Mixed-Signal Neuromorphic Processors: Quo Vadis?

Mohammad Bavandpour
ECE Department
UC Santa Barbara
Santa Barbara, CA, US
mbavandpour@ece.ucsb.edu

Mohammad Reza Mahmoodi
ECE Department
UC Santa Barbara
Santa Barbara, CA, US
mrmahmoodi@ece.ucsb.edu

Shubham Sahay
ECE Department
UC Santa Barbara
Santa Barbara, CA, US
shubhamsahay@ece.ucsb.edu

Dmitri B. Strukov
ECE Department
UC Santa Barbara
Santa Barbara, CA, US
strukov@ece.ucsb.edu

***Abstract*—This paper outlines different design options and most suitable memory devices for implementing dense vector-by-matrix multiplication operation, the key operation in neuromorphic computing. The considered approaches are evaluated by modeling system-level performance of 55-nm 4-bit mixed-signal neuromorphic inference processor running common deep learning feedforward and recurrent neural network models.**

Keywords—Nonvolatile Memory Device, Mixed-Signal Circuits, Neuromorphic Processor, Vector-by-Matrix Multiplication

I. Introduction

The growing applications of neural networks calls for the development of efficient neuromorphic computing hardware. A very promising approach to address this need is to utilize mixed-signal (MS) circuits based on emerging nonvolatile memories (NVMs), which enables dense in-memory vector-by-matrix multiplication (VMM) with low-to-medium precision, the most critical operation in inference computation.

Though the general idea is similar for many MS-VMM circuits (Fig. 1a) [1, 2], there are differences in peripheral circuitry design and how input/output signals and weight are encoded, which in turn favor the specific choice of NVM. The goal of this paper is to compare different approaches and provide examples of their use in MS neuromorphic processor.

II. Mixed-Signal Multiplier Design Options

MS-VMM circuits can be broadly classified by the type of input encoding and utilized signal amplification in the crosspoint memory cells (Fig. 1b). The choice of these options determines the optimal design of other parts VMM circuit, e.g. of the output integration (OI) and conversion (OC) circuits.

Specifically, for the fastest, **instant** (INS) encoding, the inputs are encoded by the amplitudes of the fixed-duration voltage pulses (or current pulses in gate-coupled design [3]). In the **linear** (LIN) encoding, the inputs are applied sequentially, bit by bit, using fixed-duration digital pulses, so that the total input duration (T_{in}) is proportional to the input signal precision (p) [4]. For the slowest, **exponential** (EXP) time-based encoding, the inputs are encoded in the duration of the digital pulses, with the worst-case input time scaling as 2^p with precision [5]. Digital inputs of LIN (and EXP) scheme eliminate the need for (allow to replace with more compact counters) potentially bulky DAC circuits needed in INS approach, at the cost of adding more complex clock distribution

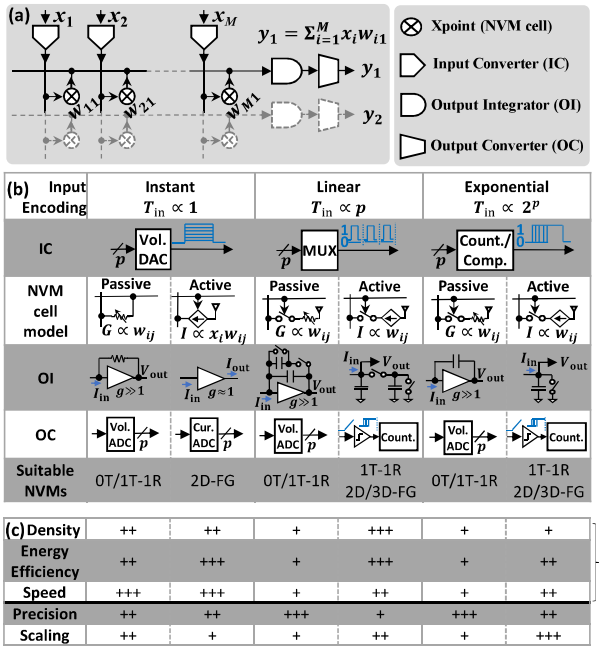
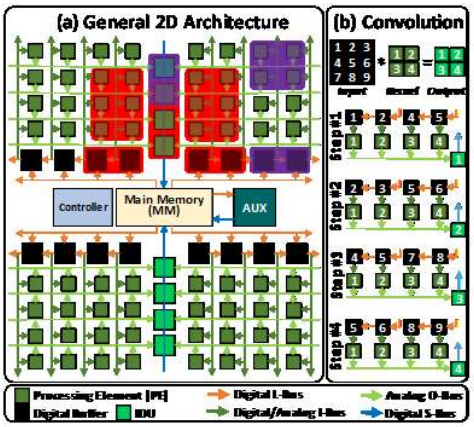


