electronics

Article

A Novel CNFET SRAM-Based Compute-In-Memory for BNN Considering Chirality and Nanotubes

Youngbae Kim [1,*], Nader Alnatsheh [1], Nandakishor Yadav [2,†], Jaeik Cho [1], Heeyoung Jo [1] and Kyuwon Ken Choi [1]

[1] DA-Lab, Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA; nalnatsheh@hawk.iit.edu (N.A.); jcho1@iit.edu (J.C.); hjo1@iit.edu (H.J.); kchoi12@iit.edu (K.K.C.)
[2] Fraunhofer Institute of Photonics Microsystems IPMS, 01109 Dresden, Germany; nkyadav.vlsi@gmail.com
* Correspondence: ykim102@hawk.iit.edu
† The author executed this work when he was working at Illinois Institute of Technology Chicago, USA.

Abstract: As AI models grow in complexity to enhance accuracy, supporting hardware encounters challenges such as heightened power consumption and diminished processing speed due to high throughput demands. Compute-in-memory (CIM) technology emerges as a promising solution. Furthermore, carbon nanotube field-effect transistors (CNFETs) show significant potential in bolstering CIM technology. Despite advancements in silicon semiconductor technology, CNFETs pose as formidable competitors, offering advantages in reliability, performance, and power efficiency. This is particularly pertinent given the ongoing challenges posed by the reduction in silicon feature size. We proposed an ultra-low-power architecture leveraging CNFETs for Binary Neural Networks (BNNs), featuring an advanced state-of-the-art 8T SRAM bit cell and CNFET model to optimize performance in intricate AI computations. Through meticulous optimization, we fine-tune the CNFET model by adjusting tube counts and chiral vectors, as well as optimizing transistor ratios for SRAM transistors and nanotube diameters. SPICE simulation in 32 nm CNFET technology facilitates the determination of optimal transistor ratios and chiral vectors across various nanotube diameters under a 0.9 V supply voltage. Comparative analysis with conventional FinFET-based CIM structures underscores the superior performance of our CNFET SRAM-based CIM design, boasting a 99% reduction in power consumption and a 91.2% decrease in delay compared to state-of-the-art designs.

Keywords: compute-in-memory; CIM; CNFET-CIM; chiral vectors; nanotube numbers; CNFET; carbon nanotube field effect transistors; SRAM

Citation: Kim, Y.; Alnatsheh, N.; Yadav, N.; Cho, J.; Jo, H.; Choi, K.K. A Novel CNFET SRAM-Based Compute-In-Memory for BNN Considering Chirality and Nanotubes. *Electronics* **2024**, *13*, 2192. https://doi.org/10.3390/electronics13112192

Academic Editor: Hyungjin Kim

Received: 17 March 2024
Revised: 19 May 2024
Accepted: 27 May 2024
Published: 4 June 2024

1. Introduction

In recent times, as AI models have grown in complexity to enhance accuracy, the supporting hardware has grown heavier and more intricate. This complexity and weight pose several limitations, including increased power consumption and reduced processing speed due to high throughput demands. Compute-in-memory (CIM) technology has emerged as a promising solution to these challenges. CIM utilizes the internal embedded memory array, such as SRAM, instead of external memory, thereby reducing unnecessary access to external memory by performing calculations internally. In AI applications, where high accuracy requires numerous continuous calculations, significant power consumption occurs due to the frequent use of external memory for each calculation. With the rising complexity of AI models and the increasing number of operations, CIM technologies are gaining attention as innovative solutions in AI research.

Additionally, carbon nanotube field-effect transistors (CNFETs) hold considerable potential to bolster these CIM technologies. Despite notable advancements in silicon semiconductor technology, CNFETs have the capability to emerge as robust contenders, offering reliability, high performance, and low power consumption. Due to their appropriate carrier

mobility and symmetrical [1] and balanced subthreshold electrical performance, carbon nanotube field-effect transistors (CNFETs) are being viewed as a possible future option for AI devices that demand low power consumption and high throughput [2–4]. These CNFET graphene sheets are rolled into single-walled carbon nanotubes (CNTs), and the graphene sheets can be either metal or semiconductor depending on the direction in which they are rolled. Semiconductor carbon nanotubes show great potential as high-performance channel materials [5] due to their high current density and easy controllability. Figure 1 illustrates how a CNFET forms a conductive path between its source and drain terminals through the utilization of carbon nanotubes arranged in a parallel configuration. As compared to a gate-body voltage-induced traditional channel, carbon nanotubes provide much greater driving currents. Unlike the numerous factors in bulk CMOS technology leading to significant subthreshold voltage variations [6,7], the threshold voltage of CNFETs is determined only by the diameter of carbon nanotubes. This dependency is linked to the following chiral vector:

$$D_{CNT} = \frac{a}{\pi}\sqrt{m^2 + n^2 + mn} \tag{1}$$

$$V_{th} \approx \frac{E_g}{2e} = \frac{aV\pi}{\sqrt{3} \times qD_{CNT}} \tag{2}$$

In this context, the parameter (denoted as q) denotes the energy differential, which stands as the basic unit of electron charge. The symbol α signifies the atomic separation, set at 2.49 for Carbon Nanotubes (CNTs). The term πV corresponds to the energy associated with the carbon π to π bonding, set at 3.033 eV according to the rigid bonding model. 'e' represents the elementary electron charge, while 'CNT D' denotes the diameter of a CNT. The dimensions of the Carbon Nanotube Field-Effect Transistor (CNFET) can be readily tailored by varying the number of tubes employed. Given that both n-type and p-type exhibit identical carrier mobility, P-CNFET and N-CNFET configurations employing an equivalent count of carbon nanotubes demonstrate comparable strength characteristics. While most Single-Walled Carbon Nanotubes (SWNTs) typically possess diameters close to 1 nm, this study employs a diameter of 1.5 nm.

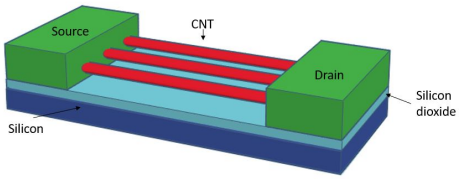

Figure 1. Structure of the CNFET.

The configuration of a Single-Walled Carbon Nanotube (SWNT) can be conceptualized as a cylindrical entity enfolding a graphene layer that is merely one atom in thickness. This wrapping pattern of graphene sheets is delineated by a distinct set of indices (n, m), visually depicted in Figure 2. The parameters n and m represent the number of unit vectors traversing along the orthogonal directions of the hexagonal crystal lattice. When m = 0, depicted in Figure 2, the nanotubes are categorized as zigzag nanotubes. Conversely, when m = n, as depicted in Figure 2, Nanotubes with armchair configuration are denoted as armchair nanotubes. Any other configuration is termed as chiral nanotubes, as shown in Figure 3.

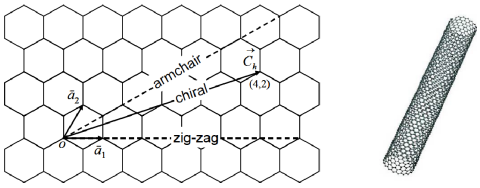

Figure 2. The lattice structure of the unfolded graphite sheet and the rolled carbon nanotube.

