Received 30 March 2022; revised 23 May 2022 and 7 June 2022; accepted 28 June 2022.
Date of publication 4 July 2022; date of current version 26 September 2022.

*Digital Object Identifier 10.1109/JXCDC.2022.3188366*

# IGZO CIM: Enabling In-Memory Computations Using Multilevel Capacitorless Indium–Gallium–Zinc–Oxide-Based Embedded DRAM Technology

**SIDDHARTHA RAMAN SUNDARA RAMAN, SHANSHAN XIE (Graduate Student Member, IEEE ), and JAYDEEP P. KULKARNI (Senior Member, IEEE)**

Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA

CORRESPONDING AUTHOR: S. R. SUNDARA RAMAN (s.siddhartharaman@utexas.edu)

This work was supported by The University of Texas at Austin.

**ABSTRACT** Compute-in-memory (CIM) is a promising approach for efficiently performing data-centric computing (such as neural network computations). Among the multiple semiconductor memory technologies, embedded DRAM (eDRAM), which integrates the DRAM bit cell with high-performance logic transistors, can enable efficient CIM designs. However, the silicon-based eDRAM technology suffers from poor retention time-incurring significant refresh power overhead. However, eDRAM using back-end-of-line (BEOL) integrated *C*-axis aligned crystalline (CAAC) indium–gallium–zinc–oxide (IGZO) transistors, exhibiting extreme low leakage, is a promising memory technology with lower refresh power overhead. A long retention time in IGZO eDRAM can enable multilevel cell functionality, which can improve its efficacy in CIM applications. In this article, we explore a capacitorless IGZO eDRAM-based multilevel cell, capable of storing 1.5 bits/cell for CIM designs focused on deep neural network (DNN) inference applications. We perform a detailed design space exploration of IGZO eDRAM sensitivity to process temperature variations for read, write, and retention operations followed by architecture-level simulations comparing performance and energy for different workloads. The effectiveness of IGZO eDRAM-based CIM architecture is evaluated using a representative neural network, and the proposed approach achieves 82% Top-1 inference accuracy for the CIFAR-10 dataset, compared with 87% software accuracy with high bit cell storage density.

**INDEX TERMS** Compute-in-memory (CIM), embedded DRAM (eDRAM), indium–gallium–zinc–oxide (IGZO), leakage, multilevel cell, read, write.

## I. INTRODUCTION

The development of neural networks has resulted in an unprecedented increase in the size of deep neural networks (DNNs). These emerging large size DNN models cannot fit within the limited on-chip memory even in the latest server CPUs [1], GPUs [2], and specialized machine learning (ML) accelerators, such as Graphcore [3]. This necessitates a massive amount of data movement from off-chip memory to on-chip computer cores in modern ML accelerators, resulting in increased energy for computation. Thus, it is important to explore technologies and algorithms that maximize capacity and further reduce the data movement for performing multiply and accumulate operations (MACs) in case of ML workloads. On the algorithms front, different low resolution networks have been proposed to reduce the data movement energy and computation cost. One such example is the usage of binary/ternary neural networks that make use of binary$+1$, $-1$/ternary$+1$, 0, $-1$ weights and activations to perform MAC operations, resulting in the reduction of data movement. These networks approximate the dot product as a simple AND gate (binary) or a combination of AND and XOR gate (ternary), further resulting in the reduction of computation energy.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

On the memory technologies front, compute-in-memory (CIM) designs performing analog computations for MACs operations for DNN applications can mitigate the "memory wall" bottleneck of latency and energy, leading to energy efficient designs. It is important to explore technologies that offer dense bit cell to store weights and to perform energy efficient dot product compute for large-scale CIM applications. The 6T SRAMs offer high performance due to its low access latencies. However, compute-in-SRAMs are limited by the bit cell variations, causing inaccurate computations and degrading CIM accuracy [4]. To overcome this issue, 8T SRAMs with decoupled read or write ports have been proposed, but they degrade the compute array density [5]. Nonvolatile memories making use of resistive random access memory (RRAMs) [6], [7], PCMs, ferroelectric field effect transistor (FeFET), and flash have been explored for they offer high densities and offer zero standby leakage. Compute in these devices primarily rely on the principle of bitline current summation for realizing dot product between input activations and weights stored onto the bit cell. However, they are susceptible to process variations, such as conductance in RRAM, limited write endurance ($10^6$), and lower write speeds (tens of ns) [8] degrading the performance further and making it difficult to be utilized in high-speed accelerators, which require frequent updates. Commodity DRAM is realized using a one transistor one capacitor (1T1C) bit cell structure, optimized for bit cell density and process cost, and supports limited number of metal layers. In general, commodity DRAM process technology is different from the logic transistor technology. Embedded DRAM (eDRAM), on the other hand, is generally implemented using the same logic transistor technology having access to high-performance logic transistors and interconnects. However, eDRAM offers smaller bit cell storage node (SN) capacitance compared with the commodity DRAM and, hence, higher refresh power overheads. However, for data intensive applications (such as DNN), eDRAM being monolithically integrate with logic blocks can mitigate the off-chip data movement energy and latency costs due to dedicated commodity DRAM accesses. Furthermore, they offer high bandwidth (>100 GB/s), high clock speed (2.6 GHz), high endurance $10^{16}$, and low pJ/bit ($\ll$1 pJ/bit), thus having the desired attributes of a CIM design [9], [10]. However, the major limitation from eDRAM becoming a potentially strong candidate for CIM applications is the leakage of the bit cell capacitor, requiring frequent refreshes. Indium–gallium–zinc–oxide (IGZO)-based eDRAM, on the contrary, make use of extreme low leakage access transistors to reduce the leakage of eDRAM bit cell.

