A Resonant Time-Domain Compute-in-Memory (rTD-CiM) ADC-Less Architecture for MAC Operations

Dhandeep Challagundla*
vd58139@umbc.edu
University of Maryland Baltimore County
Baltimore, Maryland, USA

Ignatius Bezzam
i@rezonent.us
Rezonent Inc.
Milpitas, California

Riadul Islam
riaduli@umbc.edu
University of Maryland Baltimore County
Baltimore, Maryland, USA

ABSTRACT

In recent years, Compute-in-memory (CiM) architectures have emerged as a promising solution for deep neural network (NN) accelerators. Multiply-accumulate (MAC) is considered a *d*e facto unit operation in NNs. By leveraging the minimal data movement required and inherent parallel processing capabilities of CiM, NNs that require numerous MAC operations can be executed more efficiently. Traditional CiM architectures execute MAC operations in the analog domain, employing an Analog-to-Digital converter (ADC) to digitize the analog MAC values. However, these ADCs introduce significant increase in area and power consumption, as well as introduce non-linearities. This work proposes a resonant time-domain CiM (rTD-CiM), an ADC-less architecture that reduces the power consumption of traditional CiM architectures with ADCs. The feasibility of the proposed architecture is evaluated on an 8KB SRAM memory array using TSMC 28 nm technology. The proposed rTD-CiM architecture demonstrates a throughput of 2.36 TOPS with an energy efficiency of 28.05 TOPS/W.

CCS CONCEPTS

• **Hardware** → **Static memory**; • **Computer systems organization** → Neural networks.

KEYWORDS

Static Random Access Memory (SRAM), compute-in-memory (CiM), convolution neural network (CNN), multiply-accumulate (MAC), time-to-digital converter (TDC).

ACM Reference Format:
Dhandeep Challagundla, Ignatius Bezzam, and Riadul Islam. 2024. A Resonant Time-Domain Compute-in-Memory (rTD-CiM) ADC-Less Architecture for MAC Operations. In *Proceedings of Great Lakes Symposium on VLSI (GLSVLSI '24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, Compute-in-memory (CiM) architectures have emerged. Based on the Von Neumann architecture, neural network accelerators are currently implemented in edge devices for complex tasks. These networks require substantial memory for accessing inputs and weights, creating memory wall bottlenecks that diminish the processor's performance. CiM architectures reduce the energy overhead associated with data movement by leveraging parallel computation within memory arrays. Like conventional neural networks (NNs), the multiply-accumulate (MAC) operation is a fundamental process in CiM computations. CiM architectures execute several multiplications between inputs and weights concurrently, summing the results in the current domain. Subsequently, an analog-to-digital converter (ADC) is utilized to convert the analog voltage into digital output bits. However, this analog processing accounts for a prohibitively large amount of the total power consumption in a CIM architecture [6].

Minimizing the overhead associated with ADCs is pivotal to improving the energy efficiency of CiM accelerators. Some CiM architectures achieve this by reducing the ADC precision for sparse inputs, whereas other approaches employ reduced precision ADCs with non-linear quantization techniques [12]. A primary issue in the integration of ADCs within mixed-signal design architectures, such as CiMs, is ensuring consistent ADC performance across diverse operating conditions and scaling with technology.

To overcome the non-linearity and high power consumption of ADCs, more research is focused on time-domain ADCs (TD ADCs) in which the analog voltage is converted to delay that can be processed in digital domain [9, 15]. In this work, we alleviate the ADC issues in CiM computation by introducing an ADC-less resonant time-domain CiM (rTD-CiM) architecture using a time-to-digital converter (TDC) that performs the same functionality as an ADC while mimicking the behavior of a digital circuit to perform MAC operations within the SRAM memory elements. In particular, the main contributions of this work are:

- First-ever ADC-less resonant CiM architecture for MAC operations utilizing a new TDC.
- A dedicated read-port 8T SRAM cell that enables read-disturb free bitwise multiplications.
- Functionality and robustness validation of the TDC.

2 BACKGROUND

As an emerging paradigm, SRAM-based CiM architectures have shown promising potential in significantly enhancing processing speed and energy efficiency for a wide range of computing tasks, such as MAC [5, 7, 8, 14] and, boolean logic [1, 2].

