RRAM-based Spiking Nonvolatile Computing-In-Memory Processing Engine with Precision-Configurable In Situ Nonlinear Activation

Bonan Yan[1], Qing Yang[1], Wei-Hao Chen[2], Kung-Tang Chang[2], Jian-Wei Su[2,3], Chien-Hua Hsu[3], Sih-Han Li[3], Heng-Yuan Lee[3], Shyh-Shyuan Sheu[3], Mon-Shu Ho[4], Qing Wu[5], Meng-Fan Chang[2], Yiran Chen[1] and Hai Li[1]

[1]Duke University, Durham, NC, USA [2]National Tsing Hua University, Hsinchu, Taiwan [3]Industrial Technology Research Institute, Hsinchu, Taiwan [4]National Chung Hsing University, Taichung, Taiwan [5]Air Force Research Laboratory, Rome, NY, USA

(E-mail: bonan.yan@duke.edu, mfchang@ee.nthu.edu.tw, hai.li@duke.edu)

Abstract

This work presents a hybrid CMOS-RRAM integration of spiking nonvolatile computing-in-memory (nvCIM) processing engine (PE) that includes a 64Kb RRAM macro and a novel in situ nonlinear activation (ISNA) module. We integrate the computing controller and nonlinear activation function on-chip to compute convolutional or fully-connected neural network. ISNA merges A/D conversion and activation computation by leveraging its nonlinear working region. This eliminates the need for additional circuits to realize nonlinearity and reduces area by 43.7× w.r.t. the ADC scheme. The activation precision of ISNA can be configured from 1 to 8 bits to balance throughput, accuracy and power efficiency. The measurement of 4-layer LeNet shows such optimization improves 23.1% of computing speed via compromising a 2.5% relative accuracy drop. The proposed nvCIM PE achieves 16.9 TOPS/W power efficiency and a maximum spike frequency of 99.24 MHz.

Keywords: activation precision, computing in memory, nonvolatile emerging memory, processing engine, RRAM, spiking.

Introduction

Designs of nvCIM use memory resistivity to locally store synaptic weights, significantly reducing memory accesses (Fig. 1). Multiply-accumulate (MAC) operations are conducted through analog currents, necessitating A/D conversion to interface with surrounding digital systems. Thereafter, a nonlinear activation function, e.g. tanh or rectified linear unit (relu), filters the MAC results. Previous designs use binary sense amplifiers with 1-bit activation precision between perceptrons to avoid area-consuming ADC/DAC [1]. Neuron computation in multi-bit precision is then realized by accumulating the 1-bit MAC results followed by additional digital nonlinear activation function circuits (AFC). In this work, we demonstrate a compact RRAM-based spiking nvCIM PE (Fig. 1) which comprises a 64Kb RRAM macro for synaptic weight storage and MAC computation, and ISNA which executes activation function computation on the fly, obviating the need for the additional AFC and reducing design overheads.

RRAM-based Spiking nvCIM PE Architecture

The proposed RRAM PE (Fig. 2) is an integrated solution whose function can switch between memory storage and computation. Fig. 3 depicts the 1-transistor-1-RRAM (1T1R) array structure. BLs are parallel to SLs while vertical to WLs. In the memory mode, the Read/Write (RD/WR) logic programs the pre-trained synaptic weights into the RRAM array or senses the RRAM cell resistances. In the computing mode, the macro conducts MAC operations, and ISNA performs the activation and outputs digital spikes. The SLs are tied to the ground. The input spikes are applied to WLs to control the ON/OFF status of 1T1R cells. ISNA drives the BLs to excite the computing currents. The BL current amplitude denotes the MAC result that is determined by the overall conductance of the ON cells in the column as well as the inputs. When deploying neural networks, the proposed macro can adopt binary (0, 1) or ternary (0, ±1) weights and map each weight onto one or two 1T1R cells, respectively.

