First Demonstration of Ultra-High Precision 4Kb 28nm HKMG 1FeFET-1T Based Memory Array Macro for Highly Scaled Deep Learning Applications

Sourav De [1]

[1]Fraunhofer IPMS

October 30, 2023

Abstract

This paper reports high precision, highly linear MAC operation conducted on 28nm ferroelectric (Fe) FET (FeFET) based 4Kb computing-in-memory (CIM) core with 1FeFET-1T structure. The CIM-macro consists of 4 Kbit ultra-high precision FeFET based synaptic core, ADCs, and peripheral components for data processing. The crossbar array in the synaptic core was divided into 8×8 tiles for minimizing the voltage swing variations across word-line (WL), source-line (SL) and bit-line (BL). The FeFET-based macro achieved software-comparable inference accuracy for LeNET-5 and VGG-16 networks for MNIST and CIFAR-10 datasets.

First Demonstration of Ultra-High Precision 4Kb 28nm HKMG 1FeFET-1T Based Memory Array Macro for Highly Scaled Deep Learning Applications

Sourav De[1], Franz Müller[1], Nellie Laleni[1], Taha Soliman[2], Ashish Shrivastava[1], Nandakishor Yadav[1], Sukhrob Abdulazhanov, Maximilian Lederer[1], Shaown Mojumder[1], Alptekin Vardar[1], Tarek Ali[1], Tobias Kirchner[2], Fu-Xiang Liang[3], Hoang-Hiep Le[3], Md. Aftab Baig[3], Darsen Lu[3], Konrad Seidel[1], Thomas Kämpfe[1]

[1]Center Nanoelectronic Technologies, Fraunhofer IPMS, [2]Robert Bosch GmbH, Renningen Germany

[3]IME, Department of EE, NCKU, Tainan, Taiwan, emails: sourav.de@ipms.fraunhofer.de

Abstract:-This paper reports high precision, highly linear MAC operation conducted on 28nm ferroelectric (Fe) FET (FeFET) based 4Kb computing-in-memory (CIM) core with 1FeFET-1T structure. The CIM-macro consists of 4 Kbit ultra-high precision FeFET based synaptic core, ADCs, and peripheral components for data processing. The crossbar array in the synaptic core was divided into 8×8 tiles for minimizing the voltage swing variations across word-line (WL), source-line (SL) and bit-line (BL). The FeFET-based macro achieved software-comparable inference accuracy for LeNET-5 and VGG-16 networks for MNIST and CIFAR-10 datasets.

Introduction: Recent research in hafnium oxide (HfO_2) based Fe memories have manifested its potential as next generation non-volatile memory (eNVM). Compatibility with 28nm HKMG-technology and finFETs [1-8] have also paved the way of implementing CIM-macro using FeFET based synaptic core with state-of-art CMOS technologies. However, the major drawback lies in the device-to-device variation ($\Delta I^{D2D}{}_d$) in drain current (I_d) of the FeFET cells, especially for low threshold voltage (LVT) state [1], which hinders high precision arithmetic operation. Although previous report on 1F-1R devices shows mitigation of $\Delta I^{D2D}{}_d$ in LVT FeFET [6], it was limited to memory cells. In this work, we focus on circuit level implementation of FeFET memory (Fig.1(a)). Most of the eNVM based CIM-macros till date are based on 1T-1RRAM or phase change memory (PCM) devices, which get severely affected by large feature size and low throughput. We have built 1FeFET-1T based 4 Kbit memory arrays along with peripheral circuits and analog-to-digital converters (ADCs) using 28nm HKMG technology. Fig.1(b) compares this work with other state-of-art works. $\Delta I^{D2D}{}_d$ in BL-current (I_{BL}) were mitigated by current-limiter (CL) transistor and the sneak-path issue was mitigated by deploying *WRITE*-inhibit and *READ*-inhibit operation. The I_{BL} from each column of a tile is connected to the input of 3 bit current-mode ADCs. The experimentally obtained results from the CIM-unit was statistically modeled for multi-layer perceptron (MLP) and convolutional neural network (CNN) simulation in CIMulator frameworks [7], which achieved inference accuracy above 99% and 83% for MNIST and CIFAR-10 datasets.

