator configuration, we demonstrate that wavelength range with an exemplary design of ring diameters around 1.5 μm is capable of containing 16 distinctive channels [22].

C. All-photonic SNN inference hardware

All-photonic neuromorphic computing systems can be realized with the integration of the proposed IF neuron and synaptic dot product engine. Efficient neural network operations, which relies immensely on matrix operations, desire to have computing cores with massive parallelism. Therefore, as is shown in Figure 4, the proposed single-bus resonator based dot engine is extended to multiple rows to facilitate matrix-vector multiplications (MVM). A 2D synaptic weight matrix can be mapped to the transmission of the 2D ring resonators by modifying the crystalline states of the GST components therein. Two MVM arrays are used for mapping of positive and negative weights, respectively. Input with multiple channels to such a MVM computing core will first be fed into a multiplexer, and then the WDM signal will be split evenly based on the row number and connected to the rectangular bus waveguide at each row. The signals obtained by photodetector (PD) arrays connected to the outports will be proportional to the result of MVM. In order to build an integrated synapse-neuron system, the electrical current from the PD arrays are further passed onto laser diodes so that the electrical current can be converted to optical spikes for the post-synaptic neurons.

We developed a device-to-algorithm framework for evaluating the functional performance of the proposed neuromorphic system. The transmission characteristics of the ring resonators with varying states of the GST element are taken into account to evaluate the accuracy of the dot-product operation. The error in the computation stems from the non-idealities induced

by the crosstalk between adjacent channels. At the algorithm level, we consider a fully connected SNN consisting one hidden layer. For MNIST hand written digit dataset, the network architecture is set with M=784, N=500, and P=10, where M, N, and P are the numbers of neurons in the input layer, hidden layer and output layer respectively. We adopt the approach of converting a trained ANN to obtain the synaptic weights of SNN. It is found that, with non-idealities of photonic devices included, the SNN implemented with the behavior of the proposed neuromorphic system can have less than 0.5% degradation of inference accuracy from the ideal scenario. The proposed photonic neuromorphic hardware can offer faster inference operations due to the ultrafast dynamics with 200 ps pulse width. Moreover, significant improvement in write latency can be further harvested when synaptic weights need updates, since the write pulse in photonic system is subnanoseconds while PCM devices in electrical domain have write latency around 50-100 ns.

D. Future opportunities and challenges

The demonstrated GST-on-silicon neuromorphic system suggests a promising pathway of implementing brain-inspired computing based on PCM. The benefits of NVM is retained with PCM-based photonics as the non-volatile states of GST components eliminate extra need for off-chip memory access. Moreover, compared to the popular NVM based PIM in electrical domain, the photonic approach achieves highly parallel fan-in leveraging WDM technique, and provides significant improvement in processor latency. It also offers immunity from the impact of various circuit-level non-idealities such as sneak paths and IR drop. We envision that integration of photonic device and emerging PCM may offer exciting opportunities in developing high-performance AI processors [25].

However, a few challenges remains before scalable computing hardware can be realized efficiently on such systems. At the device level, large-scale photonic system desires shrinking the physical dimension of ring resonators in order to achieve high-density integration. But smaller devices will face more fabrication difficulty and controllability/variability issue with the integration of GST component. Moreover, the parallelism of the proposed dot product engine is constrained by the FSR of ring design, based on the mechanism of synaptic connection. Increased computational error will be induced due to interference among adjacent wavelength channels, if more channels are squeezed into a limited FSR. Time-domain multiplexing in combination with the low-latency modulation of PCM, may offer some mitigation of processing MVM with large array sizes, but further reduction of the writing energy of PCM is still desirable [26]. Lastly, functional interface blocks such as analog-digital conversion and electrical-optical conversion at a matching speed (∼GHz) with the photonic components consumes significant energy. The proposed neuromorphic photonic hardware would benefit at the system level from the incorporation of low power interfaces such as PD arrays, laser diodes, and AD/DA circuitries.

IV. Brain-Inspired Computing: Algorithm, Technology, and Application-Driven Innovations

Brain-inspired computing has the potential to break the von Neumann bottleneck and build an Artificial Intelligent (AI) system. In this section, we present the latest trends in brain-inspired computing, and summarize these studies into algorithm, technology, and application-driven innovations.

