2021 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-7281-9201-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCAS51556.2021.9401798

Compute-in-eDRAM with Backend Integrated Indium Gallium Zinc Oxide Transistors

Siddhartha Raman Sundara Raman, Shanshan Xie, Jaydeep P.Kulkarni

Department of Electrical and Computer Engineering, The University of Texas at Austin

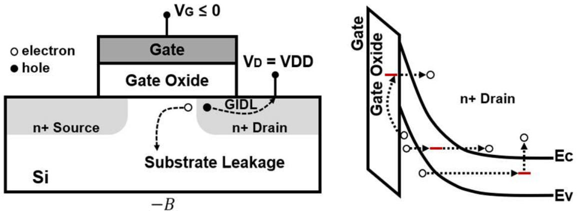
Email: s.siddhartharaman@utexas.edu, jaydeep@austin.utexas.edu

***Abstract*— With rapid growth in data intensive applications, there is an ever-increasing need for energy efficient machine learning/AI hardware accelerators. The performance and the energy efficiency of such accelerators are primarily limited due of massive amount of data movement between processing engines and the off-chip memory. This memory wall bottleneck can be mitigated by performing accelerator specific computations in the memory (CIM) array embedded with the rest of the logic blocks. Multiple embedded memory technologies are being explored to advance CIM designs. Among these, embedded Dynamic Random Access Memory (eDRAM) using backend of the line (BEOL) integrated C-Axis Aligned Crystalline (CAAC) Indium Gallium Zinc Oxide (IGZO) transistors is a promising candidate. IGZO transistor having extremely low leakage when used as an access transistor of the eDRAM bitcell can enable multi-level cell (MLC) eDRAM functionality. Moreover, higher bandwidth can be achieved by 3D stacking multiple layers of BEOL integrated IGZO devices in a monolithic manner improving the CIM performance. In this paper, we analyze various IGZO based eDRAM bitcell topologies and present an IGZO eDRAM CIM architecture. It supports 8-bit inputs/activations and 8-bit signed weights. 2-bit Flash Analog to Digital converter (ADC) is used for MLC weight bit read sensing. A representative neural network model using IGZO eDRAM and peripheral 8-b A/D converters based CIM design achieves 80% Top-1 inference accuracy for the CIFAR-10 dataset, which is within 3% of ideal software accuracy.**

Keywords*— *Accelerators, Compute in memory, Embedded DRAM, Indium Gallium Zinc Oxide, Multi-level cell

I. INTRODUCTION

Large model size and complex deep neural network architecture lead to massive amount of data movement between the processing engines and the main memory. This off-chip data access poses significant energy and latency challenges in modern ML/AI accelerators. Compute-in-Memory (CIM) is a promising approach to mitigate this memory-wall bottleneck by performing computations within the memory array [1-4]. Multiple embedded memory technologies are being explored for efficient CIM designs. In Static Random Access Memory (SRAM), CIM operations can be realized by turning on multiple word lines (WL) connected to a common 6T SRAM column [1,2]. However, WL based CIM operation in 6T SRAM suffers significantly from bitcell variations which further leads to inaccurate voltage integrated on the bitline and also increases the possibility of bit-flips. This necessitates use of larger 8T/10T SRAM bitcells [3, 4] having decoupled read and write ports incurring significant area overhead for large scale CIM designs. Embedded non-volatile memories (eNVM) such as RRAM, PCM, MRAM, FeFET and Flash are also extensively explored for CIM applications. Although, eNVMs offer zero bitcell leakage and density advantage compared to the SRAM technology, they suffer from poor write-endurance (~10^6) and lower write-speeds (10's of ns) [5]. This can limit eNVM usage in high performance, programmable CIM designs requiring fast and frequent kernel weight updates. Dynamic Random Access Memory (DRAM) has been predominantly used as the main memory technology due to its high storage density, simple 1T1C (T= Transistor, C=Capacitor) bitcell structure. Commodity DRAM can be used for CIM applications. However, this approach is susceptible to capacitor variations, charge leakage, and necessitates data duplication due to the intrinsic destructive read operation. Addition of CIM specific circuits is also prohibitive as commodity DRAM is optimized for bit density and low-cost. Furthermore, commodity DRAM utilizes limited number of lower metal layers for highly dense floorplan which severely limits available throughput. Hence, DRAM based approaches are often realized as compute-near memory [6] designs.



