Local Bit-Line Sharing Robust Dual-Port 8T SRAM With Virtual VSS for Energy-Efficient In-Memory Computing Architecture

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

30-09-2021 / 01-10-2021

CITATION

rajput, anil; Pattanaik, Manisha; Kaushal, Gaurav (2021): Local Bit-Line Sharing Robust Dual-Port 8T SRAM With Virtual VSS for Energy-Efficient In-Memory Computing Architecture. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.16704805.v1>

DOI

[10.36227/techrxiv.16704805.v1](https://doi.org/10.36227/techrxiv.16704805.v1)

Local Bit-Line Sharing Robust Dual-Port 8T SRAM With Virtual V_{SS} for Energy-Efficient In-Memory Computing Architecture

Anil Kumar Rajput, Manisha Pattanaik, Gaurav Kaushal

Abstract—The In-Memory Computing (IMC) architecture based on 6T, 8T, 10T SRAM suffers from compute-disturb, compute-failure, half-select issues, respectively, which affect the reliability of IMC operations. To overcome these problems, local bit-line sharing Dual-Port 8T (SDP8T) SRAM with Virtual- V_{SS} is proposed to improve the stability and energy-efficiency of IMC architecture. The local bit line sharing structure makes it possible for the SDP8T SRAM to use bit-interleaving architecture and address the half select issues, achieving area efficiency. The virtual- V_{SS} write-assist is used in SDP8T SRAM to improve the write-margin by 26.49% and lower the leakage power by 55.49% as compared to DP8T SRAM. Furthermore, IMC architecture is proposed using SDP8T SRAM. The remarkable feature of SDP8T-IMC architecture is that it performs IMBC operation on four operands simultaneously using all bit-line ports in a single cycle, thus doubling the throughput. The SDP8T-IMC architecture is implemented in 65-nm CMOS technology to verify the IMBC operations, which operates a wide voltage range (0.4 to 1 V) and obtain 11.04 fJ/bit average energy consumption at 1 V supply. Moreover, the proposed IMC architecture is configured as Binary Content Addressable Memory (BCAM) for searching applications which achieves 0.60fJ energy consumption per search/bit at 1 V.

Index Terms—In-Memory Computation, Compute-disturbance, SRAM, BCAM, Energy-Efficiency.

I. INTRODUCTION

THE state-of-art computing systems are based on the Von-Neumann architecture [1], which is implemented by separating data storage and processing element. Energy efficiency and performance are a priority in furthering development of data intensive application such as Artificial Intelligence (AI), machine learning and neuromorphic computing. However, the Von-Neumann architecture impose bottleneck [2] on transfer of data from memory to processing element. Energy consumption and latency are increased due to this bottleneck.

In-Memory Computing (IMC) is a new paradigm to mitigate the bottleneck of the Von-Neumann architecture-based computing system. IMC integrates ALU (computation) and storage in a single memory array, and all computation can be done within the memory array. IMC not only enhances the system performance but also improves the energy efficiency and supports massive parallelism [3]. IMC is a promising candidate for beyond Von-Neumann computing that focuses on the application ranging from high-performance artificial

intelligence to the neuromorphic computing system and AI-based Internet of Things (IoT) devices [4].

Recently, IMC concept has been widely discussed in all type of memory from emerging nonvolatile memories [5] to Static Random Access Memory (SRAM) [6]–[9]. The non-volatile memories are still in research face and suffer from low endurance and high write energy, which limit the scope of nonvolatile memory based IMC to system without the need for frequently updating the store data [4]. The SRAM based IMC provides faster write and lower write energy with a high endurance (approximately unlimited), which makes it applicable to system of small and medium capacity for wide range of applications. Implementation of IMC (integration of computation and storage at the same place) in SRAM [10] is extremely significant to improve the system throughput and energy efficiency.

Conventional 6T (C6T) SRAM has been used to perform the IMC operation [6], [7], [8], by activating two Word line (WL) simultaneously. However, C6T SRAM-based IMC suffered from compute-disturbance (flip the data during IMC) due to destructive read and inevitable low operating frequency. The read decoupled path has been proposed in 4+2T SRAM [9] to resolve the destructive read problem. It improves the Read margin and the performance of Content Addressable Memory (CAM) and IMC operation. However, 4+2T SRAM based IMC requires an extra cycle to write back the computed result, which increases the system's latency. A read computation scheme [11] has been proposed to improve the latency during IMC operation by using the 8T SRAM with a skew inverter as a sense amplifier. However, the proposed scheme suffered from compute-failure (provides false computing results) due to process variations [12]. The author in [13] has presented the sensing scheme to overcome the compute-failure in 8T SRAM for IMC operation. Apart from this, the author in [14] has used the separate write word line to perform compound Boolean logic with searching operations with the help of 8T SRAM. Further, the Transpose-8T based SRAM has been introduced to perform the programmable IMC [15]. The dual-port 8T SRAM has been proposed for current-based compute-in-memory dot product operations [16]. However, the 8T SRAM [11], [13], [15], [16] still suffers from low array efficiency and half-select issue [12] due to 6T-like write operations [17], [18]. The 10T SRAM reported in [19], [20] are free from compute-failure and compute-disturb and compute all logic operations in a single cycle. However, 10T SRAM suffers from half-select issue during IMC operation under process variation

Anil Kumar Rajput was with the Department of Information Technology, ABV-Indian Institute of Information Technology, Gwalior, India e-mail: (anil@iiitm.ac.in).