CiM architectures utilizing standard 6T SRAM cells [5, 8] perform computations by propagating information across the bitlines. This computational approach activates multiple rows simultaneously to

GLSVLSI '24, June 12–14, 2024, Tampa Bay Area, FL

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

fetch operands in one cycle but causes significant voltage changes on bitlines, potentially leading to severe read disturbances and data corruption. [7] employs a 7T SRAM cell for MAC computations by leveraging the discharge of the read bitline to eliminate read-disturb issues. However, this method may lead to reverse current flow if the voltage of the read bitline has significantly dropped, inducing non-idealities in MAC computation and affecting accuracy. [14] utilizes a 9T1C architecture for MAC operations but suffers from high latency, primarily due to the time consumed during charging the capacitor during bitwise multiplication operations. To address these issues, this work introduces an ADC-less CiM architecture that performs the MAC computation and converts the analog MAC value into a digital output through digital TDC circuitry.

Numerous studies on TDC circuits have been documented in the literature. The TDC architecture in [3] employs a delay element incorporated with a delay-locked loop and a counter to count the number of pulses resulting in a digital output. Another approach [10] utilizes two pulse-shrinking delay lines and a delay stabilization loop to convert a pulse input into digital output. Unlike the previous TDC works, this work uses a simple pulse-shrinking delay element alongside DFFs to capture the pulse count depending on the voltage input, further encoded as digital bits.

3 PROPOSED ARCHITECTURE

In a convolutional neural network (CNN), the multiply-accumulate (MAC) operation is a fundamental computation. This paper introduces a novel compute-in-memory (CiM) architecture that enhances the standard SRAM cache by enabling MAC operations within the memory structure. The design, detailed in Figure 2, integrates a conventional SRAM array with a binary-weighted capacitor array to perform analog MAC operations using charge-sharing for 1-bit inputs and 4-bit weights. A Readout circuit formed by proposed Time-to-Digital (TDC) converters is used to convert the analog MAC values to 4-bit digital values.

3.1 Proposed Multibit Multiplication

In the proposed architecture, 4-bit filter weights are loaded along the row direction in four adjacent 8T SRAM bitcells. The mapping of the filter weights inside the rTD-CiM array is shown in Figure 2. The 1-bit *IFMs* are applied as input pulse into the rTD-CiM array for MAC operation.

Figure 1(a) shows the transistor-level schematic of the 8T bitcell used for performing MAC operations. Figure 1(b) illustrates bitwise multiplication using an 8T SRAM bitcell. Here, INPUT = "1" is represented by a unit pulse, whereas INPUT = "0" indicates no pulse. When the weight (Q-value) is set to "1," and the input is also "1," a current flow occurs, discharging the RBL line by ΔV. Hence, it performs a 1-bit multiplication of stored weight and input for a single bitcell. Enabling multiple rows simultaneously lets us perform a series of these one-bit multiplications in parallel, leading to a collective discharge on the RBL that represents the total of the individual multiplication operations, as shown in Figure 1(c). This process essentially accomplishes a bitwise MAC function across the RBL for the activated bitcells.

Figure. 2 illustrates the process of layer mapping in a CNN using an 8T SRAM cell array for performing MAC operations. The 1-bit *IFM* is provided as input to the RWLs of the SRAM cells by the

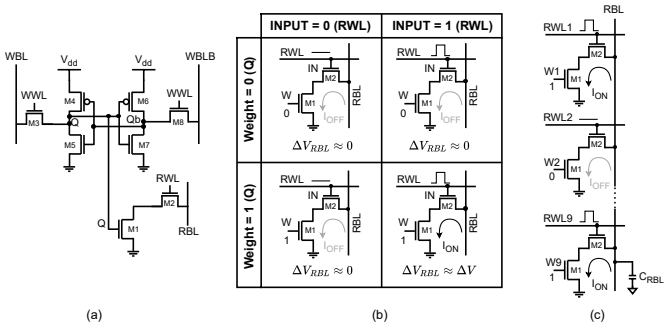

Figure 1: (a) Schematic of 8T SRAM bitcell, (b) bitwise multiplication of inputs and weights using 8T bitcells show discharge of ΔV when Input and Weight are "1," (c) multirow activation executes series of bitwise multiplications that collectively perform a bitwise MAC operation.

