# IGZO-Based Compute Cell for Analog In-Memory Computing—DTCO Analysis to Enable Ultralow-Power AI at Edge

D. Saito, J. Doevenspeck, S. Cosemans, H. Oh, M. Perumkunnil, I. A. Papistas, *Member, IEEE*, A. Belmonte, N. Rassoul, R. Delhougne, G. Kar, P. Debacker, A. Mallik, D. Verkest, *Member, IEEE*, and M. H. Na

***Abstract*— We propose, for the first time, an indium gallium zinc oxide (IGZO)-based 2T1C compute cell (IGZO-cell) for analog in-memory computing. To assess the impact of an IGZO-cell-based array including the periphery on power and accuracy, a PyTorch framework was developed to analytically modeled analog components. The results are reported for a ResNet20 network on the Canadian Institute For Advanced Research-10 (CIFAR-10) benchmark. The state-of-the-art energy efficiency of 15 peta operations per second (POPS)/W including the periphery is achieved by using our proposed IGZO-cell with CMOS compatibility. Finally, it is shown that, with a properly trained neural network model, there is no degradation of test accuracy with 10% device to device variability for the IGZO devices.**

***Index Terms*— In-memory-computing, indium gallium zinc oxide (IGZO), inference.**

## I. Introduction

ENERGY-EFFICIENT and small-footprint deep neural network (DNN) accelerators are needed to enable inference in edge devices. Hardware–software codesign has resulted in efficient low-precision digital DNN accelerators [1]. However, analog in-memory computing (AiMC) has the potential to achieve even higher efficiency and higher precision at the same time. The matrix–vector multiplication (MVM) is a key component since the dominant operations in DNNs is MVM which can be implemented with high efficiency by AiMC as demonstrated for an static random access memory (SRAM)-based AiMC accelerator in [2]. Other existing implementations [3]–[5] are also based on traditional memories, resulting in a large area and suboptimal efficiency. Also, the circuit design in this approach reaches the limit of design capability. Therefore, emerging memories such as resistive random access memory (RRAM) [6] and phase change memory (PCM) [7] are promising candidates for AiMC and device-circuit codesign (DTCO) is needed. In this work, we explore the potential of a yet unconsidered memory technology, indium gallium zinc oxide (IGZO). A DTCO approach is developed to evaluate several circuit and algorithm requirements and constraints on the IGZO technology. A novel DTCO study, developed in this work, shows that IGZO is a promising candidate to implement AiMC hardware. FETs with ultralow leakage current ($<1$ aA/$\mu$m) can be achieved when IGZO is used as an active layer which is reported in [8] with the gate length and width are 25 and 21 nm, respectively. We fabricated the IGZO FETs in our facility [see Fig. 1(a)]. Since this is a back end of line (BEOL) device, the front end of line (FEOL) is available for the periphery. Importantly, the low ON-state current (less than $\mu$A/$\mu$m) of the FETs due to the low mobility of IGZO material is a key requirement for AiMC arrays that are scalable to large DNNs and ultralow-power MVM operation [9]. Recently, we reported a circuit implementation based on a successive approximation (SAR) ADC and pulsewidth (PW) -encoded activations [9]. This article is focused on the enablement of an IGZO-based 2T1C compute cell (IGZO-cell) for AiMC with a nondestructive read. A DTCO framework is developed in PyTorch to explore the energy efficiency, latency, and accuracy of such an implementation. Also, this framework can guide future technology development of IGZO devices for AiMC to perform DNN inference.

Manuscript received June 15, 2020; revised September 11, 2020; accepted September 16, 2020. Date of publication October 5, 2020; date of current version October 22, 2020. This work was supported by the IMEC. The review of this article was arranged by Editor P. Narayanan. *(Corresponding author: D. Saito.)*

D. Saito is with Sony Semiconductor Solutions Corporation, Atsugi 243-0014, Japan (e-mail: daisukea.saito@sony.com).

J. Doevenspeck is with the IMEC, 3001 Leuven, Belgium, and also with the Department of Electrical Engineering, KU Leuven, 3001 Leuven, Belgium.

