IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 December 2023, accepted 26 December 2023, date of publication 28 December 2023, date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347913

RESEARCH ARTICLE

Hardware Efficient Transposable 8T SRAM for Orthogonal Data Access

DAIN CHON, (Student Member, IEEE), AND WOONG CHOI, (Member, IEEE)
Department of Electrical Engineering, Sookmyung Women's University, Seoul 04310, South Korea

Corresponding author: Woong Choi (woongchoi@sookmyung.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant NRF-RS-2023-00252402; in part by the Korea Evaluation Institute of Industrial Technology (KEIT) Grant funded by the Korean Government [Ministry of Trade, Industry and Energy (MOTIE)] under Grant 20009972; in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korean Government (MSIT) under Grant 2021-0-00903, Grant 2021-0-00875, and Grant IITP-2023-RS-2022-00164800; and in part by the Sookmyung Women's University Research under Grant 1-2203-2035.

ABSTRACT This paper presents a novel 8T SRAM bitcell-based transposable (TP) memory supporting both row-wise and column-wise data access. The proposed TP-SRAM enables orthogonal data access with additional diagonal word-lines and a low-complexity addressing scheme. To reduce cell array area overhead, the proposed TP-SRAM adopts a bitcell structure that can share all aspects of layout with adjacent cells like standard 6T-SRAM. We also propose a bidirectional barrel shifter based on dynamic logic gates to minimize the hardware cost required for the TP addressing scheme. In the proposed bidirectional barrel shifter, area and delay are minimized by using two complementary dynamic inverting MUXs that can balance the number of NMOS and PMOS transistors. The proposed 16Kb TP-SRAM implemented in 28nm CMOS technology has 17% reduced power, 52% faster operation delay, and 39% smaller area compared to the state-of-the-art.

INDEX TERMS Transposable, SRAM, barrel shifter, dynamic gate.

I. INTRODUCTION

There is a growing demand for transposable (TP) memory in modern neuromorphic processors [1], [2], [3], [4]. TP memory reduces design complexity and enables low-cost hardware implementation, making it an attractive option. It is also becoming popular in the field of in-memory-computing (IMC), which aims to reduce energy consumption by minimizing data movement within deep neural network (DNN) processors and simplifying vector operations (such as dot products) [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. Unlike conventional memory, where only bitcells placed in the word-line (WL) direction can be accessed, TP memory allows for selective access to both bitcells distributed in the WL and bit-line (BL) directions. This flexible access to orthogonally distributed data enables fast updating of weight values for synapse arrays during on-chip training with neuromorphic processors. Previous TP memory can be divided into two types: i) row- and column-wise data can be read and written to arbitrary addresses [1], [2], [3], and ii) only limited column-wise read feature is added to the conventional memory for IMC. In this paper, we focus on the first type of TP memory, which allows fully functional orthogonal data access [22], [23], [24], [25].

TP memories are typically implemented as static random-access memory (SRAM), which provides flexibility in data access direction for on-chip internal operations. In [1], each neuron in the synapse array is implemented as a TP-SRAM bitcell. Two access transistors are added to a standard 6-transistor (6T) SRAM cell to form two orthogonal WL and BL pairs. However, in the bitcell layout, two additional BLs must be drawn in the WL direction, which significantly increases the area due to the lack of margin in the conventional layout. Additionally, it requires additional peripheral circuitry to control the added WL and BL pair. To address the area overhead, a diagonal WL activation-based TP addressing scheme was introduced in [2] and [3]. This design approach uses traditional 6T SRAM cells but rotates the data stored in each row of the cell array by a different amount so that each bit of BL-direction data is diagonally laid out.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Hossein Moaiyeri.

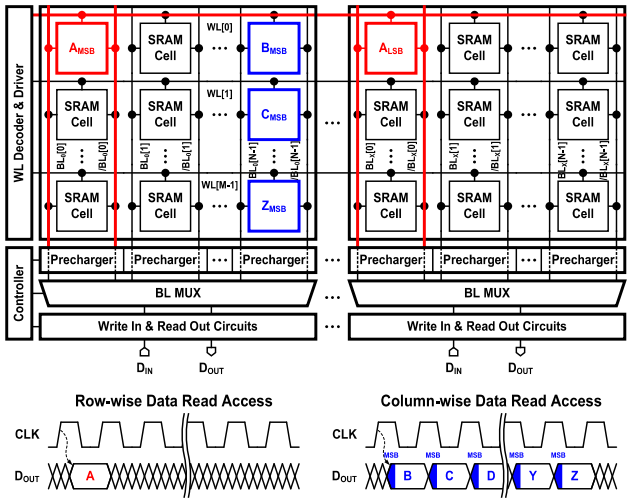
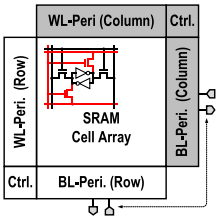
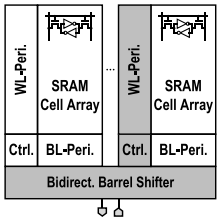


FIGURE 1. Two directional data access in the conventional column-interleaved SRAM.