To address the abovementioned issue, a promising technology that offers extreme low leakage is the usage of $C$-axis aligned crystalline IGZO (CAAC-IGZO) transistor [11], wherein the CAAC-IGZO transistor can be utilized as an access transistor to realize increased retention time of eDRAMs. These transistors can further be integrated back end of line (BEOL) with capacity of 3-D stacking, thus enabling high-performance CIM applications. The low leakage of IGZO-based eDRAM can be utilized for multilevel storage, thus improving the density of the array further. Sundara Raman *et al.* [12] explored the possibility of using eDRAM for storage of 2 bits per eDRAM bit cell, by utilizing a three transistor one capacitor (3T1C)-based IGZO eDRAM. However, the storage bit cell capacitor limits the amount of 3-D stacking in the IGZO-based capacitor. In this article, we explore the usage of capacitorless IGZO-based eDRAM for storing three levels (1.5 bits/cell). The capacitorless IGZO-based eDRAM can offer higher array density benefits as opposed to the 3T1C eDRAM-based bit cell because of 3-D stacking without loss in storage density, thus making it opportune for high-performance, high-density CIM designs. The IGZO eDRAM bit cells with dedicated read port have been explored before [1]. However, prior approaches [13], [14] consider only 1-bit/cell CIM design. However, this work explores the possibility of MLC in IGZO eDRAM and efficiently mapping ternary weights in a neural network for CIM applications considering device-circuit-architecture-level analysis. This article is organized as follows. Section II provides the case for the CAAC-IGZO eDRAM leakage mechanism, advantages of IGZO in terms of retention time. Section III describes modeling of the device. Section IV discusses the different bit cell topologies. Section V analyses the read/write timing diagram for capacitorless IGZO eDRAM bit cell. Section VI analyses the variability study for the bit cell in terms of the SN voltage for write operation. Section VII analyses the read variability study in terms of voltage at read bitline (RBL). Section VIII validates the MLC potential by studying CIM design that is capable of producing accurate MACs in case of ternary neural networks. Section IX presents the architecture-level simulations for different benchmarks and understands the trade-off between energy and latency for IGZO eDRAM over Si eDRAM. Sections X–XII present the analysis of CIFAR-10 results on a custom CNN. Section XIII concludes the key analysis results and observations from this work on IGZO-based eDRAM.

## II. CAAC-IGZO eDRAM STUDY

The CAAC-IGZO transistors are typically realized as N-type devices having a moderate on-current and are suitable for low-temperature BEOL CMOS integration. This allows for increasing the bit cell density by stacking multiple layers of IGZO access transistors and backend capacitors in a 3-D fashion. In addition to 3-D integration, IGZO-based eDRAMs can increase bit density by storing multiple bits per cell, owing to the extremely low leakage characteristics of IGZO devices, with high sense margins for resolving between multiple storage capacitor voltage levels. The SN of eDRAMs and DRAMs starts leaking 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, as shown in Fig. 1. The extent of this leakage defines the refresh times of such memories. One mechanism of reducing the subthreshold leakage (exponentially dependent on the transistor gate to source voltage) is the use negative word line voltages in the off-state. However, this negative voltage increases the electric field at the gate–drain overlap region, which leads to an increase in GIDL. The higher energy bandgap ($E_g$), higher effective mass of electron ($m_0$), and higher relative permittivity ($E_r$) in IGZO as compared with Si are the primary driving factors for reduced GIDL. This increases the retention time to more than ten days in IGZO-based eDRAMs [16]. Furthermore, low leakage enables successful retention of the bit cell contents for a longer time and, hence, enables reliable read (enough bitline

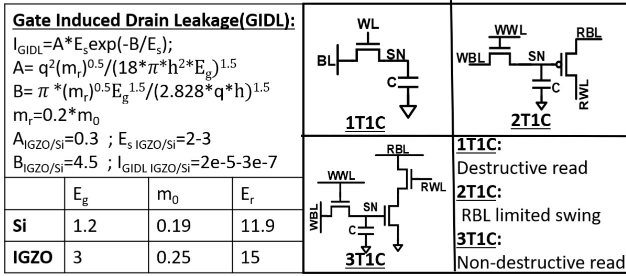



**FIGURE 1. Left: CAAC-IGZO transistor exhibiting extremely low leakage. High effective mass, high energy bandgap, and high relative permittivity for IGZO lower leakage significantly compared with the silicon-based eDRAM. Right: different eDRAM bit cell configurations with the properties of different bit cells.**

differential for successfully differentiating between different levels) after a long retention time, even in the presence of process variations.

## III. BIT CELL STUDY

Various gain-cell topologies employing dedicated read port transistors (e.g., 2T1C, 3T1C, 4T1C, and 5T1C gain cells) have been proposed to eliminate the issue of destructive read observed in the widely used 1T1C eDRAM bit cell. Despite being the most area efficient gain-cell topology, 2T1C is not the ideal choice for MLC CIM applications due to the threshold clipped voltage swing on the RBL. The 3T1C gain-cell topology, on the other hand, exhibits a full-rail voltage swing at the RBL. This helps improve the resolution of multiple data levels stored on the bit cell capacitance by monitoring the extent of RBL discharge. The 3-D integration of DRAM bit cell to improve the performance is an essential requirement for the performance of large-scale ML workloads. The scalability is severely limited by the requirement of bit cell storage capacitor [17]. Thus, the capacitorless IGZO-based eDRAM alleviates this problem by utilizing the gate capacitance of the read port transistor as the storage capacitor.

## IV. DEVICE MODELING

A compact device model for an n-type CAAC-IGZO transistor is developed using the experimentally demonstrated CAAC-IGZO of gate length 72 nm [11]. Fig. 2 shows the calibration of the model parameters with experimental data. Modeling is performed by empirically calibrating the Log(ID) versus $V_{gs}$ (gate–source voltage) characteristics corresponding to the first layer of the 3-D stack with different body bias voltages of 0, −1.5, and −3 V. The different body bias voltages translate to different threshold voltages of the transistor, by modulating the channel charge. The experimental results are demonstrated for $V_{ds}$ = 1.2 V, and the characteristics are scaled to model $V_{ds}$ = 1.3 V for SPICE simulations. The compact SPICE performing that calibrates the device characteristics for performing circuit simulations is modeled using BSIMIMG (102.9.2) [18]. BSIMIMG-based models have been used so as to efficiently capture the effect of body bias voltage on the threshold voltage. The experimental model is calibrated in such a way that it exhibits similar $I_{on}$ and $I_{off}$ characteristics as in [11]. The off