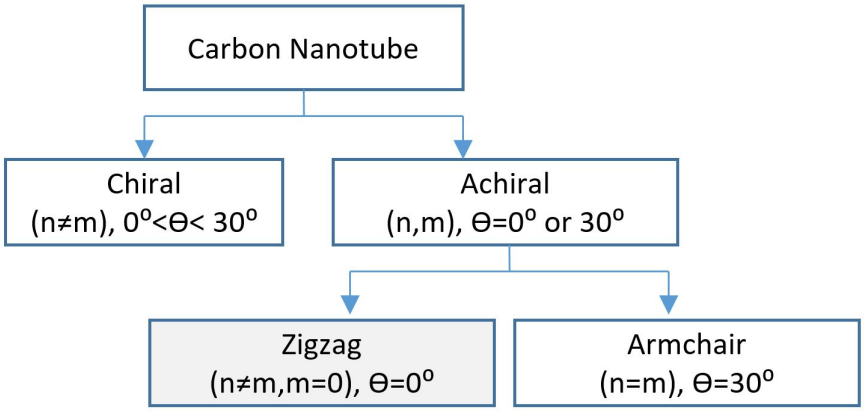

Figure 3. Classification of carbon nanotubes.

Drawing from the electrical characteristics of the CNFET device, our devised SRAM, leveraging commendable carrier mobility and robust subthreshold electrical behavior, showcases an impressive enhancement. Specifically, in comparison to the prevailing SRAM under identical conditions, our proposed SRAM manifests a remarkable 99% reduction in power consumption and a notable 97% improvement in delay. These discernible advancements in power efficiency and latency hold significant implications for AI computations within the PE Block.

2. Comparative Analysis of CNFET and CMOS

CNFETs offer notable advantages over scaled-down silicon MOSFETs, which suffer from increased power consumption and various short-channel effects resulting in reduced mobility and reduced drive current. CNFETs leverage the atomically thin nature of carbon nanotubes to demonstrate excellent scalability and enable the synthesis of nanometer-scale diameters that are ideal for advanced nanoelectronics applications. Their small size holds promise for denser, more compact integrated circuits that can handle massive data volumes in cutting-edge AI computing systems. However, challenges still remain in accurately modeling dimensional effects, series resistance, and tunneling leakage currents, which affect device performance and measurement consistency. To alleviate these problems, facilitate scaling-free device design, and comprehensively explore measurement variations and scaling effects, ensuring constant input parameters is essential. In the domain of the virtual source CNFET semi-empirical model, understanding the determinants of the current scale is of utmost importance. Switching from an inverter, which is the main operation of SRAM, to a CNFET inverter, the device utilizes gate voltage manipulation to perform transitions between states and uses P-CNFET and N-CNFET configurations based on binary inputs. Operating similarly to CMOS but with superior performance and reduced power consumption, this inverter circuit seamlessly interfaces with binary logic to provide high noise immunity and produce accurate output values corresponding to the input voltage range. CNFET-based inverters show strong potential in digital electronics by properly converting the input voltage to accurate binary while maintaining a remarkable noise margin, which is effective for SRAM-based CIM and promises good performance. This pioneering technology represents a significant advancement in the field of digital design and heralds a notable paradigm shift in the field.

3. Proposed CNFET-Based Compute-In-Memory (CIM) Design

A column is shown in Figure 4 to represent the calculation of the single neuron using the new proposed SRAM cells, while Figure 5 shows the current state of the art for the single neuron using SRAM-CIMs. Given the complete isolation between the write unit and the read unit, the novel SRAM cell architecture introduced herein boasts an autonomous structure, thereby reinforcing its reliability and performance integrity. The read unit can be

used for the multiplication of BNNs, whereas the write unit can be used to store weight values for BNNs in Q and Qb. Through this complete separation of reading and writing, we have resolved the conflicting matter concerning read and write operations, enabling the read unit to facilitate CIM calculations without requiring supplementary transistors. Table 1 compares the structural pros and cons of the proposed CNFET SRAM-CIM with those of state-of-the-art SRAM-CIM

Table 1. Comparison of structural pros and cons of proposed CNFET SRAM-CIM and state-of-the-art SRAM-CIM [8].

	Proposed CNFET SRAM-CIM	**State-of-the-Art SRAM-CIM [8]**
Pros	No disturb issue: decoupled R/W No need extra TRs No need extra Bitline: Use single RBl for Accumulate OP Saves more power consumption from single bitline	No major structural changes No disturb issue: decoupled R/W
Cons	Need minor structural changes	Need extra TRs: Two access TRs for each cell Need extra Bitline: RBL and RBLb for Accumulate OP Need more power consumption for extra bitline and TRs

Furthermore, due to the layered configuration of our read unit, the leakage power generated by BNN calculations can be significantly reduced. Moreover, employing a singular read bitline configuration, in contrast to the traditional dual-read bitline structure employed in SRAM cells, can mitigate unnecessary power consumption. Further to this, the proposed new SRAM cell is expected to provide a high yield of data in the low threshold area of its operation and can provide stable caches in an NTC (Negative Temperature Coefficient) based system. Given the singular nature of our proposed single-ended SRAM bit cell structure, it bears a notable drawback wherein the design necessitates a specialized sense amplifier owing to the singular read bitline (RBL). Nevertheless, it offers several benefits in terms of latency and power efficiency.

Figure 6 illustrates the result of multiplying Input = +1 and weight = −1, while Table 2 displays the possible four states of the proposed SRAM-CIM with binary input/weight combinations, utilizing our proposed 8T-based SRAM-CIM.

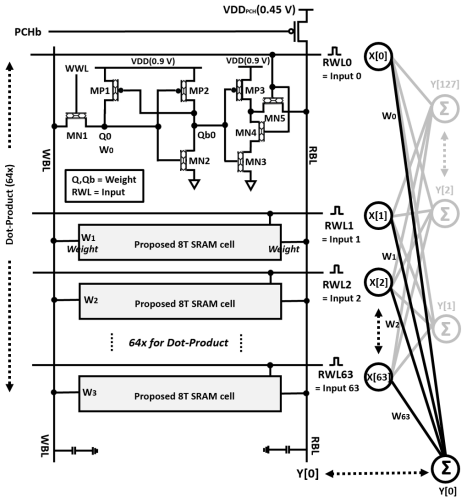

Figure 4. Proposed CNFET 8T-based SRAM-CIM for 64x dot-product cell array.

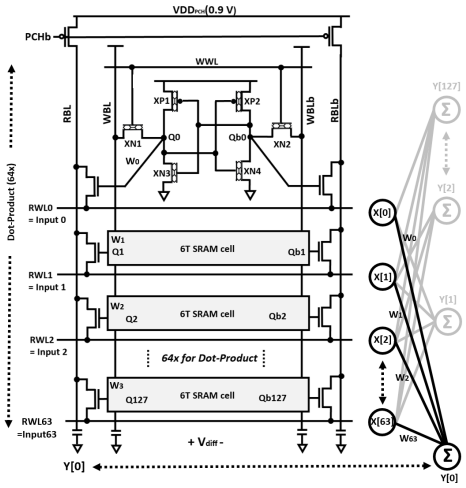

Figure 5. The-state-of-the-art 6T-based SRAM CIM [8].