Data I/O	Non-integrated I/O	Integrated I/O
References	[1]	[2], [3]
Array structure		
Cell type	Transposable 8T-SRAM Cell	Conv. 6T-SRAM Cell
Data mapping	Straightforwardly	Relocation (Diagonally)
Area overhead	Bitcell, Additional WL/BL-Peri.	Barrel Shifter, Additional WL-Peri.
Interface	Long-distant I/Os	Conventional way

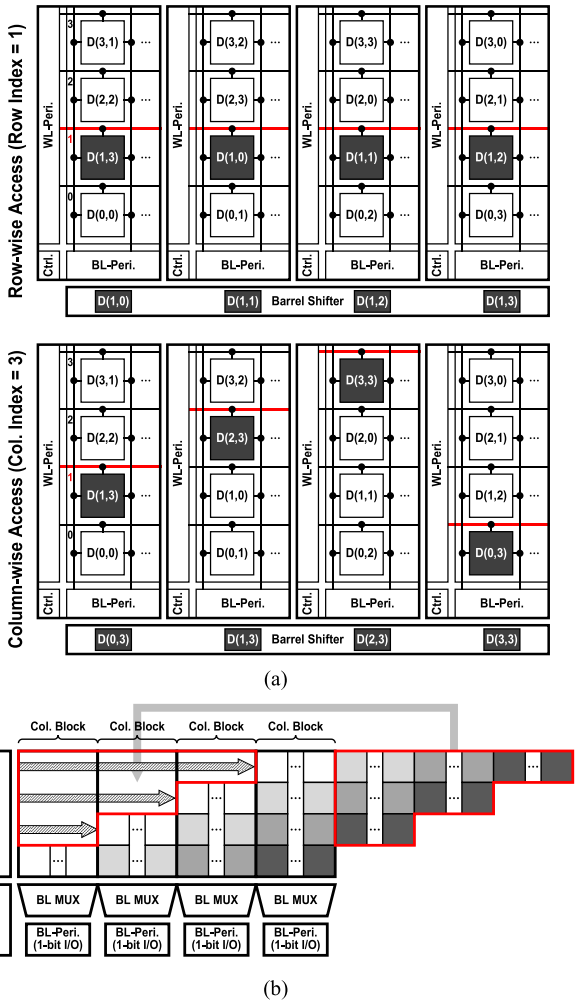
(a)

(b)

FIGURE 2. (a) Summary of the previous TP-SRAM and (b) the conventional TP addressing scheme for the integrated I/O-based approach.

By selectively activating WL in the row or diagonal direction, the TP memory access function is enabled. However, the inefficient memory structure for diagonal WL implementation introduces significant area overhead. Moreover, the increased complexity of diagonal WL address decoding and the overhead of bidirectional barrel shifters are challenges that have not been considered in previous studies.

In this paper, we propose a hardware-efficient TP-SRAM design method. To facilitate orthogonal data access, we propose an improved TP addressing scheme that simplifies the address decoding. By adopting the novel 8T SRAM bitcells and diagonal WL bridges, the excessive area overhead of the TP-SRAM is alleviated. To minimize the data reordering



(a)

(b)

FIGURE 3. (a) Operation principle of the integrated I/O based TP-SRAM, and (b) conventional TP addressing scheme in column-interleaved array structure.

cost for the TP addressing scheme, an inverting multiplexer (MUX) based bidirectional barrel shifter is also proposed. By alternating two complementary dynamic logic gates in each stage of the barrel shifter, the hardware inefficiencies of traditional domino logic gate-based designs are eliminated. This is advantageous for optimizing the entire TP-SRAM macro with low area and high speed compared to the conventional static CMOS implementations. The numerical results show that the proposed 16Kb TP-SRAM has 17% reduced power, 52% faster operation latency, and 39% smaller area compared to the state-of-the-art.

The rest of the paper is organized as follows. Section II provides an overview of the conventional TP-SRAM. The proposed TP-SRAM and bidirectional barrel shifter are presented in Section III. In Section IV, the experimental results are drawn with the comparisons with state-of-the-art. Finally, conclusions are drawn in Section V.

II. PRELIMINARIES

Fig. 1 shows the row-wise and column-wise data access in the conventional column-interleaved SRAM. In this ex-ample,

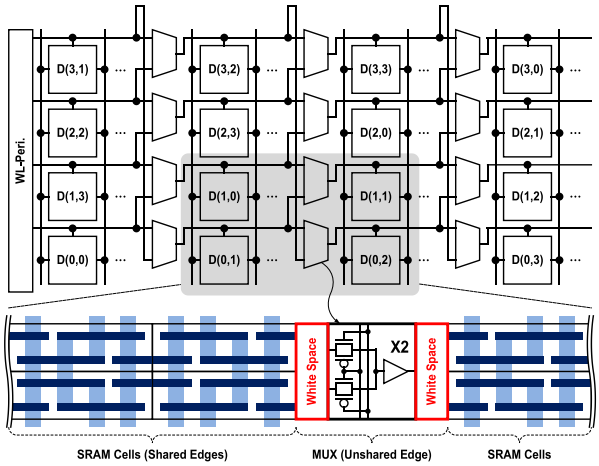


FIGURE 4. Conventional row-transition MUX-based TP-SRAM [3].

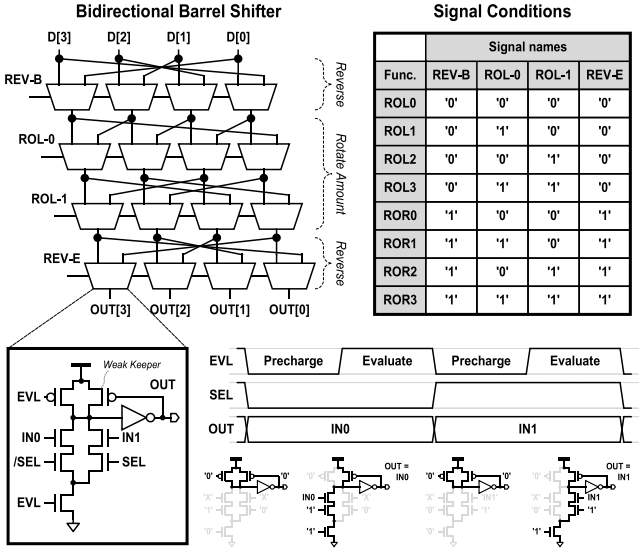


FIGURE 5. Conventional domino-logic gate-based bidirectional barrel shifter [16].