$$I_{GIDL} = A \cdot E_s exp(\frac{-B}{E_s})$$

where:

$$A = \frac{q^2 \cdot m_r^{1/2}}{18 \cdot \pi \cdot h^2 \cdot E_g^{3/2}} \text{ and } B = \frac{\pi \cdot m_r^{1/2} \cdot E_g^{3/2}}{2\sqrt{2} \cdot q \cdot h}$$

$$m_r = 0.2m_0$$

$$E_s = \frac{V_{gd} - \mathrm{E_g}}{(E_r/E_{r_{oxide}}) \cdot T_{OX}}$$

	m0	Eg	Er
Si	0.19	1.2	11.9
IGZO	0.25	3	15

	A	B	Es*	IGIDL
Ratio of IGZO/Si	0.3	4.5	2~3	2E-5~1E-7

*E_s: surface electric field, assume ratio to be in the range of 2~3 due to trap-assisted GIDL

Fig.1 Extreme low leakage in CAAC-IGZO transistor structure and oxide-semiconductor band diagram. High effective mass for electron, high band gap, high relative permittivity, lowers Gate Induced Drain Leakage (GIDL) significantly compared to the Silicon counterpart.

Embedded DRAM (eDRAM) on the other hand which monolithically integrates DRAM bitcell with high performance logic transistors and interconnects can support high performance CIM operations. Among the incumbent embedded memories, eDRAM offers the densest bitcell size, largest on-die capacity, high endurance, high-performance, high-bandwidth, low energy/bit access, and low soft error rate [7-10]. In addition, charge sharing operation intrinsic to DRAM can be efficiently harnessed to perform analog CIM computations. Although, eDRAM is a promising embedded memory technology for performing energy efficient and high-performance large scale CIM designs, it suffers from low retention time due to higher access transistor leakage and reduced storage cell capacitance compared to the commodity DRAM bitcell. Hence, alternative device/technology approaches need to devised to improve the eDRAM retention time for its adoption in high speed CIM designs. One such promising device technology which exhibits extreme low leakage is C-Axis Aligned Crystalline Indium Gallium Zinc Oxide (CAAC-IGZO) transistor [11-13]. CAAC-IGZO

978-1-7281-9201-7/21/$31.00 ©2021 IEEE

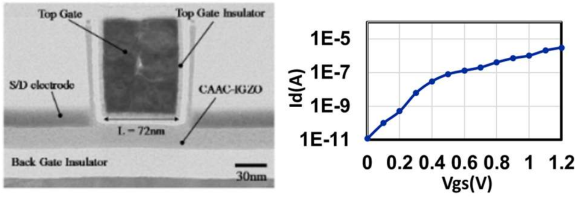


Fig. 2 72nm CAAC-IGZO transistor experimental device cross-section [12] and Log(I_D) vs V_{GS} characteristics calibrated with a BSIM Level-3 model

transistor can be utilized as the eDRAM access transistor increasing its retention time. IGZO devices can be integrated in the BEOL process steps[14], and can be stacked in a monolithic 3-D fashion to achieve higher density/bandwidth and operate with modest voltages for read-write operation.

Thus, the IGZO transistor based 3-D stacked eDRAM technology with dense bitcell, large capacity, high retention time, low leakage, high performance and ability to perform neural network computations is suitable for large scale, high performance, CIM applications. In this paper, we analyze multiple IZGO eDRAM bitcell topologies and present 3T1C gain-cell IGZO eDRAM multi-level cell for performing CIM using peripheral A/D converters. The classification accuracy for CIFAR-10 dataset using a representative neural network is also quantified.

II. Indium Gallium Zinc Oxide Transistor

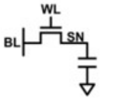
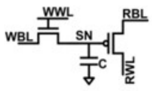
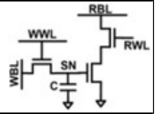
The factors affecting the refresh-time in an eDRAM/DRAM is the charge loss due to sub-threshold leakage, band-to-band tunneling and the gate induced drain leakage (GIDL) [15] of the access transistor and the storage capacitor leakage. Subthreshold leakage is exponentially dependent on the gate to source voltage and can be reduced by applying negative voltage at the access transistor gate terminal. However, this increases the electric field at the G/D interface and causes increased GIDL. It is exponentially dependent on the difference between the V_{GD} voltage as described by equation 1.