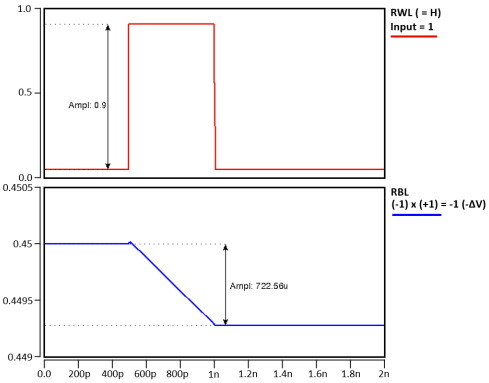

Figure 6. The result of multiplying Input = +1 and weight = −1 (**below**), and the RWL represents Input = 1 (**above**).

Table 2. Possible four states of the proposed SRAM-CIM with binary input/weight combinations.

Input (RWL) / **Weight (Q, Qb)**	**0** (RWL = L)	**1** (RWL = H)
−1 (Q = L, Qb = H)	0 (No change)	−1 (−ΔV)
+1 (Q = H, Qb = L)	0 (No change)	+1 (+ΔV)

3.1. A Column-Based Neuron Design for BNN

As illustrated in Figure 7, the proposed SRAM-CIM architecture for the 16 K fully connected Binary Neural Network (BNN) is presented herein. We have introduced a column-based neuron arrangement, incorporating a group of 128-bit cells within each column. This arrangement encompasses 64-bit cells dedicated to dot-product computation, 32-bit cells assigned for ADC reference, and an extra 32-bit cells set aside for ADC calibration. Utilizing a comparator ADC alongside ADC reference cells, the analog output derived from the dot product undergoes quantization, thus enabling sequential quantification of the output via a sense amplifier positioned at the base.

Comparators exhibit resemblances to operational amplifiers (OP amps) and are conventionally employed in scenarios where multiple signal levels are to be juxtaposed against a fixed voltage reference, such as in signal comparison tasks. Leveraging this characteristic, we employed comparators to quantize the analog summation of dot-product outcomes.

This approach aligns with the common usage of comparators as 1-bit Analog-to-Digital Converters (ADC), facilitating efficient signal processing.

$$2^{N-1}+1 \tag{3}$$

According to the equation $2^{N-1}+1$, in which N is the number of output bits if we have 64× dot-products and 32× ADC reference cells, then we require $2^{7-1}+1=33$ cycles for the ADC, as shown in Figure 8.

In the Figure 4, we can see that the weight values are retained within the dot-product cells and that the sum of 64x dot-product values might be calculated using the RBL. In the example, the sum of the dot-products of Figure 8 is +30, which can be calculated by charging or discharging the RBL.

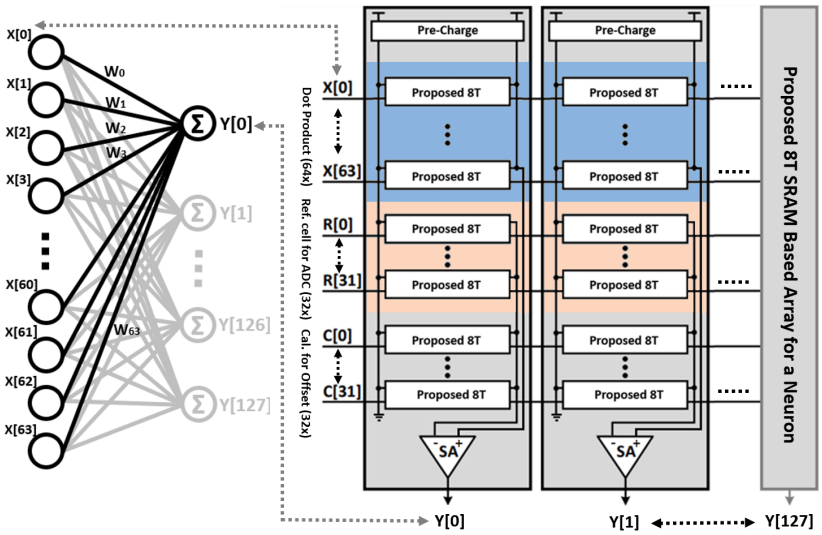

Figure 7. Proposed 8T-based SRAM-CIM for 64x dot-product cell array.

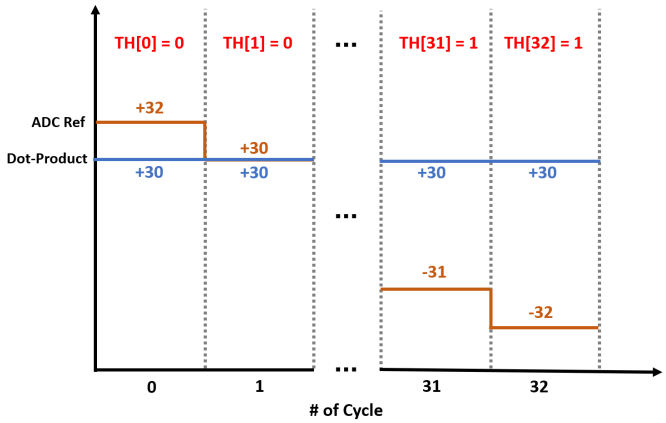

Figure 8. ADC operation (64× input for 7-bit output).

The ADC operation unfolds across 33 cycles, wherein each cycle enables the sweeping of a cell's value within the range of −32 to +32, contingent upon a step size of 2. As delineated in Figure 8, the ADC reference value undergoes sweeping (−32 ∼ +32) and is contrasted against the summation of dot-product outcomes (+30) in accordance with the fundamental operation of a comparator ADC. If the dot-product summation (+30) descends beneath the ADC reference threshold, it is discretized as 0. Conversely, if the ADC reference value surpasses or equals the measured ADC value, the quantized result is denoted as 1. To derive the quantized value TH[32:0] after 33 cycles, the resultant 1-bit outcomes from each

cycle, TH[0], TH[1], . . ., TH[32], are concatenated, yielding 33 bits comprising TH[32:0], as shown in Figure 9. Employing a one-hot coding approach [9], the initial 33-bit thermometer code produced by the signal generator undergoes conversion into a 7-bit binary output value specific to an individual neuron.