WL and BL are oriented row-wise and column-wise, respectively. Due to the WL activation-based data access, the row-wise data (A in Fig. 1) can be accessed within a single cycle. On the other hand, to process the column-wise data (from B to Z in Fig. 1), multi-cycle memory access is required while selectively buffering the target data bit. To cope with the limited data access direction of conventional memory, many studies on TP memory have been performed. As presented in Fig. 2, previous works can be classified according to the input/output (I/O) interface. The TP-SRAM with non-integrated I/O [1] is based on 8T bitcells with additional WL and BL pairs orthogonal to the conventional orientation. As mentioned earlier, the increased area of the bitcell itself and the additional peripheral circuitry introduce significant area overhead. Moreover, when orthogonal data is selectively used in a computation module, the long I/O distance between orthogonal data increases routing complexity.

To overcome the challenges of the non-integrated I/O-based design, the integrated I/O-based TP-SRAM has been studied in [2] and [3]. In this approach, the TP memory access

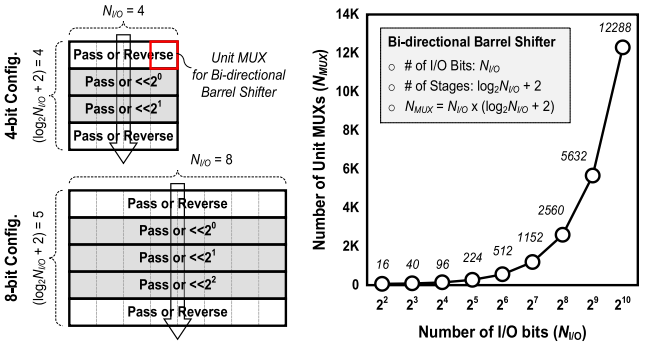


FIGURE 6. The number of required unit MUX based on the number of I/O bits of the bidirectional barrel shifter.

function is enabled by selectively accessing either row-wise distributed row data or diagonally distributed column data. An example of an integrated I/O-based TP-SRAM is shown in Fig. 3. In Fig. 3(a), each square represents a bitcell, and the numbers shown inside are the row index and column index of the data to be stored, respectively. As shown in Fig. 2(b) and Fig. 3(a), by right rotating the data in each row by increasing the amount by 1, bitcells with the same column index are placed in different columns. That is, data having the same row index are positioned in the same row, while data having the same column index are positioned diagonally. As shown in Fig. 3(a), reordered data can be selectively accessed row-wise or column-wise via bank-separated WL peripherals (WL-Peri.) and bidirectional barrel shifters. For column-interleaved SRAM structures, this data relocation is organized as shown in Fig. 3(b). Unlike Fig. 2(b), which rotates in units of 1 bit, it is rearranged in units of data bundles connected to one BL MUX. Compared to non-integrated I/O-based TP-SRAM, where the bitcell area is significantly increased by orthogonal WL and BL wiring, this approach maintains the area efficiency of a standard 6T SRAM bitcell. However, the peripheral circuitry (WL-Peri.) and controllers added to each memory bank for diagonal WL activation incur significant area overhead. To mitigate the area overhead of the multi-bank-based approach [2], TP-SRAM with diagonal WL inside a cell array is proposed in [3]. As shown in Fig. 4, row-transition MUXs are inserted into the cell array at regular intervals to selectively set the WL activation direction to row or diagonal. Through this, peripheral circuits and controllers for WL activation can be integrated, but a large area overhead still occurs because the uniformity of the SRAM cell array that maximizes layout efficiency is broken. Unlike SRAM cells, which share all sides with other cells, the added MUX creates unnecessary empty space (white space in Fig. 4) in the cell array.

For the integrated I/O-based TP-SRAM (Fig. 2(a)), a bidirectional barrel shifter is required to support the TP addressing scheme. Fig. 5 shows the conventional bidirectional barrel shifter [16]. Here, the signal names REV, ROL, and ROR indicate reverse, rotate-left, and rotate-right, respectively. Accordingly, in the 4-bit bidirectional barrel shifter (Fig. 5), each row of the 4 × 4 MUX array represents a reverse or

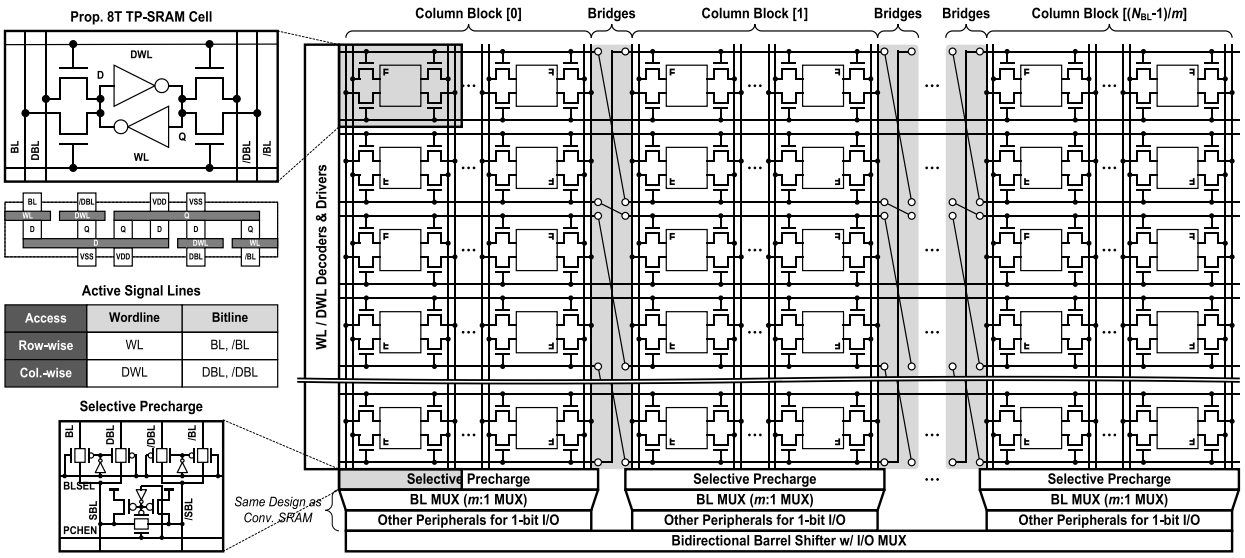


FIGURE 7. Overall array structure of the proposed 8T TP-SRAM including bitcell schematic, layout, and signal conditions.