A. DNN and SNN Acceleration

In past years, deep neural networks (DNNs) have been proved its power in a wide range of applications, such as computer vision, speech recognition, and language processing. The design principles of DNNs are borrowed from the mechanism of brain, where information is stored in neurons and passed through synapses. Compared with DNNs, spiking neural networks (SNNs) exhibit a closer scheme to the biological neuron models which attract extensive attention. Meanwhile, some resent studies show advantages of SNNs in processing sparse and noisy data. However, with the development of algorithms, the hardware demand for DNNs and SNNs increased dramatically. many studies make effort to design efficient accelerators that reduce the hardware resources and execution latency. Such that the deployment of DNNs/SNNs in the real system becomes achievable.

1) DNN Accelerators: Many works accelerate DNNs by exploiting reconfigurable computation parallelism or dataflow. For instance, Evolver [27] designs hybrid dataflows to accelerate different DNN structures with high resource utilization. There are also accelerators that explore the sparsity in processing DNNs. These accelerators design special architectures to skip operations with zero activations and weights [28]. Zero activations are produced by ReLU activation, and zero weights are caused by redundancy in DNN models.

2) Preliminary of SNNs: One of the most distinct character of SNNs is the dynamic neuron modeling that simulate the brain behaviour. Leaky integrate-and-fire is the most widely used model as Figure 5(a). Each neuron is composed by membrane potential u and spike s. Once the neuron's membrane potential is greater than a threshold th_f, it will generate a weighted spike to the connected neurons and its membrane potential is rest to rst. Otherwise, the membrane potential will decay with a factor α.

Because of the special neuron modeling, SNNs usually involve a more complex spatial information propagation. Also, the dynamic modeling demands an additional temporal axis to propagate information along time as Figure 5(b). These characters lead the inefficiency of commercial platforms to run SNNs. Thus, many studies design accelerators to boost the inference and training of SNNs.

3) SNN Accelerators: Currently, most of neuromorphic chips tend to accelerate the inference stage of SNNs. Tianjic [29] designs a hybrid architecture with unified routing infrastructure that can deploy both DNNs and SNNs.

Despite the inference accelerator, some studies design training accelerators for different SNN learning algorithms. Most

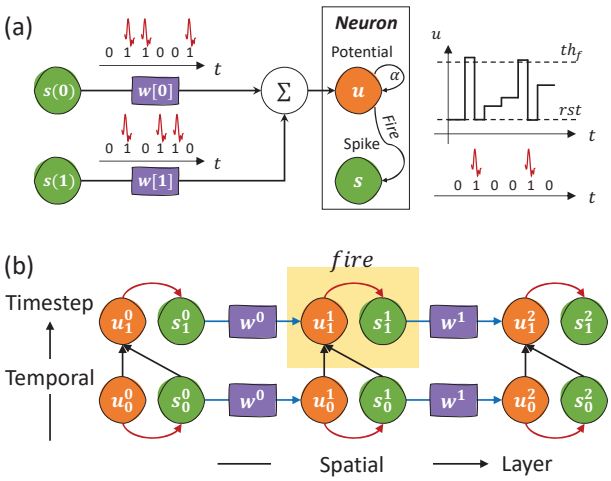

Fig. 5: Preliminary of SNNs: (a) Neuron modeling. (b) Information propagation in SNN inference.

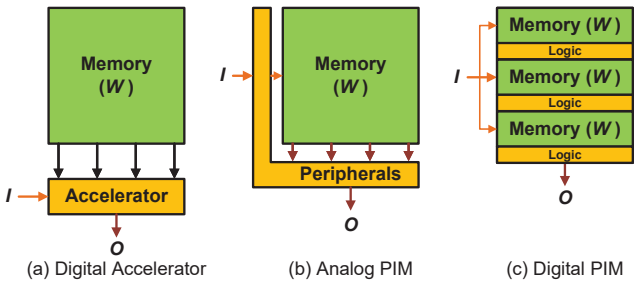

Fig. 6: Architectures for NN acceleration: (a) Digital Accelerator. (b) Analog PIM. (c) Digital PIM. (W: Weights in memory. I, O: Inputs and Outputs.)

training neuromorphic chips target on the local synaptic plasiticity learning rules such as spike time dependent plasiticity (STDP). Recently, H2Learn [30] designs a training neuromorphic chip that can support back propagation through time (BPTT) learning algorithm. Specifically, the BPTT learning algorithm can improve the model accuracy a lot and H2learn utilizes the binary input pattern and the sparsity during learning to boost the BPTT based learning efficiently.