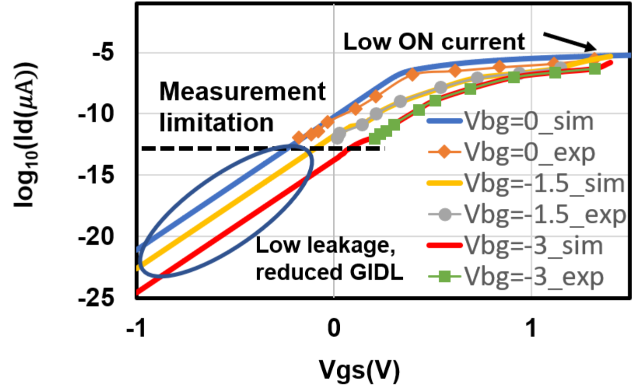
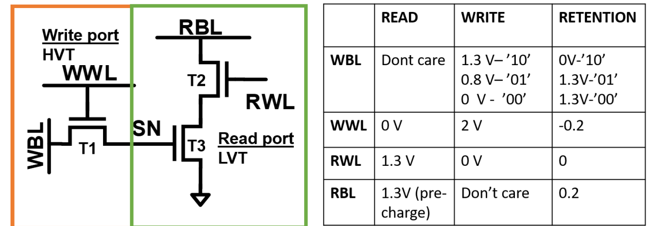



**FIGURE 2. 72-nm CAAC-IGZO transistor experimental device cross section and Log(ID) versus $V_{gs}$ (gate–source voltage) characteristics calibrated with a BSIMIMG model [11].**

characteristics of low-leakage IGZO-based eDRAMs have been obtained by carefully optimizing the device parameters, such as doping of the channel (NBODY), mobility temperature coefficient (UTL), and nonuniform doping in the lateral direction ($K0$). These parameters enable tuning of the subthreshold slope and the off-state current. on-current, which is typically in the range of $\mu$A and lower than the on-current of the Si-based transistors, is modeled with a decreased mobility value using the low field mobility coefficient parameter ($U0$).




| | READ | WRITE | RETENTION |
|---|---|---|---|
| WBL | Dont care | 1.3 V– '10'<br>0.8 V– '01'<br>0 V - '00' | 0V-'10'<br>1.3V-'01'<br>1.3V-'00' |
| WWL | 0 V | 2 V | -0.2 |
| RWL | 1.3 V | 0 V | 0 |
| RBL | 1.3V (pre-charge) | Don't care | 0.2 |

**FIGURE 3. Capacitorless IGZO eDRAM bit cell. SN is the bit cell storage node. WBL/RBL is write/read bitline. Write port transistor (orange) is of higher threshold voltage than read port transistor (green). RWL is read word line with the voltages required for read, write, and retention.**

## V. CAPACITORLESS IGZO eDRAM STUDY

Fig. 3 shows a three-transistor capacitorless eDRAM structure that has been used for circuit simulations for multilevel storage. A write port transistor ($T1$), marked in orange, has been optimized with a higher threshold voltage ($V_t$) using a larger body bias voltage so as to reduce the leakage of the bit cell. The read port transistors ($T2$ and $T3$) use low threshold voltage ($V_t$), so that the read time is optimized effectively and to have a wider bit cell swing on the RBL, thus enabling better sense margin. This allows storage of multiple levels in the same bit cell.

Fig. 4 shows the timing diagram for the capacitorless IGZO-based eDRAM configuration. Fig. 4 illustrates that the read operation is performed using the read port transistors ($T2$ and $T3$) by turning on read word line (RWL), and the write operation is performed using a write port transistor ($T1$) write word line (WWL). Both these word lines have a

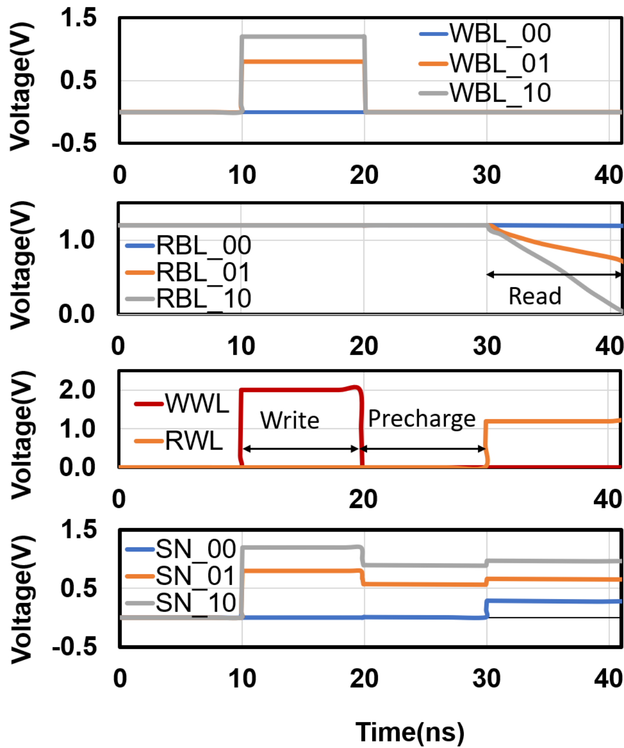



**FIGURE 4.** Simulation results for read and write operations for capacitorless IGZO-based MLC eDRAM. SN is the bit cell storage node. WBL/RBL is write/read bitline. RWL is read word line.

10-ns (0.1 GHz) pulsewidth. During write, WWL is turned on, and the voltages at write bitline (WBL) are chosen (shown in Fig. 3) to enable writing of three data levels, i.e., "00," "01," and "10." The extreme low leakage obtained using a high-$V_t$ IGZO-based transistor ($T1$) facilitates storing the bit cell SN values for extremely longer time. The read operation is performed in two steps. The first step involves precharge of the RBL to 1.3 V, followed by the evaluation step. During the evaluation step, RWL turns on, and the discharge rate of RBL determines the voltage stored on the gate of $T3$. The read bitline capacitance is assumed to be 30 fF for performing the read simulations. The overdrive voltage of $T3$ determines the effective discharge rate of RBL. The discharge rate of storing "10" discharges the $T3$ transistor faster than storing "00" and "01." Furthermore, the voltage at the SN node increases instantaneously due to the coupling effect between RWL and the intermediate node between $T2$ and $T3$. This action further causes coupling onto SN from the intermediate node, thus increasing the SN voltage. The three different data levels that are stored in capacitorless IGZO eDRAM bit cell are differentiated using a flash analog–digital conversion (ADC) based on the voltage at the RBL after a predefined time. It is important to note that the full swing of RBL enables storage of multiple levels and reliable readout, as compared with the 2T1C gain-cell structure. The capacitorless IGZO-based eDRAM enables realizing computations inside memory with increased array density as compared with 3T1C because of the absence of bit cell capacitor.