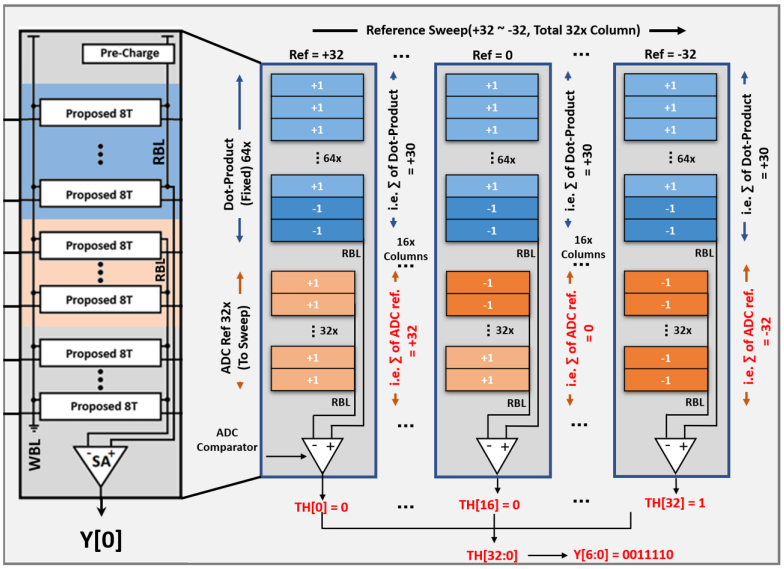

Figure 9. Structure of column ADC (64× input for 7-bit output).

The CIM architecture based on the proposed SRAM cells is shown in Figure 7 below. A network of 128× bit cells is connected vertically to a network of 64× (×[0] − ×[63]) inputs, with each column representing a neuron. Hence, this framework is designed to effectively realize the 16 K SRAM-CIM and the subsequent paper will detail the design of the test chip. The structure was designed in such a way that RBLs have been separated for the dot-product cell and RBLs have been separated for the ADC-ref for the ADC comparator, while the WBL is used in common by all cells. If we want to write the weight value to the WBL, a signal is administered through the WBL driver to the WBL, which is under the control of the WWL driver when the write operation takes place. Specifically, this means that the WWL driver controls which rows are available for writing, and sends a signal only to those rows through the WBL driver. As opposed to the current state-of-the-art structure, we utilize a singular bitline (WBL, RBL) arrangement, which results in the lack of the disturbing issue of R/W, resulting in improved stability in comparison with the current structure.

In addition, because of the autonomy of the read bitline, an additional transistor is not required for CIM calculation. Consequently, the suggested configuration demonstrates several enhanced outcomes concerning latency, power efficiency, and stability for CIM, in comparison to current structures. During pre-charge, in order to reduce the amount of unnecessary power consumed by pre-charge, the input driver and the AND gate are used to control the amount of pre-charge that is applied to the input driver. During pre-charge, in order to reduce the amount of unnecessary power consumed by pre-charge, the input driver and the AND gate are used to control the amount of pre-charge that is applied to the input driver.

3.2. Optimization of CNFET Tubes and Chiral Variants

We simulated the CNFET using the Stanford HSPICE model on a 32 nm technology node, employing the following device parameters for model stability, as shown in Table 3.

As a substitute for metal in MOSFETs and FinFETs, carbon nanotubes (CNTs) are made by rolling graphene sheets into cylindrical structures. Generally, CNTs are divided into two types [10]: single-walled nanotubes (SWNTs) and multi-walled nanotubes (MWNTs).

When the graphene sheet is rolled into a cylindrical shape, the direction of the sheet can be determined, which we call chirality. In Figure 2, chirality can be represented through a pair of indices (n, m). The number of indices is influenced by the way the graphene sheet is rolled up. The integers n and m represent the number of unit vectors along both directions of the honeycomb crystal lattice. There are two types of nanotubes: zigzag nanotubes when m = 0, and armchair nanotubes when m = n. All the other types of nanotubes are called chiral nanotubes.

Table 3. Model device parameters.

Parameter Name	Values
Lch	32.0 nm
Lgeff	100.0 nm
Lss	32.0 nm
Ldd	32.0 nm
Kox	16.0 nm
Pitch	20.0 nm

Due to the existence of no gap between their valence and conduction bands, Armchair CNTs can be considered as having metallic properties, whereas Zigzag CNTs can be considered as semiconducting due to their compact spacing between bands.

According to Figure 2, and using the chirality of the CNTs as a basis, we can categorize CNTs into two types in accordance with their chirality: chiral CNTs and achiral CNTs. In addition to Zigzag CNTs, Armchair CNTs are also classified under chiral CNTs.

A state-of-the-art 6T SRAM-based CIM structure has been shown in Figure 5. There are two access transistors in this structure, they are XN1 and XN2, respectively; the inverters that form this structure are XP1, XN3, XP2 and XN4, respectively. This circuit consists of two p-CNFETs, which are named XP1 and XP2, and all of the rest of the transistors are n-CNFETs. This storage cell is mainly formed by XP1, XN3, XP2 and XN4 and XN1 and XN2 have the main responsibility of controlling the access to the storage cell, especially during the read/write operations on the storage cell. There are two stable states possible for the storage cell (0 and 1). XN1 and XN2 transistors remain off when the standby condition is in effect. In accordance with the inputs provided by the WL terminal, the access transistors can be switched on or off based on the inputs. As long as the access transistors are on, then weight data will be stored on the Q terminal when input is given in the BL terminal during the period in which the access transistors are off. It is called the Write operation.

If the access transistors are on, at a time when the output can be observed in the BL terminal, at that time this operation is called the Read operation because the output is being observed.

As a simulation tool, HSPICE has been used for executing the simulations. This simulation involved the consideration of two different technologies: the 20 nm FinFET and the 32 nm CNFET and it was performed with a variation in the chiral vector (n) and the number of tubes for the model based on these technologies.

Based on a variation in the Chiral vector and the number of tubes, we were able to determine Average Power Consumption and Average Delay values. The average power consumption is shown in Table 4. Taking a proposed 32 nm CNFET CIM as an example, for a constant number of tubes, the power consumption would increase correspondingly with the increase in chiral vector (n) for a constant number of tubes. Additionally, tubes have been added to the channel, which has resulted in a similar change. In Table 5, the delay is shown to have decreased with the increase in the chiral vector of the CNT (n), which is shown in the chart. There is, however, no significant change in the delay caused by the addition of tubes to the circuit as a result of the increase in tubes.

For easier analysis, a comparative analysis between the values obtained due to variation in chiral vector and tube numbers has been shown in Figures 10 and 11. If the

number of chiral vectors is increased, the delay decreases; however, the amount of power consumed increases rapidly. Since we cannot ignore both power consumption and delay, we determine the optimal chiral vector and tube number by considering the correlation between power consumption and delay. As you can see from Figure 10, we can see that the power consumption increases rapidly after the chiral vector (19,0). Delay shows the highest efficiency at (23,0), but shows an incomparably sharp increase in terms of power consumption. Similarly, the number of tubes shows variations in power consumption as well as delay, similar to the chiral vector, but when the tube ratio is 3 in terms of power consumption, it is shown to be the most effective.

Table 4. Comparison of average power consumption with different chiral vectors and number of tubes in 32 nm CNFET technology.