B. PIM for NN Acceleration

Ubiquitous NN applications have motivated many NN accelerator designs in the past few years. However, as the NN model size increases, massive data movement between computing units and memory becomes a bottleneck in the computing system. Processing-in-Memory (PIM), inspired by the in-memory computing nature of our brain, is a promising hardware technology that tackles the memory bottleneck in conventional accelerators. The basic idea of PIM is placing the multiply-accumulate (MAC) units near or in the memory, thus utilizing the high bandwidth in memory to reduce data movement latency and energy. Based on the implementation logic of MAC in memory, PIM architectures can be classified into two categories: analog PIM and digital PIM, as illustrated in Fig. 7.

1) Analog PIM: In comparison to conventional digital NN accelerators with separate MAC units and memory (Fig. 7(a)), analog PIM realizes in-memory MAC based in current or voltage (Fig. 7(b)). Analog PIM is usually implemented in SRAM or non-volatile memory like ReRAM. Input digital-to-analog converters (DAC) and output analog-to-digital converters (ADC) are required in the peripheral circuits. For one NN layer ($O = I * W$), the weights (W) don't need reading out of memory for MAC computation, saving lots of data movement. For the MAC operation itself, analog computing usually consumes less power than conventional digital logic.

However, in a practical analog PIM, only a limited number of rows and columns in a cell array can be activated each cycle [31]. The number of activated rows depends on accuracy limitations, because activating too many rows will cause large accumulated analog deviation that harms NN accuracy. To save the overhead of high-resolution ADCs, usually multiple columns share one ADC in analog PIM, which limits the number of activated columns. As a result, practical analog PIM works in a smaller granularity than the entire array, called an operation unit (OU).

2) Digital PIM: As the requirement for higher accuracy and robustness arises, there is a new trend of integrating digital logic into PIM design, which is called digital PIM (Fig. 7(c)). According to the base memory devices, digital PIM can be further classified as SRAM-based and DRAM-based.

Recently, TSMC's implements a digital PIM in ISSCC'21 [32] by attaching only one NOR gate to each cell and placing accumulators at the subarray level. All the in-memory logic can be activated concurrently to achieve almost 100% array utilization, with no accuracy loss caused by PIM.

Unlike SRAM-based digital PIM, DRAM-based digital PIM targets a different problem, which performs computation in DRAM to optimize off-chip memory access. For example, Samsung's recent HBM-PIM integrates computing units deeper into the bank level of their 3D DRAM [33]. Such DRAM-based PIMs can be used to accelerate much larger scale NN models.

C. Go Beyond NN, PIM-based Bioinformatics Computing

As mentioned above, PIM brings lots of opportunities to the acceleration of NN. Actually, besides NN, a wide range of important applications can also benefit from PIM, for example graph analytics, image processing, and bioinformatics. In this subsection, we use PIM based hardware acceleration of bioinformatics as an example to demonstrate the benefits of PIM to those emerging applications. Bioinformatics is getting more and more important and it is developing rapidly, because it is helpful to, for example, wildlife conservation, understanding of human disease, and precise medical care. As an important example, during the global pandemic Coronavirus Disease 2019 (COVID-19), the Next Generation Sequencing (NGS) technology plays an crucial role during the disease characterization. Unfortunately, with the advancement of the