rotate-left stage [16], [17]. Like the binary number representation, the total amount of rotation is determined by a combination of intermediate stages that per-form rotation in power-of-two [18], [19], [20]. On the other hand, the direction of rotation is decided by the operation of the two inversion stages. Detailed signal conditions for each rotation can be found in Fig. 5. The hardware cost of the bidirectional barrel shifter is affected by the total number of stages and the total number of unit MUXs. Fig. 6 shows the total number of stages and the total number of unit MUXs required for each number of I/O bits in the bidirectional barrel shifter. When the number of I/O bits is $N_{\mathrm{I/O}}$, the total number of stages is '$\log_2 N_{\mathrm{I/O}} + 2$ (two reverse stages)', and the total number of unit-MUXs (NMUX) is multiplied by $N_{\mathrm{I/O}}$. As shown in Fig. 6, as the number of I/O bits ($N_{\mathrm{I/O}}$) increases, the total number of unit MUXs (N_{MUX}) in the bidirectional barrel shifter increases exponentially. Also, the required number of stages increases linearly. For this reason, smaller, faster domino-logic gate-based MUXs are preferred over static CMOS-based MUXs in the barrel shifter designs. However, the high-power consumption and the exponential increase in the area of the domino-logic gate-based barrel shifter are one of the important challenges for low-cost TP memory design. The following sections present a hardware-efficient TP-SRAM design method using i) a novel 8T SRAM bitcell, ii) a low-complexity TP addressing scheme, and iii) a low-cost bidirectional barrel shifter.

III. PROPOSED 8T TRANSPOSABLE SRAM

A. ARRAY STRUCTURE

Fig. 7 shows the overall array structure of the proposed TP-SRAM. The proposed TP-SRAM is based on an integrated I/O-based TP memory structure (Fig. 2(a)), so it includes additional WLs in the diagonal direction and a bidirectional barrel shifter. In addition, column-interleaving is applied for soft error immunity and the efficiency of peripheral circuit

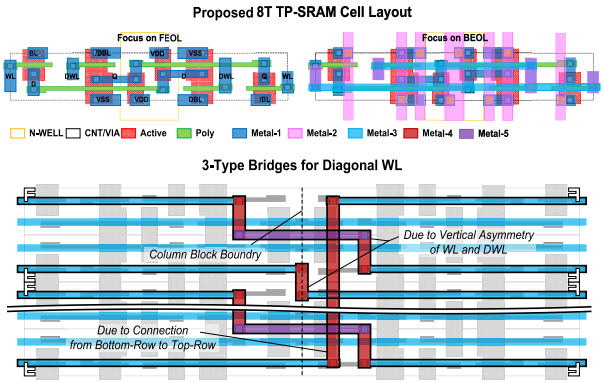


FIGURE 8. Layout of the proposed 8T TP-SRAM cell and three types of bridges.

integration [21]. As presented in Fig. 7, the proposed bitcell includes two additional transistors for diagonal WL (DWL) connection to a standard 6T SRAM cell. It has the same structure as a standard dual-port 2RW 8T SRAM cell. However, the proposed 8T TP-SRAM cell uses WL, BL, and /BL signals for row-wise access and DWL, DBL, and /DBL signals for column-wise access. In the column-interleaving structure, when columns sharing one BL MUX in a cell array are grouped into one column block, 1-bit data is accessed from one column block. As shown in the cell array in Fig. 7, DWL is connected horizontally within the column block like conventional WL but connected diagonally downward through a bridge at the column block boundary. For DWL connections at column block boundaries, three types of bridges are required due to i) vertical asymmetry of WL and DWL, and ii) bottom row and top row connections. The detailed layout of the proposed 8T TP-SRAM cell array with three types of bridges is illustrated in Fig. 8. By placing WL instead of DWL on the left and right edges of bitcells, diagonal DWL connections through bridges are possible without extra space.

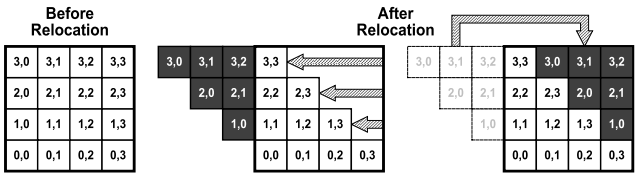


FIGURE 9. Proposed TP addressing scheme for low control complexity.

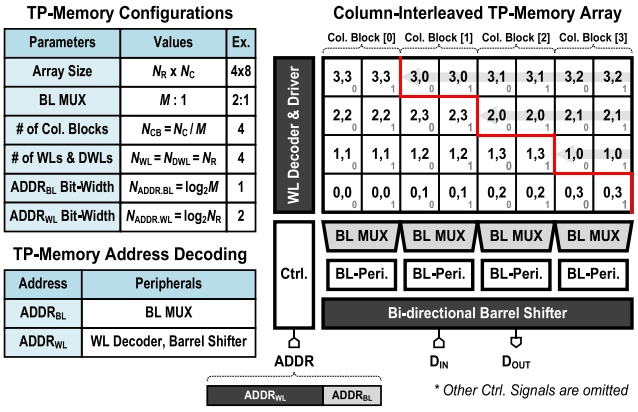


FIGURE 10. TP memory configuration and related addressing scheme.