Chiral Vector (n,m)	Diameter (nm)	Number of Tubes Ratio Power Consumption (w)				
		#1	#2	#3	#4	#5
(13,0)	1.0	1.914×10^{-11}	1.869×10^{-11}	1.725×10^{-11}	1.366×10^{-11}	6.219×10^{-12}
(15,0)	1.2	1.264×10^{-11}	4.835×10^{-12}	1.129×10^{-11}	4.417×10^{-11}	1.102×10^{-10}
(17,0)	1.3	2.098×10^{-11}	6.308×10^{-11}	1.477×10^{-10}	3.176×10^{-10}	6.575×10^{-10}
(19,0)	1.5	1.313×10^{-10}	2.842×10^{-10}	5.904×10^{-10}	1.203×10^{-9}	2.429×10^{-9}
(21,0)	1.7	4.066×10^{-10}	8.350×10^{-10}	1.692×10^{-9}	3.408×10^{-9}	6.838×10^{-9}
(23,0)	1.8	9.717×10^{-10}	1.966×10^{-9}	3.954×10^{-9}	7.930×10^{-9}	1.588×10^{-8}

Table 5. Comparison of average delay with different chiral vectors and number of tubes in 32 nm CNFET technology.

Chiral Vector (n,m)	Diameter (nm)	Number of Tubes Ratio Average Delay (s)				
		#1	#2	#3	#4	#5
(13,0)	1.0	398.41p	386.27p	371.56p	349.81p	303.99p
(15,0)	1.2	395.41p	385.82p	369.16p	348.23p	299.2p
(17,0)	1.3	393.83p	383.17p	364.83p	341.03p	280.18p
(19,0)	1.5	389.47p	371.57p	340.42p	297.69p	251.94p
(21,0)	1.7	387.05p	356.35p	297.44p	241.05p	208.99p
(23,0)	1.8	381.74p	318.91p	240.89p	175.53p	164.44p

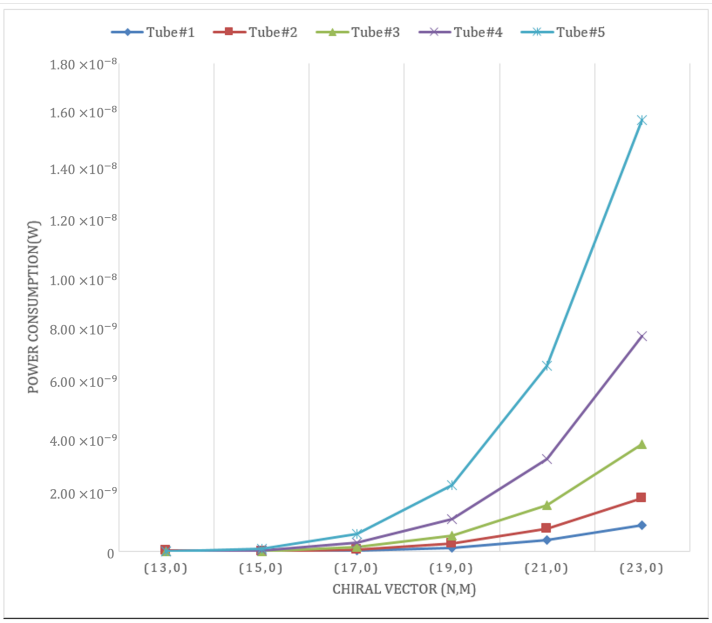

Figure 10. Comparison of average power consumption in 32 nm CNFET technology.

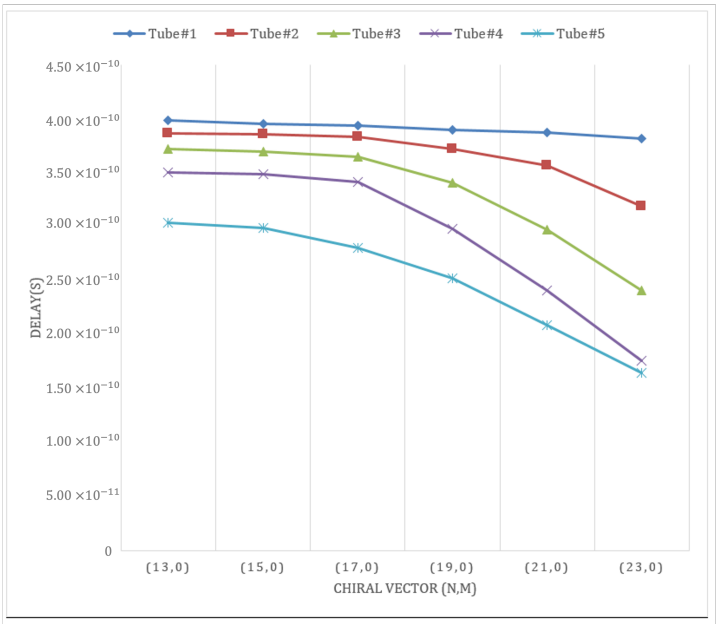

Figure 11. Comparison of average delay in 32 nm CNFET technology.

4. Performance Evaluation and Analysis

4.1. Proposed Compute-In-Memory Design

In order to uphold the robustness of the proposed CIM architecture, all parameters utilized in both the conventional CIM and the proposed CIM were meticulously maintained at parity. This encompassed consistent specifications pertaining to transistor size, capacitor size, voltage sourcing, ambient temperature, and the configuration of the test circuit across both CIM implementations. The proposed structure was verified using 22 nm FinFET and 32 nm CNFET technology nodes, and the simulations were conducted using Synopsys HSPICE and Cosmoscope tools to verify the proposed structure.

$$\Delta V = \tau \frac{I_{unit}}{C} \tag{4}$$

In order to maintain the linearity of accumulated computational tasks, the charging or discharging levels of an individual SRAM cell are precisely calibrated to 0.72 mV within our column-based neuron design. The fluctuation in the discharge level of the Read bitline (RBL) is contingent upon the number of SRAM cells undergoing charging or discharging. For instance, if two SRAM cells discharge, the RBL discharge level is computed as 1.44 mV (i.e., 0.72 mV $\times$ 2). Consequently, within our column-based framework the outcome of each cell, whether in terms of charging or discharging, contributes to the cumulative effect within the RBL, thereby aggregating the product of the input multiplied by weight for each cell into a unified neuron value. In ensuring linearity, a uniform VDD of 0.9 V is enforced across all bit cells to mitigate leakage currents, while a pre-charge voltage of 0.45 V, equating to half of 0.9 V is established for charging or discharging the RBL. As a result, if $+1 \times -1 = -1$, the RBL discharges from 0.45 V to 0 V, whereas $+1 \times +1 = +1$ triggers RBL charging from 0.45 V to 0.9 V. Guided by our multiplication methodology outlined in Table 2, the RBL discharges from 0.45 V to 0 V when $-1 \times +1 = -1$, preserving linearity throughout the computational process.

Equation (4) provides a method for calculating the range of charges or discharges. By controlling RWL's pulse width, Tau(τ) can be used to control the charge and discharge delay of RBL. When arranged in stacks of either -64 or +64, we assessed the maximum range of stacking on the Read bitline (RBL) when the input results of 64$\times$ were stacked accordingly.

As shown in Figures 12 and 13, each bit cell is stacked. At the same time, the linearity of the three-bit cells is maintained (0.72 mv).