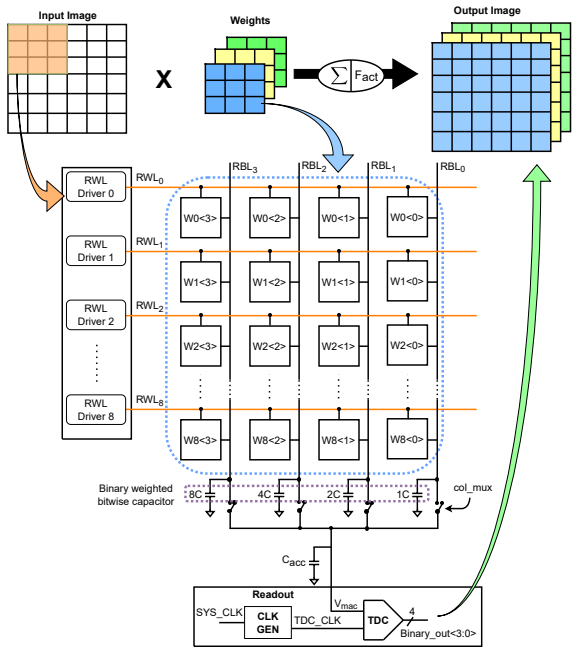

Figure 2: Layer mapping of a convolution operation performed by applying input data through RWL drivers, multiplied with fixed weights in SRAM cells performing bitwise MAC operations via binary-weighted capacitors, culminating in a multi-bit MAC output through charge sharing on the C_{acc}, and digitizing the resulting voltage V_{mac} using TDC.

RWL drivers. A logical "1" in the input feature map generates a unit pulse on the RWL, while a logical "0" results in no pulse.

The kernel weights, which are 4-bit stationary values, are stored horizontally adjacent to 8T SRAM bitcells within the same row. Each RBL is initially connected to a binary-weighted capacitor. During the multiplication phase, the initial precharged RBLs discharge incrementally, depending on the number of activated rows. The configuration depicted consists of nine wordlines corresponding to a 3×3 convolutional kernel typically used in CNN layers. Once the bitwise multiplication process is completed, we turn "ON" the *col_mux* signal that connects the binary-weighted capacitors to

the C_{acc}. This initiates a charge-sharing mechanism, leading to the C_{acc} capacitor being charged based on the collective charge from the binary-weighted capacitors resulting in the V_{mac} voltage. The accumulated voltage V_{mac} in C_{acc} represents the MAC operation's result in analog form.

3.2 Proposed TDC Architecture

Figure 3 illustrates the proposed 4-bit TDC. The overall spice simulation of the TDC circuit is shown in Figure 4. The TDC takes the analog output from the MAC operation, which is stored as a voltage (V_{mac}) in the C_{acc}, and turns it into a digital signal. The TDC also has an input pulse, TDC_CLK, generated from the system clock using a buffer-based delay circuit.

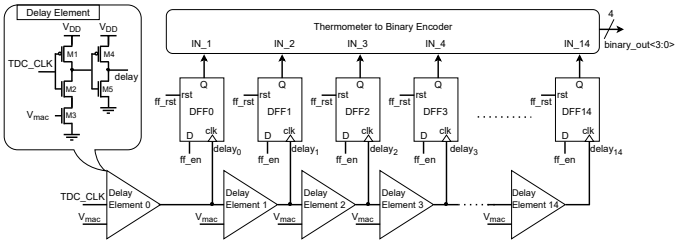

Figure 3: The 4-bit TDC architecture comprises an array of pulse-shrinking voltage-controlled delay elements along with DFFs to generate a pulse count corresponding to the analog MAC value V_{mac}, converted into digital output using a MUX-based thermometer-to-binary encoder.

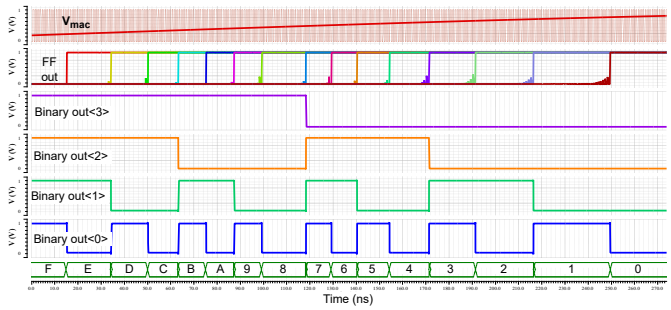

Figure 4: The functionality simulation of the TDC shows a linear decrease of the digital output corresponding to the input voltage ramp signal V_{mac}, successfully capturing the full spectrum of 4-bit values.

The transistors $M1 - M5$, shown in Figure 3, form the pulse-shrinking delay element, which consists of the voltage-controlled buffer. The propagation of the rising edge of the input pulse (TDC_CLK) is slowed down by the current-starving transistor $M3$ while the falling edge travels fast. Each delay element will shrink the input pulse by ΔT depending upon the voltage control input V_{mac} resulting from charge sharing after bitwise multiplication.