S. Cosemans, H. Oh, M. Perumkunnil, I. A. Papistas, A. Belmonte, N. Rassoul, R. Delhougne, G. Kar, P. Debacker, A. Mallik, D. Verkest, and M. H. Na are with the IMEC, 3001 Leuven, Belgium.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2020.3025986

## II. IGZO Cell Array Study and Modeling

### A. Device and Compact Model

A compact model of the IGZO FET was built using experimental data extrapolated to the dimensions of the IGZO compute cell used in this study based on our fabricated devices (see Fig. 1). The measured devices have a channel length and width of 170 and 200 nm, respectively. The design of the proposed IGZO-cell is shown in Fig. 2(a). The functionality of the weight storage of the cell [see in Fig. 2(c)] is confirmed

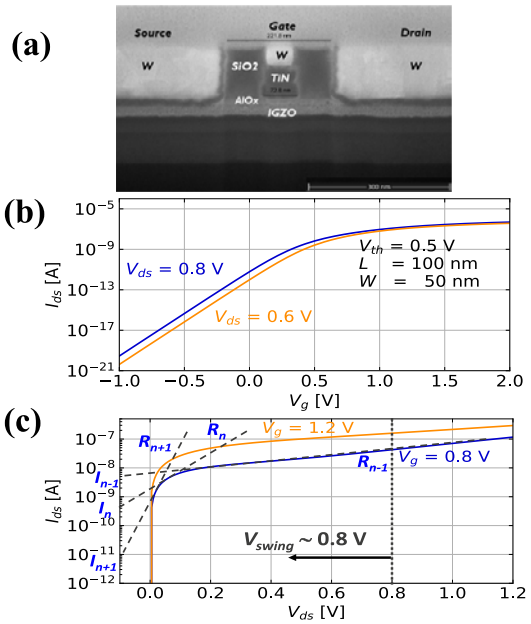
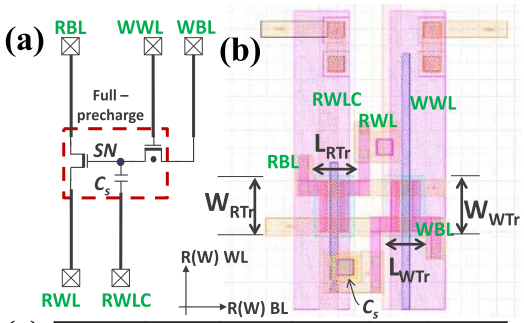



Fig. 1. (a) Cross SEM of IGZO FET and its (b) $I_{ds} - V_g$ curve and (c) $I_{ds} - V_{ds}$ curve from the compact model. The device size ($L$ = 100 nm, $W$ = 50 nm) is extrapolated from the measurement data of devices with the size $L$ = 170 nm, $W$ = 200 nm.




(c)

| CM (Vth) | WWL | WBL / SN = $V_g$ | RWL / RWLC | RBL = $V_d$ | RWL = Act |
|---|---|---|---|---|---|
| Rev1 (0.5) | 1.4 | 0.8 | GND | 0.8 | 0.8 / 0 |

Fig. 2. 2T1C IGZO-cell of (a) schematic, (b) layout, and (c) operating condition of Fig. 7(a). RBL is the Sum. line. RWL is the Act. line. SN is the storage node for storing the weight in Fig. 3(a).

by storing a charge of 0.1 fF on storage $C_s$ node (SN) in 3 ns by applying 1.4 V on the write word line (WWL).

The wire resistance was determined to be 1 Ω/cell by extracting it from a layout in a commercially available 22-nm node [see Fig. 2(b)].

## B. MVM Architecture

Fig. 3(a) shows the MVM architecture using the proposed IGZO-cell consisting of two IGZO FETs and a capacitor. Each layer of DNN network can be mapped to the MVM architecture. The multiply–accumulate (MAC) operation is performed by a complementary summation line (Sum. line). Ternary weights are encoded by using two binary IGZO-cells. Weight is stored as charge on a storage capacitor ($C_s$) of IGZO-cell [Fig. 2(a)]. A zero weight can be programmed by having no charges on $C_s$ for two binary IGZO-cells and a +1 or −1 weight is programed if only one of two IGZO-cells