Figure 1: Scheme of Circuits (a) The Proposed SDP8T SRAM based IMC architecture. (b) The schematic of local bit-line sharing Dual-Port 8T (SDP8T) SRAM with Virtual V_{SS} .

[21], and hence degrades the performance of 10T SRAM based IMC architecture. To overcome the compute-failure, compute-disturb and half-select issue, 12T SRAM [12] has been proposed for IMC, at the cost of increased area and energy consumption.

In this paper, local bit-line sharing Dual-Port 8T (SDP8T) SRAM with Virtual V_{SS} is proposed. The proposed SRAM improves write stability, leakage power, read energy, and read delay at the cost of an increase in write delay and write energy as compared to recently reported DP8T SRAM [16].

A reliable IMC architecture is proposed by using SDP8T-SRAM. The proposed IMC architecture is free from the compute-disturb, compute-failure, half-select issue during IMC operation. Further, The proposed SDP8T SRAM-based IMC architecture fully utilizes all read and write bit line ports and simultaneously performs In-Memory Boolean Computation (IMBC) which improves the latency, throughput, and energy consumption as compared to recently reported IMC in [6], [11], [12] [14]. Furthermore, SDP8T-IMC architecture is configured as a Binary Contain Address Memory (BCAM) for searching applications.

The remaining of the paper is organized as follows; Section II gives the overview of the proposed SRAM and IMC architecture structure; Section III presents the analyses of the proposed SDP8T SRAM, Section IV presents SDP8T SRAM based IMC architecture, and Section V presents the evaluation of IMC architecture. Finally, conclusion is presented in section VI.

II. THE PROPOSED SDP8T SRAM

This section describes the proposed SDP8T SRAM structure and operation, as shown in Figure 1. Figure 1(a) shows the overall IMC architecture using the proposed SDP8T SRAM. The SRAM has three main parts: read decouple bit-cells,

Figure 2: Control signal of SDP8T SRAM.

two pass-gate transistors (PGL/PGR), and one write assist transistor (WA), as shown in Figure 1(b). The proposed SDP8T SRAM has a double-ended write by using transistors (GL, PGL) and (GR, PGR), and decouples read path by using high V_{th} (HVT) transistors (RL, RR) for a differential read. The VV_{SS} , PGL, and PGR transistors are shared with low V_{th} (LVT) NMOS write-assist transistor (WA) in a block of 4 SRAM cells as shown in Figure 1(b). The VV_{SS} is used to reduce the leakage power during the hold state and also to enhance the write operation [22].

The proposed SDP8T SRAM has a lower impact on Q and QB storage nodes from RWL and RBL/RBLB. Therefore, the read noise margin is significantly improved compared to C6T SRAM. As a result, reliable multi-word activation for IMC operation is achieved. To perform the IMC operations, multi-word RWL and WWL are activated simultaneously using RWL-driver and WWL-driver. The RBL and WBL Sense Amplifier (SA) pairs are used to perform the boolean logic operation by using the four operands available on RBL, RBLB, LWBL, and LWBLB. The multiple IMBC operations (NAND, AND, NOR, OR, XOR) are obtained using the RBL-SA and WBL-SA in a single cycle. Hence, it increases the throughput

and makes IMC architecture more energy and performance efficient. The Match Line SA (ML-SA) is used to perform BCAM operation, discussed in section IV. The V_{ref} generator is used to provide the reference voltage to RBL-SA, WBL-SA, and ML-SA. The control signals are shown in Figure 2 to perform the hold, read, write, and IMC operation.

1) *Hold Operation*: RWL kept at high during hold operation, and WWL, WS, and CNTRL signals kept at low. The GND is not directly connected to the pull-down transistors (DL, DR). The logic zero is obtained at V_{VSS} by using the Low V_{TH} (LVT) WA transistor with higher width (Approximately four times then access transistor), as shown in Figure 1(b). As a result, the V_{VSS} value is achieved at 100 mV during hold operation.

2) *Read Operation*: During the read operation, RBL and RBLB are precharged to VDD, WWL keeps at GND, WS signals keep at GND, and CNTRL signals are kept high as a result, logic zero is achieved at V_{VSS} . As RWL drives from VDD to GND, the RBL and RBLB discharge according to storage value at Q and QB, respectively. The difference voltage between RBL and RBLB is sensed by using the low-offset double-tail latch-type voltage sense amplifier [23].

3) *Write Operation*: During the write operation, RWL keeps at VDD, and CNTRL signals are kept at low. The voltage of GWBL and GWBLB are set to written data by write driver as shown in Figure 1(a). The voltage of GWBL and GWBLB is transferred to storage node Q and QB, respectively, of selected bit-cell inside SDP8T SRAM by enabling WS and WWL signal. The write access paths are composed of two NMOS transistors (GL/GR) and (PGL/PGR), which increase the write delay and reduce write stability. To overcome the write stability problem, the Virtual V_{SS} write assist [24] technique is used in the proposed SRAM.

Figure 3: SDP8T SRAM array and the storage node of row and column half select cell during write '1' operation.

4) *Half-select issue*: The half-select issue is defined as the flip of the stored value at half selected cell (row, column) during the write-back operation of IMC [12]. The C6T SRAM suffers from bit line disturbance in the row-half-selected bit cells [18]. As write word-line simultaneously enable all bit-cells, including words in non-selected columns, half-selected disturbance happens in the row-half selected cells, but column-

half-selected bit cells escape from half-select issue due to inactive WWLs.