$$I_{GIDL} = A * E_s * e^{-B/Es} \quad (1)$$

where E_s is directly proportional to the difference of the gate to drain voltage (V_{GD}) and the energy band gap of the material, A is a constant inversely proportional to the bandgap of the material and B is a constant directly proportional to the bandgap of the material as shown in Fig.1. Thus, GIDL is exacerbated when the gate voltage is negative, diminishing its effectiveness in reducing the subthreshold leakage. A high band gap energy of 3eV in case of IGZO as opposed to 1.2eV in Silicon, higher effective mass (0.25 for IGZO, 0.19 for Si) and higher relative permittivity of the device further helps in the reduction of GIDL by 5-7 orders of magnitude in comparison with the Silicon DRAM [16], thus improving the retention time from milliseconds range in Silicon DRAMs to more than 10 days in IGZO DRAMs [11]. Moreover, the mobility of these devices is not affected by temperature change, unlike Silicon based DRAM [12].

In addition, CAAC-IGZO transistors are typically realized as n-type devices having moderate ON current and are amenable for low temperature backend CMOS integration. The storage density of eDRAM can be further increased by storing more than one bit in a single eDRAM bit cell [17]. However, Silicon based eDRAM's low retention time, high leakage and high coupling noise sensitivity limit it from being used as a multi-level cell (MLC) with high sense margins for differentiating between multiple storage capacitor voltage levels. This can degrade the accuracy of analog CIM computations. CAAC-IGZO access transistors having extreme low leakage current can enable MLC functionality in an eDRAM bitcell improving bit density. The bitcell density can be further increased by stacking multiple layers of CAAC-IGZO access transistors and backend capacitors. Both these attributes can help in meeting the capacity needs of large CIM designs and can sustain the eDRAM scaling while achieving long retention time and higher bit density. Compact model for CAAC-IGZO transistor which is experimentally demonstrated in [12] (Fig. 2) is developed by calibrating Log(I_D) vs V_{GS} characteristics corresponding to the first layer of the 3D layers with a body bias voltage of 0V. The device characteristics are modeled using BSIM level-3 parameters [15]. CAAC-IGZO transistor leakage current is in the order of pA [16] and the subthreshold slope parameter of the device has been modeled using the fast surface states model parameter (NFS). The on current (of the order of uA) is modeled with a lower mobility value using Ueff (effective mobility) parameter (Fig. 2).

TABLE I eDRAM bitcell configurations

	1T1C	**2T1C**	**3T1C**
Bitcell			
Area	1X	1.95X	2.14X
Access method	Destructive read	Decoupled read, write	Decoupled read, write

III. Embedded DRAM Bitcell Topologies

The 1T1C eDRAM bitcell configuration suffers from destructive read operation which can be mitigated by various gain-cell topologies using dedicated read port transistors (e.g. 2T1C, 3T1C, 4T1C and 5T1C gain-cells [18]. Among these 2T1C and 3T1C gain-cells are particularly attractive due to smaller bitcell area compared to the other gain-cell variants. For MLC CIM applications, 2T1C cannot be used because of threshold voltage clipped read bit line (RBL) swing (Vt swing around Vcc rail) making it challenging to resolve MLC under the presence of process variations. 3T1C gain-cell using N-type read and write port transistors provides a full-rail voltage swing at the RBL which can be used for accurate sensing of multiple levels stored on the bitcell storage capacitor. Table I summarizes the gain-cell eDRAM bitcell configurations and their access patterns. The timing diagram for the 3T1C multi-level cell is shown in Fig. 3. For the illustration purpose, read wordline (RWL) and write word line (WWL) pulse-width is chosen to be 10ns. For the write operation, the write bitline voltage (WBL) is set to the appropriate voltage level depending on the 2-bit data to be written. Extreme low leakage CAAC-IGZO access transistor helps in maintaining the bitcell storage node value for a long duration thus improving the gain-cell retention time.

