33.1 A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models

Weier Wan[1], Rajkumar Kubendran[2], S. Burc Eryilmaz[1], Wenqiang Zhang[3], Yan Liao[3], Dabin Wu[3], Stephen Deiss[2], Bin Gao[3], Priyanka Raina[1], Siddharth Joshi[4], Huaqiang Wu[3], Gert Cauwenberghs[2], H.-S. Philip Wong[1]

[1]Stanford University, Stanford, CA, [2]University of California, San Diego, CA, [3]Tsinghua University, Beijing, China, [4]University of Notre Dame, Notre Dame, IN

Many powerful neural network (NN) models such as probabilistic graphical models (PGMs) and recurrent neural networks (RNNs) require flexibility in dataflow and weight access patterns as shown in Fig. 33.1.1 Typically, Compute-In-Memory (CIM) designs do not implement such dataflows or do so by replicating circuits at the memory periphery such as ADCs/neurons along both the rows and columns of the memory array, leading to an overhead in operation. This paper describes a CIM architecture implemented in a 130nm CMOS/RRAM process, that delivers the highest reported computational energy-efficiency of 74 tera-multiply-accumulates per second per watt (TMACS/W) for RRAM-based CIM architectures while simultaneously offering dataflow reconfigurability to address the limitations of previous designs. This is made possible through two key features: 1) a runtime reconfigurable dataflow with in-situ access to RRAM array and its transpose for efficient access to NN weights and 2) a voltage sensing stochastic integrate-and-fire analog neuron (I&F) that is reused for correlated double sampling (CDS), stochastic voltage integration, and threshold comparison.

Figure 33.1.1 shows the top-level architecture of the neurosynaptic core (NSC) composed of 16×16 sub-cores with shared bit-lines (BLs), word-lines (WLs), and source-lines (SLs). BLs connect to RRAM top electrodes, WLs to access transistors' gate and SLs to access transistors' source. Each sub-core contains 16×16 RRAM synapses, which store the NN weights, and one I&F neuron. The neuron in the (j, k)th sub-core in the 16×16 grid connects to the (16j+k)th BL and (16k+j)th SL for *in-situ* transposability in RRAM array access. The periphery consists of BL and WL drivers along the rows, SL drivers along the columns, and a delay-line-based tunable pulse generator (PG) to generate RRAM read pulses. Linear-feedback shift register (LFSR) chains enable stochastic sampling required for PGMs and 256-bit BL and SL registers along the periphery provide I/O. Configuring the SL/BL switches connects the I&F neuron to its respective SL and BL to realize multiple types of dataflows such as forward (INFer) and reverse (GENerate) matrix-vector multiplication (MVM), and recurrent connections (REC) as shown in Fig. 33.1.5. During GEN, input pulses are applied at the BLs, weighed by the RRAM, and sampled by the I&F through the SL switch. The I&F outputs are written into SL registers through the SL switches. During INF, pulses are applied at the SLs, weighed by the RRAMs, and sampled by the I&F through the BL switches.

Figure 33.1.2 (top) shows the design of the I&F neuron which is composed of a single high-gain cascode amplifier, a latch, and switches to reconfigure the amplifier feedback loop between multiple modes of operation. Periodic offset cancellation through CDS mitigates circuit variation across the array and establishes a DC operating point for the capacitively coupled amplifier. During CDS offset compensation, first, a self-bias phase establishes the DC operating point (V_{op}) by shorting the input of the amplifier with the output. Simultaneously, a known reference (V_{ref}) is sampled onto the capacitor C_{sample}; together these sample the unknown input offset onto C_{CDS} for subsequent cancellation. The I&F supports integration of input pulse sequences over time using multiple sample-integrate (SI) cycles to enable stochastic sampling and accumulation in PGMs. Figure 33.1.2 (bottom) shows two such SI cycles. During the sample phase, the input voltage is sampled onto C_{sample}, followed by the integration phase during which the charge on C_{sample} transfers onto C_{integ}. We implement the neuron activation function, by breaking the amplifier feedback loop while simultaneously driving the bottom plate of C_{integ} by V_{th}. The open loop amplifier functions as a comparator, generating a binary decision by comparing V_{th}+V_{integ} and V_{ref}. The output is then latched and written to the BL/SL registers through the corresponding switches. Each neuron occupies an area of 1200μm^2 and measurements show operation with 63nW static power drawn from a 1.8V supply. The total static power consumption for one NSC (256 I&F neurons and biasing) is 17μW. Energy efficiency is in part derived from voltage sensing in the I&F which drives the output of the RRAM array to high impedance during MVM, thus avoiding the static current draw in current sensing configurations.