In the proposed SDP8T SRAM, one extra control signal WS is incorporated with the help of pass gate transistors PGL/PGR to resolve the half select issue at row half selected cell. In the row half select cell during the write operation in the selected cell, as shown in Figure 3. However, the storage nodes (Q1 and QB1) are maintained because the storage node is disconnected from the GWBL pair. After all, WWL[X] and WS[Y] are kept at low. Thus row-half select issues are resolved using pass-gate transistor PGL/PGR. To analyze the half select issues, the Monte-Carlo (MC) simulation are performed during write operations on the selected cell, row, and column half selected cells, and It is found from the MC simulation result as shown in Figure 4, the half select issues are not present in the proposed SDP8T SRAM.

Figure 4: Waveform of the selected cell and half-selected cell during the write operation.

III. EVALUATIONS OF SDP8T SRAM

In this section, the benchmarks circuits of C6T, Transpose-8T [15], DP8T [16], 8T [11] 10T [20] and the proposed circuit are implemented with the same 65nm CMOS technology with exact sizing and simulate at 1V supply voltage and 27 °C temperature with a typical process corner for performance comparison.

Figure 5: Variation of RSNM with a) V_{DD} . (b) Temperature.

A. Static Noise Margin

1) *HSNM*: HSNM is the noise margin of SRAM during a standby condition. The HSNM of different SRAM cells is

Table I: Performance comparison of various SRAM at 1 V supply voltage.

SRAM	Write delay(ps)	Write Energy(fJ)	Read Delay(ps)	Read Energy(fJ)	HSNM (mv)	RSNM (mV)	WSNM (mV)	RDNM (mV)	WDNM (ps)	Leakage power(nw)	Area*	FOM
SDP8T	22.09	32.36	5.16	0.187	230	325	573	999	5.6	4.58	1.49×	2.005
DP8T [16]	18.13	27.48	5.09	0.424	320	320	453.41	999	5.6	10.29	1.3×	1.80
Tr. 8T [15]	17.64	24.89	5.92	0.293	320	152	453.40	877	5.8	8.80	1.3×	0.93
10T [20]	18.12	26.13	7.50	0.505	320	320	453.42	999	6.1	9.67	1.68×	1.28
C6T	16.07	20.71	5.78	0.353	320	132	453.41	773	7.0	8.52	1.00×	1

*Area of various cell normalized with respect to 6T SRAM¹

shown in Table I. During the hold state, the V_{VSS} is connected to an off-state LVT WA transistor. Therefore, the HSNM of the proposed SRAM is 31.19% lower with respect to C6T SRAM.

Figure 6: Comparison of RSNM under different process corners.

2) *RSNM*: Table I contains the RSNM value of various SRAM cells at $V_{DD}=1V$. The proposed SRAM has read decouple path. Hence, the read current does not flow through the storage node, which improves the read stability. Therefore, the RSNM of the proposed SRAM is improved by $2.43\times$ and $2.11\times$ over C6T and Transpose 8T [15]. The use of virtual ground through WA transistor makes pull-down network slightly weaker than pull-up network. Therefore, The RSNM of the proposed SDP8T SRAM is improved by 1.6% as compared to the other read decouple SRAM (DP8T [16], and 10T [20]), as given in Table I. The RSNM improves with the increment in supply voltage, as shown in Figure 5(a). The threshold voltage of the transistor reduces with the increased temperature. As a result, the leakage current increases, thereby resulting reduction in RSNM, as shown in Figure 5(b).

The RSNM of the proposed SRAM with respect to the process corner is shown in Figure 6. The threshold voltage changes due to inter-die variation at different process corners. Therefore, the read stability changes accordingly. The difference of the mean value of RSNM between best and worst-case corners of the proposed SRAM is 60mV, which is lower than the difference of RSNM between best and worst-case corners of DP8T SRAM (64mV). Therefore, the proposed SDP8T SRAM show better process variation tolerance in RSNM as compared to DP8T SRAM [16].

3) *WSNM*: The WM (WSNM) is calculated in terms of dynamic write margin with Write Trip Voltage(WTV) to reflect

Figure 7: WM of proposed SDP8T SRAM a) without V_{VSS} write assist (b)with V_{VSS} write assist. Figure 8: Variation of WM with a) V_{DD} (b) Temperature.

the dynamic behavior of SRAM using V_{VSS} during a write operation. The WTV is defined as the difference between V_{DD} and WWL voltage at which the storage nodes Q and QB flips. The proposed SRAM has two series access transistors (PGL, GL), which will degrade the WM of SRAM, as shown in Figure 7(a). However, the use of V_{VSS} at the ground terminal during the write operation improves the WM of SRAM cell [24]. The V_{VSS} degrade the strength of the pull-down transistors, which improves the write 1 stability and degrade write 0 stability. It is found that the overall WM of the proposed SRAM improves, as shown in Figure 7(b). Therefore, The WSNM of the proposed SRAM improved by 26.29%, over C6T, as given in Table I.

The WM of the proposed SRAM improves with increment in supply voltage, as shown in Figure 8(a). The WM of the proposed SRAM is better than C6T, DP8T [16], Transpose 8T [15], and 10T [20] SRAM at all supply voltage from 0.4V to

Figure 9: Comparison of WM under different process corners.

1V. The sub-threshold current increases with the increase in temperature. Therefore, the WM is increased with temperature, as shown in Figure 8(b). Furthermore, the WM of the proposed SRAM with change in process corner is shown in Figure 9. The difference of mean value of WM between best and worst case corners is 60mV, which is smaller than the DP8T [16] SRAM cell's WM variation (101mV). Therefore, the proposed SRAM shows better process variation tolerance in term of WM as compared to DP8T [16].