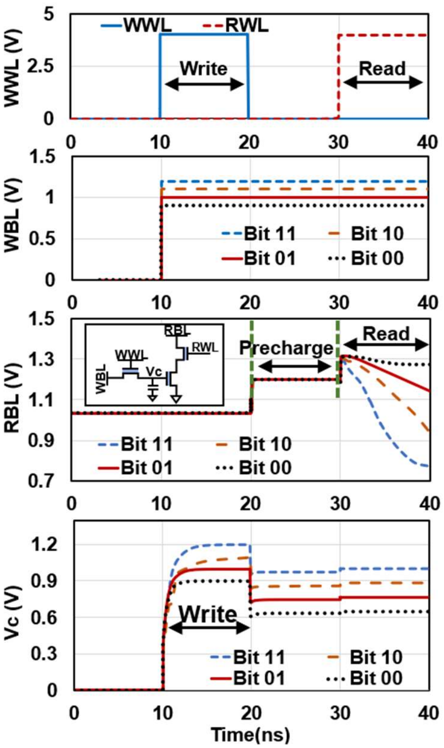


Fig. 3 Simulation results for read and write operation for 3T1C IGZO based MLC eDRAM. Vc is the bitcell storage node.

During the read operation, read bitline (RBL) is first precharged to VDD. Once the RWL is activated, the RBL discharge rate is governed by the voltage at the cell storage node which determines the overdrive voltage of the read port transistor. The multiple levels in a single gain-cell are differentiated by sensing RBL voltage after a fixed time interval by using a 2-bit flash ADC. The sense margins for differentiating between multiple levels in case of 3T1C configuration are higher compared to 2T1C bitcell due to full-rail RBL voltage swing. This helps in realizing CIM computations with minimal loss in computation accuracy. Furthermore, it is worth noting that the voltage swing at the RBL is sufficient enough to efficiently distinguish between the different levels stored in the bitcell and the voltage swing doesn't degrade much in the presence of process variations in contrast to the conventional Silicon based eDRAM approaches.

IV. CIM Architecture with 3T1C IGZO DRAM

Fig. 4 shows the block diagram of the proposed CIM architecture. The input activation and weight integer operands are 8-bits wide. The 8-bit activation values along with their 1's complement values are converted into analog voltages Va and Va_bar as shown in Fig.4 using D/A converters [19]. IGZO eDRAM array is used for storing the neural network weights in 3T1C bitcells. With MLC functionality, 8-bit kernel weights can be stored in 4 3T1C gain cells each storing 2 bits/cell. The MAC computations are performed by using charge share peripheral circuits. Fig. 5 describes the detailed circuit implementation of the 3T1C gain cell eDRAM based CIM design. Initially, the write word line is boosted so that the weights can be written using the n-type write port IGZO transistor. The RBL is pre-charged and the value stored in the bit-cell is distinguished based on the rate of discharge for different weight values during read as shown in Fig.5. The comparator circuits along with the digital encoder (3 reference voltages) each having a separate reference voltage can be used to distinguish between 2-bit values. The weight values are then fed into a charge share circuit which performs dot product between analog input (Va or Va_bar) and the weight using capacitor charge sharing operation. The zero-reference voltage is chosen to be Vcc/2 so as to accommodate signed weight operands and the signed dot products. If the weight is '0', a zero reference of 'VDD/2' is selected to be charge shared and if the weight is '1', the input activation of Va or Va_bar is selected to be charge shared. Another charge-share operation is then performed across different columns using binary scaled capacitors depending on the weight bit position to achieve accumulate and average functionality. The charge shared value is then converted into digital output using an 8-b A/D converter as shown in Fig. 5. The IGZO eDRAM based CIM design parameters are listed in Table II.

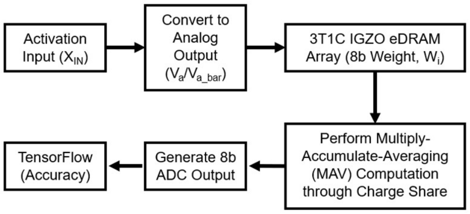

Activation Input (X_{IN})
Convert to Analog Output (V_a/V_{a_bar})
3T1C IGZO eDRAM Array (8b Weight, W_i)
Perform Multiply-Accumulate-Averaging (MAV) Computation through Charge Share
Generate 8b ADC Output
TensorFlow (Accuracy)

