Mixed-Signal Neuromorphic Inference Accelerators: Recent Results and Future Prospects

M. Bavandpour*, M.R. Mahmoodi*, H. Nili, F. Merrikh Bayat, M. Prezioso, A. Vincent, and D.B. Strukov#
UC Santa Barbara, Santa Barbara, CA 93106-9560, U.S.A.
* equal contributions, # strukov@ece.ucsb.edu

K.K. Likharev
Stony Brook University, Stony Brook, NY 11794-3800, U.S.A.
Konstantin.Likharev@stonybrook.edu

Abstract- Recent advances in dense, continuous-state nonvolatile memories have enabled extremely fast, compact, and energy-efficient analog and mixed-signal circuits. Such circuits are perfectly suited, in particular, for hardware implementations of the inference operation in advanced neuromorphic networks, which requires massive amounts of dot-product operations with low-to-medium precision. In this paper, we first review typical implementations of such mixed-signal circuits. We then describe some recent experimental demonstrations of prototype mixed-signal neuromorphic networks by our team, in particular, a mixed-signal inference accelerator with unprecedented speed and energy efficiency. The paper is concluded by outlining some urgently needed work, in particular the development of high-performance general-purpose inference accelerators, and discussing our preliminary results in this direction.

I. Introduction

The rapidly growing range of applications of machine learning for image classification, speech recognition, and natural language processing have led to an urgent need in specialized neuromorphic hardware. Of that, there is much more demand for fast, low-precision inference accelerators than for higher-precision systems for network training [1].

Though the vast majority of demonstrated accelerators from industry [1-3] and academia [4, 5] are digital, the most natural approaches, however, are based on analog and mixed-signal circuits [6-13]. Though the core principles of analog computing had been developed almost four decades ago [14, 15], its efficient implementations were enabled only recently by the appearance of novel continuous-state, nonvolatile memory devices [16] - the most crucial elements of analog circuits.

II. Mixed-Signal Circuits Using Emerging Memories

Fig. 1 shows typical mixed-signal circuits for the implementation of the vector-by-matrix multiplication (VMM), i.e. the most important operation in inference accelerators and other neuromorphic tasks, while Fig. 2 provides their qualitative comparison. Specifically, due to their superior integration density, VMMs based on passive crossbars with resistive nonvolatile devices (Fig. 1.I) [6], including metal-oxide memristors, conductive-bridge and phase-change memories, might be the most promising in the long term. Passive integration is, however, significantly more challenging, since in this case the distribution in the memristors' effective switching thresholds should be narrow enough to avoid the disturbance of already tuned devices at their half-selection (Fig. 3a,b). Additional gate lines in active crossbars with 1T1R cells (inset of Fig. 1.I) solve the half-select problem [8, 12] and allow for either higher device variations at synaptic weight tuning (Fig. 3c), or higher precision of the finite weights, or both. (The cell's selector functionality, the main advantage of the 1T1R approach for digital memories, is less important for neuromorphic inference applications, since writes are typically very infrequent.)

Though the integration density of the floating-gate (FG) circuits (Figs. 1.II and 1.III) is comparable with that of systems using 1T1R cells, the fabrication technology available for the latter approach is more scalable. The main relative advantage of the former approach is the FG cell's amplification, that relaxes the requirement for gain of sensing circuitry, and enables very compact peripheral circuits.

Note also that each of options I-III may also operate with time encoding, which allows for better computing precision, for the price of certain speed reduction.

Finally, the lack of continuously tunable devices in the switch capacitor approach (Fig. 1.IV) typically enables only 'near-memory' computing (instead of 'in-memory' computing possible with other options), and leads to inferior density and other metrics.

III. Experimental Demonstrations

Because of still immature device fabrication technology, memristor-based inference circuit demos have been limited in complexity, and/or not fully integrated [6, 8, 12]. Fig. 4 shows a recent result from our collaboration – a small-scale, one-hidden-layer perceptron classifier implemented entirely in integrated hardware. This specific network used two passive 20×20 crossbar arrays with on Pt/Al_2O_3/TiO_{2-x}/Pt memristors (Fig. 4a), board-integrated with discrete CMOS components [6]. The network was successfully trained (both in-situ and ex-situ) to perform classification of 4×4 pixel images (Fig. 4c). The successful demonstration was facilitated by improvements in memristor fabrication technology lowering device-to-device variations, and thus enabling accurate individual state tuning (Fig. 4d, e).