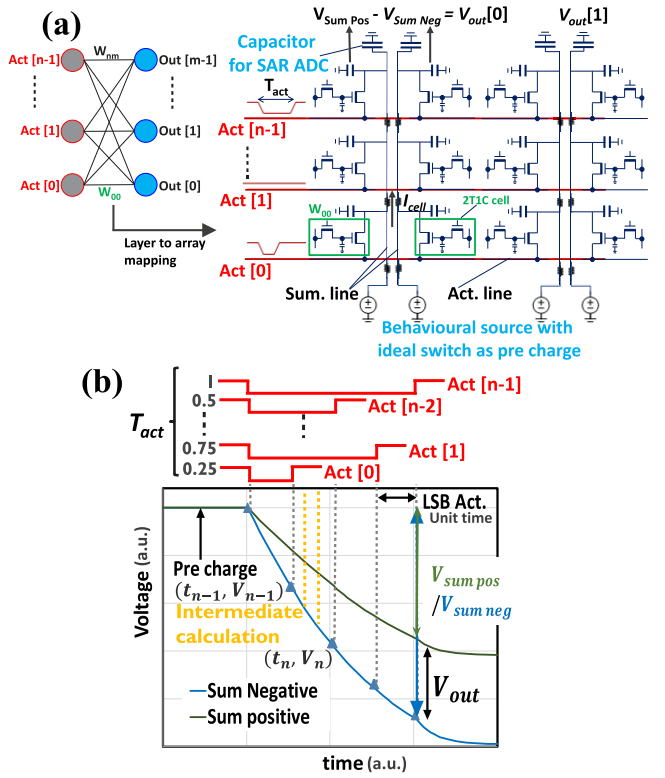



Fig. 3. AiMC architecture. (a) DNN layer to analog MVM architecture with PW-encoded signals and complementary summation line. (b) Calculated summation output from modeling equation in (1).

has no charges. Digital input activations were translated to PW-encoded signals on the activation line (Act. line) as shown in Fig. 3(a). After the Sum. lines are precharged and the electric potentials of Act. lines are matched with the Sum. lines, the discharge happens by setting Act. line to 0 V. This results in a cell charge of $Q_{\text{cell}} = I_{\text{cell}} \times T_{\text{act}}$ [see Fig. 3(a)] which emulates the operation in hardware: weight × activation. Finally, the SAR ADCs capture the differential signal [$V_{\text{out}}$ in Fig. 3(b)] at each pair of Sum. line. This architecture has the advantage that weights can be stored independent of the Act. line voltage. The IR drop is not considered on the Act. line due to the low wire resistance and small number of Sum. lines in this study.

## C. Modeling

The MVM operation in hardware was modeled by equations of the equivalent circuit in (1) as a

$$V_n = V_{n-1} \exp\left(\frac{-(t_n - t_{n-1})}{\boldsymbol{R_{norm}} * C_{line} * N_{Sum\_line}}\right)$$

$$-\alpha \frac{\boldsymbol{I^*_{n\_total}}(t_n - t_{n-1})}{\boldsymbol{C^*_{line}} N_{Sum\_line}} \quad \text{Calibration coefficient}$$

$$\boldsymbol{R_{norm}} = \frac{\boldsymbol{R_{act1w0}}}{\boldsymbol{N_{act1w0}}} + \frac{\boldsymbol{R_{act1w1}}}{\boldsymbol{N_{act1w1}}} \frac{\boldsymbol{R_{act0w0}}}{\boldsymbol{N_{act0w0}}} \frac{\boldsymbol{R_{act0w1}}}{\boldsymbol{N_{act0w1}}}$$

$$\boldsymbol{I_{n\_total}} = \boldsymbol{I_{act1w0}} * \boldsymbol{N_{act1w0}} + \boldsymbol{I_{act1w1}} * \boldsymbol{N_{act1w1}} + \boldsymbol{I_{act0w0}} * \boldsymbol{N_{act0w0}} + \boldsymbol{I_{act0w1}} + \boldsymbol{N_{act0w1}} \quad (1)$$

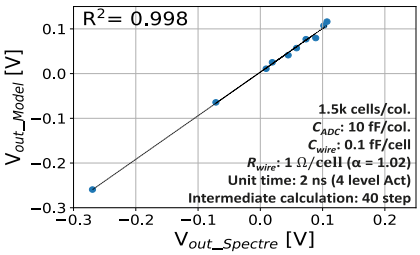



Fig. 4. Fitting result of $V_{out}$ between the Spectre simulation and our analytical model based on the concept in Fig. 1(c) and (1).