Fig.4 Block diagram for CIM architecture

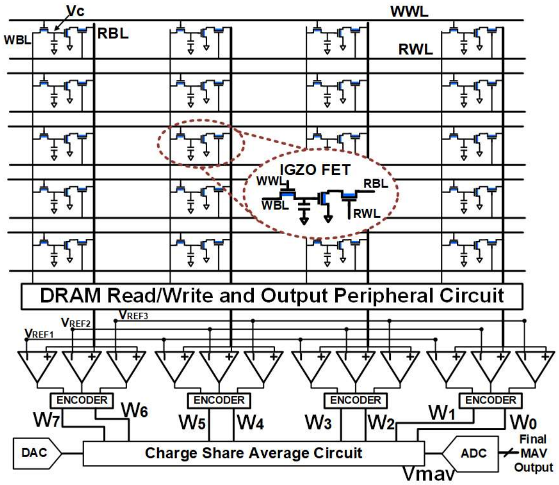


Fig. 5 Detailed circuit schematics for IGZO eDRAM based CIM: Charge Sharing circuits, flash ADC for MLC sensing in the read path

V. CIFAR-10 Results

The proposed IGZO eDRAM based CIM design efficacy is quantified using CIFAR-10 dataset with a representative convolutional neural network as shown in Fig. 6. Tensorflow [20] accuracy measurements are performed using a lambda layer as a replacement for the convolutional layers in the neural network model shown in Table II. The network has 4 convolutional layers with the first and second layer containing 32 channels each of size 32*32, the third and fourth layer containing 64 channels each of size 32*32. A 3*3 kernel has been used. The proposed design achieves 80%

Top-1 classification accuracy compared to the 83% Top-1 accuracy obtained from ideal software implementation for the same network. This analysis indicates that the CAAC-IZGO eDRAM can be a promising candidate for performing large scale, CIM designs with good accuracy.

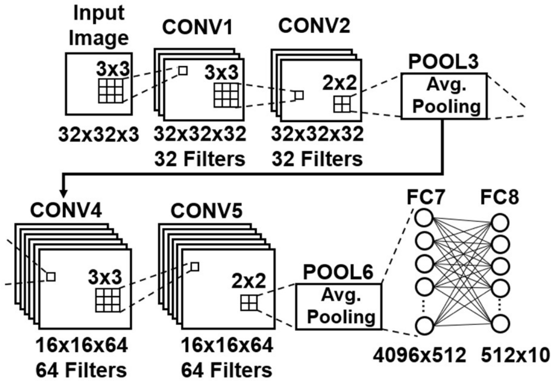


Fig. 6 Convolutional Neural Network used for evaluation of the proposed CIM design

TABLE II NEURAL NETWORK CIFAR-10 MODEL

Design parameters	Description
Neural network configuration	CONV layer – 32 3x3; ReLU CONV layer – 32 3x3; ReLU Max Pooling layer – 2x2 CONV layer – 64 3x3; ReLU CONV layer – 64 3x3; ReLU Max Pooling layer 2x2 FC layer – 512 – BN – ReLU FC layer – 10
IGZO eDRAM based CIM parameters	Inputs: 8-bit unsigned integer Weights: 8-bit signed integer Bitcell: 3T1C IGZO eDRAM MLC: 2-bits/cell Read sensing: 2-b Flash ADC Zero input voltage: VDD/2 A/D, D/A precision: 8-bit
Top-1 classification Accuracy for CIFAR-10	Software = 83% IGZO eDRAM= 80%

VI. Conclusion

Among the incumbent embedded memories, eDRAM offers the densest bitcell size, largest on-die capacity, high endurance, high-performance, high-bandwidth, low energy/bit access, low soft error rate and can be a promising candidate for large scale, high performance CIM designs. However, conventional Silicon based eDRAM suffers from leakage, and noise coupling issues. CAAC-IGZO based eDRAM on the other hand exhibits high retention time, with reduced GIDL and reduced noise coupling. Furthermore, significantly improved retention time in CCAC-IGZO eDRAM can enable MLC functionality improving the bitcell density. The backend integrated CAAC-IGZO transistors can be 3-D stacked along with backend capacitors to realize vertically stacked 3D eDRAM columns. This can improve the bit density as well as the memory bandwidth. In this paper, we analyzed multiple CAAC-IGZO eDRAM gain-cell topologies and presented a CIM design using 3T1C IGZO eDRAM gain-cell supporting MLC functionality (2 bits/cell). Classification accuracy results for CIFAR-10 dataset using a representative network shows good agreement with the ideal software accuracy. This analysis suggests that CAAC-IGZO based eDRAM technology can be a promising candidate for realizing large scale, dense, energy efficient CIM designs.

VII. Acknowledgements

This research is supported by Semiconductor Research Corporation project #2962.001 and The University of Texas at Austin. Authors would like to thank Prof. Ananth Dodabalapur, Wang Xiao and Saikat Chakraborty for useful technical discussions on IGZO device characteristics.