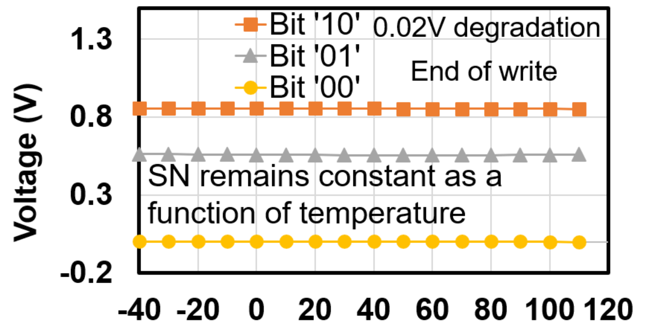



**FIGURE 5.** Difference between SN voltage at the end of write obtained for different levels remains constant with temperature.

## VI. WRITE ANALYSIS

Write stability is measured in terms of SN voltage at the end of write. The voltage at the SN modulates the effective resistance of the read port transistors, which, in turn, affects the read stability as well. Thus, it is important to note that voltage at the SN at the end of write should be large enough to differentiate between different levels during read.

The robustness of the design to write needs to be tested at 110 C, which captures the effect of worst case condition for SNs to leak and not retain the SN values. The SN values are robust to temperature variations, ranging from −40 to 110 C, as shown in Fig. 5. The average value of the SN at the end of write is identified by averaging across $10^5$ runs. Thus, the effects of process and temperature variations were captured in this analysis.

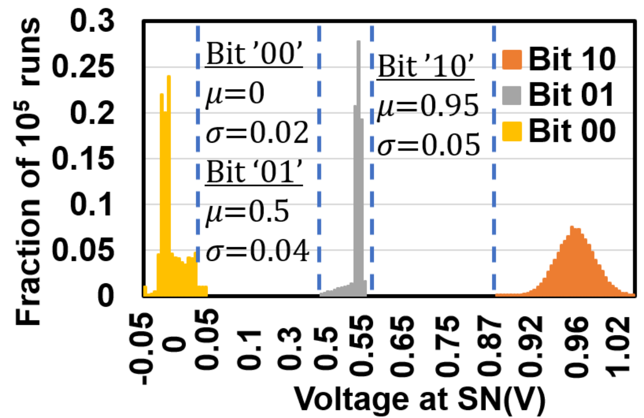



**FIGURE 6.** Histogram showing the SN values at the end of write cycle.

Fig. 6 suggests the histogram of SN at the end of write and is maximum for storing "11" even in the presence of process variations. The capacitive coupling between the voltage at the WWL and the SN further aids in the decrease of SN value at the end of write. The abovementioned analysis suggests that the voltage at SN is fairly resilient to temperature variations. Fig. 6 depicts the range of SN values for different levels in the presence of process variations at the end of a write cycle. This analysis captures the effect of capacitive coupling at the end of a write cycle to understand the separation between different levels stored in the bit cell. There are $10^5$ Monte

Carlo simulations performed with 1$\sigma$ $V_t$ of 30 mV to capture the effect of process variations. The mean and sigma values for different levels are depicted in Fig. 6. Thus, the voltage at the SN is separated by 0.5 V across different levels even in the presence of process variations.

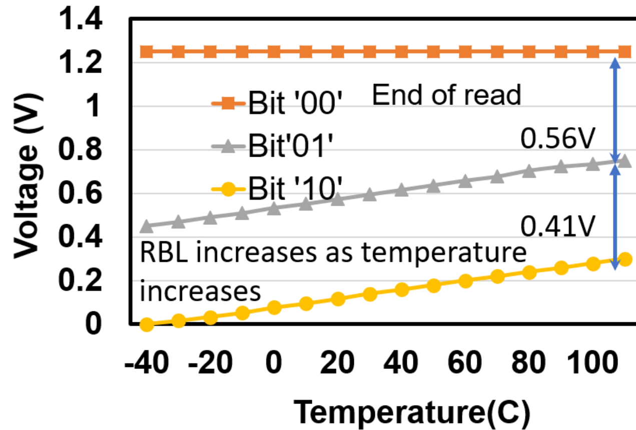



FIGURE 7. Difference between RBL voltage average obtained at the end of read for different levels (with temperature variation) remains sufficient for reliable read.

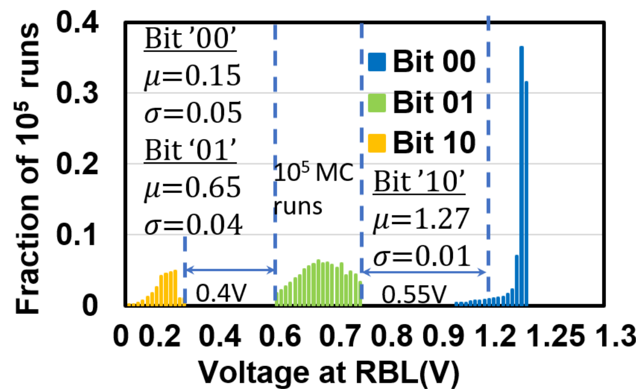



FIGURE 8. Histogram showing the RBL values at the end of read cycle when read is performed immediately after write.

