Towards 10000TOPS/W DNN Inference with Analog in-Memory Computing – A Circuit Blueprint, Device Options and Requirements

Conference Paper · December 2019

DOI: 10.1109/IEDM19573.2019.8993599

8 authors, including:

Stefan Cosemans
axelera AI
81 PUBLICATIONS **957** CITATIONS
SEE PROFILE

Bram-Ernst Verhoef
imec
36 PUBLICATIONS **721** CITATIONS
SEE PROFILE

Peter Debacker
imec
74 PUBLICATIONS **636** CITATIONS
SEE PROFILE

Arindam Mallik
imec
72 PUBLICATIONS **831** CITATIONS
SEE PROFILE

Towards 10000TOPS/W DNN Inference with Analog in-Memory Computing – A Circuit Blueprint, Device Options and Requirements

S. Cosemans[1], B. Verhoef[1], J. Doevenspeck[1,2], I.A. Papistas[1], F. Catthoor[1,2], P. Debacker[1], A. Mallik[1], D. Verkest[1]
[1]imec, Leuven, Belgium, email: stefan.cosemans@imec.be ; [2]KU Leuven, Leuven, Belgium

***Abstract*—**This paper presents a blueprint for a 10000TOPS/W matrix-vector multiplier for neural network inference based on Analog in-Memory Computing (AiMC), an energy efficiency at least 10x beyond ultimate digital implementations. The presented analysis connects circuit design with technology options and requirements. A compute array using pulse-width encoded activations and precharge-discharge summation line is used as circuit blueprint, key device requirements for this compute array are derived and 3 suited device options are discussed: SOT-MRAM, IGZO-based 2T1C DRAM gain cell, and projection PCM with separate write path.

I. Introduction

Deep learning constitutes the state-of-the-art in AI, from image processing to translation and speech recognition. Energy-efficient inference is key to bringing these capabilities to edge devices. Both for convolutional neural networks (CNN) and recurrent networks such as LSTMs (Long Short-Term Memory), the dominant operations for inference are low-precision matrix-vector multiplications (MVM) of a weight matrix with an input vector. This paper focusses on implementing MVM using Analog in-Memory Computing (AiMC). See [1,2] for an overview of recent progress in the field. Complete circuit implementations have been presented, mostly based on capacitive summation or averaging [3,4,5]. Pulse-width modulation has been applied in prior work, e.g. [6], and [7] uses SRAM sub-threshold current and WL pulse width encoding to realize binary weighted weights. Previous device-centered work has projected significant gains, e.g. [8], but periphery assumptions are rather implicit.

II. Setting the Energy and Accuracy Bar

An optimized Multiply-Accumulate (MAC) (3-bit weight x 4-bit activation, 16-bit running sum) in a 16nm node consumes only 10fJ/MAC [9]. Further optimizations and scaling will still improve the energy efficiency of digital implementations, but an AiMC MVM achieving 0.1fJ/operation (10000TOPS/W) will certainly be an attractive option for many applications, especially if it can be implemented in a more accessible technology node. This paper provides estimates for periphery for a 22nm node. Note that 1 MAC is counted as 2 operations.

For many applications, AiMC will only be a viable option if it provides close to the state-of-the-art accuracy. Notice that there is no need to achieve equivalent accuracy on any given network topology though – it is acceptable to e.g. increase the network size to recover accuracy if the overall power, performance, area and cost is significantly better.

Analog computation is likely only attractive if one can tolerate low precision operations, some variation and some systematic errors, and still achieve high accuracy. Low precision quantization is also a key enabler for efficient digital implementation, and important progress has been made. 2-bit quantized networks (both weight and activations) now achieve good accuracy on ImageNet [10]. If variations and errors are included during training, NNs seem robust to a limited amount of noise [11]. The exact amount of variation that can be tolerated depends not only on the problem and the network, but also on the quantization and training approach. Fig. 1 shows that for a character-prediction LSTM, 5 to 10% σ/LSB variation on the weights can be tolerated, but 20% significantly degrades the result. This and other similar results lead us to consider only device concepts that can achieve σ/range below 10%.