Figure 33.1.3 (top) characterizes the linearity of the SI operation over multiple input pulses. Sizing C_{integ} to be 6× C_{sample} extends the linear accumulation range of the I&F. Figure 33.1.3 shows the measured linear dependence between the I&F threshold and 1) the number of input pulses (top left), and 2) the pulse amplitude (top right). Spatially uncorrelated, controlled Bernoulli PRNs are required for PGM operation; these are generated through two counterpropagating LFSRs whose outputs are modulated to be V_{ref}+V_n and V_{ref}-V_n and applied to the I&F via SLs. The accumulation of multiple noise pulses smoothens the sharp decision point of the comparator to a more sigmoidal function as measurements show in Fig. 33.1.3 (bottom left). Figure 33.1.3 (bottom) shows the measured relationship between the number of stochastic pulses and the sigmoidal characteristic of the I&F.

We characterize the MVM performance of the system by programming analog-valued weights in each RRAM device (device TEM shown in Fig. 33.1.4 bottom) using a write-verify scheme similar to [1]. The RRAMs are programmed to ±2μS of their targeted conductance (program success rate = 99.4%). Figure 33.1.4 (top) shows the RRAM conductance distribution measured immediately after programming and 1 day after programming. We apply random input vectors to collect the output distribution of each neuron at different nominal MVM output values, testing MVM operation. The sharp transition between -1 to +1 in the neuron output in Fig. 33.1.4 (bottom right) demonstrates robustness to RRAM resistance drift over time.

To illustrate MVM capabilities in both forward and reverse directions, as well as temporal accumulation and probabilistic sampling in the neuron, we use a generative Restricted Boltzmann Machine (RBM) to reconstruct MNIST digit images, shown in Fig. 33.1.5. RBMs are PGMs consisting of a set of fully-connected visible neurons and hidden neurons. During inference, binary inputs sampled from the gray-level pixels are presented to the visible neurons. MVM and Gibbs sampling are repeated back-and-forth between visible and hidden neurons. We subsample MNIST digit images to 15×15 gray images and implement an RBM using 225 visible neurons (15×15 pixels) and 60 hidden neurons. We use 4 RRAM cells (2-row by 2-column) per weight for differential encoding in both horizontal and vertical directions to implement positive/negative weights in both INF and GEN. Figure 33.1.5 (bottom) shows the reconstruction performed on all 10 digits with a mean square reconstruction error of 1.91 per image, comparable to software implementations with 7 level quantized weights.

The NSC along with the periphery measures 1.79 mm^2 in 130nm CMOS with a single sub-core occupying 1849μm^2. Energy breakdown for the NSC operation is in Fig. 33.1.6 (top), with energy dominated by WL switching. Scaling clocking frequency of the I&F linearly scales throughput and power (from 50μW to 1.5mW), with 74 TMACS/W energy efficiency over that range. A comparison of the key metrics with state-of-the-art RRAM-based CIM designs is given in Fig. 33.1.6 (bottom), showing that our design achieves the highest reported computational energy-efficiency among RRAM-based CIM implementations. This work is a system-level demonstration of a high-efficiency, fully-integrated CMOS-RRAM CIM architecture with core-level support for a diverse set of NN architectures and dataflows including probabilistic graphical models. Multi-core extensions of this work open the door to advances in artificial intelligence at a large scale with extreme energy-efficiency and integration density.