Precision-Configurable In Situ Nonlinear Activation

4-bit or higher activation precision is needed in many neural networks with binarized weights to attain high accuracy [2]. To achieve multi-bit A/D conversion, ISNA (Fig. 4) includes a current amplifier (CA) and an integrate & fire circuit (IFC). IFC generates spikes at a maximum frequency of 99.24 MHz (Fig. 5(a)). In this way, ISNA reshapes MAC results into time-domain spike numbers. The OTA in CA holds a stable BL voltage at 300±1.0 mV in the load range of 0.3~60 KΩ determined by the RRAM characteristics (Fig. 6(a)). The phase compensation accelerates the convergence of the step response (Fig. 6(b)). Our ISNA occupies 47.2% less area than the prior compact spiking-based neuron circuit [3] and is 43.7× smaller than ADC [4]. Activation precision is reconfigurable through scaling the output spike collecting time (Fig. 7). This flexibility allows a fabricated RRAM nvCIM PE chip to compute with various precisions in different NN layers for better design tradeoffs.

ISNA module goes beyond ADC by filtering MAC results with nonlinear activation function (Fig. 5(b)). We call it "in situ nonlinear activation." In Fig. 5(a), the saturation region originates from the intrinsic delay of the buffer in IFC and is purposely tuned by adjusting V_{ref} and V_{th} in Fig. 4. By controlling WL width parallelism (numbers of ON RRAM cells in an BL, Fig. 8) to match ISNA input load, both the linear and saturation region can be used during computing. Their combination provides a designated shape (tanh or clipped-relu) of nonlinear A/D conversion. Our ISNA is 1.99× faster than the existing scheme [5] by avoiding the additional AFC latency.

Measurement Results

We tested our prototype by deploying 1-layer (MLP-1), 2-layer (MLP-2) perceptrons and 4-layer LeNet (CNN-1) on MNIST and 5-layer LeNet (CNN-2) on CIFAR-10. The accuracies of MNIST and CIFAR-10 are respectively 98.1% and 95.9% compared with the "binarized" case (Fig. 9(a)) from software inference with binarized weights. In CNN-1, conv2 layer occupies 9% of stored synapses but 51% of overall latency due to its large number of MAC operations (Fig. 9(b)). To further improve the performance, we reconfigure the ISNA activation precision from 8-bit to 5/6/7-bit in the designated layers. Hence, the proposed PE reduces the overall latency by up to 2.11× with a slight accuracy drop (Fig. 9(c)). In general, we tune the activation precision to optimize the overall computing latency, power and power efficiency of the PE (Fig. 10). Fig. 11 presents the die photo and chip summary. Table I compares this work to state-of-the-art CIM macro and PE designs.

Acknowledgements This work is supported by AFRL FA8750-15-1-0176 and ITRI-NTHU joint project.

Reference

[1] R. Mochida, VLSI, 2018, pp.175. [2] I. Hubara, JMLR, 18, 187, 2016. [3] N. Qiao, Front. Neurosci., 9, 141, 2015. [4] K. Ohhata, VLSI, 2018, pp.95. [5] A. Amravati, ISSCC 2018, pp.124. [6] W.-H. Chen, ISSCC, 2018, pp. 494. [7] W.-S. Khwa, ISSCC, 2018, pp.496. [8] A. Biswas, ISSCC, 2018, pp.488. [9] S. K. Gonugondla, ISSCC, 2018, pp.490. [10] P. A. Merolla. Science, 345, 668, 2014.

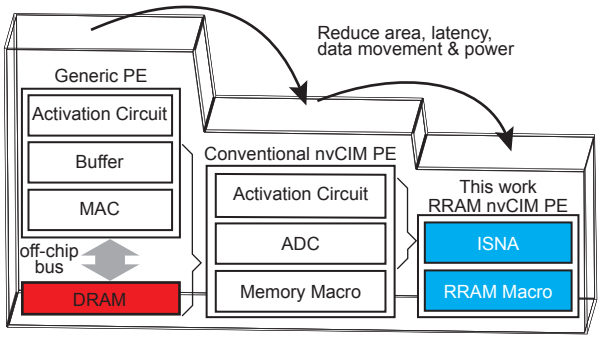


Fig. 1 PE comparison between the generic design [5], the conventional nvCIM PE [6] and this work.