B. Dynamic Noise Margin

1) *WDM*: Write Dynamic Noise Margin (WDM) is an essential parameter for analyzing SRAM's noise handling capability during the active condition. WDM is defined by the minimum pulse width of the write word line (WWL) to reach the bit-cell storage node into the switching threshold [25]. It is observed from Table I, The WDM of the proposed SRAM is 20%, 3.44%, and 8.19% better than C6T, Transpose 8T [15], and 10T [20] SRAM, respectively.

2) *RDNM*: The Read dynamic Noise Margin (RDNM) is defined by the voltage difference of storage node Q and QB when read bit line reach to sensing voltage during read operation [26]. The RDNM of the proposed SRAM is 29.23% and 13.91% better than C6T and Transpose 8T [15] SRAM, respectively, as observed from Table I

Figure 10: Variation of leakage power with (a) VDD. b) Temperature.

C. Leakage Power

The power consumes by the cell during standby conditions measure as leakage power. Sub-threshold leakage is a major component of leakage power in the hold condition. The V_{VSS} offers a stacking effect that reducing sub-threshold leakage

[22] in hold condition. Therefore, the leakage power of the proposed SRAM is improved by 46.24%, 55.49%, 54.38%, 47.95%, and 52.63% as compared to C6T, DP8T [16], 8+T [11], Transpose-8T [15], and 10T [20] SRAM, respectively as observed from Table I. At lower supply, the sub-threshold leakage power decrease due to reduction on drain-induce barrier lowering (DIBL), gate-induce drain leakage (GIDL), and gate tunneling current voltage [22]. Therefore, the leakage power is reduced with VDD reduction, as shown in Figure 10(a). The leakage power is measured at different temperature values. Figure 10(b) shows that the proposed SRAM has the least possible change in leakage current w.r.t. temperature, which shows that the proposed SRAM has remarkable temperature tolerance compared to C6T, DP8T [16], Transpose-8T [15], and 10T [20] SRAM.

Figure 11: Layout of (a) Proposed SDP8T SRAM (b) DP8T SRAM.

D. Cost Comparison

The cost of the SRAM is compared in terms of read/write energy and read /write delay of SRAM.

1) *Read/Write Delay*: The proposed SRAM uses a series of transistors for the write operation. Therefore, the write delay of the proposed SRAM is 21% larger compared to DP8T SRAM [16]. The proposed SRAM uses a high V_{th} transistor for the read operation. Therefore the read delay of the proposed SRAM is reduced by 1.37% compared to DP8T SRAM [16] as given in Table I.

2) *Read/Write Energy*: The proposed SRAM is designed by a high V_{th} transistor for the read operation. Therefore the read energy of the proposed SRAM is improved by 47.02% compared to DP8T SRAM [16], as observed from Table I. The write energy in the proposed SRAM is relatively higher because of floated V_{VSS} and two series transistors for the write operation.

3) *Area*: Figure 11 shows the layout of the proposed SDP8T SRAM and DP8T SRAM [16] using 65nm UMC technology. The layout of three extra transistors PGL, PGR and WA are laid at left, right and bottom, respectively as shown in Figure 11(a). The extra transistor lead to area overhead of 20% in the proposed SDP8T SRAM (80.44 μm^2

($6.41\mu\text{m} \times 12.55\mu\text{m}$) when compared to DP8T SRAM [16] ($66.91\mu\text{m}^2$).

E. Figure of Merit

There are different parameters (SNM, delay, power, and area) to characterize the performance of the SRAM. These parameters are application-dependent and have a trade-off between them. Thus, The Figure of Merit (FOM) is given in equation 1 [27].

$$FOM = \frac{RSNM_N * WSNM_N * RDNM_N * HSNM_N}{Area_N * Energy_N * WDNM_N * Delay_N * Leakage_N} \quad (1)$$

Where $RSNM_N$, $WSNM_N$, $RDNM_N$, $HSNM_N$, $Area_N$, $Energy_N$, $WDNM_N$, $Delay_N$, $Leakage_N$ are normalized value w.r.t 6T SRAM. It is observed from Table I, the FOM of proposed SRAM is $2.005\times$, $1.11\times$, $2.20\times$, and $1.56\times$ better as compared to C6T, DP8T [16], Transpose 8T [15], and 10T [20] SRAM, respectively.

IV. SDP8T BASED IMC ARCHITECTURE

The overall IMC architecture using SDP8T SRAM is shown in Figure 1(a), which performs both data storage and computational inside memory array. The proposed IMC architecture works in two modes IMBC and In-memory CAM mode. The control signal CAMEN is used to select the mode. In the SDP8T-IMC architecture, multiple rows are activated simultaneously to perform IMBC operations. The SA compute different Boolean logic operation in a single clock cycle, and obtained data is stored in the same column. The Asymmetric SA is used in [11] to sense bit-wise NAND and NOR operation, which reduces process variation and requires longer discharging time (Bit-line sensing delay) to develop the difference between bit line voltage which reduces the max operating frequency and degrades the performance of IMC architecture [28]. The symmetric SA with low offset voltage [23] is used in the proposed SDP8T-IMC architecture to sense bit-wise NAND and NOR operations of A and B. Therefore, the proposed IMC architecture operates at a higher maximum operating frequency as compared to [11], [28].