Acknowledgements:
Work supported in part by NSF Expeditions in Computing (Penn State, Award #: 1317470), Stanford SystemX Alliance, Stanford NMTRI, Beijing Innovation Center for Future Chips, and Office of Naval Research.

References:

[1] X. Zheng et al., "Error-Resilient Analog Image Storage and Compression with Analog-Valued RRAM Arrays: An Adaptive Joint Source-Channel Coding Approach," *IEEE IEDM*, pp. 71-74, 2018.

[2] W.-H. Chen et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with Sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, Feb. 2018.

[3] R. Mochida et al., "A 4M Synapses Integrated Analog ReRAM Based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," *IEEE VLSI*, pp. 175-176, 2018.

[4] C.-X. Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," *ISSCC*, pp. 388-389, Feb. 2019.

[5] F. Cai et al., "A Fully Integrated Reprogrammable Memristor–CMOS System for Efficient Multiply–Accumulate Operations," *Nature Electronics*, vol. 2 (7), pp. 290-299, 2019.

 978-1-7281-3205-1/20/$31.00

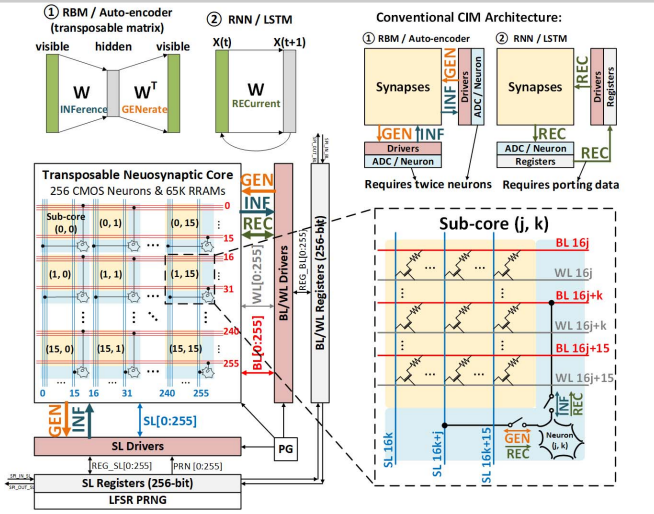


Figure 33.1.1: Chip architecture with dynamically reconfigurable neurosynaptic core (NSC); the *in-situ* transposability is shown through the connectivity for a general sub-core (j,k).

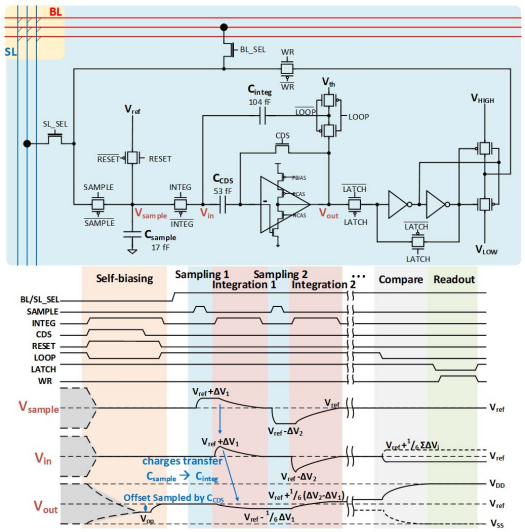


Figure 33.1.2: Integrate-and-fire (I&F) neuron schematic and operation timing diagram.