## VII. READ ANALYSIS

Read stability as a function of temperature is measured in terms of the voltage at the RBL at the end of read. The voltage difference should be sufficiently large enough to read different levels accurately in the presence of process and temperature variations. RBL voltage at the end of read (with worst case scenario of process variation) is shown in Fig. 7. As the temperature increases from −40 to 110 C, the devices tend to be slower because of lowered mobility of the devices, and this leads to a lower voltage difference across different levels. The effect of mobility decrease compensates the effect of reduced threshold voltage at higher temperature. This leads to lesser discharge rates of the RBL node, leading to decrease in the voltage difference between levels. However, it is important to note that, even at increased temperatures, 0.4–0.5-V difference is observed across different levels. To further study the robustness of the design to process variations, the analysis of RBL voltage at the end of read cycle at 110 C was performed. The $10^5$ Monte Carlo simulations assuming 1$\sigma$ $V_t$ of 30 mV for the read port and access transistors are used. This analysis is performed with a read operation performed immediately after a write operation (i.e., with lesser amount of retention/standby time). This result captures the effect of capacitive coupling degrading the SN postwrite and the increase in SN value during read. In the case of reading "10," the voltage at RBL is close to 0 V, the voltage at RBL for "01" is close to 0.65 V, and the RBL for "10" is close to 1.3 V. Thus, there is a difference of 0.4 V in RBL voltage at the end of read between "10" and "01" and 0.55 V between "01" and "00" in worst case scenario. This suggests that there is sufficient difference in the voltages for reading the levels "00," "01," and "10" efficiently, as shown in Fig. 8.

Fig. 9 demonstrates the effectiveness of the bit cell to be able to read multiple levels efficiently. This is performed assuming the read is performed long (10 s) after it has been written. This exploration captures the effect of the capacitive coupling degrading the storage node observed postwrite, SN degrading because of leakage and the slight increase in SN observed during read. Histograms corresponding to levels "10" and "01" are shifted to the right in contrast to Fig. 8, while "00" histogram is shifted to the left. This is because the SN while storing 0 V increases over a period of time. This analysis explains the feasibility of successfully differentiating between different bit cell contents.

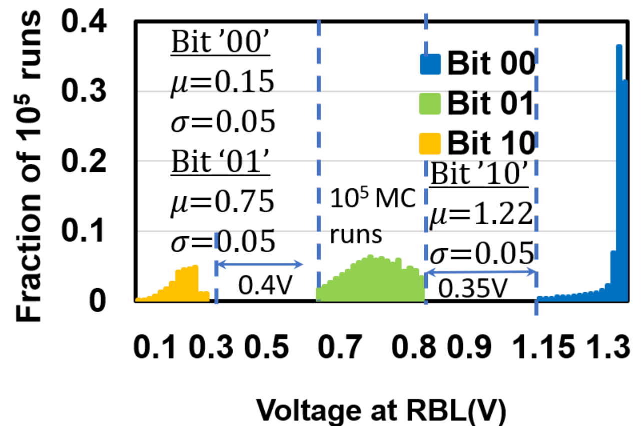



FIGURE 9. Histogram showing the RBL values at the end of read cycle when read is performed long after write, with SN voltage degradation.

## VIII. RETENTION ANALYSIS

Fig. 10 shows the degradation of SN voltage as a function of time in seconds. This analysis is performed once a write operation is performed on the bit cell and left unaccessed for a long time. The leakage current is a strong function of the voltage difference between WBL and SN. Thus, the voltage on the WBL is conditioned to capture the worst case leakage for each of the levels. The initial voltage at SN is assumed to be the voltage after the capacitive coupling onto the SN node from WWL. WBL is conditioned to 1.3 V, when the bit cell is programmed to "00" or "01," and WBL is conditioned to 0 V when the bit cell is programmed to "10." The degradation of SN is least for "01," because the voltage difference between SN and WBL is 0.8 V as compared with 1.3 V for "00" or

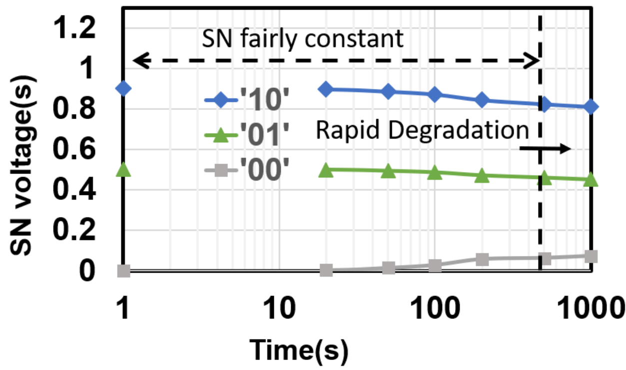



FIGURE 10. Waveform showing SN degradation as a function of time.

| Parameter (#cycles) | IGZO eDRAM | Si eDRAM |
|---|---|---|
| tRCD | 6 | 4 |
| tCL | 5 | 5 |
| tWL | 4 | 4 |
| tWTR | 3 | 3 |
| tRTP | 3 | 3 |
| tRP | 6 | 4 |
| **Energy corresponding to different operations** | | |
| Operation | Energy | |
| Write | '10' – 0.11pJ; '01' - 0.09pJ; '00' - 0.03pJ | |
| Read | '00' – 0.02pJ ; '01' – 0.04pJ ; '10' – 0.11pJ | |

FIGURE 11. Comparison between different simulation parameters for the IGZO eDRAM and DRAM followed by energy per bit for different operations of IGZO eDRAM bit cell.

"01." Typically, a negative voltage is applied to the WWLs to reduce leakage. For the high $V_t$ transistor $T1$, with a back bias of $-0.3$ V, leakage current of $10^{-18}$ A is observed at a gate voltage of $-0.25$ V. This is the predominant leakage current component for the bit cell.

The retention time reported is for the worst case scenario of bit cell storing "10." Retention time is quantified as the time required for the SN voltage to drop by 0.1 V post the write process for "10" and is observed to be approximately 1000 s. The voltage at the SN storing "00" increases post the write process, because the WBL charges the SN, and the charging process happens at a similar rate as compared to the discharge of "10."