III. Compute Array with Pulse-Width Encoding

Fig. 2 depicts a generic AiMC architecture. The weight matrix is stored in the compute cells. Digital inputs are converted to analog signals on the "activation lines". Each cell contributes to its summation line a current or charge proportional to the product of its activation and weight. The resulting voltage swing on each summation line is then proportional to the desired result. An ADC quantizes the result.

Fig. 3 shows a popular implementation: programmable resistors store the weights as conductance, activations are encoded as voltage levels, the summation line is clamped to a fixed reference level, and each cell contributes a current $I_{cell} = V_{act}/R_{cell} \propto activation \cdot weight$. Notice that even though the evaluation operation does not require selectors, the write operation typically does. This approach has some significant issues: clamping the summation line is energetically expensive if feasible at all, a voltage DAC that can source significant current is required, and IR voltage drops on both activation line and summation line affect the result. It might be possible to drop the clamp. The result will not be an exact MVM, but if the summation line swing is kept small, it will likely be possible to compensate for this during training.

The approach taken in this paper (Fig. 4) avoids these issues. Weights are encoded in the current level of current source-like elements (CSE) – in practice transistors operating in saturation or subthreshold with sufficiently large V_{ds}. Flash transistors provide a direct physical implementation, but similar behavior can be obtained with other cells, e.g. the IGZO 2T1C DRAM cell discussed below. Resistive elements (RE) can also be used, but the result will not be an exact MVM.

Activations are pulse-width (PW) encoded. Assuming the array time constants are sufficiently large, many activation levels can easily be provided at low cost – e.g. 16 levels: "off", 100ps, 200ps, ..., 1.5ns. Using the digital-to-time converter from [5], even 256 levels could be affordable. Summation lines are not clamped, but simply precharged, and then discharged by cell charge $Q_{cell} = I_{cell}(weight) \cdot T_{act} \propto weight \cdot act$. The result is captured by a regular voltage-input ADC. This approach requires circuit design that copes with all transient effects.

Two important effects place an upper bound on RE conductance and CSE current: IR drop on the summation line (SL), and the speed of operation. Fig. 5 shows IR drop for an example scenario – 1024 cells per SL, 1Ω wire resistance per cell (representative for 22nm node). Although several mitigation options can be considered, this suggests that cells with R_{LRS}<1MΩ or I_{on}>1μA come with a penalty. Fig. 6 shows the time it takes the cells to discharge the SL capacitance in a precharge-discharge scheme. For PW-encoded activations, multiple levels must be placed within this time window. The shorter the available time, the harder this is. With 0.1fF/cell capacitance and 25% active cells, a 1ns target corresponds to approximately 5MΩ or 0.2μA. Adding additional capacitance on the SL increases the time constants but also increases energy consumption. In a SL clamp scheme, the consequences of this speed constraint are different – if the periphery cannot keep up, DC current flows longer than necessary, wasting energy.

IV. Simple Energy and Throughput Estimates

Fig. 7 lists the key contributions to MVM energy consumption: DAC, activation lines, summation lines and ADC. The table assumes a fairy-tale device, performing MVM at logic supply voltage (0.8V), with capacitance including wiring comparable to a 22nm logic transistor (0.1fF on activation and summation line). We believe all 3 cell types discussed in section V can eventually meet or exceed these assumptions. Assuming a 100fJ 5-bit ADC, 15fJ 5-bit DAC and MVM size of N=1024 inputs and M=1024 outputs, leads to 0.2fJ/MAC, or 10000TOPS/W. Some key parameters affect this value: (1) MVM size: with fewer inputs N, ADC energy is amortized over fewer MACs. (2) Activation and summation line capacitance. (3) Operating voltage on these lines. (4) Required ADC properties. Some authors [12] assert that high ENOB ADCs are required for accurate results on more complicated tasks such as imageNet, but that seems at odds with both our own work on quantization [13] and with the 2-bit quantized results from [10].

The lowest energy AiMC architecture keeps the weights fully stationary. This reduces endurance requirements and relaxes write energy and time constraints, but it does require that each weight of each network of the application is stored in a compute cell, hence cell area is critical. Fig. 8 suggests that even a fairly large 0.1μm^2 cell enables state-of-the-art applications. For even smaller cells, periphery area dominates.

A 100MHz 1Kx1K array provides 200 TOPS. Assuming 0.1μm^2 effective area per compute cell, that is 2000TOPS/mm^2. Fig. 9 situates this relative to state-of-the-art implementations.