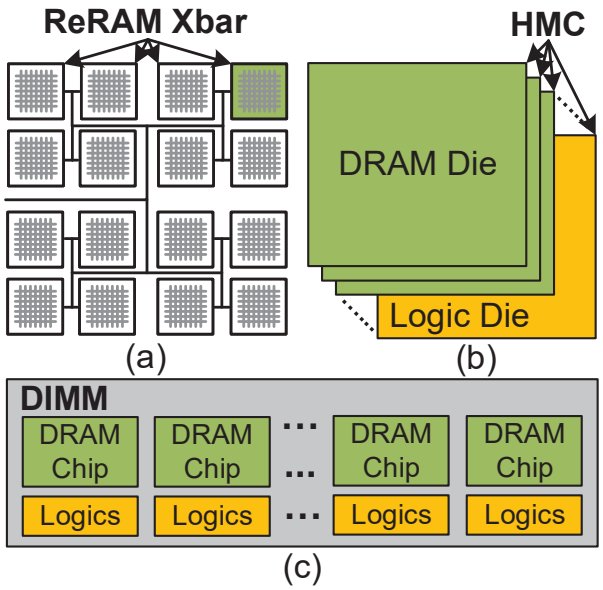

Fig. 7: PIM architecture for bioinformatics. (a) Emerging memory (ReRAM) based PIM architecture. (b). HMC based PIM architecture. (c). DIMM based PIM architecture.

NGS technology, bio-data grows exponentially, putting forward great challenges for data processing in bioinformatic.

Due to the importance of bioinformatics, many hardware approaches have been explored to accelerate different applications in bioinformatics. Most of those approaches are compute-centric, i.e., based on CPU/GPU/FPGA. However, the space for performance improvement is limited for compute-centric accelerators, because many of their target applications in bioinformatics are memory-bounds. To address the key issues of hardware acceleration for bioinformatics from the memory perspective, many researchers propose PIM solutions for bioinformatics, which can be divided into two major categories:

1) Emerging Memory based Architectures: Many researchers leverage emerging memory technologies, mainly ReRAM, to build domain-specific accelerators for bioinformatics. As shown in Fig. 7 (a), those architectures store the DNA data within the ReRAM cells and perform parallel computation/comparison within the ReRAM array in place. Compared with the previous compute-centric accelerators, those emerging memory based PIM accelerators for bioinformatics can achieve significant performance improvement and energy reduction due to the features of high density, low power consumption, and ability to perform parallel operations in ReRAM [34].

2) DRAM based Architectures: Although emerging memory technologies bring significant performance improvement, those technologies are relatively long-term and cannot be adopted in the foreseeable future [35], [36]. To address this issue of the emerging memory based approaches, DRAM based PIM architectures for bioinformatics are proposed. Those DRAM based PIM architectures for bioinformatics can also be divided into two categories:

- 3D-Stacking DRAM: Hybrid Memory Cube (HMC) has been leveraged in previous work to accelerate bioinformatics. As shown in Fig. 7 (b), those HMC based accelerators place processing elements on the logical die of HMC and leverage the high bandwidth of Through-Silicon Vias (TSV) to access data in the DRAM dies.
- Dual-Inline Memory Module (DIMM): As shown in Fig. 7 (c), DIMMs based architectures, such as MEDAL and NEST [35], [36], insert processing elements into the PCB board of each DIMM, leaving the cost-sensitive DRAM dies on the DIMM untouched. Compared with the above 3D-stacking memory based approaches, those DIMM based solutions are more cost-effective and practical due to their non-invasive designs.

To summarize, besides NN, many different applications involve huge amount of data and the memory, instead of the computation, becomes their bottlenecks. We use an bioinformatics as an example application to demonstrate the new design opportunities brought by PIM and the possible explorations we can do with PIM.

V. Conclusion

The inherent limitations in the existing von Neumann architectures in which memory communications form a profound bottleneck for data-centric application largely increase the need for novel computing paradigms. The journey to achieve that starts from the underlying technology in which novel beyond-CMOS devices are required. However, such innovations in technology need to be combined with novel architectures. Otherwise, neuromorphic computing cannot be efficiently realized. Most importantly, hardware/software co-design is, in fact, a key to overcome the inherent reliability challenges that come with novel beyond-CMOS devices. In this special session paper, we have provided a comprehensive overview on how neuromorphic photonics can be implemented using PCM technologies. We have also discussed how we can implement reliable BNNs that use unreliable FeFET-based NVM. Finally, we discussed the latest trends in brain-inspired computing, and summarized these studies into algorithm, technology, and application-driven innovations.

Acknowledgment

This paper has been supported by German Research Foundation (DFG) project OneMemory (405422836) and project “ACCROSS: Approximate Computing aCROss the System Stack”, and as part of the Collaborative Research Center SFB 876 ”Providing Information by Resource-Constrained Analysis” (project number 124020371), subproject A1 (http://sfb876.tu-dortmund.de).