A. In Memory Boolean Computation

Figure 12(a) shows the simulation setup used to realize basic logic operations like NAND/AND, OR/NOR, and XOR/XNOR through proposed IMC architecture with BL and BLB capacitance of 100fF for all simulations. Four sense amplifiers, SA1, SA2, and SA3, SA4, are used to realize the basic logic operations on two pairs of different operands A, B, and C, D, respectively. The RBL and RBLB are connected to the positive and negative terminals of SA1 and SA2, respectively. The LWBL and LWBLB are connected to the negative and positive terminals of SA3 and SA4, respectively. The optimal V-REF value of $0.9V_{DD}$ is used as one terminal of sense amplifiers in SA1, SA2, SA3, and SA4. For performing the basic logic operations on A and B, both RBL and RBLB are pre-charged to VDD. After that, two RWLs, RWL0 and RWL1, are connected to the ground. The voltage level for RBL, RBLB, and LWBL and LWBLB for IMBC operations

Figure 12: (a) Simulation setup of SDP8T-IMC with bit-line and bit-line bar capacitance (Cbl) and (Cblb) of 100 fF (b) Discharging condition of RBL, RBLB, LBL, and LBLB for different A, B, and C, D.

Figure 13: Timing waveform of IMBC operation on different operands (a) A = 0, B = 0 (b) A = 0, B = 1 (c) A = 1, B = 0 (d) A = 1, B = 1.

at different operands (A and B, C and D) are shown in Figure 12(b).

If both the operands A and B are at 0. The pre-charged RBL line retains its pre-charged state, and the pre-charged RBLB line starts discharging from VDD. The discharging of the RBLB is indicated through ↓, as shown in Figure 12(b). When RBLB discharge below the reference voltage, then, SEN signal is activated to SA, and RWL signals are deactivated. After that, the SA1 senses the difference between RBL voltage and V-REF and correspondingly achieves 1 and 0 at the positive terminal and negative terminal of SA1, respectively. The timing waveform for the same is shown in Figure 13(a).

If one of the operand A or B is at 1, both the RBL or RBLB discharges. After that, the SA1 senses the difference between RBL voltage and V-REF and correspondingly achieves 0 and 1 at the positive terminal and negative terminal of SA1, respectively. The timing waveform for the same is shown in

Figure 13(b-c).

Suppose both the operands A and B are at 1. The pre-charged RBLB line retains its pre-charged state, and the pre-charged RBL line starts discharging. After that, the SA1 senses the difference between RBL voltage and V-REF and correspondingly achieves 0 and 1 at the positive terminal and negative terminal of SA1, respectively. The timing waveform for the same is shown in Figure 13(d).

From Figure 12(a) and Figure 13(a-d), It is found that the positive and negative terminal of SA1 represents NOR and OR operations, respectively. RBLB is connected to the complementary nodes. Therefore, RBLB represents the AND/NAND operation of A and B through sense amplifier SA2. The LWBL and LWBLB perform the basic logic operation on operands 'C' and 'D' through SA3 and SA4 in a similar way as described for operands 'A' and 'B'. The additional NOR gate is used between two sense amplifiers (SA1 and SA2) to perform the XOR operation of A and B, as shown in Figure 12(a). The additional NOR gate is used between two sense amplifiers (SA3 and SA4) to perform the XOR operation of C and D, as shown in Figure 12(a).

Figure 14: Monte-Carlo Simulation at different write word line access time (T_{WWL}) (a)The 16.6% data flip when $T_{WWL} = 300ps$, (b) No data flip when $T_{WWL} = 200ps$.

To perform a boolean logic operation between C and D, WWL[2] and WWL[3] are activated simultaneously, which seems scenario similar to C6T, and there is the chance for compute-disturbance. Figure 14[a-b] shows a 1K Monte-Carlo simulation result of compute-disturbance versus varied write word line(WWL) activation time. It is observed from Figure 14(a), there is no compute disturbance when WWL activation time is 200ps under process variation. When WWL is turned on for 300ps, compute-disturbance occurs, and 16.6% of data is flipped, as shown in Figure 14(b). In proposed IMC architecture, WWL[2] and WWL[3] are activated for the duration of

200ps for reliable sensing(no-compute-disturbance) of boolean logics [28].

Figure 15: BCAM Operation:RBL/RBLB configure as SL/SLB and RWL configure as ML.

B. In Memory CAM operation

The CAM can be divided into BCAM and Ternary CAM(TCAM) according to search accuracy. The cross-coupled inverter structure is used to store the data, and two decoupled read paths corresponding to two match lines are used for comparison.

1) *BCAM Search*: The proposed SDP8T-IMC architecture is also configured as binary contains address memory(BCAM). The CAMEN control signal is used to configure the proposed architecture as BCAM mode. Figure 15 shows the BCAM configuration for a 2-bit search operation using SDP8T SRAM. In BCAM configuration, the RBL, RBLB, and RWL are configured as select line (SL), select line bar(SLB), and Match Line (ML), respectively, when CAMEN signal at 1. The one terminal of the sense amplifier is connected to ML, and another terminal is connected to V_{ref} , similar to conventional BCAM [29]. In BCAM mode, the search data 1 and 0 are applied on SL[0] and SL[1], and ML is precharged to VDD. The ML remains at initial value VDD if store data matches with the select line otherwise, ML discharges. The row-wise sense amplifier has the same circuit structure as the column-wise small differential sense amplifier used for logic operations. It is observed from Figure 15, the row[0] data match with the search data string; as a result, ML[0] will remain at precharged value VDD, and the sense amplifier will be at the value logic 1 at the output for row[0]. The search data do not match from the store value of row[1], row[2], and row[3], however in the case of row[1], row[2], and row[3], the search data string values do not match with the store value; as a result, the ML[1], ML[2]and ML[3] discharge from initial precharged value to zero value and sense amplifier will be at the value of logic 0.