Each pulse-shrinking delay element is connected to the clock pin of a positive edge-triggered DFF. The DFF latch a "1" using the propagated pulse until the pulse disappears, which results in the following DFF retaining a "0." The Q from the $DFFs$ is subsequently provided as input to a MUX-based thermometer-to-binary encoder. This converter takes the string of "1s" and "0s" and converts them into a 4-bit binary number representing the original analog voltage V_{mac}. This 4-bit digital output is latched and stored in the SRAM array for computing the next layer in a CNN using Resonant Write Driver [4].

4 EXPERIMENTAL RESULTS

4.1 Experimental Setup

An 8KB SRAM memory instance is designed and simulated using 28nm TSMC technology. The SRAM memory array comprises 256×256 8T bitcells, implemented using Cadence Virtuoso, and simulations were performed utilizing the Cadence Spectre simulator at 1 GHz clock frequency. The bitwise LSB capacitor value is set to $1C = 4fF$, and the accumulation capacitor that stores the analog MAC output V_{mac} is set to $32fF$.

4.2 TDC Characterization and Comparison

The full voltage range of the proposed TDC circuit is from 200 mV to 800 mV. The input offset can be corrected by performing the differential nonlinearity (DNL) and integral nonlinearity (INL) analysis, shown in Figure 5.

Figure. 5 (a) illustrates the DNL characteristics of the proposed TDC. The DNL is a measure of the deviation from the ideal step size between consecutive codes expressed in least significant bits (LSBs). If the DNL value ever reaches -1 LSB, it would indicate a missing code, a severe nonlinearity error in the TDC. However, the graph shows that all of the DNL values are within ±0.5 LSB, which is considered tolerable bounds for TDC characterization. Figure 5 (b) shows the INL plot for the proposed TDC. INL is a measure of the converter's linearity, indicating the maximum deviation from the ideal function mapping of input to output over the full range of the converter. The INL plot exhibits an initial positive deviation, signifying that the early output codes from the TDC are larger than expected for an ideally linear system. As the input value increases, the INL plot trends downward, eventually falling below the zero level, which indicates that the output codes from the TDC are incrementally lower than what would be predicted by a linear model.

Table 1 compares various ADC architectures across several design references. The resolution of the TDC used in this work is 4 bits at a sampling rate of 1GS/s. The voltage supply is 1V for the TDC. The TDC exhibits an SNDR of 19.45 dB and an SFDR of 22.4 dB. The TDC utilizes V_{mac} as its input, eliminating the need for the voltage-to-time converters typically required in conventional TDC architectures. This approach reduces the overall power consumption of the TDC framework. The proposed TDC achieves 71% lower power consumption than [15] and 45.6% lower power consumption

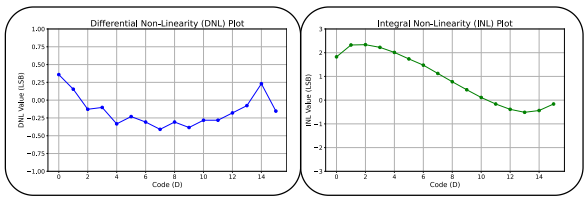

Figure 5: (a) The DNL characteristics plot of TDC demonstrates a tolerable deviation of +0.3/-0.4 LSBs, (b) The INL characteristics plot shows a maximum deviation of +2.2 LSBs and a minimum deviation of -0.5 LSBs

Table 1: Comparison of the TDC architecture with prior TD and SAR ADCs shows lower power consumption of 45.6% compared to [15], and 96% compared to [11].

Reference	[15]	[11]	[13]	[9]	This Work
Architecture	TD ADC	SAR ADC	SAR ADC	TD ADC	**TDC**
Technology (nm)	28	28	28	65	**28**
Resolution (bits)	9	6	8	8	**4**
Fs (GS/s)	0.5	1.4	8.8	1	**1**
Supply (V)	0.9	0.9	1.5	1	**1**
SNDR (dB)	54.69	67	38.4	45	**19.45**
SFDR (dB)	55.16	NR	48.9	60.3	**22.4**
Power (mW)	4.27	32	83.4	2.3	**1.25**
FoM (fJ/conv.step)	19.29	143.2	139.5	18.7	**162.8**

Table 2: Comparison of the proposed rTD-CiM architecture with prior CiM works showcases 32% higher throughput compared to [14] and 4.17× higher throughput than [5].