The situation is much better for mixed-signal circuits based on floating-gate crosspoint devices, due to the availability of advanced industrial-grade flash-memory technologies. Our team has recently designed, fabricated, and tested a prototype mixed-signal, 28×28-binary-input, 10-ouput, 3-layer neuromorphic network (Fig. 5a) based on embedded nonvolatile FG cell arrays, redesigned from a commercial 180-nm NOR flash memory [13]. Each array performs a very fast and energy-efficient analog VMM operation. All functional components of the prototype circuit, including 2 synaptic crossbar arrays with 101,780 floating-gate synaptic cells, 74 analog neurons, and peripheral circuitry for weight adjustment and I/O operations, have a total area below 1 mm^2. Its testing on the MNIST benchmark set has shown a classification fidelity of 94.65%, close to the 96.2% obtained in simulation (Fig. 5b). Most importantly, the classification of one pattern takes time less than 1 μs (Fig. 5c) and energy below 20 nJ – both numbers at least $10^3\times$ better than at a digital implementation of the same task, with similar fidelity, fabricated using a much more advanced process [3].

Moreover, there are still many reserves for improving the performance and energy efficiency of such circuits. For example, Fig. 6 shows preliminary results for a much larger network-specific inference accelerator with more advanced circuitry. This chip was designed and fabricated in a 55-nm process, adapted for analog computing applications [9].

IV. Future Work and Summary

For the memristor-based approach, the most important goal is the development of foundry-grade, highly uniform fabrication technology, which would allow for monolithic integration of much larger, denser crossbars with CMOS circuits. Hopefully, this work would piggyback on the recent industrial efforts toward digital resistive memories.

For the FG-based approach, the preliminary experimental results for the chip-to-chip statistics, long-term drift, and temperature sensitivity of the 55-nm [9] and 180-nm [13] prototypes showed no evident showstoppers towards much more complex deep neuromorphic networks. This is why the major focus of future work in this direction may be on the system-level design. In this context, while ASICs have important application niches, general-purpose inference accelerators [2, 3, 18] may be more useful at the moment, in part due to the continuing evolution of neuromorphic algorithms and architectures.

Fig. 7a shows one such architecture, currently being developed by our group. The core of this design is four $M\times N$ rectangular blocks of $K\times K$ VMM crossbar arrays, with front-end digital-to-analog converters (DAC) and back-end sensing circuitry. The array outputs can be connected, via programmable analog buses, to implement larger-size VMMs. Other components of the accelerator include an instruction memory, a controller for decoding instructions and orchestrating the data flow, a small memory buffer for keeping frequently-used data close to the processing unit, and the main memory based on embedded DRAM for storing input, output, and intermediate data.

The performance of the proposed processor was simulated for three representative neural network architectures [19-21] (Fig. 7b). The results show that the mixed-signal VMM blocks take the largest fraction of the chip area, while communications, i.e. sending data across the VMM blocks, often dominates its energy consumption. This fact highlights the importance of in-memory computing using very dense memories for storing weights, as well as of an efficient design of peripheral VMM circuits and configurable busses, which allows fine-grain mapping of network models.

Our preliminary estimates show (Fig. 7c) that general-purpose mixed-signal inference accelerators may retain at least the same large advantage, in speed and energy efficiency, over their digital counterparts, that has been demonstrated in our first, network-specific experiments. The experimental verification of these estimates, as well as the refinement of cons and pros of various approaches to this key task of neuromorphic computing are very important goals for the nearest work.

Acknowledgment

This work was supported by DARPA's UPSIDE program under contract HR0011-13-C-0051UPSIDE via BAE Systems and NSF grant CCF-1528502. Authors are grateful to B. Chakrabarti, X. Guo, I. Kataeva, and M. Klachko for useful discussion and technical support.

References

[1] NVIDIA Corp. Investor day presentation (2017).

[2] ARM ML processor (https://www.arm.com/products/processors/machine-learning); Intel Mobileye (https://www.mobileye.com/en-us/); Google Edge TPU (https://cloud.google.com/edge-tpu/).

[3] P. A. Merolla *et al.* A million spiking-neuron integrated circuit with a scalable communication network & interface. *Science* **345** 668 (2014).

[4] Y. H. Chen *et al.* Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *ISSCC* 262 (2017).

[5] B. Moons *et al*. Envision: A 0.26-to-10 TOps/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI. *ISSCC* 246 (2017).