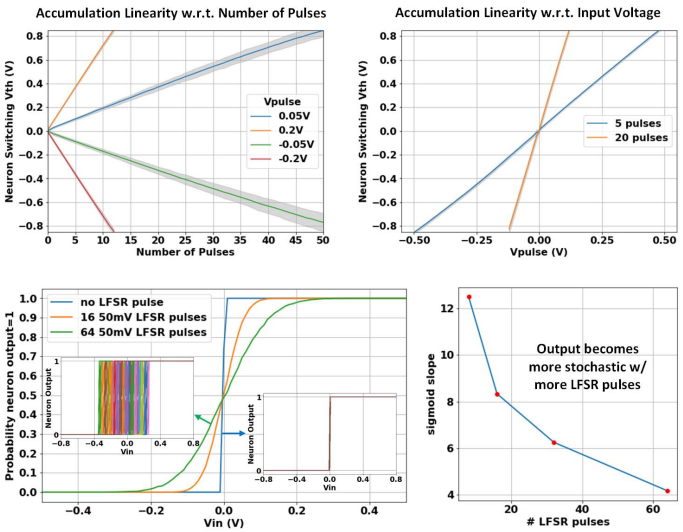


Figure 33.1.3: Characterization of neuron accumulation linearity and neuron sigmoidal characteristics of stochastic sampling.

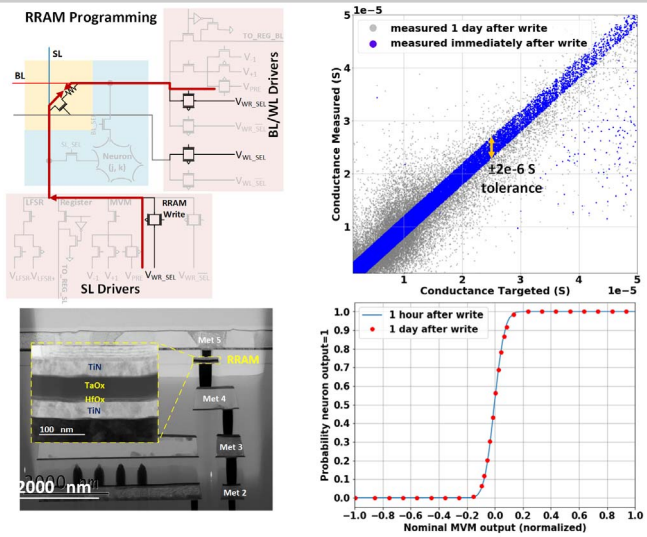


Figure 33.1.4: Characterization of RRAM analog-programming, RRAM resistance drift and its effect on in-memory MVM with neuron activation.

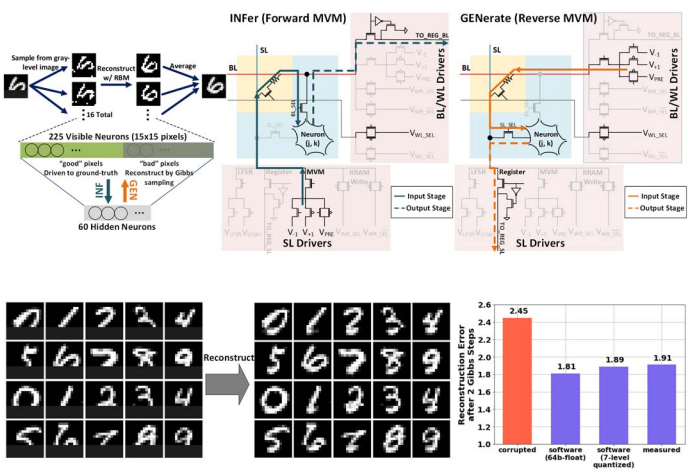
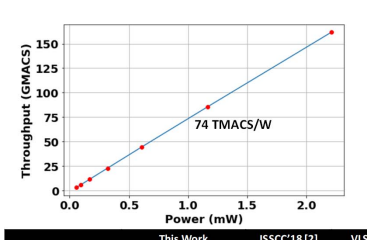
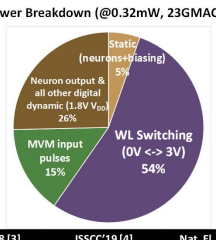


Figure 33.1.5: Demonstration of MNIST image reconstruction using Restricted Boltzmann Machine (RBM) stochastic sampling implemented with transposable NSC.