In the proposed bitcell with two WLs (WL and DWL), only one WL can be active, so BLs of the same polarity (BL-DBL and /BL-/DBL) can be hardwired to two different BL pairs (BL pair, DBL pair). However, this configuration increases the BL capacitance, which increases the BL switching power and reduces the stability of the read operation. For this reason, the proposed TP-SRAM uses selective precharge, including MUX for BL and DBL pairs, as shown in Fig. 7. The following MUX for column-interleaving and peripheral circuits for read/write can be designed in the same way as the conventional SRAM. For this reason, the proposed 8T TP-SRAM cell is similar to the conventional 6T SRAM cell in terms of stability, power consumption, and delay.

B. IMPROVED TRANSPOSABLE ADDRESSING SCHEME

In the conventional TP addressing scheme, the WL index for row-wise access matches the row index, but the DWL index for column-wise access does not match the column index, as shown in Fig. 3(a). This increases the complexity of address decoding to find the index of the DWL that needs to be activated. Fig. 9 shows the proposed TP addressing scheme. Unlike the conventional approach (Fig. 2(b)), when the data rotation direction is changed to the left, the row and column indices of the data in the leftmost column have the same value. This reduces DWL decoding complexity by making the decoding circuit configuration of WL and DWL identical.

To describe the WL decoding method for the TP addressing scheme, memory configuration parameters are defined as shown in Fig. 10. In the column-interleaved TP-SRAM cell array, bitcells that belong to the same row and column block are assigned the same row and column index if they are simultaneously active. Here, the gray number marked on the bottom right of each square (bitcell) means the selection

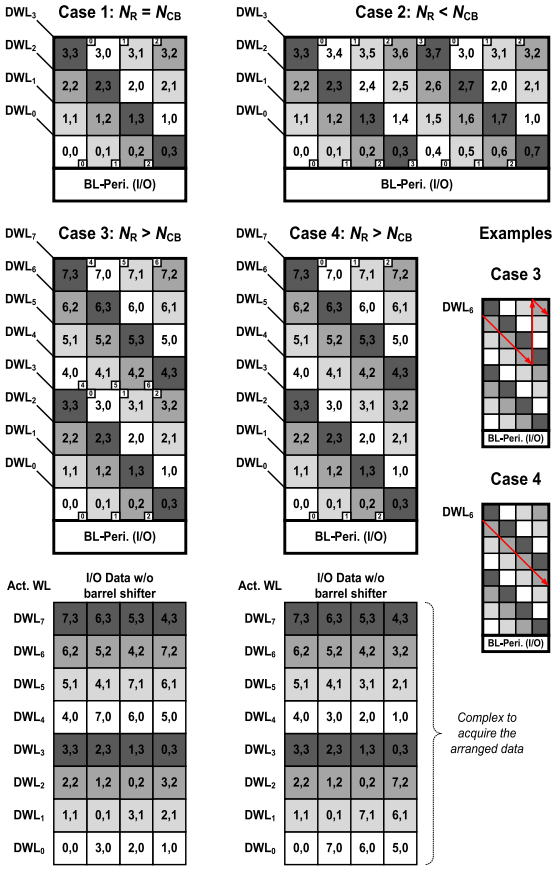


FIGURE 11. Proposed data relocation in cell arrays classified into three types according to horizontal and vertical lengths.

signal of BL MUX. This notation is used because bitcells with the same row and column index are activated together by WL or DWL, and their distinction is determined solely by the BL MUX. Therefore, the BL address ($ADDR_{BL}$) within the memory address controls the BL MUX regardless of the access direction. This property is used to simplify the column-interleaved TP-SRAM for various array sizes, as shown in Fig. 11, by omitting the BL address and related parts. In Fig. 11, the numbers in the small white squares indicate the DWL addresses corresponding to the DWL connections from the bottom row to the top row. For row-wise access, the WL address for activating bitcells with the same row index is the same as in conventional SRAM. This row-wise address decoding is independent of cell array size, as the number of physical WLs and the number of row indices are always the same. On the other hand, in column-wise access, the number of physical DWLs (N_R) and the number of column indices (N_{CB}) may differ depending on the array size. Because of this mismatch, TP memory may require virtual DWL addresses.

These features are common in the TP addressing scheme but have not been addressed in previous studies. Thanks to the low complexity of the proposed TP addressing scheme, DWL address decoding is clearly organized as follows. When N_R is equal to N_{CB} (Case 1 in Fig. 11), DWL address decoding is the same as in WL. On the other hand, when N_R is smaller than N_{CB} (Case 2), the DWL address is decoded using only

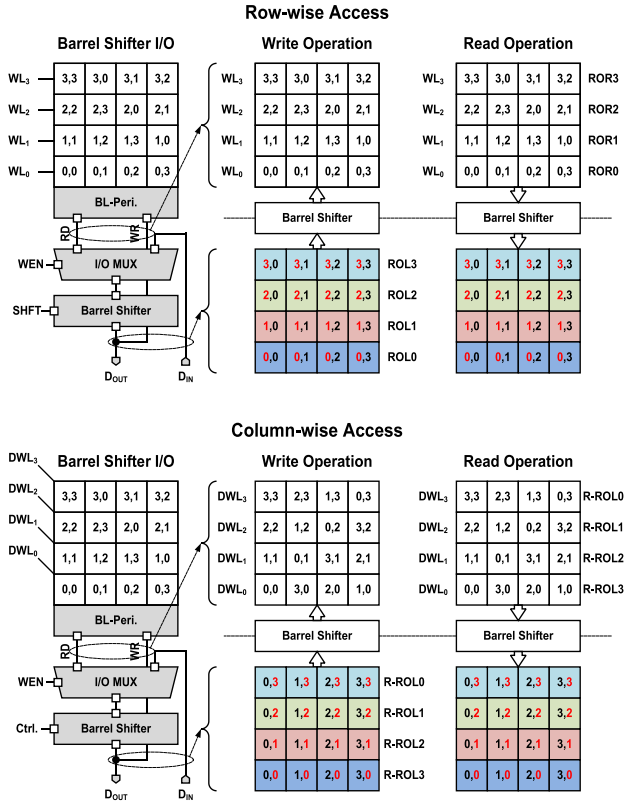
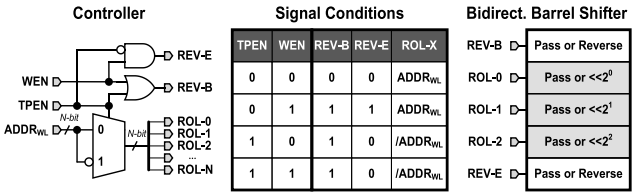


FIGURE 12. Operating principle of the bidirectional barrel shifter for the proposed TP addressing scheme.