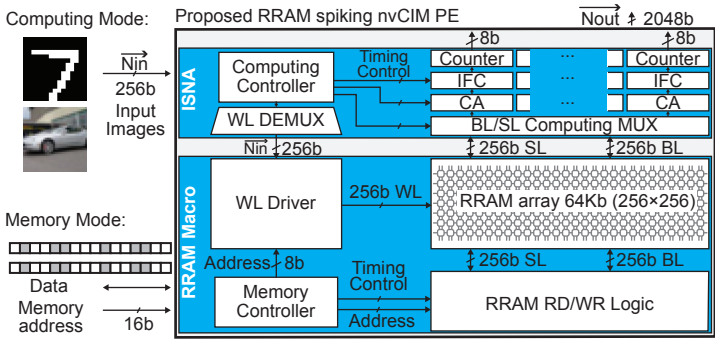


Fig. 2 Proposed PE diagram. ISNA: in situ nonlinear activation; IFC: integrate & fire circuit; CA: current amplifier; RD/WR: read/write.

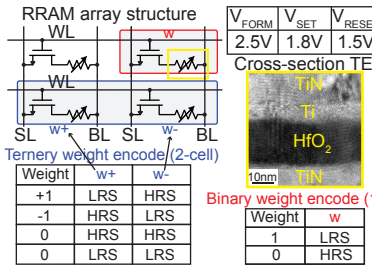


Fig. 3 RRAM array structure, device characteristics and encoding schemes.

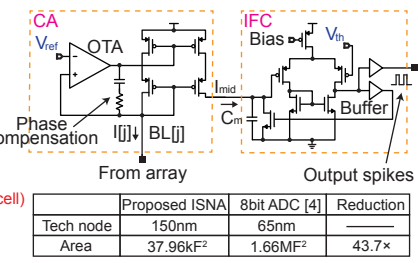


Fig. 4 ISNA design diagram: (a) CA+IFC diagram; (b) proposed ISNA vs. latest ADC.

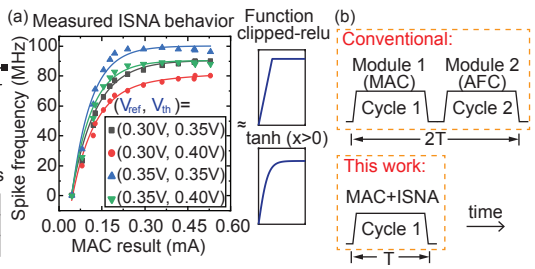


Fig. 5 (a) Measured ISNA behavior (points: measured, curves: fitting); (b) combined MAC & ISNA.

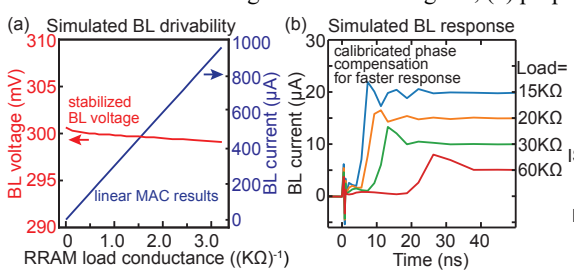


Fig. 6 Simulated CA functions: (a) stabilize BL voltage; (b) reduce response time with low input currents.

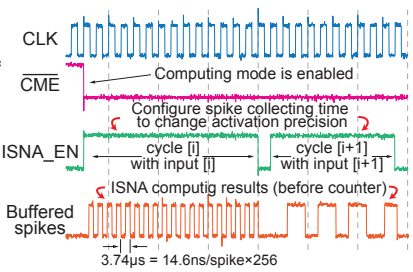


Fig. 7 Measured waveforms of real-time configuring activation precisions.

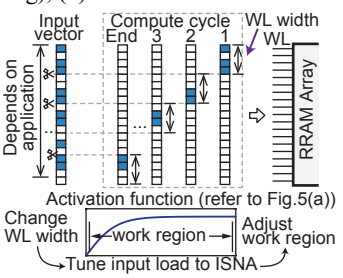


Fig. 8 WL width parallelism for configuring activation precision.