The research was funded in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, Vannevar Bush Faculty fellowship, National Science Foundation, Army Research Laboratory and Sandia National Laboratory. This work was supported in part by NSF 1725447 and 1816833.

REFERENCES

[1] M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, *et al.*, "A 28nm hkmg super low power embedded nvm technology based on ferroelectric fets," in *2016 IEEE International Electron Devices Meeting (IEDM)*, pp. 11–5, IEEE, 2016.

[2] A. A. Sharma *et al.*, "High speed memory operation in channel-last, back-gated ferroelectric transistors," in *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 18.5.1–18.5.4, 2020.

[3] H. Akinaga and H. Shima, "Resistive random access memory (reram) based on metal oxides," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2237–2251, 2010.

[4] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[5] Z. Cheng, C. Ríos, W. H. Pernice, C. D. Wright, and H. Bhaskaran, "On-chip photonic synapse," *Science advances*, vol. 3, no. 9, p. e1700160, 2017.

[6] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[7] M. Yayla, S. Buschjäger, A. Gupta, J.-J. Chen, J. Henkel, K. Morik, K.-H. Chen, and H. Amrouch, "Fefet-based binarized neural networks under temperature-dependent bit errors," *IEEE Transactions on Computers*, 2021.

[8] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in neural information processing systems*, pp. 4107–4115, 2016.

[9] E. Sari, M. Belbahri, and V. P. Nia, "How does batch normalization help binary training?," *arXiv:1909.09139*, 2019.

[10] S. Koppula, L. Orosa, A. G. Yağlikçi, R. Azizi, T. Shahroodi, K. Kanellopoulos, and O. Mutlu, "Eden: Enabling energy-efficient, high-performance deep neural network inference using approximate dram," in *International Symposium on Microarchitecture*, MICRO '52, p. 166–181, Association for Computing Machinery, 2019.

[11] A. Gupta, K. Ni, O. Prakash, X. S. Hu, and H. Amrouch, "Temperature dependence and temperature-aware sensing in ferroelectric fet," in *Proceedings of the IEEE 58th International Reliability Physics Symposium (IRPS'20), Dallas, Texas, U.S.*, 2020.

[12] K. Ni, A. Gupta, O. Prakash, S. Thomann, X. S. Hu, and H. Amrouch, "Impact of extrinsic variation sources on the device-to-device variation in ferroelectric fet," in *Proceedings of the IEEE 58th International Reliability Physics Symposium (IRPS'20), Dallas, Texas, U.S.*, 2020.

[13] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-sram: Enabling in-memory boolean computations in cmos static random access memories," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4219–4232, 2018.

[14] X. Chen, X. Yin, M. Niemier, and X. S. Hu, "Design and optimization of fefet-based crossbars for binary convolution neural networks," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1205–1210, 2018.

[15] S. Buschjäger, J.-J. Chen, K.-H. Chen, M. Günzel, C. Hakert, K. Morik, R. Novkin, L. Pfahler, and M. Yayla, "Margin-maximization in binarized neural networks for optimizing bit error tolerance," *DATE '21*.

[16] M.-L. Wei, M. Yayla, S. Ho, C.-L. Yang, J.-J. Chen, and H. Amrouch, "Binarized snns: Efficient and error-resilient spiking neural networks through binarization," in *IEEE/ACM 40th International Conference on Computer-Aided Design (ICCAD)*, 2021.

[17] M.-L. Wei, H. Amrouch, C.-L. Sung, H.-T. Lue, C.-L. Yang, K.-C. Wang, and C.-Y. Lu, "Robust brain-inspired computing: On the reliability of spiking neural network using emerging non-volatile synapses," in *2021 IEEE International Reliability Physics Symposium (IRPS)*, pp. 1–8, 2021.

[18] C. Lee, G. Srinivasan, P. Panda, and K. Roy, "Deep spiking convolutional neural network trained with unsupervised spike-timing-dependent plasticity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 3, pp. 384–394, 2018.

[20] I. Chakraborty, G. Saha, A. Sengupta, and K. Roy, "Toward fast neural computing using all-photonic phase change spiking neurons," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[19] W. H. Pernice and H. Bhaskaran, "Photonic non-volatile memories using phase change materials," *Applied Physics Letters*, vol. 101, no. 17, p. 171101, 2012.