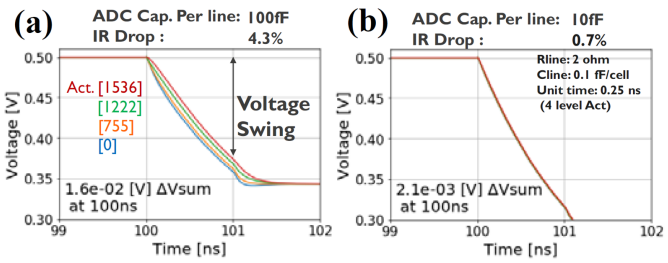



Fig. 5. IR drop along one summation line after performing the summation in Fig. 3(a) for the IGZO-cells corresponding to the activation address from Act [0] to Act [1536]. The result is shown in case of (a) 100 and (b) 10 fF for ADC capacitance. IR drop is defined as the ratio between the value of voltage drop and the voltage at the least voltage drop node.

combination of a current source and a resistor incorporating the $I_{ds}$–$V_{ds}$ curve [Fig. 1(c)] considering weight and activation patterns. The MVM results can be readout as the voltage swing on the Sum. line. The first term shows the $RC$ decay component due to the discharge and the second term represents the current source-like behavior [Fig. 1(c)]. The voltage at the node of the IGZO-cell on the Sum. line ($V_n$) at the time, $t_n$, is shown in Fig. 3(b). $V_n$ is calculated from the voltage, $V_{n-1}$, at the time, $t_{n-1}$, equivalent to one LSB duration ago ($t_n - t_{n-1}$ = LSB Act.). $R_{norm}$ is the total resistance of the IGZO FETs in IGZO-cells on the Sum. line. The resistance of each IGZO FET is determined by the activation state and the weight state that are applied to each IGZO FET such as $R_{act1w0}$ having the activation (ON or OFF-activation: act1 or act0) and the weight state of the IGZO FET (charged on SN or not charged: w1 or w0). $N_{act1w0}$ is the number of IGZO FETs which are applied to the ON-activation and no charge at SN. $C_{line}$ is the wire capacitance per IGZO-cell. $N_{Sum\_line}$ is the total number of IGZO-cell on the Sum. line. $I_{n\_total}$ is the total current of the current source components ($I_n$) of each IGZO FET in IGZO-cell on the Sum. line at $V_n$ calculated from the $I_{ds}$–$V_{ds}$ curve in Fig. 1(c). $I_n$ is determined by the activation state and the weight state that are applied to each IGZO FET such as $I_{act1w0}$ which is the current source component of IGZO FET having the ON-activation and no charge at SN. The calibration coefficient, $\alpha$, is adjusted to take IR drop into account. An intermediate calculation for $V_n$ in LSB Act. shown in Fig. 3(b) is also available.

Fig. 4 shows that our analytical model matches well with the Spectre result. Tuning the ADC capacitance allows to mitigate the IR drop impact. The 10 fF of ADC capacitance reduces the IR drop due to distributed currents to the IGZO-cells and a lower wire resistance relative to the current on the Sum. line (see Fig. 5). Note that the activation with four levels (2-bit) is used to simplify our study. The same results are shown with 4-bit activations discussed in Section III.

## III. DTCO Framework Result

This DTCO framework captures the impacts of device imperfections and system level optimization applicable for any DNN networks [10]. ResNet 20 on Canadian Institute For Advanced Research-10 (CIFAR-10) is used for the study [Fig. 8(a)]. The model for inference is quantized for Acts, MACs, and weights of 4-bit, 4-bit, and ternary, respectively. The model was trained with Gaussian noise ($\sigma$) that is a percentage of the quantization interval. Noise of weights

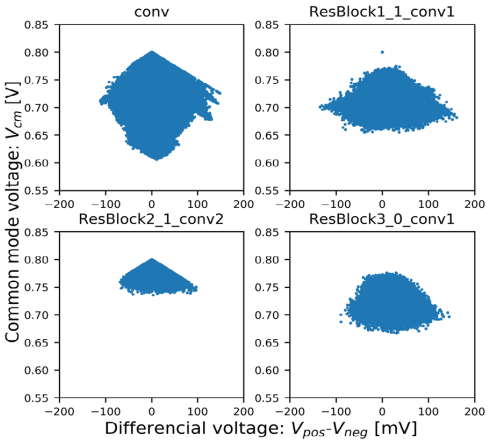



Fig. 6. Output signal study of the summation line with operating conditions of Fig. 7(a). The relationship between $V_{out}$ and the average swing voltage of $V_{pos\_sum}$ and $V_{neg\_sum}$ in Fig. 3(b) (common mode voltage: $V_{cm}$).

and MACs is corresponding to device conductance variations and is randomly distributed in MACs. Noise is assumed to change time to time as a random drift. The model including $\sigma_{Acts}$, $\sigma_{MACs}$, and $\sigma_{weights}$ of 10%, 50%, and 10%, respectively. This lower precision and noise resilience model were trained by using teacher aware training technique as reported in [11]–[13]. The circuit parameters for this DTCO study are shown in Fig. 8(b) and we assume that each layer of ResNet 20 network is mapped to a different AiMC array with the array size corresponding to each layer size.