[6] F. Merrikh Bayat *et al.* Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. *Nature Comm*. **9** 2331 (2018).

[7] M. J. Marinella *et al*.. Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator. *JETCAS* **8** 86 (2018).

[8] C. Li *et al.* Analogue signal and image processing with large memristor crossbars. *Nature Electron.* **1** 52 (2018).

[9] X. Guo *et al.* Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells. *CICC* 1 (2017).

[10] M. Bavandpour *et al*. Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond. *arXiv*:1711.10673 (2017).

[11] E. H. Lee *et al*. A 2.5 GHz 7.7 TOps/W switched-capacitor matrix multiplier with co-designed local memory in 40nm. *ISSCC* 418 (2016).

[12] G. W. Burr *et al*. Experimental demonstration and tolerancing of a large-scale neural network using phase-change memory as the synaptic weight element. *TED* **62** 3498 (2015).

[13] X. Guo *et al.* Fast, energy-efficient, robust, and reproducible mixed signal neuromorphic classifier based on embedded NOR flash memory technology. *IEDM* 6.5.1 (2017).

[14] C. Mead, *Analog VLSI and Neural Systems* (1989).

[15] J. Hasler *et al.* Finding a roadmap to achieve large neuromorphic hardware systems. *Front. Neurosci*. **7** 118 (2013).

[16] H. S. P. Wong *et al*. Memory leads the way to better computing. *Nature Nanotechnol.* **10** 191 (2015).

[17] M. R. Mahmoodi *et al.* Breaking POp/J barrier with analog multiplier circuits based on nonvolatile memories. *ISLPED* 124 (2018).

[18] A Shafiee *et al.* ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *Computer Architecture News* **44** 14 (2016).

[19] C. Szegedy *et al.* Going deeper with convolutions. *CVPR* 1 (2015).

[20] K. He *et al.* Deep residual learning for image recognition. *CVPR* 770 (2016).

[21] Y. Wu *et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144* (2016).