	This work	[8]	[5]	[14]	[7]
Technology (nm)	**28**	22	65	28	65
Cell Type	**8T**	6T	6T	9T1C	7T
Array Size	**8KB**	6.25KB	128KB	32KB	4KB
Precision (input/weight)	**1/4**	1/1	1/1	4/4	4/4
Supply Voltage (V)	**1**	0.8	0.6	0.9	1
Frequency (GHz)	**1**	1	0.138	0.100	2
Throughput (GOPS)	**2368**	133.12	567	1638.4	212.9
Energy Efficiency (TOPS/W)	**28.05**	2492	156	646.6	28.9

than [9], which are time-domain (TD) ADC. Additionally, the proposed TDC demonstrates 98% lower power consumption than [13] and 96% lower power consumption than [11], which are SAR ADCs, typically employed in conventional CiM architectures. The TDC achieves a Walden Figure of Merit (FoM) of 162.8 fJ/conversion step, which is 12% higher than [11] and 14% higher than [13]. This FoM, which is directly correlated to the bit resolution, can be reduced by increasing the number of output bits of the TDC.

4.3 MAC Comparison with Previous Works

Table 2 compares the proposed rTD-CiM architecture with existing CiM architectures capable of performing MAC operations. The proposed architecture can perform MAC operations for 1-bit inputs and 4-bit weights using 8T bitcells at 1 GHz clock frequency with 1V supply voltage. The transition to 8-bit weights can be achieved using more scaled bitwise capacitor banks and careful TDC design.

The rTD-CiM achieves a throughput of 2368 $GOPS$ with an energy efficiency of 28.05 $TOPS/W$. The proposed rTD-CiM employs a dedicated read port to ensure read stability and achieves 17× higher throughput.The proposed architecture achieves 7.2× higher frequency, resulting in 4.17× higher throughput than [5]. In [14], the 9T1C cell effectively eliminates read-disturb at the cost of higher latency during dot-product operations. The proposed architecture achieves 10× frequency, resulting in 31% higher throughput compared to [14]. The proposed architecture eliminates this reverse charging current by utilizing 8T bitcells and achieves an 11× higher throughput than [7].

ACKNOWLEDGEMENTS

This work was supported in part by the Rezonent Inc. under Grant CORP0061, National Science Foundation (NSF) award number: 2138253, and the UMBC Startup grant.

5 CONCLUSION

This paper presents a rTD-CiM architecture designed for low-power ADC-less in-memory computation. The architecture performs bitwise multiplications using 8T SRAM bitcells and utilizes a capacitor array to perform MAC computations in the analog domain. A TDC converts analog MAC results into digital values, effectively reducing the area and power overhead typically associated with ADCs in conventional CiM architectures. The proposed TDC achieves a 1 GS/s sampling frequency and 1.25 mW power consumption, with an SNDR of 19.45 dB and a Walden FoM of 162.8 fJ/conv.-step. The overall rTD-CiM architecture achieves a throughput of 2368 GOPS with an energy efficiency of 28.05 TOPS/W.

REFERENCES

[1] Dhandeep Challagundla, Ignatius Bezzam, and Riadul Islam. 2024. Architectural Exploration of Application-Specific Resonant SRAM Compute-in-Memory (rCiM). In *Proc. of the Design Automation Conference (DAC) 2024, WIP Track*. To appear.

[2] Dhandeep Challagundla, Ignatius Bezzam, Biprangshu Saha, and Riadul Islam. 2023. Resonant Compute-In-Memory (rCIM) 10T SRAM Macro for Boolean Logic. In *2023 IEEE 41st International Conference on Computer Design (ICCD)*. 110–117. https://doi.org/10.1109/ICCD58817.2023.00026

[3] Poki Chen, Shepz-Iuan Liu, and Jingshown Wu. 1997. A low power high accuracy CMOS time-to-digital converter. In *1997 IEEE International Symposium on Circuits and Systems (ISCAS)*, Vol. 1. 281–284 vol.1. https://doi.org/10.1109/ISCAS.1997.608704

[4] Riadul Islam, Biprangshu Saha, and Ignatius Bezzam. 2021. Resonant Energy Recycling SRAM Architecture. *IEEE Transactions on Circuits and Systems II: Express Briefs* 68, 4 (2021), 1383–1387. https://doi.org/10.1109/TCSII.2020.3029203