### A. Summation Output Study

The summation output study of MVM as shown in Fig. 3(a) was also done to compare with the ADC requirements (see Fig. 6). The graph shows the common mode voltages ($V_{cm}$) as a function of $V_{out}$ ($V_{pos} - V_{neg}$) shown in Fig. 3(b) for some layers of DNN network described in Fig. 8(a). The ADC can work at >0.35 V for $V_{cm}$ and 11 mV for LSB of $V_{out}$ in our design. Since the 4-bit precision of MACs is considered, the dynamic range of $V_{out}$ needs to be more than 175 mV. All the results of $V_{out}$ from the layer of DNN network (some layers shown in Fig. 6) were satisfied with these requirements. Therefore, IGZO-cell can be operated within the specification of our designed ADC with 4-bit MACs.

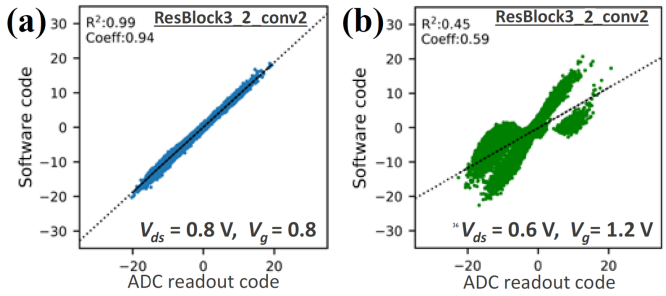



Fig. 7. Impact of transistor operation of (a) $V_d = 0.8$ V, $V_g = 0.8$ V and (b) $V_d = 0.6$ V, $V_g = 1.2$ V to ADC readout ($V_{out}$) variation compared to software code. $V_d$ is the precharge voltage in Fig. 3(b) at RBL in Fig. 2(a). $V_g$ is at SN in Fig. 2(a).

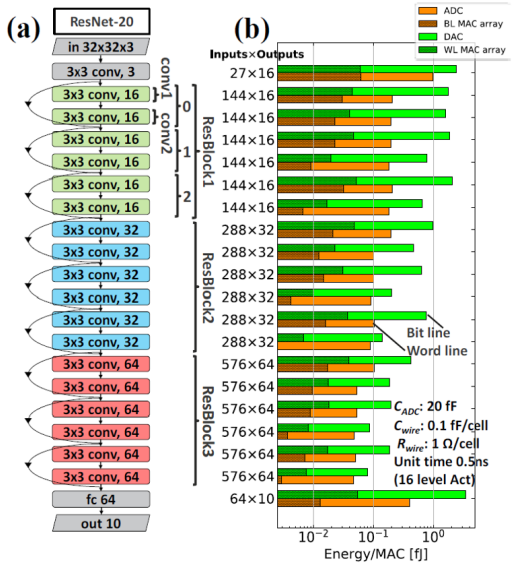



Fig. 8. (a) Network used in this article ResNet20 CNN with rectified linear units (ReLUs) and Batch normalization. The first layer of each ResBlock is 1 × 1 convolution (not shown in the figure) of which the stride in ResBlock1, ResBlock2, and ResBlock3 is 1, 2, and 2, respectively. The zero padding for the feature map is used. (b) Energy per MAC of MVM ("WL MAC array" and "BL MAC array" are for the energy coming from the Act. line and the Sum. line, respectively) with periphery ("DAC" and "ADC").

## B. Effect of IGZO FET Behavior

The effect of the IGZO FET $I$–$V$ behavior study is in Fig. 7. A larger $V_g$ and a smaller $V_d$ (precharge voltage) show a worse correlation coefficient (0.4) between $V_{out}$ and the software output code [Fig. 7(b)] due to the nonlinearity effect [an ocher line in Fig. 1(c)] and the reverse current during the discharging from the IGZO-cell having the OFF-activation and the charge at SN (weight state is "1": IGZO FET is charged at SN).