V. EVALUATION OF IMC ARCHITECTURE

Reliability remains the primary concern in existing SRAM-based IMC. The reliability of IMC architecture is assured by avoiding compute-disturb, compute-failure, and half-select issue during the IMBC operation. To verify the reliability of SDP8T-IMC, 2000-point Monte-Carlo simulation across process/mismatch variation and 3-sigma variations in nominal

Table II: Performance comparison of IMC architecture based on various SRAM

IMC Architecture	SDP8T-IMC	C6T-IMC [6]	6T-CSRAM [28]	8T-IMC [11]	Tr.8T-IMC [15]	10T-IMC [20]	12T-IMC [12]	4+2T-IMC [9]
CER (%)	0	18.62	0	8.79	9.5	0	0	0
Energy (fJ/bit)	11.04	16.29	15.13	27.67	22.5	27.94	17	31.8
Throughput ²	2	1	1	1	1	1	2	1
Latency ³	1	2	1	1	1	1	1	2
Area ¹	1.49×	1.00×	1.00×	1.35×	1.3×	1.68×	2×	1.15×
FOM	0.1215	0.02489	0.0660	0.0311	0.0618	0.0213	0.0588	0.0273
Compute-Disturb	No	Yes	No	No	Yes	No	No	No
Compute-Failure	No	Yes	No	Yes	No	No	No	No
Half-select Issue	No	Yes	Yes	Yes	Yes	Yes	No	No

¹ Area of various cell normalized with respect to 6T SRAM.

² Number of computation in single cycle.

³ Number of cycle require to perform all computation.

VDD is performed. Reliability remains the primary concern in existing SRAM-based IMC.

Figure 16: 1K MC simulation of storage node Q(A,B) of row[0] and row[1] of SDP8T SRAM during IMC.

A. Mitigating Disturbance and Failure

The proposed SDP8T contains a read decouple path, which removes sneak path current [28] and eliminates compute-disturbance. It is found from Figure 16 that there is no compute-disturbance(data flip) and no compute-failure in SDP8T-IMC architecture during IMC operation. Further, the half-select stability of SDP8T-IMC architecture is tested under process variations. The MC simulation is performed during write operations on row-half selected cells on different supply voltage, VDD to analyze the half select issues. The proposed SDP8T SRAM uses the pass gate transistors, PGL, and PGR to resolve the half select issues. Therefore, the proposed SRAM has zero failure probability for all supply voltages (0.3 V to 0.7 V), as shown in Figure 17. However, the failure probability in DP8T SRAM [16] increases with a decrease in supply voltage. Hence, the SDP8T SRAM-based IMC architecture is able to mitigate compute-disturb, compute-failure, and half-select issue, as observed in the previous IMC proposal.

B. Energy Benefits

The proposed SDP8T SRAM based IMC architecture has low bit line swing and decouple read path which eliminate

Figure 17: The Failure probability due to half-select issue.

sneak path current and virtual ground condition lead to significantly reduction in energy consumption. When the work in C6T [6], 6TCSRAM [28] 8+T [11], 8T [14] and 10T [20], 12T [12], and 4+2T [9] are imported to 65nm, they will consume 16.29fJ/bit, 15.13fJ/bit, 27.67fJ/bit, 22.5fJ/bit, 27.94fJ/bit, 17fJ/bit and 31.8 fJ/bit (Energy $\propto Technology^2$ [20]) energy, respectively, which is 32.22%, 27.03%, 60.10%, 50.93%, 60.48%, 35.05% and 65.28% higher as compared to proposed work as observed from Table II. The proposed SDP8T-IMC architecture performs reliable IMC operation without any disturbance and half select issue from 1 V to below 0.4 V. Therefore, the energy-efficiency of proposed IMC architecture can be futhure improved by scaling the supply voltage to subthreshold regions. The energy consumption of proposed IMC architecture is improved by 76% when supply voltage scale down from 1V to 0.4 V, as shown in Figure 18.

C. Figure of Merits

The performance of C6T [6], 6TCSRAM [28] 8+T [11], 8T [14] and 10T [20], 12T [12], and 4+2T [9] SRAM based IMC architectures are used to evaluated the IMC architecture in term of compute error rate (CER), energy, throughput, latency, and area in same 65nm CMOS technology. CER is defined by

Figure 18: Simulation Frequency and energy consumption of IMC and READ operation.

the percentage of MC simulation failed during IMBC operations as described in [12]. Energy is defined by the average energy consumption per bit IMBC operation. Throughput is measured in terms of the number of operands computed in a single clock cycle. The number of cycles required to perform read, compute(IMBC), and store operation is defined as latency. The above performance parameters have a trade-off between them. Therefore, Figure of Merit (FOM) [12] is described in equation 2 to compare IMC architecture.

$$FOM_{IMC} = \frac{(1-CER)*Throughput}{Energy*Area*Latency} \quad (2)$$

It is observed from Table II, The IMC architecture based on SDP8T SRAM improve the performance by 4.88×, 1.84×, 3.90×, 1.96×, 5.70×, 2.06× and 4.45×, when compared to C6T [6], 6TCSRAM [28], 8+T [11], 8T [14] and 10T [20], 12T [12], and 4+2T [9] SRAM based IMC architecture, respectively

Figure 19: Array Efficiency with no column interleave and different column interleave structure.