## IX. ARCHITECTURE STUDY

This section understands the architectural-level details for performing CIM and for performing conventional workloads by assuming that the compute can potentially be performed in the IGZO based eDRAM array, without going into the details of the way compute is performed in eDRAM array (discussed in Section X). Furthermore, the section analyses the trade-off between performance and energy of IGZO-based eDRAM over Si-based DRAM in terms of performance and energy for different benchmarks using Ramulator [19], a cycle-accurate simulator capable of modeling different DRAM standards ranging from DDR3, DDR4, and LPDDR, using different memory technologies, such as DRAM, PCM, and spin transfer torque magnetoresistive random access memory (STT-MRAM). Ramulator has been carefully calibrated for analyzing the performance parameters that are of interest, with eDRAMs being utilized for $L3$ cache for non-CIM workloads. Different parameters of IGZO-based eDRAM and Si-based eDRAM parameters used for simulation are mentioned in Fig. 11. Si-based eDRAM parameters are obtained after scaling the DRAM parameters specified by Micrometer in [20]. The eDRAM memory organization is assumed to be consisting of a hierarchy of channel, modules, rank, chip, and bank, and that each channel is responsible for data transfer between the DRAM chips and the memory controller as part of the CPU core. Each bank consists of an array of memory cells, and a row of sense amplifiers are called row buffers. CIM workloads have been simulated using the benchmarks listed in [21]. The architecture is assumed to be made of one channel, one rank, and $4 \times 16$ chips per rank for simulation in case of 2-D IGZO-based eDRAM. The different commands that are part of the DRAM interface are activation (turning ON DRAM) for write/read, precharge command, refresh, and row buffer read; 1.5-bit storage in IGZO-based eDRAM is modulated by assuming that the size of IGZO-based eDRAM is 1.5 times the size of DRAM. The different timing/delay parameters used are as follows.

1) tRCD is a measure of read time and captures the effect of a bit cell read and row buffer read out. Assuming 400-MHz clock, the read latency is roughly five cycles (as shown in timing diagram) for IGZO-based eDRAM as compared with four cycles in eDRAM.
2) tCL is the time taken to read the data once it is latched onto the row buffer. tWL and tWTR impose constraints on the successive commands of the row buffer and are independent of the underlying memory technology [22]. tRTP is a measure of the data stability in the cross coupled inverters that feed into write drivers of the memory array and are independent of memory technology.
3) tRP is an indication of time taken between a successful precharge and completion of activate command for write in the same bank. Thus, it is a measure of the write latency of memory array [22] and is roughly four cycles at 400-MHz clock.

tRCD and tRP in case of IGZO-based eDRAMs are higher, because the on-current is lower in contrast to IGZO-based eDRAMs. Furthermore, a refresh time of 300 $\mu$s has been assumed for eDRAMs, and a refresh command has been utilized every 300 $\mu$s, which leads to of performance degradation as this is a dead cycle from a memory cycle point of view. The abovementioned delay parameters are used for calibrating DRAM for simulating IGZO-based eDRAM. A CPU trace-driven approach, where instructions are directly read from the proposed benchmarks and simulates a simplified CPU core model that performs nonmemory instructions and accesses memory for load store instructions, is used for running these benchmarks. Few of the CPU SPEC2006 benchmarks have been chosen to simulate to capture the effect of performance. Fig. 12 captures the effect of the IPC

of IGZO-based eDRAM normalized to Si-based eDRAM for both non-CIM workloads and CIM workloads. The system architecture for non-CIM workloads makes use of $L3$ cache made of IGZO-based eDRAM and gets accessed in case of cache misses. In case of CIM workloads, the architecture involves just a processor interacting with eDRAM compute array. The number of cache misses in hmmer and bzip2 are relatively lower in contrast to mcf, lbm, and astar, which are memory intensive workloads with larger $L1/L2$ cache misses. Thus, the cache misses lead to larger number of DRAM accesses (in this case, $L3$ cache), and because Si-based eDRAM has better read and write latency, there is an increase in the number of instructions processed by the core. In case of CIM workloads, the instructions per cycle (IPC) is less as compared with Si eDRAM due to lesser on-currents and increased read access latency.

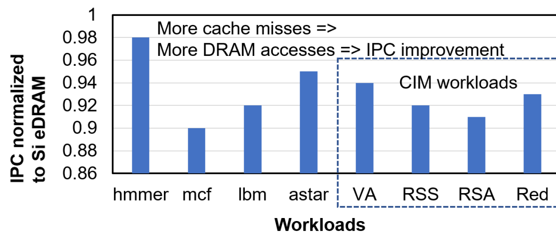



**FIGURE 12. IPC normalized to Si eDRAM for different non-CIM and CIM workloads. More cache misses lead to increased eDRAM accesses. Si eDRAM has better IPC because of faster read accesses in both these types of workloads.**

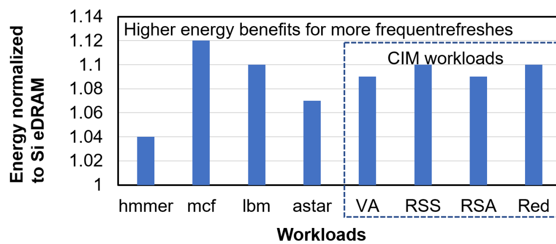



**FIGURE 13. Energy normalized to Si eDRAM for different non-CIM and CIM workloads. IGZO eDRAM has better IPC because of lesser refresh power in both these types of workloads.**

The energy analysis includes the energy for WBL, WWL, and RBL based on the operation performed. The write energy is maximum for storing bit "10" because of the large voltage applied at WBL for storing bit "10" in comparison with the other. Read energy for bit "10" is highest because of the greater RBL swing in contrast to the levels. Fig. 13 suggests the energy levels for different levels. At an array level, the energy benefits of IGZO-based eDRAM come primarily because of refresh every 300 $\mu$s in case of Si-based eDRAM, and there are less frequent refresh commands required in case of IGZO-based eDRAM due to its extreme low leakage. Thus, IGZO-based eDRAM is energy efficient at the expense of performance in case of CIM designs.

## X. COMPUTE-IN-MEMORY

This section describes the usage of the capacitorless IGZO-based eDRAM for low-resolution neural networks.

