Invited
Achieving PetaOps/W Edge-AI Processing

Gomony, Manil Dev; Ahn, Bas; Luiken, Rick; Biyani, Yashvardhan; Gebregiorgis, Anteneh; Laborieux, Axel; Zenke, Friedemann; Hamdioui, Said; Corporaal, Henk

DOI
10.1145/3649329.3689623

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 61st ACM/IEEE Design Automation Conference, DAC 2024

Invited: Achieving PetaOps/W edge-AI processing

Manil Dev Gomony
TU Eindhoven

Bas Ahn
TU Eindhoven

Rick Luiken
TU Eindhoven

Yashvardhan Biyani
TU Delft

Anteneh Gebregiorgis
TU Delft

Axel Laborieux
Friedrich Miescher Institute

Friedemann Zenke
Friedrich Miescher Institute

Said Hamdioui
TU Delft

Henk Corporaal
TU Eindhoven

ABSTRACT

Artificial Intelligence (AI) supported by Deep Artificial Neural Networks (ANNs) is booming and already used in many applications, with impressive results, and we are still its infancy. For many sensing applications it would be advantageous if we could move AI from cloud to Edge. However this requires huge improvements in energy-efficiency. The CONVOLVE project (convolve.eu) aims at enabling smart edge devices through a concerted effort at all layers of the design stack. This ranges from using much more efficient models and mappings, like exploiting Spiking Neural Networks (SNNs), to new processing architectures, like compute-in-memory (CIM), use of approximation, and using new device technology, like memristors. However these latter changes make HW more susceptible to noise and other disturbances. Online continuous learning (i.e. adapting weights) may alleviate these problems. This paper shows several CONVOLVE developments in the crucial areas of CIM architectures, SNN accelerators and online learning.

1 INTRODUCTION

With the rise of smart applications powered by AI in almost every edge device, there is a pressing need for an ultra-low-power (ULP) edge AI System-on-Chips (SoC) or Smart Edge Processors (SEP) that offloads computing closer to the source of data generation. This is necessary to address the limitations of using the cloud, such as privacy, latency and bandwidth. According to current projections, the SEP market is expected to grow by about 40% per year, reaching beyond 60 billion USD by 2028. In comparison to cloud computing, SEP hardware is significantly more constrained in terms of energy consumption. This is due to the fact that it is mostly battery powered and/or restricted by thermal dissipation.

Figure. 1 shows the energy-efficiency of state-of-the-art SEP chips optimized for Neural Network (NN) models of both ANNs and SNNs. The trends indicate that ANN chips are more mature, approaching energy-efficiency close to 1 fJ/Op. Especially, CIM based architectures have high-potential to go beyond 1fJ/Op, by reducing the data-movement energy between memory and compute units. Although SNN chips have high potential, their energy-efficiency are currently at best in the 100 fJ/Op range.

DAC '24, June 23 – 27, 2024, San Francisco, CA, USA

ACM ISBN 979-8-4007-0601-1/24/06.
https://doi.org/10.1145/3649329.3689623

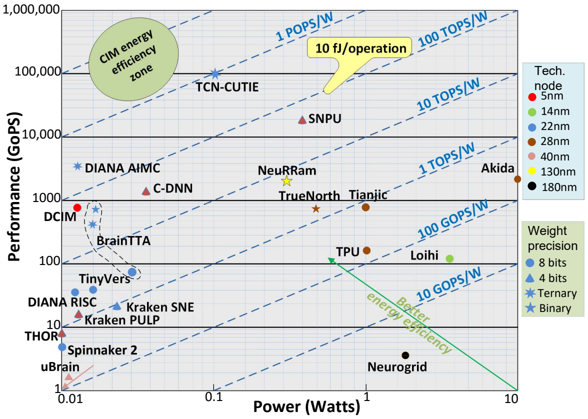


Figure 1: Energy-efficiency of SOTA edge-AI μProcessors. Note that efficient processors are typically far less flexible/programmable.