[21] S.-Y. Kim, S. J. Kim, H. Seo, and M. R. Kim, "Variation of the complex refractive indices with sb-addition in ge-sb-te alloy and their wavelength dependence," in *Optical Data Storage'98*, vol. 3401, pp. 112–115, International Society for Optics and Photonics, 1998.

[22] I. Chakraborty, G. Saha, and K. Roy, "Photonic in-memory computing primitive for spiking neural networks using phase-change materials," *Physical Review Applied*, vol. 11, no. 1, p. 014063, 2019.

[23] C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "Integrated all-photonic non-volatile multi-level memory," *Nature photonics*, vol. 9, no. 11, pp. 725–732, 2015.

[24] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," *Journal of Lightwave Technology*, vol. 32, no. 21, pp. 4029–4041, 2014.

[25] C. Ríos, N. Youngblood, Z. Cheng, M. Le Gallo, W. H. Pernice, C. D. Wright, A. Sebastian, and H. Bhaskaran, "In-memory computing on a photonic platform," *Science advances*, vol. 5, no. 2, p. eaau5759, 2019.

[26] E. Gemo, J. Faneca, S. G.-C. Carrillo, A. Baldycheva, W. Pernice, H. Bhaskaran, and C. Wright, "A plasmonically enhanced route to faster and more energy-efficient phase-change integrated photonic memory and computing devices," *Journal of Applied Physics*, vol. 129, no. 11, p. 110902, 2021.

[27] F. Tu, W. Wu, Y. Wang, H. Chen, F. Xiong, M. Shi, N. Li, J. Deng, T. Chen, L. Liu, S. Wei, Y. Xie, and S. Yin, "Evolver: A deep learning processor with on-device quantization–voltage–frequency tuning," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 2, pp. 658–673, 2021.

[28] A. Gondimalla, N. Chesnut, M. Thottethodi, and T. Vijaykumar, "Sparten: A sparse tensor accelerator for convolutional neural networks," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 151–165, 2019.

[29] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, *et al.*, "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.

[30] L. Liang, Z. Qu, Z. Chen, F. Tu, Y. Wu, L. Deng, G. Li, P. Li, and Y. Xie, "H2learn: High-efficiency learning accelerator for high-accuracy spiking neural networks," *arXiv preprint arXiv:2107.11746*, 2021.

[31] T.-H. Yang, H.-Y. Cheng, C.-L. Yang, I.-C. Tseng, H.-W. Hu, H.-S. Chang, and H.-P. Li, "Sparse reram engine: Joint exploration of activation and weight sparsity in compressed neural networks," in *Proceedings of the 46th International Symposium on Computer Architecture*, ISCA '19, (New York, NY, USA), p. 236–249, Association for Computing Machinery, 2019.

[32] Y.-D. Chih, P.-H. Lee, H. Fujiwara, Y.-C. Shih, C.-F. Lee, R. Naous, Y.-L. Chen, C.-P. Lo, C.-H. Lu, H. Mori, *et al.*, "An 89tops/w and 16.3 tops/mm 2 all-digital sram-based full-precision compute-in memory macro in 22nm for machine-learning edge applications," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, pp. 252–254, IEEE, 2021.

[33] Y.-C. Kwon, S. H. Lee, J. Lee, S.-H. Kwon, J. M. Ryu, J.-P. Son, O. Seongil, H.-S. Yu, H. Lee, S. Y. Kim, *et al.*, "A 20nm 6gb function-in-memory dram, based on hbm2 with a 1.2 tflops programmable computing unit using bank-level parallelism, for machine learning applications," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, pp. 350–352, IEEE, 2021.

[34] W. Huangfu, S. Li, X. Hu, and Y. Xie, "RADAR: a 3d-ReRAM based DNA alignment accelerator architecture," in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2018.

[35] W. Huangfu, X. Li, S. Li, X. Hu, P. Gu, and Y. Xie, "Medal: Scalable dimm based near data processing accelerator for dna seeding algorithm," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 587–599, 2019.

[36] W. Huangfu, K. T. Malladi, S. Li, P. Gu, and Y. Xie, "Nest: Dimm based near-data-processing accelerator for k-mer counting," in *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pp. 1–9, IEEE, 2020.