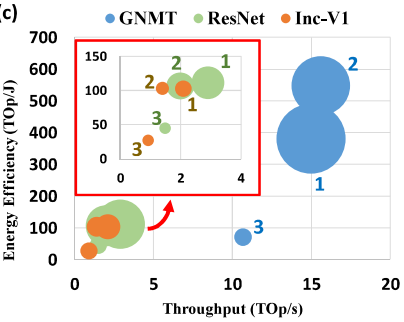
Fig. 1. (a) Top level schematics for NVM-based mixed-signal VMM circuitry. (b) Design options. The first two / last four columns are also reffered as current-mode / charge-based designs. (c) Performance comparison. The last two rows show potentials for increasing input/compute precision and improving all performance metrics due to memory/periphery feature scaling. 3D-NAND with INS encoding is also possible though only after substantial array redesign.

networks and other synchronization circuits. Higher switching activity in LIN, though, could result in higher dynamic energy.

All floating gate memories in MS-VMM circuits are typically biased in subthreshold mode, thus providing signal amplification in the cell [2,3,5,6]. A subthreshold-based transistor in 1T1R cells can be also used as adjustable current source, by tuning resistance of source-connected memristor [4,7]. Alternatively, 1T1R (and 0T1R) can be used as purely passive cells.

The lack of cell's signal amplification generally means using more expensive active peripheral circuits. Indeed, the biggest advantage of active cells is high input/output array impedance, which greatly cuts OI and IC overheads [2]. On the other hand, the overhead of OI in passive cells circuit is typically quite heavy due high-gain sense amplifiers, and especially bad for LIN and EXP due to additional requirement of high bandwidth. A potential strength for active OI is superior precision due to better control of an input current. It should be





(d) NVM	2D-NOR			1T-1R			3D-NAND		
Network	Inc-V1	ResNet	GNMT	Inc-V1	ResNet	GNMT	Inc-V1	ResNet	GNMT
AE (MB/mm²)	**0.17**	**0.26**	**0.27**	**0.26**	**0.37**	**0.38**	**4.3**		
Area (mm²)	37.9	147	293	24	102	209	18.4		
NVM (%)	24.6	36.8	39.3	37.5	50.9	53.3	2.95		
Periphery (%)	46.9	44.6	43.7	27.6	36.1	37.2	59.1		
Other (digital) (%)	28.5	18.7	17.0	34.8	13.0	9.5	37.9		
EE (TOp/J)	**103**	**111**	**380**	**103**	**107**	**548**	**27.4**	**44.5**	**70.4**
Computation (%)	29.3	32.3	69.3	19.8	25.8	55.8	67.5	73.2	94
Communication (%)	32.2	43.4	18.4	41.5	50.8	26.5	22.2	17.0	3.9
Memory access (%)	38.5	24.3	12.3	38.7	23.3	17.7	10.3	9.7	2.3
Throughput (TOp/s)	**2.1**	**2.9**	**15**	**1.4**	**1.98**	**15.6**	**0.9**	**1.5**	**10.7**
Inference time (ms)	2.3	6.5	0.16	3.4	9.5	0.16	5.2	12.6	0.23

Fig. 2. (a) The main components of aCortex architecture including centralized eDRAM-based memory, a configurable chain of digital buffers, 2D/3D arrays of MS-VMM blocks (PE), an array of neurons (IDU), and a digital auxiliary unit (AUX) used for infrequent pooling/addition/vector-vector multiplication operations. (b) Example of data movement during convolution operation on 2D-aCortex. The whole computation is perfomed one layer at a time by enabling specific PE blocks. (c) System-level estimates of major performance metrics and (d) thier breakdown. In panel c, the circle size represents the chip area, which is reported assuming minimum resources (mainly PEs) needed to run specific model. Note that due to possibility of disabling not utilized components, EE is almost the same when mapping smaller models on the largest, GNMT-compatible processor [2]. All circuit/memory assumptions are similar to cited references.

also noted that in EXP approach, OI overhead could be significant for large p due to large integration times and hence larger capacitors. This is less of an issue for passive LIN design due to efficient successive integration and division scheme [4].