To beat the fJ/Op energy-efficiency barrier while being flexible enough [1] to deal with future NN models necessitates combining innovations from all levels of the design stack, including AI deep learning models, online learning, compilers, architecture, microarchitecture, circuits, and devices. The CONVOLVE project [2] proposes a single framework that ties together the innovations from different levels, as shown in Figure 2. CONVOLVE aims to achieve 100x improvement in energy-efficiency and quickly implement SEPs combining innovations from different levels of the stack for a given application by reducing design time significantly. This paper focuses on the state-of-the-art, challenges and opportunities in three key areas in CONVOLVE project: 1) SRAM based digital, and RRAM based Analog CIM architectures; 2) SNN accelerators, reducing the efficiency gap with ANNs, and 3) Online learning strategies by getting rid of complex back-propagation in NN training.

2 CIM ARCHITECTURE

CIM is proposed as a promising computing paradigm as it integrates the computation and storage in the same physical location. Such integration overcomes the data transfer bandwidth challenge of conventional architectures and unlock new potential for efficient computing. CIM can be realised using both conventional memories (such as SRAM, DRAM and Flash), or (emerging) non-volatile devices [3] such as resistive random access memory (RRAM), magnetic (STT-MRAM), phase change (PCM), or even ferroelectric field transistor (FeFET). SRAM is primarily used as on-chip cache and has enjoyed the benefit of scaling; hence, its advantage for CIM is not only being embedded memory by nature, but also its commercial availability at the latest technology node; this is not the case for DRAMs. However, both SRAM and DRAM suffer from leakage

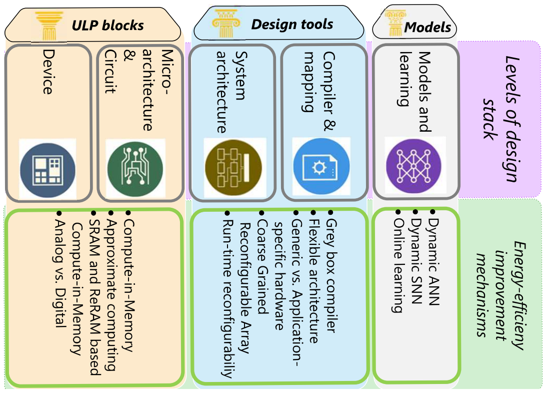


Figure 2: The three pillar design methodology of CONVOLVE project; energy improvements are addressed at all layers of the stack.

power; which negatively impacts battery-powered edge devices. Non-volatile memories are therefore more competitive in such cases. In addition, they retain the stored value when they are turned off. Moreover, they consume less silicon area than SRAM, and they have the potential to deliver multi-bit per cell resulting in even higher integration density. Among emerging non-volatile memories, RRAM is widely popular for CIM due to its larger Ron/Roff ratio than STT-MRAM, less power consumption than PCM, and more foundry available than FeFET. Therefore, CONVOLVE focuses on SRAM and RRAM based CIM (presented in Sections 2.1 and 2.2), as they are the most promising technologies for the near future.

2.1 SRAM based CIM

Digital CIM architectures using SRAM, as shown in Figure 3, suffers from a power hungry adder tree which may consume over 70% of power consumption. Innovations in digital CIM, such as integrated approximate computing, promise substantial improvements to address the challenge of the power hungry adder tree. Approximate computing reduces computational resources (and power consumption) while maintaining acceptable accuracy, particularly advantageous in NNs where minor errors are tolerable. Retraining mitigates the impact of noise introduced by approximate computing, ensuring minimal performance compromise (Section 4). Here we focus on approximate computing in CIM by encoding the weights of NN in Fibbinary format [4], which prohibits values with consecutive ones in the binary representation of the weights. This in combination with the support for multi-bit multipliers allows increased throughput and flexibility. Of course this may impact accuracy. Therefore, we aim to loosen this weight constraint to allow consecutive ones, as long as those bits are not accessed by the same multiplier. This modification increases the number of possible weights by 12.5%, 47.2%, and 154% for 4b, 8b, and 16b, respectively. By modifying the weights, the CIM architecture can benefit from custom multi-bit multipliers that are significantly smaller than standard multi-bit multipliers (76%) and use less power (66%). It further benefits from simplifying the intermediate bandwidth, leading to fewer required adders in the adder tree. Approximating the adder tree will be our next research topic.