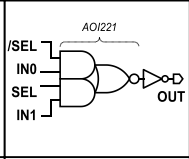
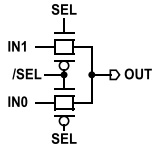
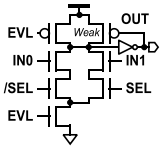


* **WEN**: Write Enable Negative (0: Write, 1: Read), * **TPEN**: Transposable Enable (0: Row-wise Access, 1: Col.-wise Access)

FIGURE 13. Dedicated controller and signal conditions for the proposed TP addressing scheme.

the least significant bits, as many as $\log_2 N_R$ in the column address. For example, if the column-wise address is 7 (111_2), the value of $\log_2 N_R$ is 2, so the DWL address is decoded into the value 3 (11_2) corresponding to the least significant 2 bits. In the same way, column-wise address 3 (011_2) is decoded to the same DWL address. Therefore, the two data assigned to column-wise addresses 3 and 7 are accessed simultaneously when the DWL address is 3. In this case, by adding a MUX stage that uses bits not selected in the column-wise address as selection signals, only the data of the desired column index can be accessed. Meanwhile, when N_R is greater than N_{CB} (Case 3 and Case 4), the DWL address is decoded by extending the column address by $\log_2(N_R/N_{CB})$ bits. For example, the DWL address for column address 3 (11_2) becomes 3 (011_2) and 7 (111_2) by extending the column address by 1 bit. This means that multiple DWL accesses are required to access all bits of a particular column index. Additionally, for this array size ($N_R > N_{CB}$), bridges for DWL connectivity at column block boundaries can be configured in

TABLE 1. Various types of unit MUX for bidirectional barrel shifters.

Type	Static CMOS	Transmission Gate	Domino Logic
Scheme			
Pros	Low Power	Compact Size	Fast Operation
Cons	Large Area	No Self Driving	High Power

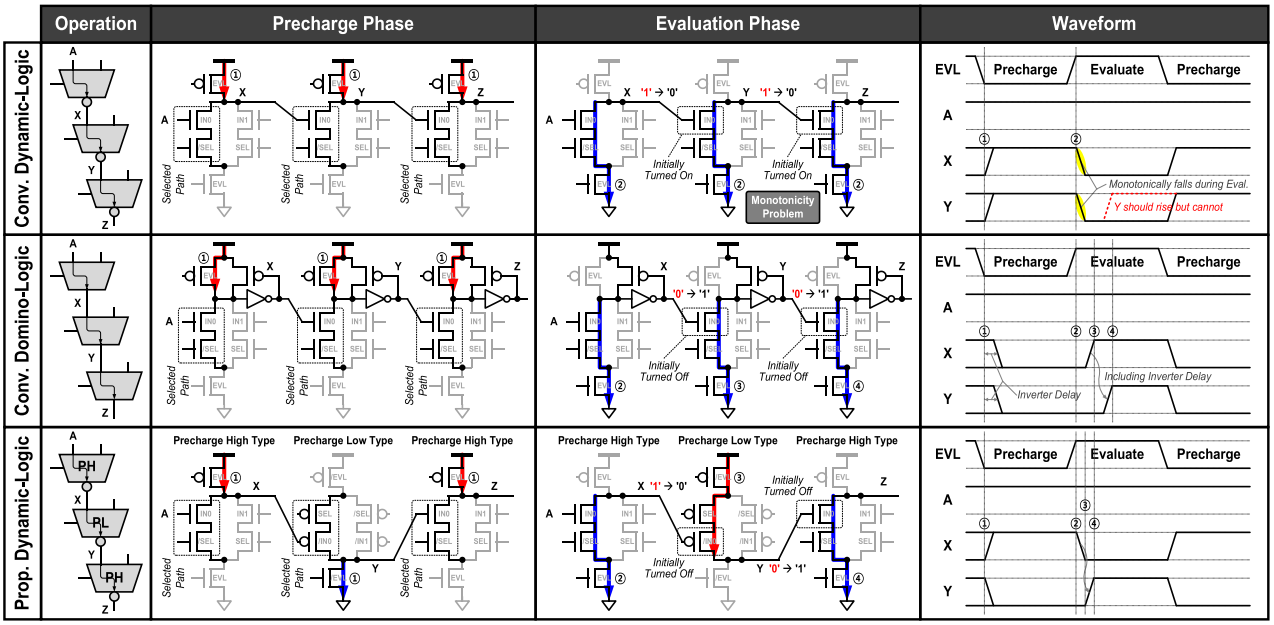
two cases (Case 3 and Case 4). Before rotation by the barrel shifter, the access data for each DWL address is shown below each case in Fig. 11. When the DWL bridge is connected to an array partitioned into squares as in Case 3, data having the same column index can be accessed in a more aligned form com-pared to Case 4. In conclusion, the proposed TP addressing scheme makes the decoding of WL and DWL addresses identical regardless of the array size by simply adjusting the WL address applied to the TP memory.

