Mixed-Signal In-Memory Multi-bit Matrix-Vector Multiplication

Kévin Hérissé, Benoit Larras, Bruno Stefanelli, Antoine Frappé, Andreas Kaiser

Mixed-Signal In-Memory Multi-bit Matrix-Vector Multiplication

Kévin Hérissé, Benoit Larras, Bruno Stefanelli, Antoine Frappé, Andreas Kaiser
Univ. Lille, CNRS, Centrale Lille, Junia, Univ. Polytechnique Hauts-de-France, UMR 8520 - IEMN, Lille, France
kevin.herisse@junia.com

Abstract

The applications for artificial intelligence are wide and cover multiple domains including industry, health, home automation, consumer electronics, automotive, and smart cities. Application-specific integrated circuits performing tiny machine learning at ultra-low power and high accuracy are needed. The Von Neumann wall forces us to shift the processing elements closer to the memory to prevent data movement and therefore reduce energy consumption. Matrix-Vector Multiplication (MVM) can be achieved with many approaches that perform well with binary weights but not with multi-bit multiplications. This paper tries to highlight the advantages of using current sources to perform in-memory computing, improving further the energy consumption to perform multi-bit MVM.

1. Introduction

Matrix-Vector Multiplication (MVM) is used in all State-of-the-Art (SoA) neural network implementations, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), or Long Short-Term Memory (LSTM). MVM performs Multiplication and Accumulation (MAC) of input and weight vectors. In a Von Neumann architecture [1], we can see that data is fetched 4 times from the memory to perform a MAC operation, as shown in Fig. 1. By estimating the register access cost at 50fJ/byte, the total access cost reaches 200fJ/byte and therefore clamps the system to maximum efficiency of 10TOPS/W (1 MAC = 2 Operations). This is known as the Von Neumann memory wall. In-Memory Computing (IMC) proposes to shift the processing elements inside the memory and therefore gain energy by reducing memory access. Mixed-signal IMC has been studied extensively, but a majority of research focuses on binary (or ternary) solutions. This paper addresses mixed-signal architectures for multi-bit operation.

Section 2 provides an overview of SoA mixed-signal multi-bit solutions for IMC. Section 3 evaluates the potential of switched current sources and their advantages, and Section 4 concludes the document.

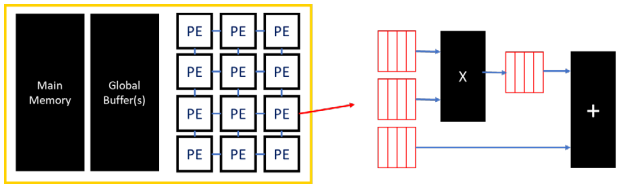

Figure 1 – MAC operation in a Von Neumann architecture

2. Overview of Mixed-Signal IMC for MVM

Fig. 2 highlights the principle of mixed-signal IMC for MVM. A digital input vector is broadcast across multiple accumulation lines (AL) where it is multiplied by the weight vectors stored in the memory array. The accumulation is done in the current or voltage domains and an ADC converts the result at the bottom of each AL.

In this configuration, the ADC usually sets the upper limit in energy efficiency. However, the energy cost of the ADC is shared with multiple MACs across an AL. ADC-limited efficiencies of 200TOPS/W are envisioned [1] for large arrays (>100 MACs). This efficiency value can only be reached if the process of multiplication of input and weight vectors can be realized at a fraction of the ADC energy cost. The next subsections detail and compare several approaches for this multiplication operation.

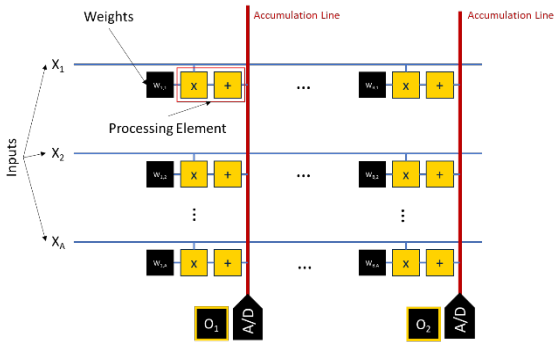

Figure 2 – Principle of mixed-signal IMC for MVM

2.1. ReRAM approach

In the resistive approach, memristors are used to perform the multiplication. The digital input vector is sent as an analog voltage across all the rows of the array. The weights are stored as conductance values of the memristors, modulating the current through the AL. This approach requires DACs with large output currents, and a current comparator, as highlighted in [2].