2.2 RRAM based CIM

RRAM based CIM comes with its own challenges, particularly arising from the parallel configuration of RRAMs in the crossbar. If given enough size, the output crossbar currents, whose magnitude represents the output of the VMM operation, may accumulate to orders of magnitude higher than individual currents generated within a RRAM, leading to high power consumption. Moreover, the impact of RRAM non-idealities such as variability of conductance states, non-zero conductance in ’off’ state and read disturb can significantly degrade the accuracy [5].

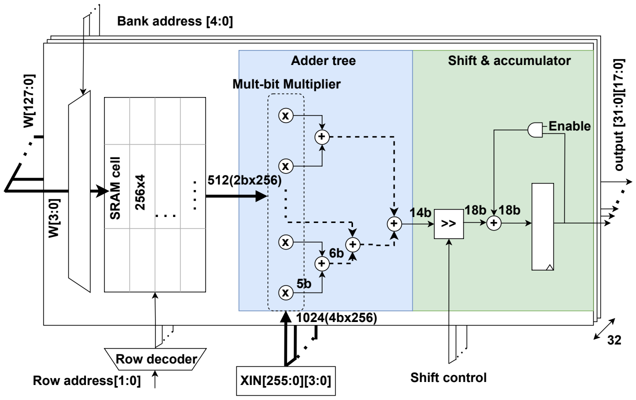
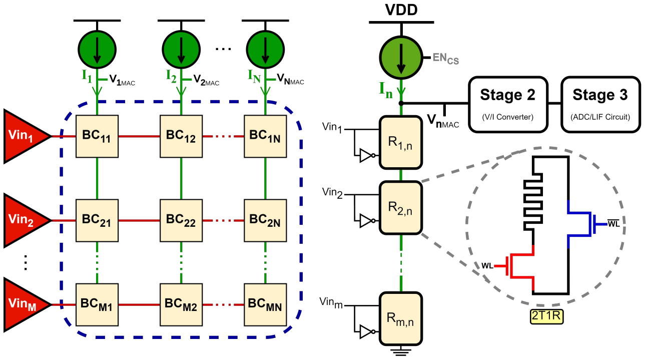


Figure 3: Instantiation of a digital 32Kb SRAM based CIM architecture showing the bitcell array, multiplier, adder tree and bitshift & accumulator.

While taking the above into consideration, CONVOLVE proposes a novel RRAM-based CIM architecture, referred to as C3CIM (Constant Current Crossbar CIM), as illustrated in Figure 4a [6]. The architecture has three unique features. First, it uses a constant current reading to enable low-power computation i.e. irrespective of the number of selected operands (rows), the value of the read current per column is constant and fixed by design. Thus, the output voltage per column is in *linear proportion* to the number of selected cells with high resistance for the entire range of operation. Second, as shown in Figure 4b), it uses 2T1R bit cell (1T1R in parallel with 1T); this is required to use the constant current as first operand, depending upon the path selected and its resistance. Apart from enabling CIM using this approach, it also has inherent advantage of compensating for the non-zero resistance of the access transistor by having complementary inputs and maintaining a constant number of access transistors in the current path, which can be addressed as constant offset during post-processing. Third, it not only supports various Neural Network (NN) flavors with the selection of a suitable periphery; but it can also leverage any emerging non-volatile memory technology, even though it was developed for RRAM to begin with. By incorporating CMOS-based non-volatile memories like FeFET, it can also compensate for non-zero low resistance states of these memories, leading to further improvement in accuracy.

The preliminary evaluation of the C3CIM reveals impressive results. A custom BNN as well as SNN model was developed over MNIST dataset to benchmark the proposed architecture against state-of-the-art [7, 8]. An energy efficiency of 30 fJ/MAC operation (for 64 RRAMs per column, 1-bit per RRAM) is realized; normalized to 1-bit by 1-bit MAC, this translates to around 0.5fJ per operation. In addition, for the BNN model, having a topology of 784 (Input)X3136 (Hidden layer)X10 (Outputs), the average energy consumption was 14.6 nJ per inference. Similarly, for the SNN model, with a topology of 324 (Input)X81 (Hidden layer)X10 (Outputs) and time steps=60, the average energy consumption was 11.24 nJ per inference. Going by the performance results, this architecture can be used to drive various edge applications such as Classification, abnormal detection, de-noising audio signals, object detection, etc.