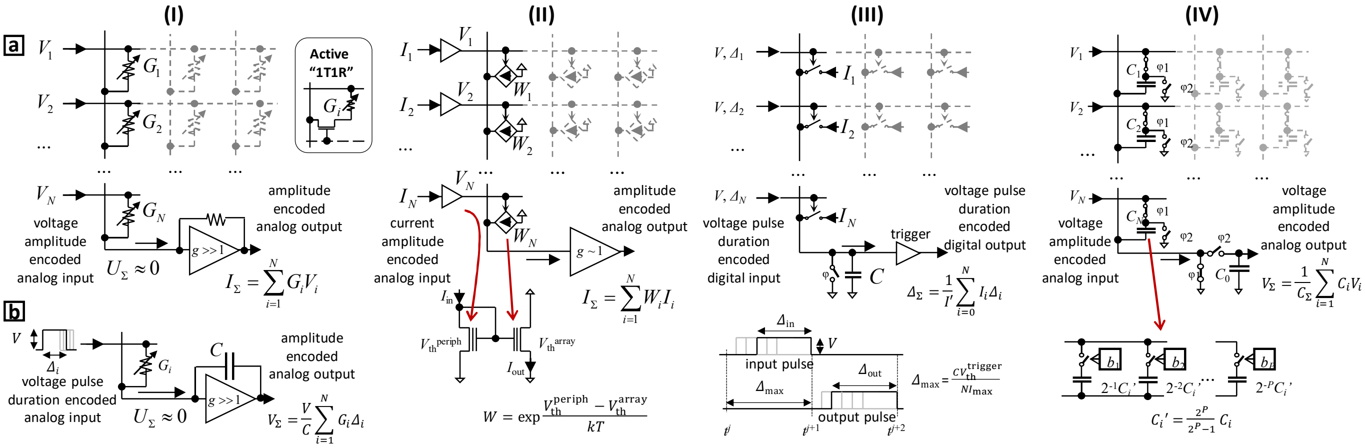

(I)
(II)
(III)
(IV)
a
V_1
V_2
G_1
G_2
V_N
G_N
Active "1T1R"
G_i
voltage amplitude encoded analog input
$U_\Sigma \approx 0$
$g >> 1$
amplitude encoded analog output
$I_\Sigma = \sum_{i=1}^{N} G_i V_i$
b
V
Δ_i
voltage pulse duration encoded analog input
C
$U_\Sigma \approx 0$
amplitude encoded analog output
$V_\Sigma = \frac{V}{C}\sum_{i=1}^{N} G_i \Delta_i$
I_1
I_2
I_N
W_1
W_2
W_N
current amplitude encoded analog input
$g \sim 1$
amplitude encoded analog output
$I_\Sigma = \sum_{i=1}^{N} W_i I_i$
I_{in}
V_{th}^{periph}
V_{th}^{array}
I_{out}
$W = \exp\frac{V_{th}^{periph} - V_{th}^{array}}{kT}$
V, Δ_1
V, Δ_2
V, Δ_N
voltage pulse duration encoded digital input
trigger
voltage pulse duration encoded digital output
$\Delta_\Sigma = \frac{1}{I'}\sum_{i=0}^{N} I_i \Delta_i$
Δ_{in}
input pulse
Δ_{max}
Δ_{out}
output pulse
$\Delta_{max} = \frac{CV_{th}^{trigger}}{NI_{max}}$
t^j
t^{j+1}
t^{j+2}
C_1
C_2
C_N
φ1
φ2
C_0
voltage amplitude encoded analog input
voltage amplitude encoded analog output
$V_\Sigma = \frac{1}{C_\Sigma}\sum_{i=1}^{N} C_i V_i$
b_1
b_2
b_P
$2^{-1}C_i'$
$2^{-2}C_i'$
$2^{-P}C_i'$
$C_i' = \frac{2^P}{2^P - 1} C_i$

Fig. 1. Major types of mixed-signal VMM circuits. In (I), the matrix elements ('synaptic weights') are represented by continuous states of adjustable nonvolatile resistive devices (e.g., memristors), while the input signals are encoded with either (a) amplitudes, or (b) durations of voltage pulses. The top right inset shows an active ('1T1R') cell, which may be also used in circuits (a, b). In (II, III), each weight is stored in subthreshold-mode floating-gate (FG) cells, implemented as either (II) a current mirror pair formed by peripheral and array FG transistors, or (III) a voltage-gated current source. In (III), both inputs and outputs are encoded by the duration of pulses, generated within the corresponding time frame *t*, as shown at the bottom of panel III. In the switch capacitor approach (IV), *P*-bit weights are typically stored in binary-weighted fixed-value crosspoint capacitors, and the computation is performed by controlling the capacitor charge/discharge, using the switches φ_1 and φ_2.

Fig	Xpoint		Input/output	Density	Precision	Speed	Energy Efficiency	Maturity	Ref
Ia	0T	R	amp/amp	++	+	+++	++	-	[6]
Ib		R	time/amp	+++	++	++	+++	-	[7]*
Ia	1T	R	amp/amp	+	++	++	+	+	[8]
Ib		R	time/amp	++	+++	+	++	+	[7]*
II	FG		amp/amp	+	++	++	+	++	[9]
III	FG		time/time	++	+++	+	++	++	[10]
IV	C		amp/amp	-	-	+	++	+++	[11]

Fig. 2. Approximate comparison of features of the VMM approaches outlined in Fig. 1: '+++' - the best, '-' – the worst. The score for precision is based on a combination of the input, weight, and computing accuracies. The scores for density, speed, and energy efficiency (EE) reflect contributions from both the arrays and the peripheral circuits. Besides the maturity, all scores are for the expected level of each technology after it has been matured, rather then for its current state-of-the-art. *Ref. [7] describes, in particular, the additional circuitry for the conversion to the time-encoded output signals.

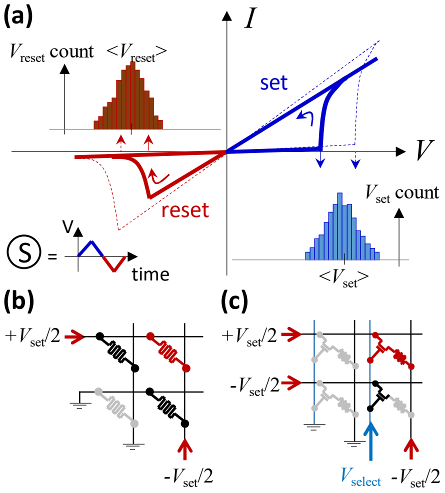


Fig. 3. Switching threshold variations in memristors: (a) A typical hysteretic dc *I*-*V* curve, for a symmetric voltage sweep (lower bottom inset). The inset histograms show typical variations in the switching voltages at which the effective conductance changes by more than a certain amount. (b, c) Four-device crossbar fragments with (b) passive '0T1R' and (c) active '1T1R' cells. The applied voltages show a specific example of the "half-biasing" technique for increasing the conductance of the selected device (shown in red). Solid black lines show the half-selected memristors, while the gray color is used to show unselected devices, with no applied voltage.