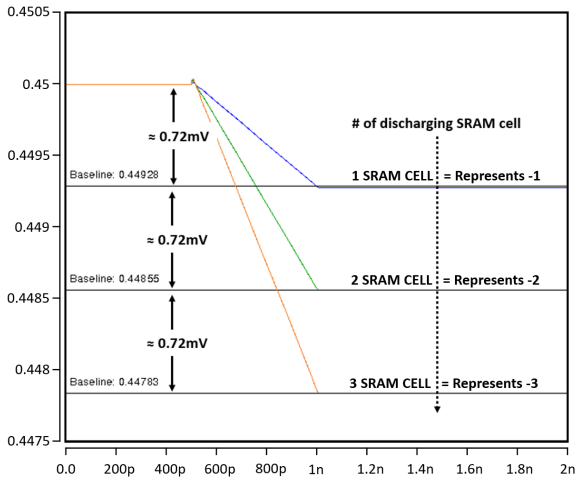

Figure 12. Discharging range with a three SRAM bit cell.

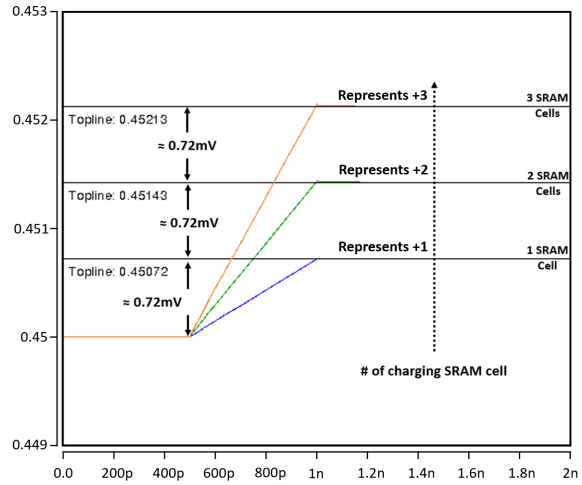

Figure 13. Charging range with a three SRAM bit cell.

Due to the fact that we use a single-bitline for our CIM technology, it maintains and reduces power consumption more effectively than a conventional method because there is only one bitline involved. Due to the fact that our new bit cells have decoupled reading and writing, our new bit cells have a lower power consumption because they are stacked together in the read unit, and our bit cells have better readability and writing properties than their predecessors. According to Table 6, traditional SRAM-CIM and proposed SRAM-CIM show different dynamic power consumption values. The proposed CIM structure, as depicted in Figure 4, does not require extra transistors and bitlines for multiplication, so is able to save a significant amount of power, as a result of the low number of transistors and bitlines needed. Compared to the state-of-the-art SRAM-CIM, in operation for bitline charging, the power consumption has been reduced by 92.34% and for discharging, it was reduced by 98.71%. Furthermore, we have achieved a power efficiency of up to 99.96% in operation for -1×0 and $+1 \times 0$ calculations. As can be seen in Figure 14, the results of multiplication have been accumulated. Due to the fact that our CIM has 64$\times$ inputs ranging from X[0] to X[63], the sum of binary multiplication results cannot exceed -64 or +64. A single cell can discharge or charge within a range of 0.72 mv, which can indicate either a negative or a positive charge as shown in Figure 15. Hence, the cumulative result of the multiplication for 64$\times$ cells amounts to 0.00072 V (0.72 mV) multiplied by 64, yielding 0.046 V. This cumulative voltage can either be charged or discharged at the reference point of 0.45 V, corresponding to +64 or -64, respectively. This means that when we have a value of -64, it will equal 0.45 v $-$ 0.046 v = 0.404 v while when we have a value of +64, it will

equal 0.45 v + 0.046 v = 0.495 v. Illustrated in Figure 14, our simulation outcomes can be meticulously scrutinized, leading to the inference that our findings consistently maintain the linearity of accumulated results across all simulations. Figures 16 and 17 show the overall structure of our proposed CNFET 16 K SRAM-CIM (64× input and 128× neurons) utilizing these 64× SRAM bit cells.

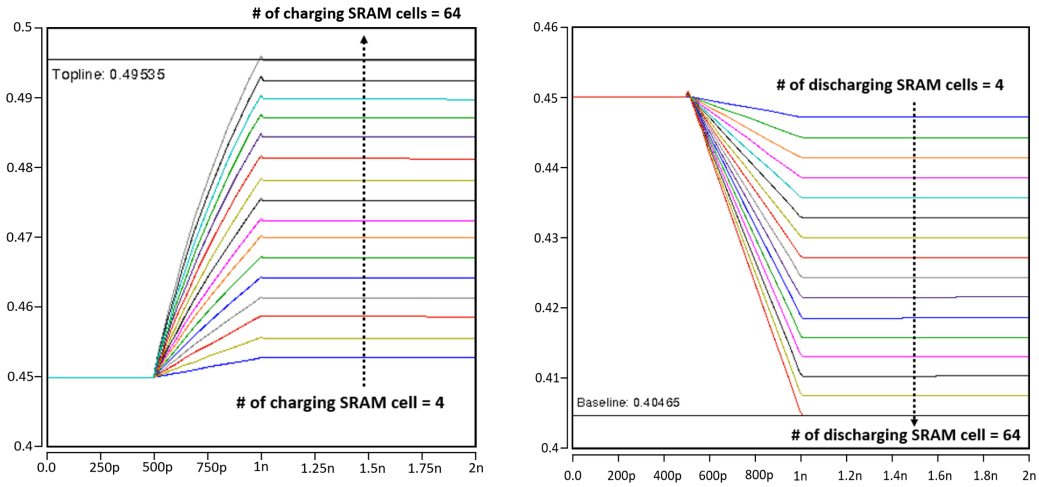

Figure 14. Charging (**left**) or Discharging (**right**) range with 64× SRAM bit cells.

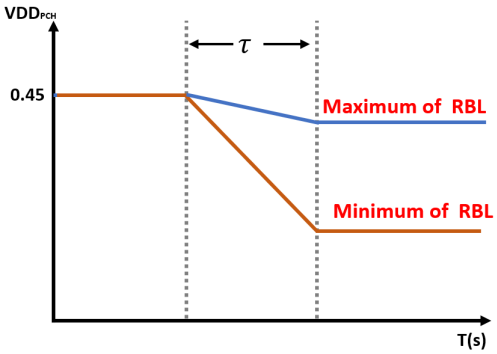

Figure 15. Maximum and minimum discharging rate of RBL with a Tau.

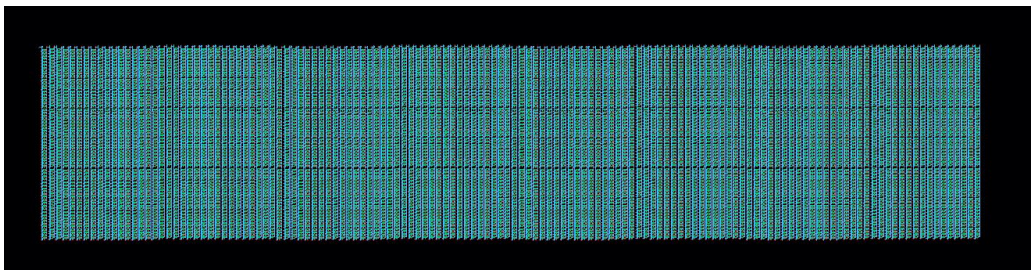

Figure 16. Proposed CNFET 16 K SRAM-CIM test-chip architecture for 64× input and 128× neurons.