Experiments and Results: The experiments began with characterization of standalone FeFETs (Fig.2(a)) into two splits. The first split had single FeFET cells, and the second split had a CL attached to drain terminal of the FeFET cell. The FeFET cells were programmed and erased by 500ns pulses of amplitudes 4.5V and -5V. Fig.2((b), (c)) shows the program-erase characteristics. Significant $\Delta I^{D2D}{}_d$, especially for LVT state, is observed for split:1. The array-level multiply and accumulate (MAC) operations get significantly affected even by small $\Delta I^{D2D}{}_d$. The statistics of I_d for CL embedded FeFETs (Fig.2(d, e)) show significant improvement in $\Delta I^{D2D}{}_d$ for LVT states. The device level benchmarking was followed by design and fabrication of CIM-macro. The CIM unit is composed of input-output drivers, FeFET-crossbar, and data converters. The gate, drain and source of the FeFETs in the crossbar, are connected to a WL, BL, and SL. The WL receives feature map of the images as input and activates corresponding cell. The BL and SL are connected to access transistors with CL feature. The access transistors across BL and SL determine the mode of operation of the crossbar. The crossbar operates in one of the following four modes (i). *WRITE* (ii). *READ* (iii). *WRITE-inhibit* (iv). *READ-inhibit*. A *RESET* pulse of 0V is applied to the inhibit switches prior to programming. The inhibit switches for adjacent columns are biased at 1V to prevent the change of state. The crossbar was block wise erased by -5V at WL, bit-wise programmed and *READ* operation of I_{BL} was performed column wise. The active BL was biased at 100mV, and the bulk was biased at 500mV. The other BLs were kept inactive by operating in *READ-inhibit* mode. Fig.3(a), demonstrating the MAC operation from a single tile, shows negligible leakage in I_{BL} with the biasing scheme. Fig.3(b). shows highly linear and V_{WL} independent MAC operation. The MAC operation was performed over 20 different tiles for statistical modelling. Fig.3(c) shows stable MAC operation over 20 different tiles from the crossbar array with a maximum standard deviation of 5% from the mean value for any state. The column of the single tile is terminated with a current-mode ADC as input. The ADC used in this work is 3-bit low-precision current mode ADC with a reference current (I_{ref}) value of 100nA. While I_{BL} is smaller than I_{ref}, the first current mirror in ADC maintains high V_{out}. As, I_{BL} rises above I_{ref}, V_{out} is dropped to lower value (Fig.3(d)). The ADC is followed by an encoder for generating the binary output for MAC operation. The linearity of ADC operation has been shown in fig.3(e), where it is shown that a proper choice of reference-bias voltage minimizes the non-linearity error in ADC operation. Finally, the performance of this system is evaluated in CIMulator platform for CNN and MLP (Fig.4(a)) operation. The inference accuracy in presence of $\Delta I^{D2D}{}_d$ for MNIST and CIFAR-10 was 99% and 83% respectively. The CIM-unit also achieved 83.4% inference accuracy for the white blood cell quaternary classification problem (WBC) (Fig.4(b)). Finally, this work is benchmarked with other *state of the art* CIM chips in Table:1.

Conclusion: Ultra-high precision 28nm HKMG technology based CIM-macro with FeFET based synaptic core has been demonstrated for the first time. High precision, linear MAC operation is conducted on chip and the neural network verification yields inference accuracy above 98% for MNIST and 83% CIFAR-10 datasets.

Acknowledgements: This work is funded by ECSEL Joint Undertaking project TEMPO in collaboration with the European Union's H2020 Framework Program and National Authorities, under grant agreement number 826655. We thank Globalfoundries for the provision of 28nm technology FeFET.