D. Array Efficiency

The bit-interleaving (or termed as column selection) architecture that is universally employed in SRAM to achieve array efficiency [18], is found disabled in the DP8T SRAM cell due to the half-select issue. A non-bit-interleaving architecture required a sense amplifier for each column [17]. Further, the array efficiency of the SRAM array is estimated by using the [17] equation 3 given below, where TTA means the total number of transistor counts in cell array and TTP mean

total transistor count in peripheral circuits (Decoder, drivers, precharge ckt, and sense amplifiers).

$$AE = \frac{TTA}{TTA+TTP} \quad (3)$$

It is found that the non-bit-interleaving architecture has lower array efficiency due to the increased peripheral circuits (require a sense amplifier for each column). However, the proposed SDP8T SRAM support bit interleave structure. Therefore, the proposed SDP8T SRAM has higher array efficiency when compared to the DP8T SRAM [16] as shown in Figure 19. The array efficiency of SDP8T SRAM can be further improved by increasing the size of column interleaving(Col. int).

E. Performance of IMC architecture

Table III shows the comparison of the proposed scheme with other available methods for logic/BCAM operations. Figure 18 shows the simulated max frequency and average energy consumption of IMC operation between two operands in two rows and read operation at 27 °C. The difference between read and IMC maximum operating frequency is only 12% over a wide supply voltage range, as shown in Figure 18. The proposed SDP8T-IMC architecture can be operated on a maximum operating frequency of 1050MHz in the TT process corner at VDD = 1 V, as shown in Figure 18. The proposed architecture is 2.23× faster than the work reported in [14] using 65nm technology and 3.15× faster than the work reported in [11] using 45nm technology as observed in Table III. The IMBC operation consumes approximately twice the energy consumed by the read, as shown in Figure 18. This is because the operands of the IMBC operations are two words. In terms of BCAM, the proposed architecture consumes 0.60fJ/search/bit energy at 1 V.

Table III: Comparison with the reported IMC architecture for logic/BCAM works

		This Work	TVLSI'20 [14]	JSSC'18 [9]	JSSC'16 [6]	TCAS'18 [11]
Function		SRAM Logic BCAM	SRAM Logic CAM	SRAM Logic CAM	SRAM Logic CAM	SRAM Logic
Technology(nm)		65nm	65nm	55nmDDC	28 nm	45nm
SRAM Cell Type		8T	8T	4+2T	6T	8T
SRAM Array Size		32X32 (1kb)	128X128 (16kb)	128X128 (16kb)	64X64 (4kb)	NA
Logic	Freq. (MHz)	1050(1V)	469(1V)	230(0.8V)	594(1V)	333(1V)
	Avg. Energy (fJ/bit)	11.04(1V)	16.6(0.8V)	18.5(0.8V)	NA	17.25
BCAM Energy ¹ (fJ/bit/search)		0.602(1V)	0.85(1.2V)	0.45(0.8V)	0.41(0.75V)	NA
Number of Inputs		Four-input Logic operation	Four-input Logic operation	Two-input Logic operation	Two-input Logic operation	Two-input Logic operation

¹ Divided by array size.

VI. CONCLUSION

A local bit-line sharing Dual-Port 8T (SDP8T) SRAM with Virtual V_{SS} for energy-efficient IMC architecture is proposed with 26.49% and 52.63% improvement in write margin and leakage power, respectively, and 62.97% reduction

in read energy, at the cost of 31.19% reduction in HSNM when compared to 10T SRAM. The FOM of the Proposed SRAM is $2.005\times$, $1.11\times$, $2.20\times$, and $1.56\times$ higher when compared to recently reported C6T, DP8T, Tr.8T, and 10T SRAM, respectively. The half-select issue is resolved in the proposed SDP8T based IMC architecture by using pass-gate transistors. Therefore, the proposed SDP8T SRAM can be operated on a wide dynamic supply voltage from 0.4 V to 1 V and obtain higher array efficiency as compared to DP8T SRAM. A reliable IMC operation is performed through SDP8T-IMC architecture without any compute disturbance. The Proposed SDP8T-IMC architecture has $2\times$ improvement in throughput and 60.48% reduction in average energy consumption per bit IMC operation with similar latency when compared to 10T SRAM-based IMC architecture. The FOM of the proposed SDP8T-IMC architecture is improved by $4.88\times$, $1.84\times$, $3.90\times$, $1.96\times$, $5.70\times$, $2.06\times$ and $4.45\times$, when compared to recently reported C6T, 6TCSRAM, 8+T, 10T, 12T and 4+2T SRAM based IMC architecture, respectively. The proposed SDP8T-IMC architecture shows excellent potential for developing a low-power and reliable computing system.