[5] Hyunjoon Kim, Taegeun Yoo, Tony Tae-Hyoung Kim, and Bongjin Kim. 2021. Colonnade: A Reconfigurable SRAM-Based Digital Bit-Serial Compute-In-Memory Macro for Processing Neural Networks. *IEEE Journal of Solid-State Circuits* 56, 7 (2021), 2221–2233. https://doi.org/10.1109/JSSC.2021.3061508

[6] Sangyeob Kim, Sangjin Kim, Soyeon Um, Soyeon Kim, Kwantae Kim, and Hoi-Jun Yoo. 2023. Neuro-CIM: ADC-Less Neuromorphic Computing-in-Memory Processor With Operation Gating/Stopping and Digital–Analog Networks. *IEEE Journal of Solid-State Circuits* 58, 10 (2023), 2931–2945. https://doi.org/10.1109/JSSC.2023.3273238

[7] Dinesh Kushwaha, Aditya Sharma, Neha Gupta, Ritik Raj, Ashish Joshi, Jwalant Mishra, Rajat Kohli, Sandeep Miryala, Rajiv Joshi, Sudeb Dasgupta, and Anand Bulusu. 2022. A 65nm Compute-In-Memory 7T SRAM Macro Supporting 4-bit Multiply and Accumulate Operation by Employing Charge Sharing. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. 1556–1560. https://doi.org/10.1109/ISCAS48785.2022.9937908

[8] Jie Lou, Florian Freye, Christian Lanius, and Tobias Gemmeke. 2023. Scalable Time-Domain Compute-in-Memory BNN Engine with 2.06 POPS/W Energy Efficiency for Edge-AI Devices. In *Proceedings of the Great Lakes Symposium on VLSI 2023* (Knoxville, TN, USA) *(GLSVLSI '23)*. Association for Computing Machinery, New York, NY, USA, 665–670. https://doi.org/10.1145/3583781.3590220

[9] Kenichi Ohhata. 2019. A 2.3-mW, 1-GHz, 8-Bit Fully Time-Based Two-Step ADC Using a High-Linearity Dynamic VTC. *IEEE Journal of Solid-State Circuits* 54, 7 (2019), 2038–2048. https://doi.org/10.1109/JSSC.2019.2907401

[10] E. Raisanen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara. 1995. A low-power CMOS time-to-digital converter. *IEEE Journal of Solid-State Circuits* 30, 9 (1995), 984–990. https://doi.org/10.1109/4.406397

[11] Lucas Moura Santana, Ewout Martens, Jorge Lagos, Piet Wambacq, and Jan Craninckx. 2023. A 70MHz Bandwidth Time-Interleaved Noise-Shaping SAR Assisted Delta Sigma ADC with Digital Cross-Coupling in 28nm CMOS. In *ESSCIRC 2023- IEEE 49th European Solid State Circuits Conference (ESSCIRC)*. 389–392. https://doi.org/10.1109/ESSCIRC59616.2023.10268759

[12] Utkarsh Saxena, Indranil Chakraborty, and Kaushik Roy. 2022. Towards ADC-Less Compute-In-Memory Accelerators for Energy Efficient Deep Learning. In *2022 Design, Automation Test in Europe Conference Exhibition (DATE)*. 624–627. https://doi.org/10.23919/DATE54114.2022.9774573

[13] X. Shawn Wang, Chi-Hang Chan, Jieqiong Du, Chien-Heng Wong, Yilei Li, Yuan Du, Yen-Cheng Kuan, Boyu Hu, and Mau-Chung Frank Chang. 2018. An 8.8-GS/s 8b Time-Interleaved SAR ADC with 50-dB SFDR Using Complementary Dual-Loop-Assisted Buffers in 28nm CMOS. In *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*. 88–91. https://doi.org/10.1109/RFIC.2018.8429007

[14] Kanglin Xiao, Xiaoxin Cui, Xin Qiao, Jiahao Song, Haoyang Luo, Xin'an Wang, and Yuan Wang. 2023. A 28nm 32Kb SRAM Computing-in-Memory Macro With Hierarchical Capacity Attenuator and Input Sparsity-Optimized ADC for 4b Mac Operation. *IEEE Transactions on Circuits and Systems II: Express Briefs* 70, 6 (2023), 1816–1820. https://doi.org/10.1109/TCSII.2023.3234620

[15] Yutong Zhao, Fan Ye, and Junyan Ren. 2022. A 500-MS/s 9-Bit Time-Domain ADC Using a Nonbinary Successive Approximation TDC. In *2022 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. 209–212. https://doi.org/10.1109/APCCAS55924.2022.10090395