C. BIDIRECTIONAL BARREL SHIFTER

For the proposed TP addressing scheme, the operating principle of the bidirectional barrel shifter is presented in Fig. 12. In the simplified block diagram on the left side of Fig. 12, the two dotted circles represent the data path before and after passing through the bidirectional barrel shifter. In addition, each I/O data (D_{IN} and D_{OUT}) for the write and read operations is marked with different colors. In the case of row-wise access, sorting is based on the row index, whereas in the case of column-wise access, sorting is based on the column index. For the row-wise write operation, D_{IN} data passed through the I/O MUX is applied to the BL peripheral circuit (write driver) after being rotated to the left (ROL) by the row index. On the other hand, in the case of a row-wise read operation, D_{OUT} data read through the BL peripheral circuit (sense amplifier) is output after being rotated to the right (ROR) by the row index. In column-wise access, data is input and output through the same path, but in a bidirectional barrel shifter, a different form of rotation operation is required. Regardless of the operation type, the barrel shifter rotates to the left by the inverted binary value of the column index for the reversed input value (R-ROL). For example, data for column index 2 (10_2) (colored green in Fig. 12) is reversed and rotated to the left by 1 (01_2), which is an inverted value of 10_2. The left rotation operation after the reversal (R-ROL) of the bidirectional barrel shifter can be accomplished simply by activating only one reverse stage (REV-B stage in Fig. 5). Fig. 13 depicts the dedicated controller and signal conditions for the proposed TP addressing scheme. The control signals for each stage of the bidirectional barrel shifters (REV-B/E, ROL-X) responsible for reversing and rotating are constructed by performing simple logic operations on the WEN and TPEN signals.

The unit MUX constituting the bidirectional barrel shifter can be composed of various types of logic gates. The detailed

TABLE 2. 3-stage 2:1 MUXs using various dynamic circuits.



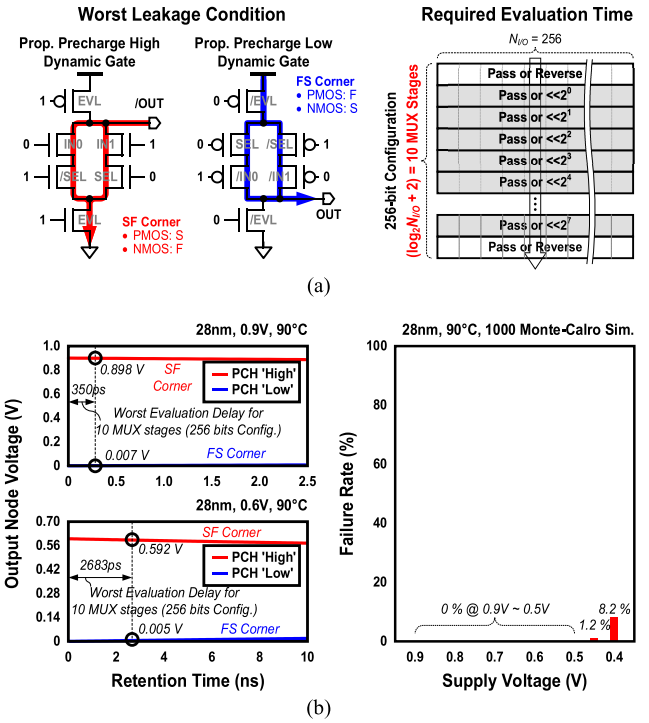


FIGURE 14. (a) Leakage worst condition and required evaluation time. (b) Output node voltage and operational failure rate for the proposed dynamic circuits.

circuitry and pros and cons of each logic gate are summarized in Table 1. In many previous studies of barrel shifters [16], [17], [18], [19], [20], domino logic-based implementations have been favored due to their relatively high speed and small size compared to static CMOS logic. However, the precharge operation that must be preceded every cycle causes high power consumption in the domino logic gate. In the case of the transmission gate, it is not suitable for use in barrel shifters that require multiple MUX stage configurations due to its lack of self-driving capability. To reduce the hardware cost of TP memory, we propose an inverting MUX-based bidirectional barrel shifter. Table 2 shows the 3-stage 2:1 MUXs using various dynamic circuits. Unlike the conventional domino logic gate, the dynamic logic gates are based on an inverting MUX. As shown in the table, dynamic circuits require two-phase operation consisting of precharge and evaluation. During the precharge phase, the output of each MUX stage is fixed at a value of '0' or '1', and one of the two paths in the middle is selected by the MUX selection signal. In this example, the left path is simply selected, and the operating principle is the same in the opposite case. During the evaluation phase, the input signal connected to the selected path is propagated through each MUX stage.

In the conventional dynamic logic-based configuration (1st row in Table 2), the output values of each stage initially turn on the selected path of the next MUX stage from the beginning of the evaluation phase. Therefore, the outputs of all stages fall monotonically regardless of the inputs. This is a monotonicity problem as it prevents the output Y node from rising when input A of the first MUX stage is '1'. In the conventional domino logic-based configuration (2nd row in Table 2), a static CMOS inverter is added to the dynamic gate to solve the monotonicity problem. The main difference, in this case, is that the output of each stage formed in the precharge phase deactivates the selected path of the next MUX stage at the beginning of the evaluation phase. Accordingly, input A of the first MUX stage is sequentially propagated through each MUX stage. In the proposed dynamic logic-based configuration (3rd row in Table 2), two types of MUXs with duality are alternately arranged. As with