V. Device Options and Requirements

We believe that the following are the key parameters:

- High Rcell (>1MΩ) or low cell current Icell (<1μA)
- Low variation: σ/range<10%. Write-verify acceptable.
- Evaluation operation at core logic-compatible voltage.
- Small area; but even 0.1um^2 cell enables applications.

Fig. 10 shows two popular cell types that don't meet these specifications. Variation in typical filamentary ReRAM is too high when operated at sufficiently high resistance levels. STT-MRAM cannot have a high resistance level, as this would make it impossible to write the cell. Flash could be an interesting option, but the high write voltage poses daunting challenges.

Fig. 11 shows a 2T1C DRAM gain cell. The weight is stored on the capacitor. The PW encoded activation is applied either on the second capacitor terminal, or on the transistor source. The readout transistor acts as CSE. A cell based on regular transistors would not work, as the analog cell state would leak away too fast, and the readout current would be too high except if the transistor is operated in sub-threshold, which increases variations drastically. Using IGZO transistors [14,15] solves both problems – the extremely low leakage maintains the state for a sufficiently long time, and the lower mobility reduces the cell current to the desired level. This cell can be stacked in a 3D monolithic way in the BEOL, leaving the FEOL available for peripheral circuits, enabling a very small footprint.

In SOT-MRAM [16] (Fig. 12), the write current does not flow through the MTJ, and hence a high readout resistance can be obtained by using a thick M_gO layer. This device option is analyzed in [17], where it is shown that the low on/off ratio of the cell is a disadvantage but not a showstopper.

PCM with projection layer (Fig. 13) [18,19] solves part of the resistance drift that plagues normal PCM, at the cost of a strongly reduced on/off window. Claims of 256 levels and more have been made for training. Only limited data has been presented regarding drift variation and retention, but based on this data, our assessment is that, for inference, 8 meaningful levels should be feasible. A weak point of this cell is still the write selector. An OTS selector is not feasible for AiMC summation due to its high hold current. A transistor selector would increase cell area and capacitance, as write requires high voltage and rather high current. Fig. 14 depicts a novel solution to this problem: by separating the write path (heater) from the read path (projection PCM), an OTS selector can be used on the write path, and the read path can either be used without selector, or with a minimal-size core transistor.

VI. Summary and Conclusion

This paper describes a blueprint for a 10000TOPS/W analog matrix-vector multiplier for DNN inference (5-bit inputs, 5-bit ADC, number of weight levels depending on memory element). Low precision quantization and training with variation and circuit errors in the loop is mandatory. Three memory device concepts are discussed that meet the requirements imposed by the presented blueprint: IGZO-based 2T1C DRAM, SOT-MRAM and projection PCM with separated write path.

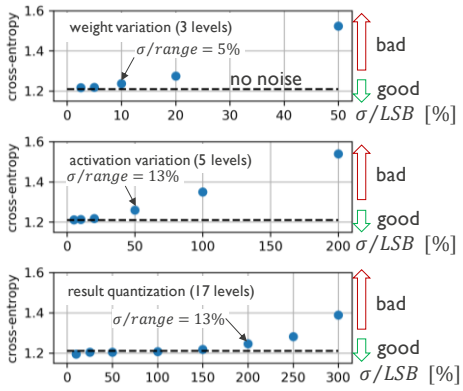


Fig 1. If properly trained with variation, DNNs can tolerate some variation on weights, activations and result quantization, but for this example, results degrade already at 5% to 13% $\sigma/range$. (Example: LSTM for WARPEACE dataset)