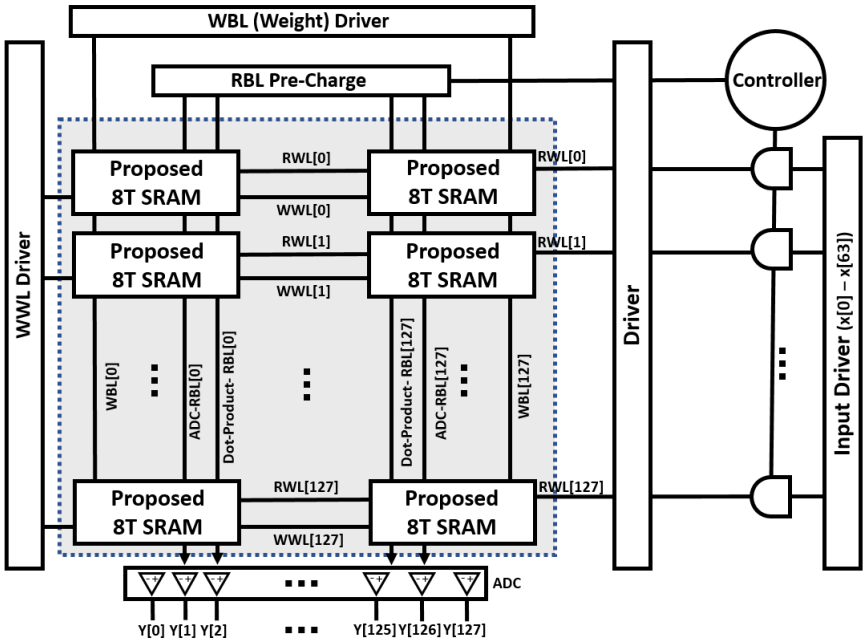

Figure 17. Overall architecture of the proposed CNFET 16 K SRAM-CIM (64× input and 128× neurons).

Table 6 displays comparison results of dynamic power consumption with the state-of-the-art 8T SRAM-based CIM for the possible four states with binary input/weight combinations, while Table 7 shows the average delay with different chiral vectors and the number of tubes in 32 nm CNFET technology.

Table 6. Comparison of dynamic power consumption with state-of-the-art 8T SRAM-based CIM [8] for possible four states with binary input/weight combinations.

Weight (Q, Qb) \ Input (RWL)	0 (RWL = L)			1 (RWL = H)		
	0 (No change)			−1 (−ΔV)		
−1 (Q = L, Qb = H)	State-of-the art [8] (20 nm FinFET)	Proposed CIM (20 nm FinFET)	Proposed CIM (32 nm CNFET)	State-of-the art [8] (20 nm FinFET)	Proposed CIM (20 nm FinFET)	Proposed CIM (32 nm CNFET)
	1.582×10^{-5} (W)	1.084×10^{-8} (W)	1.782×10^{-10} (W)	2.432×10^{-5} (W)	3.453×10^{-6} (W)	2.308×10^{-6} (W)
	0 (No change)			+1 (+ΔV)		
+1 (Q = H, Qb = L)	State-of-the art [8] (20 nm FinFET)	Proposed CIM (20 nm FinFET)	Proposed CIM (32 nm CNFET)	State-of-the art [8] (20 nm FinFET)	Proposed CIM (20 nm FinFET)	Proposed CIM (32 nm CNFET)
	1.582×10^{-5} (W)	1.094×10^{-8} (W)	2.737×10^{-9} (W)	2.432×10^{-5} (W)	1.145×10^{-7} (W)	4.629×10^{-8} (W)

Table 7. Comparison of average delay with different chiral vectors and number of tubes in 32 nm CNFET technology.

Structure	[11]	[8]	[12]	Proposed Work	
Technology node	20 nm FinFET	20 nm FinFET	20 nm FinFET	20 nm FinFET	32 nm CNFET
bit cell Type	6T	8T	10T	8T	8T
Supply Voltage(V)	0.9	0.9	0.9	0.9	0.9
Weight Bit Precision	4	4	4	4	4
Input Bit Precision	4	4	4	4	4
Power Consumption per MAC(W)	1.188×10^{-4}	2.331×10^{-5}	9.779×10^{-6}	1.955×10^{-7}	5.210×10^{-8}
Latency per MAC(S)	382.21p	392.19p	388.64p	399.45p	398.7p

4.2. Proposed SRAM Bit Cell Design

The switching power consumption of each bit cell [13–15] in SRAM consumes a greater portion of power than the other power consumption, and thus cannot be ignored when

considering the power consumption of the bit cells [16]. An improved SRAM bit cell that eliminates the switching of weak inverters was proposed in order to reduce switching power consumption. The bit cell also solves the disturbing issue of read-write conflict as described in the previous section by ensuring that read and write operations are completely separated during the operation process [17]. A key benefit of decoupled R/W operation is that it is highly effective in reducing power consumption and also in terms of reducing the delay taken for the CIM calculations to be completed. There is one notable difference between our access transistors MN1 and MN5, which is that they are separate from each other for reading and writing operations. The MN1 transistor is used for write operation and the MN5 transistor is used for read operation. That is, the read operation and write transistors are completely separated. These separate read and write operations are controlled through an enable signal. During the write operation, four TRs, which are required for the read operation are turned off, and during the read operation, four TRs required for the write operation are turned off.

In other words, out of a total of eight TRs for the SRAM cell, only four are always used for operation. In order to achieve read and write stability for the cell, we have set the size of the driver transistors MN2 and MN3 to the maximum size possible, and we have set the size of the load transistor to the smallest as well. In order to account for the bitline voltages of the bitlines, the access transistor is set at a smaller size than the driver transistor, much smaller than the driver transistor itself. Table 8 shows the transistor size and the ratio of this transistor for bit cell configurations. Based on the values of the elements shown in Table 5, we can conclude that by optimizing the transistor size, ratio, and structural characteristics of the writing part of the bit cell we can improve the write operation speed by up to 91.2%. With the improved write delay, the CIM will have the ability to store weight values faster and will benefit from a faster storage capability. Moreover, aside from the enhanced power efficiency and accelerated write speed exhibited by our proposed bit cell in comparison to conventional counterparts, it also demonstrates exceptional stability.

Table 8. Transistor size and ratio for bit cell configuration.