## C. Energy, Latency, and Inference Accuracy Estimation

Fig. 8(b) shows the energy per MAC in each layer of Fig. 8(a) with an ADC capacitance of 20 fF and an IGZO-cell capacitance of 0.1 fF. A 4-bit DAC and a 4-bit ADC were used in the periphery. DAC and ADC energies per conversion were estimated to be 25 and 40 fJ, respectively, from our design. The MAC array itself has a high energy efficiency of 50 peta operations per second (POPS)/W [Fig. 9(a)]. Considering the periphery, the average energy efficiency is 2.3 POPS/W. For larger layers, the energy efficiency reaches up to 15 POPS/W such as ResBlock3_2_ conv2 [see Fig. 8(b)]. In these larger layers, the periphery energy is amortized over a larger MAC array (see Fig. 9).

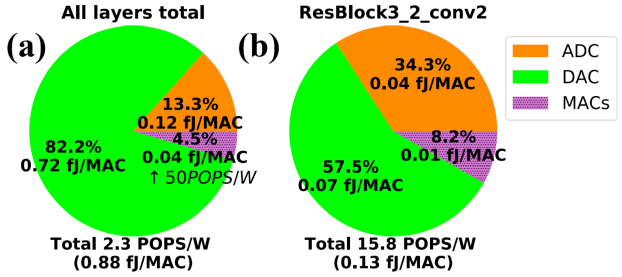



Fig. 9. Energy ratio between the MAC array and periphery (DAC and ADC) for (a) all layers and (b) ResBlock3 (576 × 64).

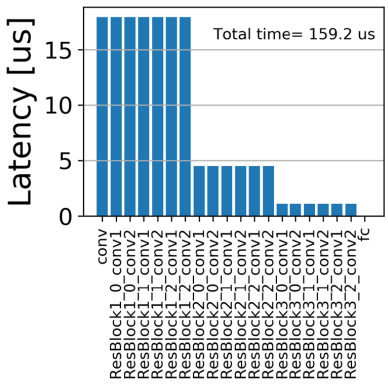



Fig. 10. Latency estimation for each layer of the ResNet20 network used in this work shown in Fig. 8(a).

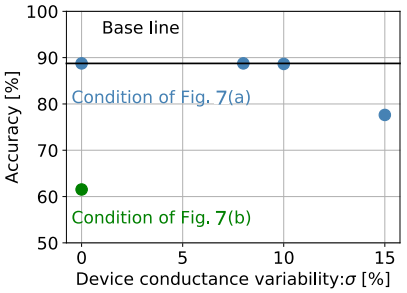



Fig. 11. Inference accuracy as a function of device-to-device conductance variability with 10 000 test samples. The conditions for the simulation are corresponding to the conditions in Fig. 7(a) and (b).

The estimated latency is 160 $\mu$s per inference assuming the precharged duration is 10 ns and the unit time in Fig. 3(b) is 0.5 ns (see Fig. 10). The inference accuracy for the condition of Fig. 7(a) shows noise resilience to device conductance variability up to 10% (Fig. 11 blue points), which is the same accuracy of inference without noise. If the device is operated as in Fig. 7(b), this results in a lower accuracy (<70%) as shown in Fig. 11 with green points.

Table I shows the benchmarking results with device metrics relevant for AiMC used for DNN inference systems. Our proposed IGZO-cell offers higher precision and has an advantage of energy efficiency and throughput even in smaller MAC