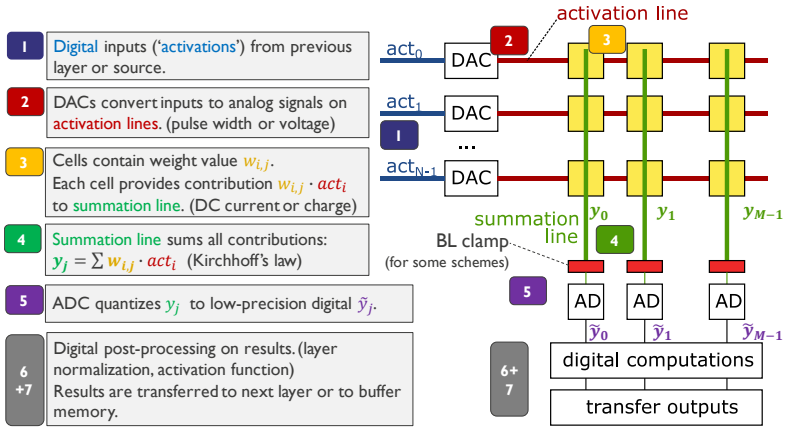


Fig. 2. General concept of Matrix-Vector Multiplier (MVM) using Analog in-Memory Computing (AiMC).

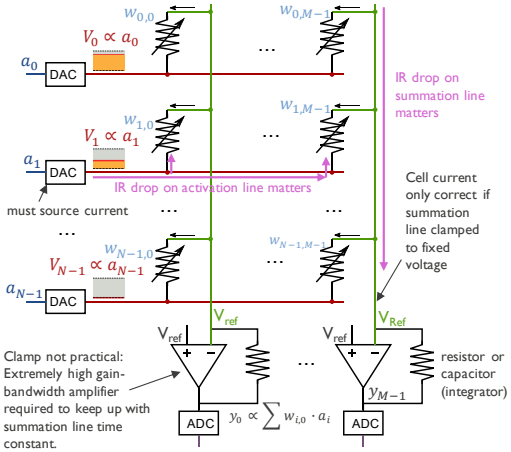


Fig. 3. A popular implementation: weights stored as 'memristor' conductance, voltage-encoded activations and summation line clamped to a fixed reference level.

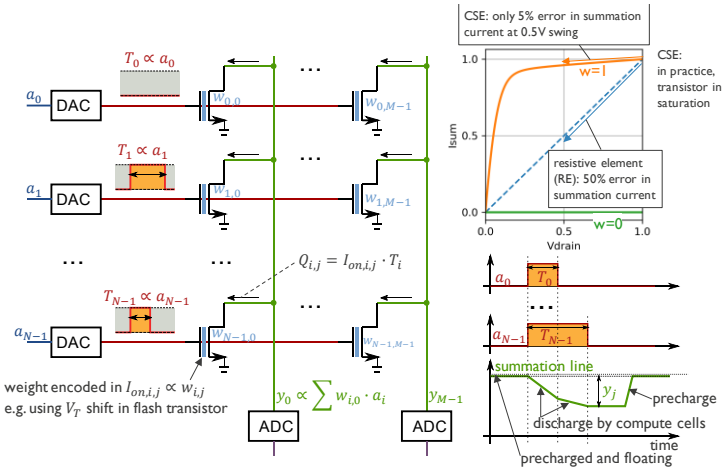


Fig. 4. MVM using current source elements (CSE), PW modulated activations and a precharge-discharge scheme on the summation lines.

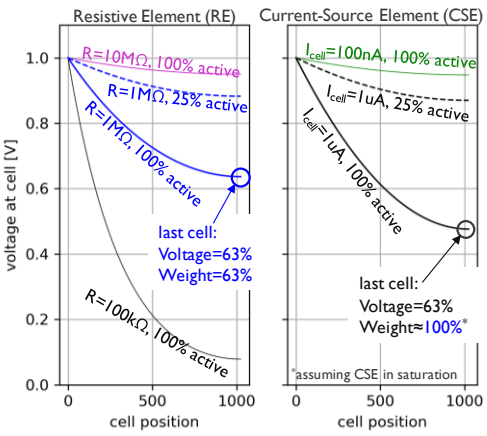


Fig. 5. Wire IR drop on summation line (SL) for 1024 cells per SL, 1Ω wire resistance per cell, for different cell resistance levels and cell current levels. inactive cells have R=∞ or I=0uA. In these graphs, SL is clamped to 1V at cell position 0. For CSE elements, voltage drop on the SL has limited impact on the effective weight level (cell current) if the device remains in saturation.