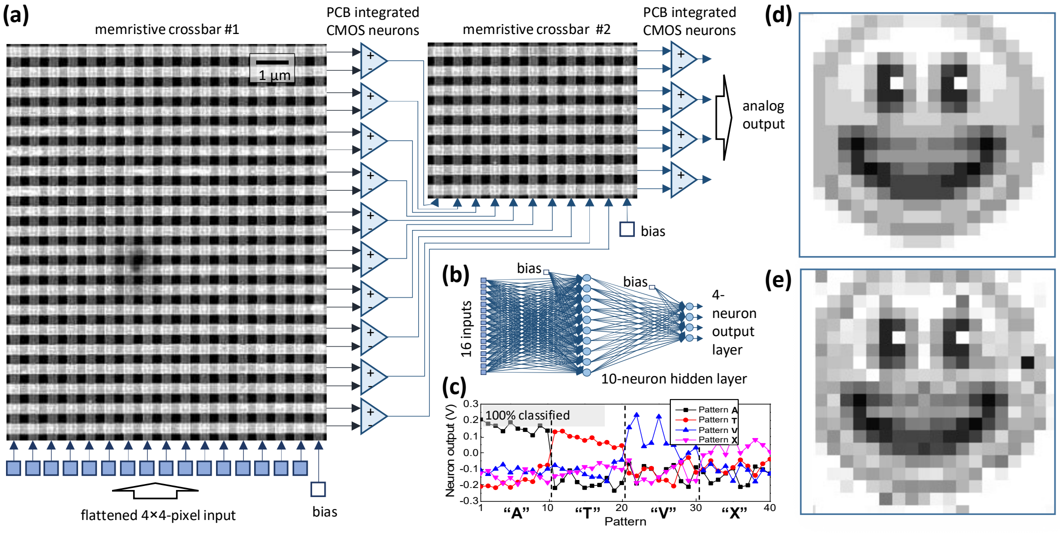


Fig. 4. MLP classifier demo based on passively integrated metal-oxide memristors [6]: (a) A perceptron diagram showing (as SEM images) the crossbar portions used in the experiment; (b) The implemented network's graph; (c) An example of measured output voltages for the ex-situ-trained network, tested on a set of 4 stylized 4×4-pixel letters; (d, e) An example of memristor tuning, showing (d) the desired 'smiley face' pattern, quantized to 10 gray levels, and (e) the actual resistance values measured after tuning all devices in a 20×20 memristive crossbar with the nominal 5% accuracy, using an automated tuning algorithm. The white / black pixels correspond to effective resistances 96.6 / 7.0 kΩ, measured at 0.2 V.