VIII. References

[1] S. K. Gonugondla, M. Kang, and N. Shanbhag, “A 42pJ/Decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier with On-Chip Training,” IEEE International Solid State Circuits Conference-(ISSCC), pp. 490–492, 2018

[2] A. Biswas and A. P. Chandrakasan, “Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-based Machine Learning Applications,” IEEE International Solid-State Circuits Conference-(ISSCC), pp. 488–490, 2018.

[3] J. Zhang, Z. Wang, and N. Verma, “In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array,” IEEE Journal of Solid-State Circuits, vol.52,no.4,pp.915–924, 2017

[4] X. Si et al., "24.5 A Twin-8T SRAM Computation-In-Memory Macro for Multiple-Bit CNN-Based Machine Learning," IEEE International Solid-State Circuits Conference - (ISSCC), pp. 396-398, 2019

[5] S. Mittal, J. S. Vetter, and D. Li, “A Survey Of Architectural Approaches for Managing Embedded DRAM and Non-Volatile On-Chip Caches,” IEEE Transactions on Parallel and Distributed Systems, vol. 26, no. 6, pp. 1524–1537, 2014.

[6] S. Aga et al, “Co-ML: A Case for Collaborative ML Acceleration using Near-Data Processing,” in Proceedings of the International Symposium on Memory Systems, pp. 506–517, 2019

[7] F. Hamzaoglu et al., "A 1 Gb 2 GHz 128 GB/s Bandwidth Embedded DRAM in 22 nm Tri-Gate CMOS Technology," IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 150-157, Jan. 2015

[8] G. Fredeman et.al “A 14nm 1.1Mb Embedded DRAM Macro with 1ns Access,” IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 230–239, 2015.

[9] Y.-P. Fang, B. Vaidyanathan and A. S. Oates, "Soft error rate cross-technology prediction on embedded DRAM," IEEE International Reliability Physics Symposium, pp. 925-928, 2009

[10] N. Kurd et al., "Haswell: A Family of IA 22 nm Processors," IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 49-58, Jan. 2015

[11] T. Atsumi et al “DRAM Using Crystalline Oxide Semiconductor for Access Transistors and Not Requiring Refresh for More Than Ten Days,” in 2012 4th IEEE International Memory Workshop, pp. 1–4, IEEE, 2012.

[12] Oota, Masashi, et al. "3D-Stacked CAAC-In-Ga-Zn Oxide FETs with Gate Length of 72nm." 2019 IEEE International Electron Devices Meeting (IEDM). IEEE, 2019.

[13] Y. Kurokawa et al. "CAAC-IGZO FET/Si-FET hybrid structured analog multiplier and vector-by-matrix multiplier for neural network." Japanese Journal of Applied Physics 59.SG (2020): SGGB03.

[14] A.Belmonte et al.”Capacitor-less, long retention(>400s) DRAM Cell paving the way towards low-power and high-density monolithic 3D DRAM”, 2020 IEEE International Electron Devices Meeting(IEDM),2020

[15] Chauhan, Yogesh Singh, et al. FinFET modeling for IC simulation and design: using the BSIM-CMG standard. Academic Press, 2015.

[16] M. Murakami, K. Kato, K. Inada, T. Matsuzaki, Y. Takahashi and S. Yamazaki, "Theoretical examination on a significantly low off-state current of a transistor using crystalline In-Ga-Zn-oxide," 19th International Workshop on Active-Matrix Flatpanel Displays and Devices, pp. 171-174, 2012

[17] T. Furuyama et al., “An experimental 2-bitcell storage DRAM for macrocell or memory-on-logic application,” IEEE J. Solid-State Circuits, vol. 24. pp. 388-393, Apr. 1989.

[18] P. Meinerzhagen, A. Teman, R. Giterman, N. Edri, A. Burg, and A. Fish, “Gain-Cell Embedded DRAMs for Low-Power VLSI Systems-on-Chip”, Springer book, ISBN 978-3-319-60402-2

[19] S.Xie,et al, ”eDRAM-CIM: Compute-in-memory design with reconfigurable embedded-dynamic-memory-array realizing adaptive data converters and charge-domain computing”, IEEE International Solid-State circuits conferernce, 2021

[20] [Online] https://www.tensorflow.org