TABLE I
AiMC COMPARISON TO STATE OF THE ART

| | NAND flash Bavandpour [14] | PCM Giannopoulos [7] | FeFET Obradovic [14] | RRAM Xue [6] | SRAM Bankman [5] | SRAM Valavi [3] | IGZO This work |
|---|---|---|---|---|---|---|---|
| MACs Inputs/Outputs | 1000/1000 | 256/256 | - | 256/512 | 1024/64 | 4608/512 | 576/64 |
| Precision | | | | | | | |
| Weights [bit] | 4 | 8 | 4 | 3 | 1 | 1 | Ternary |
| Inputs [bit] | 4 | - | 1 | 1 | 1 | 1 | 4 |
| Outputs [bit] | 4 | - | 1 | 3+sign | 1 | 1 | 4 |
| Peak performance | | | | | | | |
| Energy efficiency [TOPS/W] | 100 | - | - | 53.2 | 772 | 866 | 15000 |
| Throuhput [TOPS] | 10.7 | - | - | - | 0.48 | 18.9 | 2.1 |
| Device | | | | | | | |
| On state R [Ohm] | 6 ~ 16 M | 0.2 M | 33 k | > 1 M | Capacitive sense | Capacitive sense | > 150 M |
| Leakage [A/μm] | - | - | 1 p ~ 1 n | - | ~ 1 p | ~ 1 p | < 1 a |
| Bit cell | 4T | - | 2T2R | 1T1R | > 9T | 9T | 2T1C |
| Operating $V_{dd}$ [V] | 1.2 ~ 5 | < 1 | < 0.7 | e | 0.53 ~ 0.8 | 0.68 ~ 1.2 | < 0.8 |
| Technology | 55nm | <65nm | - | 55nm | 28nm | 65nm | 22nm |

arrays. The area of bit cell is more than four times smaller than SRAM cell for the same technology node. Further area improvements can be achieved by using 3-D integration of IGZO-cell in BEOL.

## IV. CONCLUSION

The IGZO-cell has a strong potential to be implemented in AiMC. We demonstrated DTCO framework for ResNet20 on CIFAR-10.

1) Energy efficiency of 15 POPS/W at >500 cells per Sum. line is achieved including the periphery.
2) Simple periphery implementation due to the compatibility with conventional technology is ensured.

The nonlinearity behavior of IGZO-cell can be used in this DTCO framework not only for DNN inference, as studied in this work, but also for the training of DNN as a future work.

## REFERENCES


[1] B. Moons, D. Bankman, L. Yang, B. Murmann, and M. Verhelst, "Binareye: An always-on energy-accuracy-scalable binary CNN processor with all memory on chip in 28nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, Apr. 2018, pp. 246–247.

[2] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.

[3] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing CNN accelerator employing charge-domain compute," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, Jun. 2019.

[4] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE J. Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, Jun. 2020.

[5] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8 $\mu$ J/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, Jan. 2019.

[6] C.-X. Xue *et al.*, "Embedded 1-Mb ReRAM-based computing-in- memory macro with multibit input and weight for CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 203–215, Jan. 2020.

[7] I. Giannopoulos *et al.*, "8-bit precision in-memory multiplication with projected phase-change memory," in *IEDM Tech. Dig.*, Dec. 2018, pp. 27.7.1–27.7.4.

[8] H. Kunitake *et al.*, "High thermal tolerance of 25-nm c-axis aligned crystalline In-Ga-Zn oxide FET," in *IEDM Tech. Dig.*, Dec. 2018, pp. 13.6.1–13.6.4.

[9] S. Cosemans *et al.*, "Towards 10000TOPS/W DNN inference with analog in-memory computing—A circuit blueprint, device options and requirements," in *IEDM Tech. Dig.*, Dec. 2019, pp. 22.2.1–22.2.4.

[10] J. Doevenspeck *et al.*, "SOT-MRAM based analog in-memory computing for DNN inference," in *Proc. Symp. VLSI Technol. Circuits*, 2020, p. 26, Paper JFS4.1.

[11] B.-E. Verhoef *et al.*, "FQ-conv: Fully quantized convolution for efficient and accurate inference," 2019, *arXiv:1912.09356*. [Online]. Available: http://arxiv.org/abs/1912.09356

[12] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," 2019, *arXiv:1902.08153*. [Online]. Available: http://arxiv.org/abs/1902.08153

[13] T. Gokmen, M. J. Rasch, and W. Haensch, "The marriage of training and inference for scaled deep learning analog hardware," in *IEDM Tech. Dig.*, Dec. 2019, pp. 22.3.1–22.3.4.

[14] B. Obradovic *et al.*, "A multi-bit neuromorphic weight cell using ferroelectric FETs, suitable for SoC integration," 2017, *arXiv:1710.08034v1*. [Online]. Available: https://arxiv.org/abs/1710.08034v1

[15] M. Bavandpour, S. Sahay, M. Reza Mahmoodi, and D. B. Strukov, "3D-aCortex: An ultra-compact energy-efficient neurocomputing platform based on commercial 3D-NAND flash memories," 2019, *arXiv:1908.02472*. [Online]. Available: http://arxiv.org/abs/1908.02472