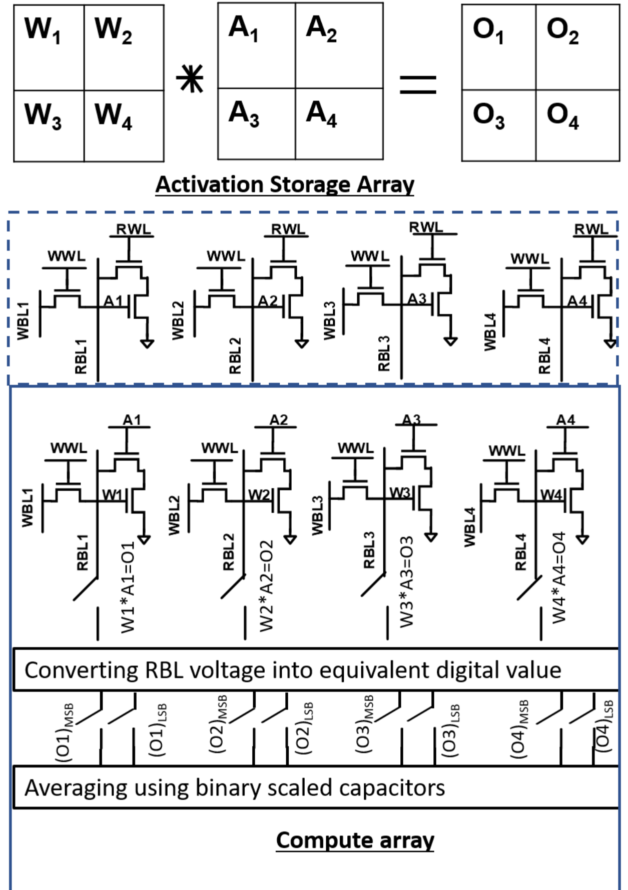



**FIGURE 14. CIM architecture using capacitorless IGZO eDRAM array showing weight and activation storage array and compute array.**

Ternary neural networks are an example of a low-resolution neural network where the weights and activations can take the value +1, 0, −1. A block diagram highlighting the features of the proposed CIM architecture is shown in Fig. 14. The proposed IGZO eDRAM can be efficiently used for ternary neural networks because of the following.

1) In conventional CIM designs, activations are stored in a separate storage array, and the activations are transferred to the compute array where the weights are stationary to perform computations. In case of conventional TNN CIM designs, with weights and activations encoded as 2 bits, 2-bit cells are needed to store the activations/weights or perform computation. This proposal utilizes a single bit cell for storing the activations, and the computation can be performed in a single bit cell.
1) Multiplication in case of ternary neural networks is usually realized by a combination of AND and XOR gate [23]. This proposal takes advantage of the multilevel storage in eDRAM to perform the computation to get rid of additional XOR gate.

The proposed CIM architecture operates on 1.5-bit wide input activations and weights and is efficient in terms of storage and compute. The data flow for performing CIM operation is described as follows.

*1) Mapping Phase:* The 1.5-bit weights/activations (0, −1, +1) are encoded as "00," "10," and "11." The weights

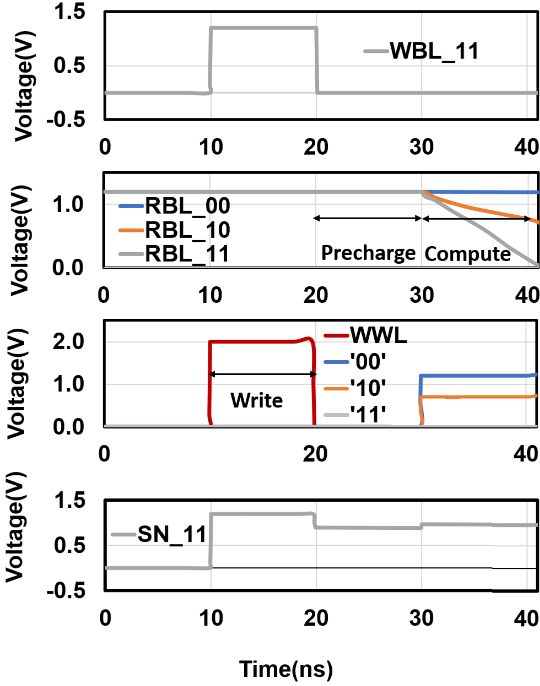



FIGURE 15. Timing diagram for performing CIM using capacitorless IGZO eDRAM.

are assumed to be stationary, located in the compute array. The activations typically are moved from activation storage array into compute array for activation. Fig. 14 shows the convolution operation performed between 2 × 2 filters and input activations. Each of the weights is stored in a single row, and the activations are mapped onto the RWL in a serial fashion. The RWL voltage is modulated to realize three activation levels, as shown in Fig. 15.

*2) Dot Product Compute Phase:* The 1.5-bit weight initially is written onto the compute array by turning on the WWL and storing onto SN. The write cycle in Fig. 15 is indicative of the weight storage onto the bit cell. During the next cycle, RBL is precharged to 1.3 V, marked as precharge. In the compute cycle, activations are mapped onto the RWL, and the discharge rate of RBL is different for different RWL voltages. Fig. 15 shows that for an activation of “11,” RBL discharges completely in contrast to activation of “10” where the bitline is partially discharged. In case of activation of “00,” the weights are not discharged. Thus, reading RBL at the end of a predetermined compute time is a characteristic measure of the dot product between the activations and the weights.

*3) Accumulation Phase:* The dot product is converted into an equivalent 1.5-bit value, as shown in Fig. 14, and the activations and then charge shared using binary-scaled capacitors to realize the accumulation operation

## XI. PROCESS VARIATION STUDY

It is important to note that the discharge rate of RBL is subject to process variations. Hence, a detailed study was conducted to understand the effect of process variations on dot product computations. The experiment setup involved running $10^5$ Monte Carlo simulations with $1\sigma$ $V_t$ of 30 mV. The analysis suggested a trend similar to that of the RBL graph shown in Fig. 8. Furthermore, the design was also tested across different temperatures to realize the temperature impact on the dot product computation. The design was extremely resilient to temperature variations, as observed in the read analysis.

## XII. CIFAR-10 RESULTS

The proposed multilevel cell IGZO-based eDRAMs CIM design efficiency is quantified using CIFAR-10 dataset with a representative convolutional neural network that has been trained for effectively utilizing 1.5-bit weights and activations as shown. The network has four convolutional layers, with the first and second layers containing 32 channels each of size 32 × 32 and the third and fourth layers containing 64 channels each of size 32 × 32. A 3 × 3 kernel has been used. The proposed design achieves 82% Top-1 classification accuracy, compared with the 87% accuracy obtained from ideal software implementation for the same network. The difference in accuracy stems from quantization loss. The design specifications of the proposed CIM design are specified in Table 1. This analysis indicates that the CAAC-IZGO eDRAM can be a promising candidate for performing large scale, ternary CIM designs with good accuracy.