(a) Novel approach

(b) Implementation

Figure 4: Overview of C^3CIM architecture

In summary, CIM based architectures have huge potential for achieving PetaOps/W processing. RRAM based CIMs using constant current reading and 2T1R cells enables reliable operation and achieve high computational accuracy. While approximate computing based Digital CIM can reduce the power consumption of adder-tree significantly.

3 SNN ARCHITECTURE

Recently, SNNs have been proposed as an alternatives to ANNs for low-power processing of sensory data. In contrast to ANNs, SNNs consist of neurons having state, while communicating using sparse binary spikes. This sparsity can be exploited by neuromorphic architectures for energy-efficient processing of SNNs. However, SNNs often achieve lower accuracy than ANNs on the same task. Recently *deep* SNNs can be efficiently trained, without converting trained ANNs, avoiding costly back-propagation through time (BPTT). This allows SNNs to bridge the accuracy gap with ANNs (see Figure 1).

There are neuromorphic architectures that can run large SNNs. E.g., SpiNNaker 2 [9] is a scalable digital neuromorphic architecture consisting of many ARM M4 cores. The SpiNNaker 2 chip contains 6 chip-to-chip communication links, which allow to form a multichip network. Due to this scalability, SpiNNaker 2 is able to run deep convolutional SNNs. However, because of its high flexibility, its energy efficiency is lower than most neuromorphic edge architectures. Similarly, Loihi 2 by Intel is a scalable architecture with 128 dedicated neuromorphic cores per chip. It contains 6 chip-to-chip communication links, which can scale to 1000s of cores. The neuromorphic cores in Loihi 2 are programmable, allowing the user to implement different neuron models. However, like SpiNNaker 2, Loihi 2 is not very energy efficient. A more energy efficient neuromorphic architecture is Sparse Neural Engine (SNE). SNE performs convolutions is an event-based manner. By performing event-based convolutions, the number of operations SNE performs is proportional to the number of input spikes. While having less flexibility than both SpiNNaker 2 and Loihi 2, SNE only consumes 0.91 pJ per Synaptic OPeration (SOP). However, SNE has not been shown to efficiently run deep convolutional SNNs.

Efficient neuromorphic architectures running deep (convolutional) SNNs require an appropriate memory hierarchy, especially since SNNs have additional neuron state. Most neuromorphic architectures include one level of memory, which can only fit small SNN. However, when an SNN does not fully fit in the on-chip memory, layers need to be swapped out to an external memory, which dominates the power consumption. We introduce Mega, a scalabale architecture with a memory hierarchy, as shown in Figure 5, consisting of 9 PE clusters, input spike (L1) buffer, weight (L1) buffer, neuron state (L1) buffer, and a global (L2) memory. The small low power memories near PEs allow for reusing states and weights, which prevents accesses to the larger memories or to external memory. Moreover, a memory hierarchy also allows for flexibility in scheduling. By choosing the right schedule, weights and state can be reused more, increasing the energy efficiency up to 0.292 pJ/SOP for a single output channels; it is expected to improve a lot when adding many (32) output channels.

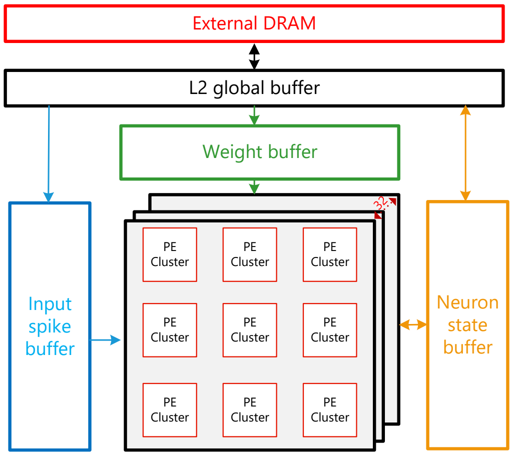


Figure 5: Mega: SNN architecture with a memory hierarchy uses 32x3x3 PE clusters. The use of small memories allow reuse of states/weights, and adds flexibility in scheduling.