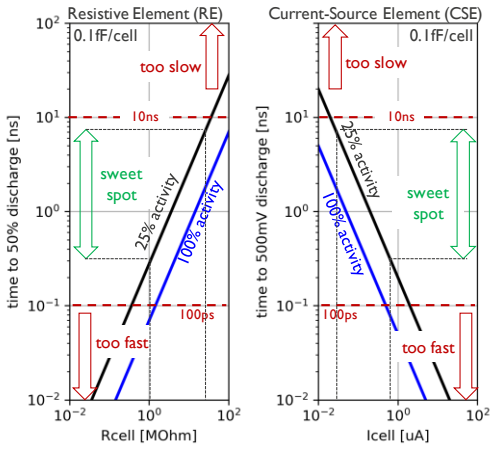


Fig. 6. Time it takes the cells to discharge the summation line capacitance by 50% (RE) or 500mV (CSE). Only array capacitance of 0.1fF/cell is considered. If discharge is too fast, control circuitry cannot keep up, making operation impossible for precharge-discharge summation line schemes, or flushing current longer than necessary for clamp-based schemes.

Assumptions		
vdd	0.8 [V]	
5-bit PWM DAC*	15 [fJ]	
5-bit ADC*	100 [fJ]	0.1fF corresponds to 22nm minimal FET or 0.5um wire
activation line C	0.1 [fF]	
activation line activity	80%	
activation line swing	0.8 [V]	
summation line C	0.1 [fF]	
average summation line swing	0.4 [V]	

*estimated based on a similar 22nm design (post-layout)

Energy estimates for precharge-discharge scheme with PW encoded activations				
dot product size		large	medium	small
# inputs (N)		4096	1024	128
# outputs (M)		4096	1024	128
# parallel MACs		16M	1M	16K
energy all DACs	[pJ]	61	15	1.9
energy all ADCs	[pJ]	410	102	13
energy activation lines	[pJ]	859	54	0.8
energy summation lines	[pJ]	537	34	0.5
control and timing	[pJ]	2	2	2
total	[pJ]	1869	207	18
energy per MAC	[fJ/MAC]	0.11	0.20	1.10
energy per operation	[fJ/op]	0.06	0.10	0.55
energy efficiency	[TOPS/W]	17954	10131	1812

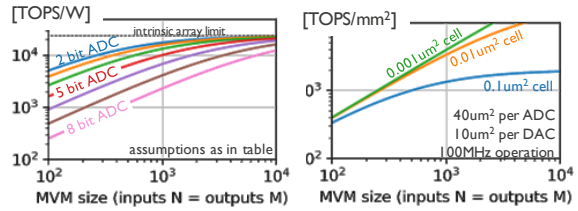


Fig.7. Energy estimates for MVM using scheme from Fig. 4, assuming a fairy-tale memory compute device. We believe that the 3 cell types discussed in section V can eventually meet or exceed these assumptions.

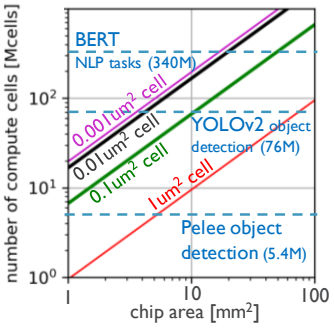


Fig. 8. Number of stored weights as function of chip area, assuming periphery area of 50um² per 1000 cells. Even relatively large 0.1um² cells might enable real-world applications at an affordable cost. (Indicated networks have not yet been mapped on AiMC)

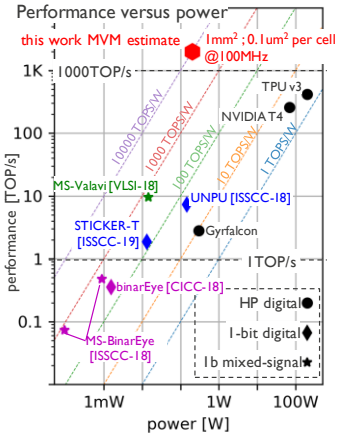


Fig.9. Situating the estimates relative to published implementations. Note the differences in operation precision, and that other datapoints include not only the MVM but also other operations and data transfers.