No.	Transistors	Ratio	# of Tubes (CNFET)	# of Fins (FinFET)
1	MN1	2	2	2
2	MN2	4	4	4
3	MN3	4	4	4
4	MN4	4	4	4
5	MN5	2	2	2
6	MP1	1	1	1
7	MP2	1	1	1
8	MP3	1	1	1

5. Conclusions

To meet the demands of complex CNN models, general von Neumann architectures have encountered physical constraints in enhancing power efficiency, primarily due to the escalating complexity of AI models and hardware density. This has emerged as a significant barrier in augmenting AI computing performance, necessitating the exploration of new architectures for AI computations. Numerous CIM approaches for AI computation have been suggested by researchers, yet substantial power consumption is required to achieve high throughput. However, in diverse AI applications like autonomous vehicles, security systems, and the Internet of Things (IoT), power consumption remains a critical factor that cannot be overlooked.

For these reasons, in this paper, we propose a new CNFET-based Compute-in-Memory (CIM) design using 8T SRAM bit cells. To attain the optimal CNFET model, we conducted an analysis of power consumption and delay by varying the number of carbon nanotube

(CNT) chiral vectors and tubes when integrating the CNFET model into our proposed CIM framework.

To facilitate the computation of vector-matrix multiplication involving binary weight and input, we employ dot products based on voltage-mode accumulation to construct the requisite multiplication framework. In our simulation setup, each of the 128-bit cells comprises 64-bit cells dedicated to dot-product computation, alongside 32-bit cells designated for ADC reference. Additionally, 32-bit cells have been allocated for offset calibration, culminating in a total of 128-bit cells. Regarding the quantized value of a single neuron, our row-by-row ADC can output 1–7 bits. Moreover, the implementation of a decoupled read and write unit, along with a single bitline (RBL), effectively mitigates operational disturbances associated with read and write operations.

According to the simulation findings, the proposed CIM exhibits a noteworthy reduction in power consumption, achieving savings of up to 99.7% when juxtaposed with the prevailing state-of-the-art CIM tasked with computing vector-matrix multiplication of binary weight and input. Furthermore, through the application of our optimized CNFET model, we effectively enhanced power efficiency and augmented throughput.

Author Contributions: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing—original draft, Y.K.; Writing—review and editing, N.A., N.Y., J.C., H.J. and K.K.C.; Supervision, Funding acquisition, Project administration, K.K.C. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by the Technology Innovation Program (20018906, Development of autonomous driving collaboration control platform for commercial and task assistance vehicles) funded By the Ministry of Trade, Industry & Energy (MOTIE, Republic of Korea).

Data Availability Statement: Data are contained within the article.

Acknowledgments: We thank our colleagues from KETI and KEIT, who provided insight and expertise that greatly assisted the research and greatly improved the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ali, M.; Abrar Ahmed, M.; Chrzanowska-Jeske, M. Logical Effort Framework for CNFET-Based VLSI Circuits for Delay and Area Optimization. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2019**, *27*, 573–586. [CrossRef]
2. Paul, A.; Pradhan, B. Effect of Chiral Vector on Voltage Transfer Characteristics of CNTFET Inverter. In Proceedings of the 2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Kolkata, India, 17–18 January 2020; pp. 1–5. [CrossRef]
3. Lin, S.; Kim, Y.B.; Lombardi, F. Design of a CNTFET-Based SRAM Cell by Dual-Chirality Selection. *IEEE Trans. Nanotechnol.* **2010**, *9*, 30–37. [CrossRef]
4. Kim, Y.; Tong, Q.; Choi, K.; Lee, Y. Novel 8-T CNFET SRAM cell design for the future ultra-low power microelectronics. In Proceedings of the 2016 International SoC Design Conference (ISOCC), Jeju, Republic of Korea, 23–26 October 2016; pp. 243–244. [CrossRef]
5. Yu, Z.; Chen, Y.; Nan, H.; Wang, W.; Choi, K. Design of a novel low power 6-T CNFET SRAM cell working in sub-threshold region. In Proceedings of the 2011 IEEE International Conference on Electro/Information Technology, Mankato, MN, USA, 15–17 May 2011; pp. 1–5. [CrossRef]
6. Deng, L.; Li, G.; Han, S.; Shi, L.; Xie, Y. Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey. *Proc. IEEE* **2020**, *108*, 485–532. [CrossRef]
7. Chen, Z.; Appenzeller, J.; Knoch, J.; Lin, Y.; Avouris, P. Impact of the nanotube diameter on the performance of CNFETs. In Proceedings of the 63rd Device Research Conference Digest, DRC'05, Santa Barbara, CA, USA, 20–22 June 2005; Volume 1, pp. 237–238. [CrossRef]
8. Yu, C.; Yoo, T.; Kim, T.T.; Tshun Chuan, K.C.; Kim, B. A 16K Current-Based 8T SRAM Compute-In-Memory Macro with Decoupled Read/Write and 1-5bit Column ADC. In Proceedings of the 2020 IEEE Custom Integrated Circuits Conference (CICC), Boston, MA, USA, 22–25 March 2020; pp. 1–4. [CrossRef]
9. Raajitha, K.; Meenakshi, K.; Rao, Y.M. Design of Thermometer Coding and One-Hot Coding. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; pp. 601–609. [CrossRef]

10. Chek, D.C.Y.; Tan, M.L.P.; Hashima, A.M.; Arora, V.K. Comparative study of ultimate saturation velocity in zigzag and chiral carbon nanotubes. In Proceedings of the 2010 International Conference on Enabling Science and Nanotechnology (ESciNano), Kuala Lumpur, Malaysia, 1–3 December 2010; pp. 1–2. [CrossRef]
11. Zhang, J.; Wang, Z.; Verma, N. A machine-learning classifier implemented in a standard 6T SRAM array. In Proceedings of the 2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits), Honolulu, HI, USA, 15–17 June 2016; pp. 1–2. [CrossRef]
12. Biswas, A.; Chandrakasan, A.P. Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications. In Proceedings of the 2018 IEEE International Solid—State Circuits Conference—(ISSCC), San Francisco, CA, USA, 11–15 February 2018; pp. 488–490. [CrossRef]
13. Maroof, N.; Kong, B.S. 10T SRAM Using Half- V_{DD} Precharge and Row-Wise Dynamically Powered Read Port for Low Switching Power and Ultralow RBL Leakage. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2017**, *25*, 1193–1203. [CrossRef]
14. Bhaskar, A. Design and analysis of low power SRAM cells. In Proceedings of the 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 21–22 April 2017; pp. 1–5. [CrossRef]
15. Mishra, J.K.; Srivastava, H.; Misra, P.K.; Goswami, M. A 40nm Low Power High Stable SRAM Cell Using Separate Read Port and Sleep Transistor Methodology. In Proceedings of the 2018 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Hyderabad, India, 17–19 December 2018; pp. 1–5. [CrossRef]
16. Khaddam-Aljameh, R.; Francese, P.A.; Benini, L.; Eleftheriou, E. An SRAM-Based Multibit In-Memory Matrix-Vector Multiplier with a Precision That Scales Linearly in Area, Time, and Power. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2021**, *29*, 372–385. [CrossRef]
17. Joshi, R.V.; Kanj, R.; Ramadurai, V. A Novel Column-Decoupled 8T Cell for Low-Power Differential and Domino-Based SRAM Design. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2011**, *19*, 869–882. [CrossRef]