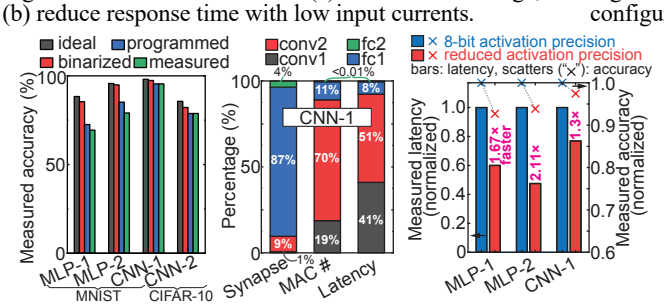


Fig. 9 Measured accuracy, measured hardware resource distribution and measured latency reduction by re-configuring activation precision. MLP-1: 1-layer perceptron, MLP-2: 2-layer perceptron, CNN-1: 4-layer LeNet, CNN-2: 5-layer LeNet.

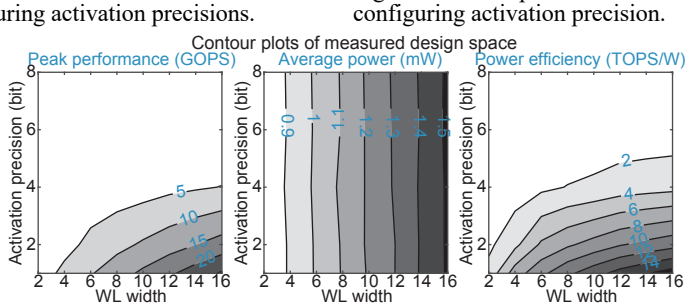
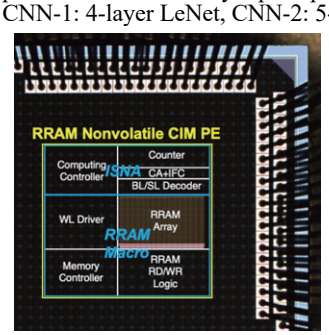


Fig. 10 Measured design space contour vs. different configurations for activation precisions: peak performance, average chip power and arithmetic power efficiency.



Technology	0.15μm CMOS + HfO RRAM
Speed	200ns / layer @ 8-bit activation
Cell size	1.66 μm²
MAC Energy	0.257 pJ/MAC
Clock	50 MHz
Operation type	Spiking

Fig. 11 Die photo and chip summary.

TABLE I COMPARISON WITH RECENT WORK

Type		CIM Macro				CIM PE		Digital Processor
Work		[1]	[6]	[7]	[8]	**This work†**	[9]	TrueNorth [10]
Technology		180nm	65nm	65nm	65nm	**150nm**	65nm	28nm
Synapse		1T1R RRAM	1T1R RRAM	6T-SRAM	10T-SRAM	**1T1R RRAM**	6T-SRAM	SRAM
Nonvolatility		Yes	Yes	No	No	**Yes**	No	No
Standby current		~zero	~zero	high	high	**~zero**	high	high
Spiking NN		No	No	No	No	**Yes**	No	Yes
Capacity		2M	1M	4K	16K	**64K**	128K	256M
Cell area [F²]		—	59	124	968	**74**	~256	—
Normalized die area		12×	—	—	30×	**1×**	11×	~17240×
Chip power [mW]		15.8	—	—	—	**1.52**	—	204.4
Activation precision		1 bit	3 bit	1 bit	7 bit	**1~8 bit**	8 bit	1 bit
Power efficiency [TOPS/W]	MAC only	20.7	16.95	55.8	28.1	—	—	—
	MAC +Activation	—	—	—	—	**16.9**	3.125	0.4
On-chip Activation Function Integration		No	No	No	No	**Yes (clipped-relu)**	Yes (relu)	Yes (relu)
FoM*		—	0.86	0.45	0.20	**1.83**	0.098	—

* We introduce a figure of merit that measures the power efficiency at the maximum activation precision on unit cell area: FoM=power efficiency×maximum activation precision/cell area.

† **A real-time handwritten digit recognition demonstration on our PE chip is available online: https://bit.ly/AICHIP.**