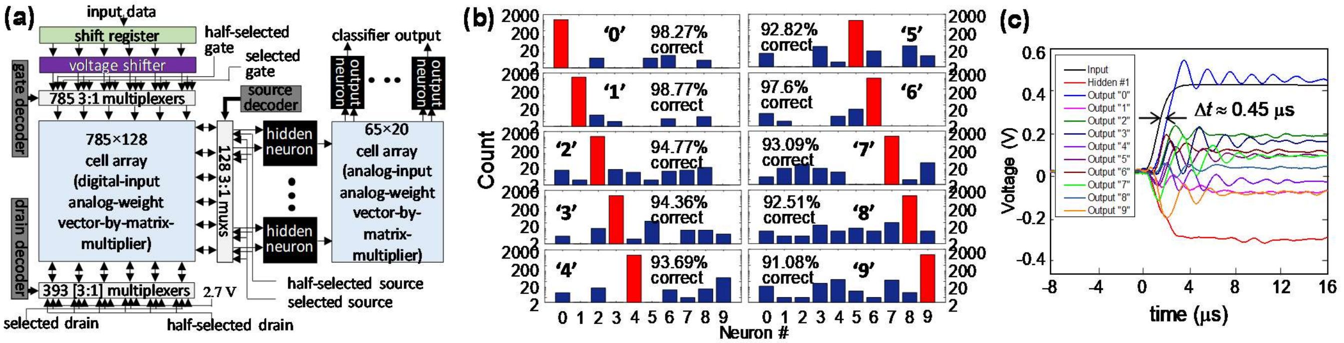


Fig. 5. MLP classifier demo in the 180 nm ESF1 process with 100+K FG cells [13]: (c) High-level architecture, designed for classification of MNIST benchmark images. (Weight tuning circuitry for the 2nd array is not shown for clarity). (b) Histograms of the experimentally measured largest output voltages from the ex-situ trained network for 10,000 MNIST test set patterns, showing that the correct outputs (red bars) always dominate. (c) Typical signal dynamics after an abrupt turn-on of the voltage shifter's power supply, measured simultaneously at the network input, at the output of a sample hidden-layer neuron, and at all network's outputs.

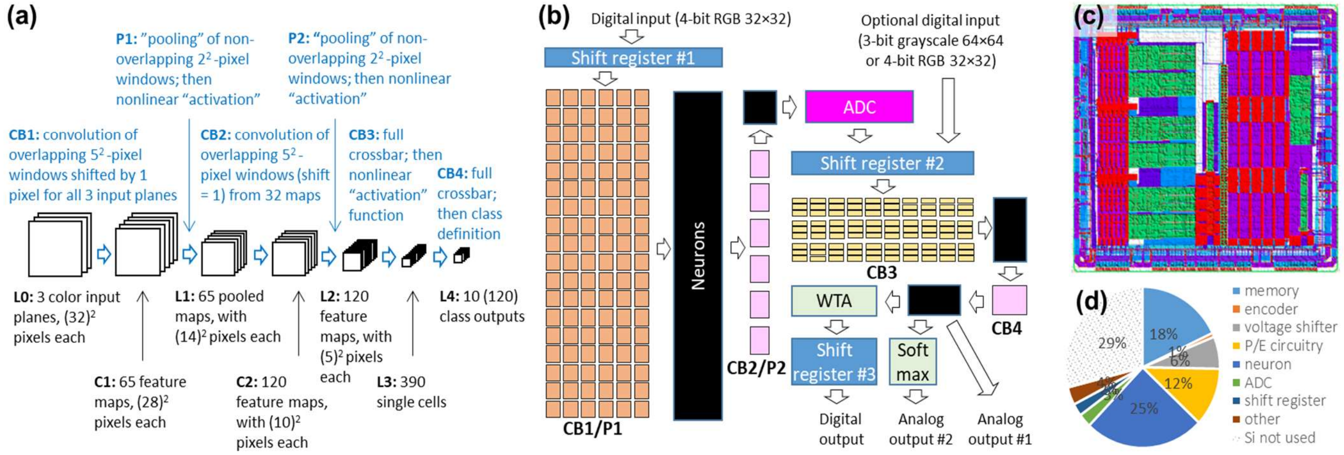
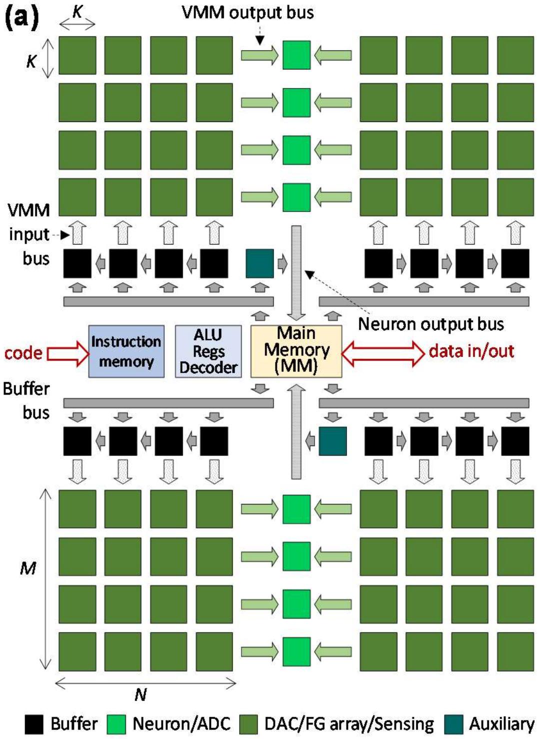