Filamentary ReRAM (OxRAM, CBRAM)

Variation too high when operated at sufficiently high resistance

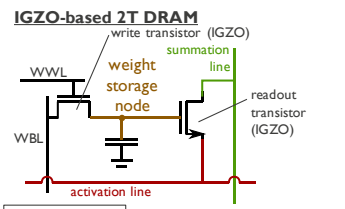
Write selector penalizes ReRAM

- High-voltage FET per cell → larger area and energy
- 1S1R selector → higher evaluation voltage → higher energy

STT-MRAM

Resistance must be low to enable the write operation → Resistance too low for AiMC MVM

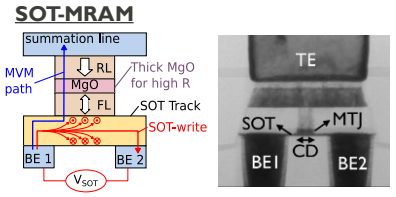
Fig. 10. Typical filamentary ReRAM and STT-MRAM don't meet the specifications required for an efficient AiMC MVM implementation.



normal Si FET not suited:
- state leaks away too fast
- readout current too high (or subthreshold)

IGZO-based
+ low leakage → analog retention
+ low readout current
+ FEOL available for periphery
+ multi-level weights possible
+ current-source like
- relatively large cell (~0.1um² in 22nm node)
- not non-volatile, refresh needed

Fig. 11. IGZO-based 2T DRAM.



- Relatively large cell (but <0.1um² in 22nm)
- selector occupies FEOL
- Low on/off ratio
- 2 states only

+ high R_{LRS}, good σ
+ logic-compatible write V
+ high endurance

Fig. 12. In SOT-MRAM [16], write and read path are separated. This allows achieving the desired high resistance level (1—10MΩ) by increasing MgO thickness.

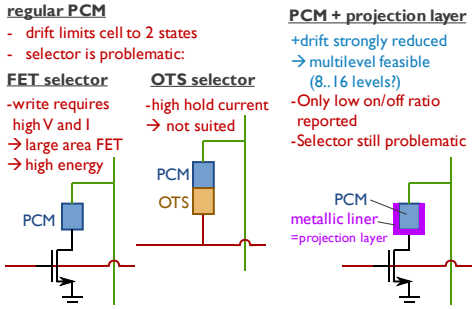


Fig. 13. Regular PCM requires a high-V, high-I selector for write, and supports only 2 cell states because of drift. PCM with projection layer solves drift but does not improve the selector problem.

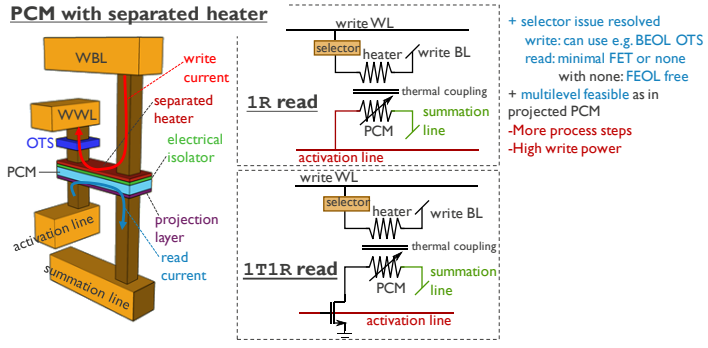


Fig. 14. Novel PCM with separated write path resolves the selector issue.

[1] H. Tsai et al., Journal of Physics D: Applied Physics, Jun. 2018
[2] D. Ielmini and H.-S. P. Wong, Nature Electronics, Jun. 2018
[3] D. Bankman et al., ISSCC 2018
[4] H. Valavi et al., JSSC Jun. 2019
[5] A. Biswas and A. P. Chandrakasan, JSSC Jan. 2019
[6] H. Jiang et al., ISCAS 2018
[7] M. Kang et al., JSSC Feb. 2018
[8] M. Hu et al., DAC 2017
[9] W. J. Dally et al., VLSI circuits 2018
[10] J. Choi et al., SysML Conference 2019
[11] Z. He et al., DAC 2019
[12] A. Rekhi et al., DAC 2019
[13] B. Verhoef et al., under revision
[14] H. Kunitake et al., IEDM 2018
[15] J. Mitard et al., submitted for IEDM 2019
[16] K. Garello et. al., VLSI 2019
[17] J. Doevenspeck et al., submitted for IEDM 2019
[18] I. Giannopoulos et al., IEDM 2018
[19] W. Kim et al., VLSI 2019