Architecture of 28nm HKMG Technology Based FeFET Memory Array

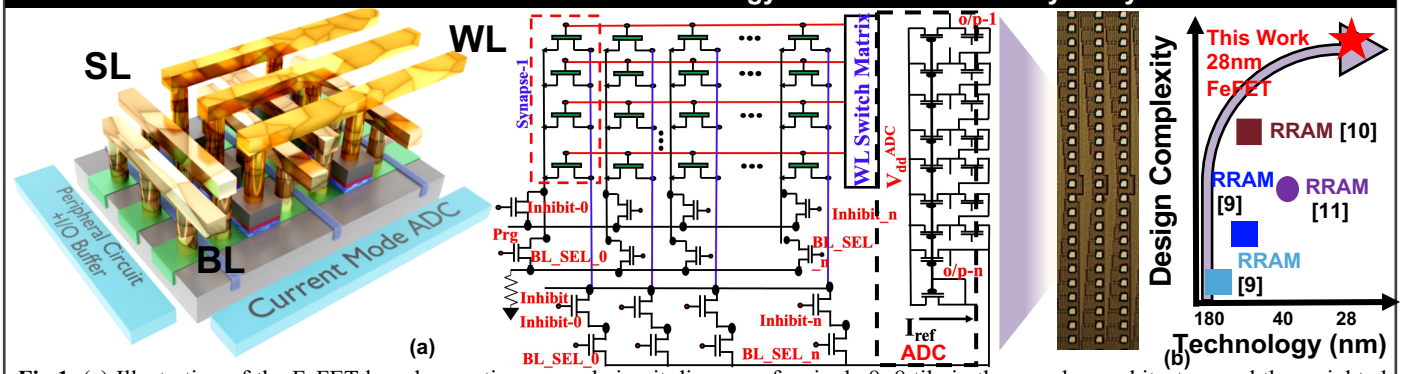

Fig.1. (a) Illustration of the FeFET based synaptic core and circuit diagram of a single 8×8 tile in the crossbar architecture and the weighted sum operation. The inset shows optical image of the crossbar. The tile-based operation minimizes the IR drop and delay across WL,BL and SL. (b) Technology node versus estimated energy efficiency shows the advantage of this CIM-macro with other *state-of-the-art* processes.

Characterization FeFET Crossbar Array

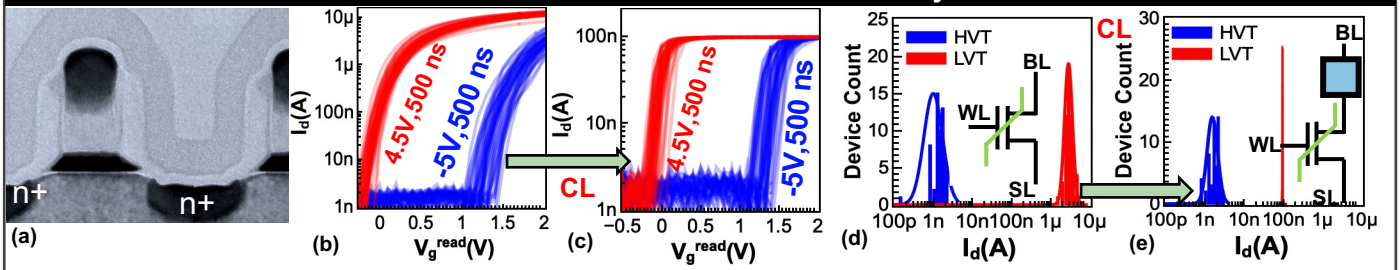

Fig.2. (a) TEM image of the 28nm current limited FeFETs used in this work. 500 ns pulses of amplitude 4.5V and -5V has been used for program-erase operations. (b) Program-erase operation of 50 FeFET samples without CL shows high variation in I_d for LVT-state. (c) Although embedding a CL at the drain terminal reduces I_{ON}/I_{OFF} ratio, high I_{ON} also induces large variability for LVT. (d,e) The distribution plot of I_d for HVT and LVT states prove that CL reduces the $\Delta I^{D2D}{}_d$, which allows high precision MAC operation in large arrays.