In summary, SNN architectures can exploit the sparse nature of these spikes for energy-efficient computing. However, there exists an accuracy gap between state-of-the-art DNNs and SNNs. To match the accuracy of DNNs, we can build deeper SNNs similar to DNNs. While existing shallow SNNs can be mapped in the local memory of state-of-the-art neuromorphic accelerators, deeper SNNs require larger memories and efficient memory hierarchy.

4 ONLINE LEARNING

Incorporating online learning in neuromorphic edge AI is highly desirable as it would endow systems with the capability to learn and adapt continuously in non-stationary environments, akin to the human brain. Moreover, online learning increases the system's resilience and robustness by conferring robustness to the substrate heterogeneity and drift; this is crucial when using Analog RRAM based CIMs and ULP hardware. However, conventional training algorithms based on back-propagation through time (BPTT) and back-propagation of error (BP) are a poor fit for implementing online learning on edge AI systems due to their associated costs. The main problems are due to backward-locking and their high memory and compute-requirements due to the need for storing intermediate activation maps which makes it difficult to operate them on long sequential data.

Many tasks relevant for edge AI require temporal processing of sequential data. To avoid BPTT is thus one prime desiderata. While real-time recurrent learning (RTRL) avoids the aforementioned problems of backward-locking and its memory requirements, it is computationally even more costly than BPTT. Fortunately, bio-inspired gradient approximations can be obtained in recurrent networks with slowly evolving neuronal variables like encountered in spiking neural networks (SNNs). In this setting, effective diagonal approximations of RTRL based on eligibility traces exist [10, 11]. Their computational cost is comparable to BP, albeit with a vastly reduced memory footprint and avoiding backward-locking in time.

Unfortunately, the situation is different for BP through space and addressing the spatial credit assignment problem requires alternative approaches. Recent studies focus on dynamic algorithms in the flavor of equilibrium propagation (EP) that use the intrinsic network dynamics to implement credit assignment. However, classic EP is sensitive to substrate noise which leads to biased gradient estimates. In recent work we proposed holomorphic EP that partially addresses this issue [12] and gives a glimpse of what algorithmic advances can offer. While holomorphic EP alleviates the noise sensitivity (Fig. 6a,b,c), it relies on weight symmetry and complex numbers. While the former issue can be addressed through suitable homeostatic learning objectives [13] (Fig. 6d), a major future goal is to provide a real-numbered version of the algorithm, for instance, by mapping complex information onto oscillations using phase modulation.

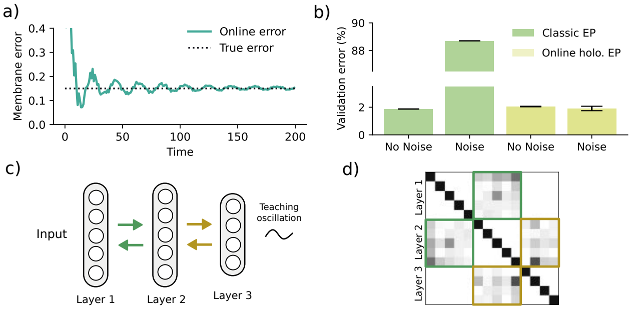


Figure 6: a) In holomorphic EP, the neuronal error required for spatial credit assignment is computed online in a noise-robust manner through temporal integration of oscillations. b) Validation error on MNIST digit recognition task, reproduced from [12, 13]. c) Topology of dynamical network with bidirectionally connected layers of neurons. d) Jacobian of the network near the prediction, whose symmetry encouraged by the homeostatic loss.

Another possible strategy that avoids costly back-propagation is to use greedy local learning rules that operate at different layers independently. Despite exciting progress on local learning rules that draw ideas from self-supervised learning (SSL) and extend to SNNs [14], there still remains a gap to end-to-end optimized networks which needs to be addressed before these methods mature into a viable option.

Finally, heterogeneity, device mismatch, and noise pose considerable challenges for neuromorphic hardware. Once more, online learning algorithms running on-device may offer a way out by effectively allowing the circuit to self-tune through learning. While initial work has demonstrated such capabilities for SNNs trained with surrogate gradients on analog neuromorphic hardware [15], this was accomplished through in-the-loop training; an on-chip demonstration is still pending.