The OIs' outputs are turned into digital signals with ADCs or by converting output voltage into pulse duration and using time-to-digital converter. The latter approach is more practical for active LIN and EXP schemes, due to their negligible IC's static energy consumption [4-7]. Also, the digital circuits in OI, IC, and OC are generally more efficient and conducive to aggressive technology scaling, compared to the analog ones.

Main metrics of considered approaches are outlined in Fig. 1c. (Note that Fig. 1 omits some more "digital" options, e.g. of using LIN with digital integration [8].)

III. Case Studies for Neuromorphic Processor

Three representative designs based on active cell 2D-NOR [2, 3], 1T-1R [4, 7] and 3D-NAND [6] were evaluated by modeling inference performance of 4-bit aCortex (Fig. 2a), a multi-purpose neuromorphic inference processor [2, 6], for popular deep learning models, such as GNMT-1024 recurrent network (with input sequence of 10), and image classifiers ResNet-152 and Inception-V1. The system-level estimates are based on simulations in 55-nm process, and, e.g., included line/device parasitics, leakages, and overheads of buses and tuning circuitry.

As expected, area-efficiency (AE), which is defined as the weight capacity normalized to processor area, is the best for 3D-NAND approach due to very dense memory cells (Fig. 2c, d). The second best is 1T-1R design due to relatively small cell area and very compact periphery. Energy efficiency (EE) closely follows AE for the first two approaches, while worse for the 3D-NAND design due to much larger parasitics, i.e. high capacitance word planes/bit select lines and high pass voltages. The instant encoding and low operating currents of the 2D-NOR approach leads to faster VMM operation (Fig. 1c) and hence the highest system-level throughput for smaller networks. However, due to very compact periphery for both sensing and front-end and back-end conversion circuits, 1T-1R design has the best throughput at the system level. The superior memory density and relatively low assumed computing precision also help achieving high throughput for larger models in the 3D-NAND approach.

Though the performance is generally much better compared to purely digital implementations, there are still many reserves for improvements. For example, shrinking the cell area in 1T-1R design would improve performance. 3D-NAND approach would benefit from more compact, previously demonstrated capacitor implementations, while its AE can be further improved by sharing peripheral circuitry. Finally, let us note that the considered version of aCortex is optimized for EE. A better throughput can be achieved by sacrificing EE at the circuit and architecture levels. Understanding such tradeoffs is important future goal. Also, though preliminary results for sensitivity of performance to device and circuit non-idealities are encouraging [9], more extensive experimental verifications are needed.

References

[1] G.W. Burr *et al.*, "Neuromorphic computing using non-volatile memory", *Advances in Physics: X*, vol. 2, pp. 89-124, 2019.

[2] M. Bavandpour *et al.*, "Mixed-signal neuromorphic inference accelerators: recent results and future prospects," in: *Proc. IEDM'18*, San Francisco, CA, Dec. 2018, p. 20.4.1.

[3] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells", in: *Proc. CICC'17*, Austin, TX, Apr.-May 2017, p. 1.

[4] M. Bavandpour *et al.*, "Efficient mixed-signal neurocomputing via successive integration and division," unpublished, June 2019.

[5] M. Bavandpour, M.R. Mahmoodi, D. Strukov, "Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond," *IEEE TCAS-II*, 2019 (early access).

[6] M. Bavandpour, S. Sahay, M. R. Mahmoodi, and D. Strukov, "3D-aCortex: An ultra-compact energy-efficient neurocomputing platform based on commercial 3D-NAND flash memories," submitted to *MICRO'19*, 2019.

[7] S. Sahay, M. Bavandpour, M.R. Mahmoodi, and D. Strukov, "Time-domain mixed-signal vector-by-matrix multiplier exploiting 1T-1R array," arXiv:1905.09454, 2019.

[8] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *Computer Architecture News*, vol. 44, pp.14-26, 2016.

[9] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in: *Proc. IEDM'17*, San Francisco, CA, Dec. 2017, p. 6.5.1.