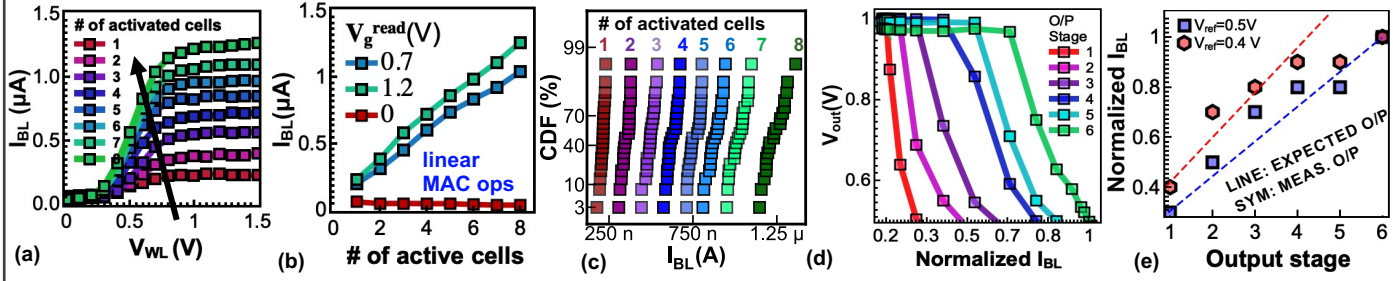

Fig.3. (a) **MAC Operation:** BL current measured *w.r.t* WL voltage from an 8×8 tile of the crossbar array. All the transistors in the tile are pre-programmed to LVT state by 4.5V pulse at WL. The *READ* of BL current was performed 2 seconds after *WRITE* operation by applying 100mV at BL. During *READ* operation the FeFETs were activated sequentially to observe current weighted sum operation through BL. (b) The current-limiter embedded at the end of each tile generates super-linear MAC operation with high precision. (c) The CDF plot of BL current variation shows stable array level operation. (d) Output Voltage vs Input current of 3-bit thermometer code ADC. (e). Linearity in ADC plays important role in neuromorphic applications. Comparison between actual o/p from the desired one shows only a small deviation.

Neuromorphic Computing Application and Performance Benchmarking

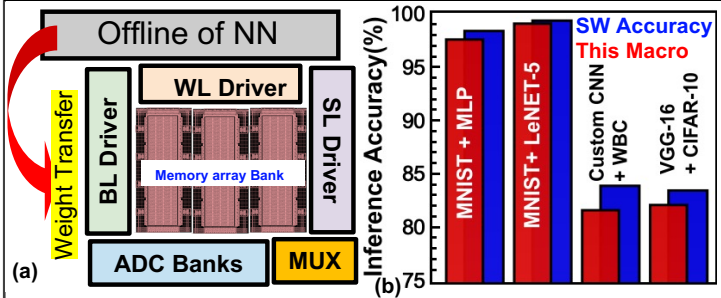

Table:I Benchmarking

	Structure	Technology Node (nm)	WRITE Voltage (V)	Ferroelectric Material	Array Size
[12]	1T-1C	130	2.5	$Hf_{0.5}Zr_{0.5}O_2$	64 Kbit
[13]	1T-1C	130	2	$Hf_{0.5}Zr_{0.5}O_2$	64 Kbit
[14]	1T-1C	130	2.5	Si:HfO_2	16 Kbit
This Work	1 FeFET-1T	28	4.5	Si:HfO_2	4 Kbit

Fig.4. (a) NN simulation performed in CIMulator platform to validate the performance of memory array for system-level applications. The platform has several NN architectures optimized for different datasets. The performance of the macro was tested for all such combinations. (b) The inference accuracy obtained after offline-training of the neural network shows accuracies comparable to software benchmark.

References: [1]. Lyu et al., IEDM, 2019, pp. 15.2 [2] Sharma et al., IEDM, 2020, pp.18. [3]. Zhou et al., IEDM, 2020, pp.18.6 [4] T. Ali et al., IEDM 2019, pp. 28.7 [5] T. Ali et al., IEDM 2020, pp. 18. [6].T. Soliman et al. IEDM, 2020, pp. 29.2.1-29.2.4 [7]. De, Sourav, et al. Frontiers in Nanotechnology: 108. [8]. S. De *et al.*, 2021 VLSI, pp. 1-2. [9] R.Mochida *et.al.*,VLSI 2018.[10]. W.Wan *et.al., ISSCC 2020.* [11]. Yoon *et al.*, ISSCC 2021 [12] J.Okuno *et al.*, VLSI 2020[13]. J.Okuno *et al.*, IMW 2021 [14] T.Francois *et al.*, IEDM 2021