TABLE 1. IGZO eDRAM based CIM parameters.

| Design parameters | Description |
|---|---|
| Neural network configuration | CONV layer - 32 3*3; ReLU CONV layer - 32 3*3; ReLU Max pooling layer - 2*2 CONV layer - 64 3*3; ReLU CONV layer - 64 3*3; ReLU Max pooling layer - 2*2 FC layer - 512-BN-ReLU FC layer-10 |
| IGZO eDRAM based CIM parameters | Input activations - 1.5 bit Weights - 1.5 bit, Bitcell - Capacitorless IGZO eDRAM, MLC - 1.5bits/cell |
| Top-1 classification accuracy for CIFAR-10 | Software = 87% IGZO eDRAM= 82% |

## XIII. CONCLUSION

In this article, we make use of the extreme low leakage CAAC-IGZO-based eDRAM to perform CIM operation for ternary neural networks. The low leakage, high retention time of IGZO can be leveraged to enable multilevel cell functionality, which further increases the storage density. We present a detailed study involving comparison between leakage of IGZO and Si-based eDRAM, different available topologies, and shortcomings of each of the topologies. We utilize the capacitorless IGZO-based eDRAM for storing 1.5 bits/cell. Architecture-level simulations comparing IPC and energy between DRAM and IGZO-based eDRAM is also presented. The feasibility of this proposal has been qualified by performing Monte Carlo simulations for read, write, and retention. Monte Carlo simulations suggest that the multilevel bit cell is not prone to bit cell variations and offers retention time of 1000 s for the given modeled device. The storage of 1.5 bits/cell allows efficient mapping of ternary weights onto a single bit cell. Overall architecture of the compute array along with charge share for performing dot product compute has been presented. A validation of this approach is obtained by performing compute for a custom neural network on a CIFAR-10 dataset with the compute array showing good accuracy. The susceptibility of CIM design to process variations is investigated and detailed in the process variations section.

## REFERENCES


[1] *Intel Xeon Platinum 8380 HL Processor*, Intel, Santa Clara, CA, USA, 1999.

[2] *NVIDIA Tesla V100 GPU Architecture*, NVIDIA.

[3] Z. Jia, B. Tillman, M. Maggioni, and D. P. Scarpazza, "Dissecting the graphcore IPU architecture via microbenchmarking," 2019, *arXiv:1912.03413*.

[4] S. K. Gonugondla, M. Kang, and N. Shanbhag, "A 42 pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 490–492.

[5] X. Si *et al.*, "24.5 A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

[6] B. Zhang *et al.*, "90% yield production of polymer nano-memristor for in-memory computing," *Nature Commun.*, vol. 12, no. 1, pp. 1–11, Dec. 2021.

[7] S. Sagar, K. U. Mohanan, S. Cho, L. A. Majewski, and B. C. Das, "Emulation of synaptic functions with low voltage organic memtransistor for hardware oriented neuromorphic computing," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Dec. 2022.

[8] S. Mittal, J. S. Vetter, and D. Li, "A survey of architectural approaches for managing embedded DRAM and non-volatile on-chip caches," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 6, pp. 1524–1537, Jun. 2015.

[9] G. Fredeman *et al.*, "17.4 A 14 nm 1.1 Mb embedded DRAM macro with 1 ns access," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.

[10] N. Kurd *et al.*, "Haswell: A family of IA 22 nm processors," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 49–58, Jan. 2015.

[11] M. Oota *et al.*, "3D-stacked CAAC-In-Ga-Zn oxide FETs with gate length of 72 nm," in *IEDM Tech. Dig.*, Dec. 2019, pp. 2–3.

[12] S. R. S. Raman, S. Xie, and J. P. Kulkarni, "Compute-in-eDRAM with backend integrated indium gallium zinc oxide transistors," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.

[13] A. Belmonte *et al.*, "Capacitor-less, long-retention (> 400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM," in *IEDM Tech. Dig.*, Dec. 2020, pp. 2–28.

[14] A. Belmonte *et al.*, "Tailoring IGZO-TFT architecture for capacitor-less DRAM, demonstrating >$10^3$s retention, >$10^{11}$ cycles endurance and $L_G$ scalability down to 14 nm," in *IEDM Tech. Dig.*, Dec. 2021, pp. 6–10.

[15] Y. S. Chauhan *et al.*, *FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard*. New York, NY, USA: Academic, 2015.

[16] T. Atsumi *et al.*, "DRAM using crystalline oxide semiconductor for access transistors and not requiring refresh for more than ten days," in *Proc. 4th IEEE Int. Memory Workshop*, May 2012, pp. 1–4.

[17] A. Belmonte *et al.*, "Capacitor-less, long-retention (>400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM," in *IEDM Tech. Dig.*, Dec. 2020, p. 28.

[18] C. Hu *et al.*, "BSIM-IMG: A turnkey compact model for fully depleted technologies," in *Proc. IEEE Int. SOI Conf. (SOI)*, Oct. 2012, pp. 1–24.

[19] Y. Kim, W. Yan, and O. Mutlu, "Ramulator: A fast and extensible DRAM simulator," *IEEE Comput. Archit. Lett.* vol. 15, no. 1, pp. 45–49, Jan. 2016.

[20] *512Mb DDR2 SDRAM Component Data Sheet: MT47H128M4B6-25*, MICRON.

[21] J. Gómez-Luna, I. E. Hajj, I. Fernandez, C. Giannoula, G. F. Oliveira, and O. Mutlu, "Benchmarking a new paradigm: Experimental analysis and characterization of a real processing-in-memory system," *IEEE Access*, vol. 10, pp. 52565–52608, 2022, doi: 10.1109/ACCESS.2022.3174101.

[22] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable dram alternative," in *Proc. 36th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2009, pp. 2–13.

[23] X. Yang *et al.*, "An in-memory-computing charge-domain ternary CNN classifier," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2021, pp. 1–2.