(a)
P1: "pooling" of non-overlapping 2²-pixel windows; then nonlinear "activation"
P2: "pooling" of non-overlapping 2²-pixel windows; then nonlinear "activation"
CB1: convolution of overlapping 5²-pixel windows shifted by 1 pixel for all 3 input planes
CB2: convolution of overlapping 5²-pixel windows (shift = 1) from 32 maps
CB3: full crossbar; then nonlinear "activation" function
CB4: full crossbar; then class definition
L0: 3 color input planes, (32)² pixels each
L1: 65 pooled maps, with (14)² pixels each
L2: 120 feature maps, with (5)² pixels each
L4: 10 (120) class outputs
C1: 65 feature maps, (28)² pixels each
C2: 120 feature maps, with (10)² pixels each
L3: 390 single cells
(b)
Digital input (4-bit RGB 32×32)
Shift register #1
Neurons
ADC
Optional digital input (3-bit grayscale 64×64 or 4-bit RGB 32×32)
Shift register #2
CB3
CB1/P1
CB2/P2
WTA
CB4
Shift register #3
Soft max
Digital output
Analog output #2
Analog output #1
(c)
(d)
18%
1%
6%
12%
25%
29%
memory
encoder
voltage shifter
P/E circuitry
neuron
ADC
shift register
other
Si not used

Fig. 6. Network-specific mixed-signal image classifier accelerator: (a) The architecture of the implemented deep convolutional neural network, (b) its block diagram, (c) the 55-nm CMOS chip layout, with ~ 13M embedded ESF3 NOR flash cells, and (d) the corresponding area breakdown. Some circuitry (e.g., for testing and cell tuning) is not shown for clarity. An advanced design [17] has enabled a reduction of the neuron circuits to ~6% of the chip area, while the FG arrays take ~30% of it.



(b)	INC-V1	ResNet	GNMT
Network specifications			
# parameters	7.2e06	1.1e07	1.3e08
# operations	5.2e09	2.0e10	2.6e09
Architecture specifications			
K	64		
M (top/bottom)	16/18	44/48	64
N	38	80	128
MM capacity (KB)	1024		
# MM R/W	3.3e05	8.1e05	1.7e03
MM util. (%)	47.8	59.8	5.07
FG array util. (%)	7.88	44.92	100
Area breakdown (%)			
MM	18.1	4.53	2.2
Sensing	15.5	23.3	25.1
FG arrays	24.2	36.5	39.3
DACs	4.5	6.8	7.3
Neurons + ADCs	0.06	0.04	0.03
P/E	26.3	14.7	11.3
Others	11.4	14.2	14.8
Energy breakdown (%)			
MM	38.8	23.9	8.3
Sensing	16.2	11.4	23.8
FG arrays	3.03	2.13	4.45
DACs	2.22	1.56	3.3
Neurons + ADCs	0.70	0.90	0.57
Buses	31.6	41.3	12.4
Leakage	4.4	17.4	46.7
Others	3.0	1.5	0.48
Performance summary			
Area (mm^2)	35.4	142	293
Power (mW)	14.9	19.8	16.1
Latency (ms)	3.1	8.75	0.59
EE (TOp/J)	114	120	283
Throughput (TOp/s)	1.69	2.37	4.54

(c)	Chip area (mm^2)	Latency (ms)	EE (TOp/J)
Digital CMOS	321	607	0.35
Mixed-signal FG	293	0.59	283

Fig. 7. aCortex, a multi-purpose mixed-signal inference accelerator: (a) the block-level architecture, and (b) the simulated performance metrics for three representative neuromorphic tasks - image classification and natural language translation, based on two deep convolutional neural networks (INC-V1 [19] and ResNet [20]) and a deep recurrent network GNMT [21]. Though the results are reported for smallest hardware resources required for each network, the use of configurable buses enables achieving, for smaller networks (INC-V1 and ResNet), essentially the same performance by their mapping on the largest architecture (used for GNMT). (c) Preliminary comparison of key metrics of the mixed-signal and digital versions of the aCortex for the GNMT network. In the 4-bit digital version, synaptic weights are stored on-chip in an SRAM (~92% of the chip area), while the mixed-signal processing arrays are replaced with arrays of 64×64 4-bit digital VMM units (~6% of the area). The dot product in each digital VMM is implemented with array-based multipliers and parallel tree adders. All performance and area numbers are for a 55 nm CMOS process. All latencies are per one classification / translation task. Note that both designs were optimized for the energy efficiency, while being suboptimal in terms of latency / throughput. (Faster operation may be achieved by performing more operations in parallel, with some energy efficiency degradation.)