	This Work	ISSCC'18 [2]	VLSI'18 [3]		ISSCC'19 [4]	Nat. El. 19 [5]
CMOS Technology	130nm	65nm	180nm	40nm	55nm	180nm
Synaptic Device	HfOx/TaOx RRAM	Unipolar RRAM	TaOx RRAM		RRAM	WOx RRAM
Connectivity (data-flow direction)	Forward, Reverse, Recurrent	Forward	Forward		Forward	Forward, Reverse
Supported Neural Network Architecture	ANNs: RNN, MLP, CNN PGMs: RBM, Bayesian	CNN, MLP	MLP		CNN, MLP	MLP, Sparse Coding
Neuron/ADC type	I&F spiking, deterministic/stochastic	3-b sense-amp	1-b sense-amp		4-b sense-amp	13-b ADC
Synaptic array dimension	256 x 256	512 x 256	–	–	256 x 512	54 x 108
Synapses/mm²	0.59M	–	0.16M	1.48M	–	0.04M
Power (mW)	0.05 – 2.2	–	15.8	9.9	–	64.4
Precision (input, weight, output)	(1b, analog, 1b)	(1b, ternary, 3b)	(1b, analog, 1b)		(1b, 3b, 3b) (2b, 3b, 3b)	(4b, analog, 13b)
Energy efficiency (TMACS/W)*	74	–	10.35	33.25	26.58 10.95	0.89

* 1 MAC counted as 2 operations

Figure 33.1.6: Measured energy efficiency of the NSC with power consumption breakdown, performance summary and comparison table.

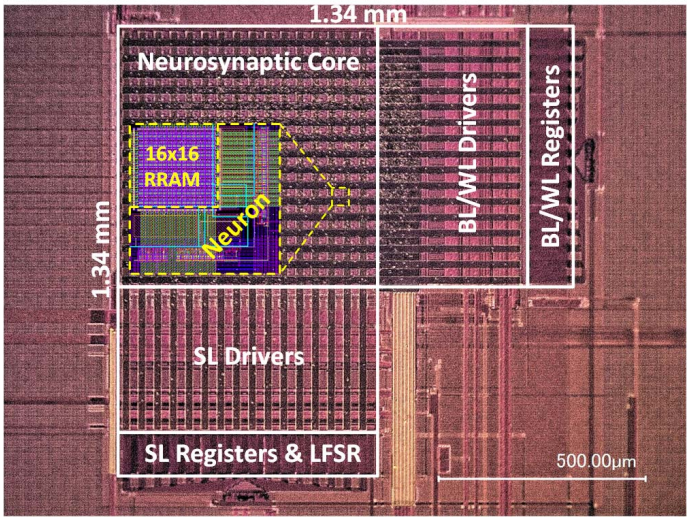


Figure 33.1.7: Die micrograph and neurosynaptic sub-core layout.

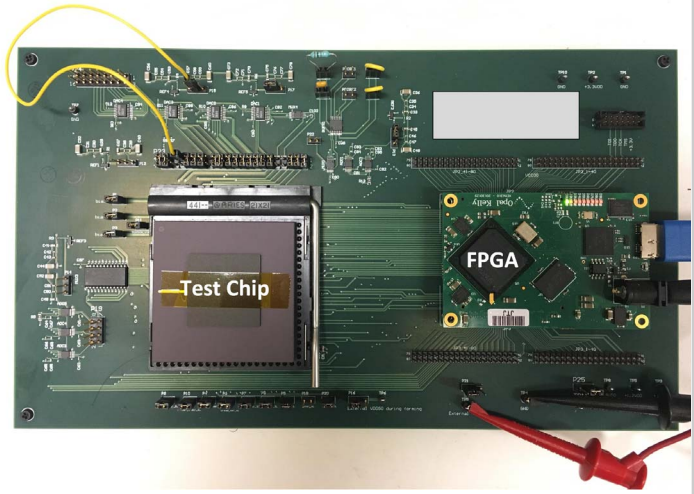


Figure 33.1.S1: Measurement setup with a test chip receiving/sending data from/to an FPGA.

978-1-7281-3205-1/20/$31.00 ©2020 IEEE