It is possible to perform multi-bit operations with this approach by using up to 16 states of conductance [3]. However, process variations prevent reaching a higher number of bits. Furthermore, in terms of energy dissipation, writing necessitates high current spikes (around 6µA in [4]), while inference costs approximately 250fJ/MAC, according to [5]. It is to note that a trade-off exists between energy consumption and variability.

2.2. Switched Capacitor approach

In the capacitive approach, the multiplication is based on sharing charges or currents on a capacitive line, usually using XNOR gates for binary multiplication [6]. Unit capacitors are charged according to each binary multiplication and charges are redistributed across all capacitors on an AL, resulting in a voltage to be converted by the ADC [7]. To perform a multi-bit operation with switched capacitors, [5] shows a topology similar to a digital multiplier using one AL for each bit. However, this method needs additional circuitry to combine all the line's results, which is adequate for a reduced number of lines (<5) but dominates the power consumption for higher numbers of bits.

3. Switched Current Sources advantages

By using current sources to charge or discharge a capacitive line, multi-bit operations can be performed without additional circuits at low energy consumption. Current sources are distributed over the AL and are switched by the activation lines with Pulse Width Modulated (PWM) signals. Recent CMOS technologies allow driving small currents precisely enough to perform MVM with good accuracy. The main principle is to take advantage of the time to charge the capacitor with ultra-small currents. This solution is specifically recommended for applications with a throughput lower than 50GOPS, such as Key Word Spotting (KWS), Voice Activity Detection (VAD), or sound recognition. For example, for an AL composed of 100 multi-bit MACs, a unitary 100pA current, and an AL capacitance equivalent to 100fF, approximately 5µs are needed to charge a capacitor to 0.5V (allowing for signed operation in a 1V supply). The equivalent energy consumption for the computation of the MVM is then calculated to 0.5fJ/MAC (4000 TOPS/W). Using this configuration, the system's core consumption is small enough so that the total energy will be eventually dominated by the ADC and reaches 10fJ/MAC (200TOPS/W) for an 8-bit SoA ADC [1].

Furthermore, another advantage of switched current sources is the configurability, since the PWM period or the current level can be tuned to address many applications. Changing the current value modifies the gain of the multiplication so that the number of MACs per AL can be adjusted. [5] also showed that mixed-signal IMC is suitable for bit resolutions up to 8 bits and that fully digital solutions should be considered for wider bit width.

	Memristor	**Switched Capacitor**	**Switched Current Source**
Configurability	-	-	+
Technology	Non-standard (CMOS + RRAM)	Standard CMOS	**Standard CMOS**
Typical energy efficiency	21.9 TOPS/W [8]	10 TOPS/W [5]	**200 TOPS/W**
Max number of bits	3 – 4	8	**8**

Table 1 – Comparison of computation methods for multi-bit MVM

4. Conclusion

Switched current sources offer great opportunities for configurable, low-energy multi-bit MVM. Enabled by ultra-low currents in advanced CMOS technologies, this contribution proposes to trade-off time to reduce energy. It allows reaching the theoretical ADC-limited consumption of 200TOPS/W for applications with low throughput requirements.

References

[1] B. Murmann, "Mixed-Signal Processing Opportunities for AI," *IEEE ESSCIRC 2020.*

[2] S. Cosemans *et al.*, "Towards 10000TOPS/W DNN Inference with Analog in-Memory Computing – A Circuit Blueprint, Device Options and Requirements," in *IEDM*, 2019

[3] E. R. Hsieh *et al.*, "Four-Bits-Per-Memory One-Transistor-and-Eight-Resistive-Random-Access-Memory (1T8R) Array," *IEEE Electron Device Lett.*, 2021

[4] N. C. Dao and D. Koch, "Memristor-based Reconfigurable Circuits: Challenges in Implementation," in *ICEIC*, 2020

[5] B. Murmann, "Mixed-Signal Computing for Deep Neural Network Inference," *TVLSI.*, 2021

[6] P. C. Knag *et al.*, "A 617-TOPS/W All-Digital Binary Neural Network Accelerator in 10-nm FinFET CMOS," *JSSC*, 2021

[7] D. Bankman and B. Murmann, "An 8-bit, 16 input, 3.2 pJ/op switched-capacitor dot product circuit in 28-nm FDSOI CMOS," in *A-SSCC*, 2016

[8] C.-X. Xue *et al.*, "24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," in *ISSCC*, 2019