While online learning holds great promise for building ULP and highly adaptable neuromorphic edge AI systems, essential questions remain open. We must channel our future efforts onto lightweight learning rules with the following properties. First, any online learning algorithm must be robust to noise and heterogeneity to be compatible with ultra-low-power (ULP) devices. Second, developing strong alternatives to BP is imperative. To that end, theoretically motivated algorithms like holomorphic EP are promising, but we must formulate them as practical, real-numbered implementations applicable to streaming data. Finally, we need to simultaneously advance the efficiency of local learning rules that further close the gap between greedy learning and end-to-end learning. Crucially, advancing the state-of-the-art on the above points and improving online learning for edge AI will only be possible through concerted efforts in joint algorithms and hardware developments.

5 CONCLUSION

The CONVOLVE project aims at enabling smart edge devices, running Deep Learning models directly on edge sensing devices. This requires energy-efficiency improvements at all levels of the design stack. This paper addresses several problems and potential solutions. In particular it highlights our efforts in 3 areas: 1) Designing CIM architectures, both digital SRAM and analog RRAM based (having different pros and cons), enabling sub-fJ per MAC operation; 2) How to efficiently support highly promising SNN models, closing the current gap with ANNs accelerators, and 3) Showing recent efforts in enabling continuous online learning, by getting rid of the extremely expensive back propagation (in time and in space). The latter is important to make ULP AI-architectures robust for various disturbances and changing environments.

ACKNOWLEDGMENT

CONVOLVE project is funded by EU's Horizon Europe research and innovation programme under grant agreement No. 101070374.

REFERENCES

[1] Maarten J. Molendijk and *et.al.* BrainTTA: A 28.6 TOPS/W Compiler Programmable Transport-Triggered NN SoC. In *ICCD*, 2023.
[2] M. Gomony and *et.al.* CONVOLVE: Smart and seamless design of smart edge processors. *arXiv preprint:2212.00873*, 2023.
[3] Daniele Ielmini and *et.al.* In-memory computing with resistive switching devices. *Nature Electronics*, 2018.
[4] William Andrew Simon and *et.al.* Exact neural networks from inexact multipliers via fibonacci weight encoding. In *DAC*, 2021.
[5] Abhairaj Singh and *et.al.* Low-power memristor-based computing for edge-ai applications. In *ISCAS*, 2021.
[6] Yashvardhan Biyani and *et.al.* Dradnats: An ultra-low power, memristor-based, computation-in-memory architecture using a serial current driver approach (Patent Reference no.: P6114616NL). Octrooicentrum Nederland, 2023.
[7] Aayush Ankit and *et.al.* Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference. In *ASPLOS*, 2019.
[8] Abhairaj Singh and *et.al.* A 115.1 tops/w, 12.1 tops/mm2 computation-in-memory using ring-oscillator based adc for edge ai. In *AICAS*, 2023.
[9] Sebastian Höppner and *et.al.* The spinnaker 2 processing element architecture for hybrid digital neuromorphic computing. *arXiv preprint arXiv:2103.08392*, 2021.
[10] Friedemann Zenke and Emre O. Neftci. Brain-Inspired Learning on Neuromorphic Substrates. *Proceedings of the IEEE*, 2021.
[11] Nicolas Zucchet, Robert Meier, Simon Schug, Asier Mujika, and Joao Sacramento. Online learning of long-range dependencies. *Advances in Neural Information Processing Systems*, 2023.
[12] Axel Laborieux and Friedemann Zenke. Holomorphic Equilibrium Propagation Computes Exact Gradients Through Finite Size Oscillations. *Advances in Neural Information Processing Systems*, 2022.
[13] Axel Laborieux and Friedemann Zenke. Improving equilibrium propagation without weight symmetry through jacobian homeostasis. In *The Twelfth International Conference on Learning Representations*, 2024.
[14] Manu Srinath Halvagal and Friedemann Zenke. The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *Nature Neuroscience*, 2023.
[15] Benjamin Cramer and *et.al.* Surrogate gradients for analog neuromorphic computing. *Proceedings of the National Academy of Sciences*, 2022.