REFERENCES

- [1] S. Petrenko, A. Asadullin, and A. Petrenko, "Evolution of the von neumann architecture," *Protect. Inf. Inside*, vol. 2, no. 74, pp. 18–28, 2017.
- [2] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L. Chen, B. Zhang, and P. Deaville, "In-memory computing: Advances and prospects," *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43–55, 2019.
- [3] Y. Ma, Y. Du, L. Du, J. Lin, and Z. Wang, "In-memory computing: The next-generation AI computing paradigm," in *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, 2020, pp. 265–270.
- [4] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang, and M.-F. Chang, "Challenges and trends of SRAM-based computing-in-memory for AI edge devices," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1773–1786, 2021.
- [5] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic RAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 3, pp. 470–483, 2018.
- [6] S. Jeloka, N. B. Akes, D. Sylvester, and D. Blaauw, "A 28 nm configurable memory (TCAM/BCAM/SRAM) using push-rule 6T bit cell enabling logic-in-memory," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, 2016.
- [7] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, 2017.
- [8] C. Eckert, X. Wang, J. Wang, A. Subramaniyan, R. Iyer, D. Sylvester, D. Blaauw, and R. Das, "Neural cache: Bit-serial in-cache acceleration of deep neural networks," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018, pp. 383–396.
- [9] Q. Dong, S. Jeloka, M. Saligane, Y. Kim, M. Kawaminami, A. Harada, S. Miyoshi, M. Yasuda, D. Blaauw, and D. Sylvester, "A 4+ 2T SRAM for searching and in-memory computing with 0.3-V VDDmin," *IEEE Journal of Solid-State Circuits*, vol. 53, pp. 1006–1015, 2018.
- [10] S. Srinivasa, A. K. Ramanathan, J. Sundaram, D. Kurian, S. Gopal, N. Jain, A. Srinivasan, R. Iyer, V. Narayanan, and T. Karnik, "Trends and opportunities for SRAM based in-memory and near-memory computation," in *2021 22nd International Symposium on Quality Electronic Design (ISQED)*, 2021, pp. 547–552.
- [11] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling in-memory boolean computations in CMOS static random access memories," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4219–4232, Dec 2018.
- [12] N. Surana, M. Lavania, A. Barma, and J. Mekie, "Robust and high-performance 12-T interlocked SRAM for in-memory computing," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2020, pp. 1323–1326.
- [13] A. K. Rajput and M. Pattanaik, "Implementation of boolean and arithmetic functions with 8T SRAM cell for in-memory computation," in *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020, pp. 1–5.
- [14] Z. Lin, H. Zhan, X. Li, C. Peng, W. Lu, X. Wu, and J. Chen, "In-memory computing with double word lines and three read ports for four operands," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 5, pp. 1316–1320, 2020.
- [15] J. Wang, X. Wang, C. Eckert, A. Subramaniyan, R. Das, D. Blaauw, and D. Sylvester, "A 28-nm compute SRAM with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, 2019.
- [16] C. Yu, T. Yoo, T. T.-H. Kim, K. C. Tshun Chuan, and B. Kim, "A 16k current-based 8T SRAM compute-in-memory macro with decoupled read/write and 1-5bit column ADC," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, 2020, pp. 1–4.
- [17] M. E. Sinangil, N. Verma, and A. P. Chandrakasan, "A 45nm 0.5V 8T column-interleaved SRAM with on-chip reference selection loop for sense-amplifier," in *2009 IEEE Asian Solid-State Circuits Conference*, 2009, pp. 225–228.
- [18] L. Wen, X. Cheng, K. Zhou, S. Tian, and X. Zeng, "Bit-interleaving-enabled 8T SRAM with shared data-aware write and reference-based sense amplifier," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 7, pp. 643–647, 2016.
- [19] W. A. Simon, Y. M. Qureshi, A. Levisse, M. Zapater, and D. Atienza, "Blade: A bitline accelerator for devices on the edge," in *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, 2019, pp. 207–212.
- [20] A. Biswas and A. P. Chandrakasan, "Conv-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan 2019.
- [21] N. H. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective*. Pearson Education India, 2015.
- [22] A. Islam and M. Hasan, "Leakage characterization of 10T SRAM cell," *IEEE Transactions on Electron Devices*, vol. 59, no. 3, pp. 631–638, 2012.
- [23] D. Schinkel, E. Mensink, E. A. Klumperink, A. J. M. van Tuijl, and B. Nauta, "A low-offset double-tail latch-type voltage sense amplifier," in *18th Annual Workshop on Circuits, Systems and Signal Processing, ProRISC 2007*. STW, 2007, pp. 89–94.
- [24] J. Park, T. W. Oh, and S.-O. Jung, "Pmos pass gate local bitline SRAM architecture with virtual Vss for near-threshold operation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 4, pp. 1079–1083, 2020.
- [25] A. K. Rajput and M. Pattanaik, "Energy efficient 9T SRAM with r/w margin enhanced for beyond von-neumann computation," in *2020 24th International Symposium on VLSI Design and Test (VDATE)*. IEEE, 2020, pp. 1–4.
- [26] P. Singh and S. K. Vishvakarma, "Ultra-low power, process-tolerant 10T (PT10T) SRAM with improved read/write ability for Internet of Things (IoT) applications," *Journal of Low Power Electronics and Applications*, vol. 7, no. 3, p. 24, 2017.
- [27] P. Sanvale, N. Gupta, V. Neema, A. P. Shah, and S. K. Vishvakarma, "An improved read-assist energy efficient single ended ppn based 10T SRAM cell for wireless sensor network," *Microelectronics Journal*, vol. 92, p. 104611, 2019.
- [28] J. Chen, W. Zhao, Y. Wang, and Y. Ha, "Analysis and optimization strategies toward reliable and high-speed 6T compute SRAM," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1520–1531, 2021.
- [29] A. T. Do, C. Yin, K. S. Yeo, and T. T.-H. Kim, "Design of a power-efficient CAM using automated background checking scheme for small match line swing," in *2013 Proceedings of the ESSCIRC (ESSCIRC)*. IEEE, 2013, pp. 209–212.