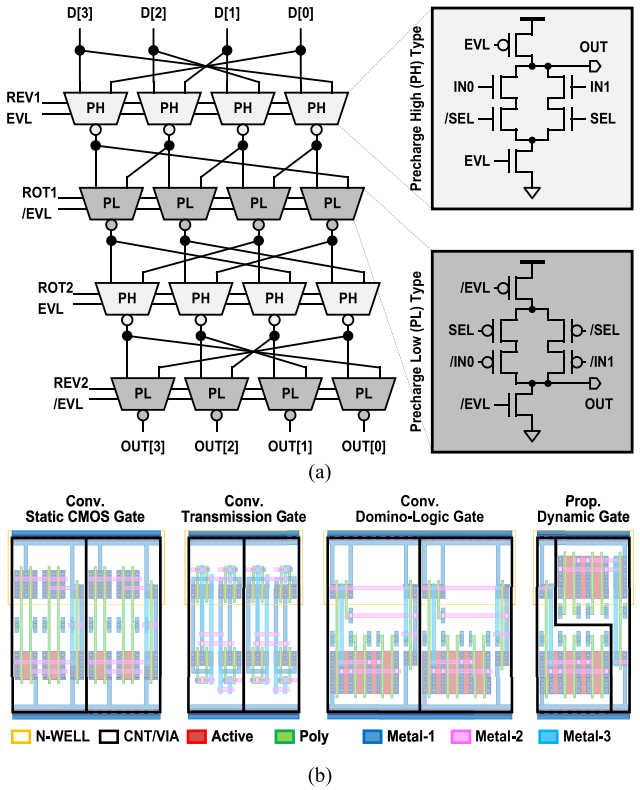


FIGURE 15. (a) Proposed inverting MUX-based bidirectional barrel shifter. (b) Layout comparison of two-stage MUX.

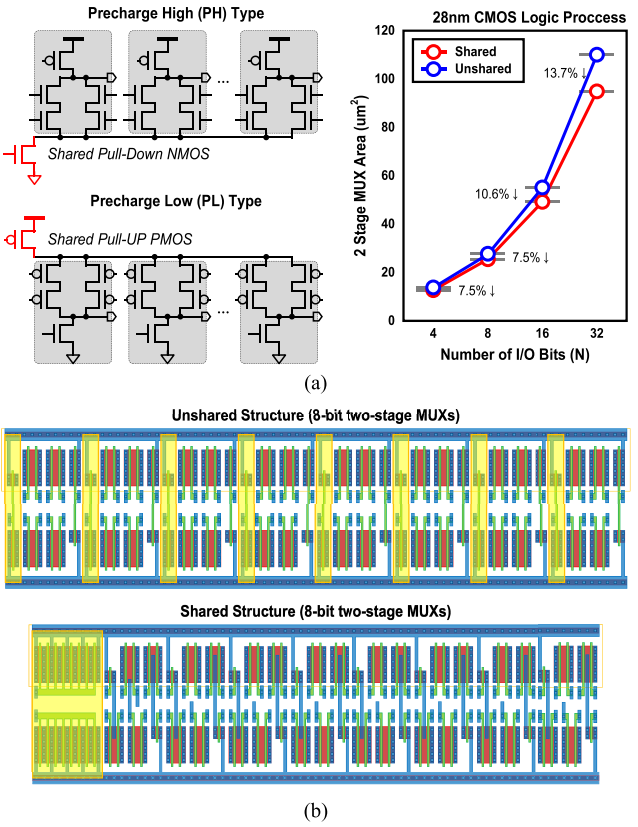


FIGURE 16. (a) Shared pull-up and pull-down path in the proposed dynamic gate and (b) comparison of the 2 MUX stages layout.

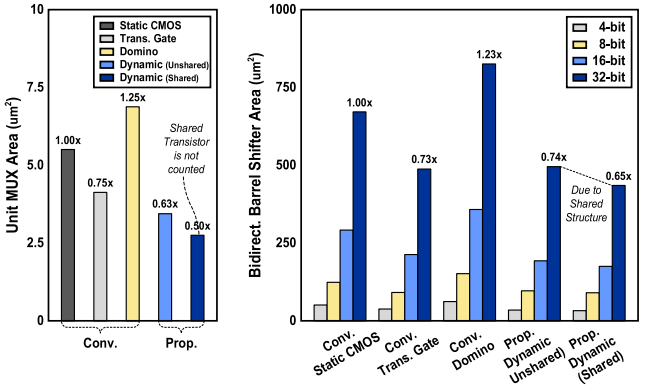


FIGURE 17. Area comparison of the unit MUX and bidirectional barrel shifter.

domino logic gates, it prevents the output of each M UX stage from falling monotonically early in the evaluation phase.

Compared to the conventional domino-logic gate, the proposed dynamic logic gate does not have a keeper, but the effect of leakage current is insignificant due to its fast-operating speed and three-stacked transistors in the leakage path. Fig. 14 presents the worst leakage condition and the corresponding output node voltage and operational failure rate of the MUX. The worst-case leakage in the proposed dynamic gate is when only one transistor is turned on in the middle MUX select paths during the evaluation phase. Here, SF and FS corners are applied to the proposed precharge high and low dynamic gates, respectively. As shown in Fig. 14(b), as a result of Monte-Carlo simulation to determine the impact of process changes, it was confirmed that operation errors begin to occur below the supply voltage of 0.45V. In addition, the required evaluation time of the bidirectional barrel shifter was investigated to find out the effect of the worst leakage in the evaluation phase on the operational stability. As mentioned in Fig. 6, the total number of stages is determined by the number of input bits of the bidirectional barrel shifter, which requires 10 stages for 256 bits. As shown in Fig. 14(b), under the worst evaluation delay conditions (SS corner, $-10°C$), the output voltage change of the proposed dynamic gate is negligible. Since the total number of stages required for a bidirectional barrel shifter increases gradually on a logarithmic scale, the proposed dynamic gate can easily construct multiple MUX stages with short evaluation times.

Fig. 15(a) shows the overall structure of the proposed dynamic logic gate-based bidirectional barrel shifter. By alternately arranging two types of dynamic gates with duality, the number of PMOS and NMOS is balanced compared to conventional gates. Therefore, the proposed structure improves the area efficiency by facilitating the L-shaped layout of the two-stage MUX as shown in Fig. 15(b). To further maximize area efficiency, the proposed dynamic gate can share transistors placed in the pull-up or pull-down path as presented in Fig. 16(a). The layout difference before and after sharing is shown in Fig. 16(b), and the area reduction rate according to the number of bits is shown in the right graph of Fig. 16(a). In the shared structure, the area reduction rate

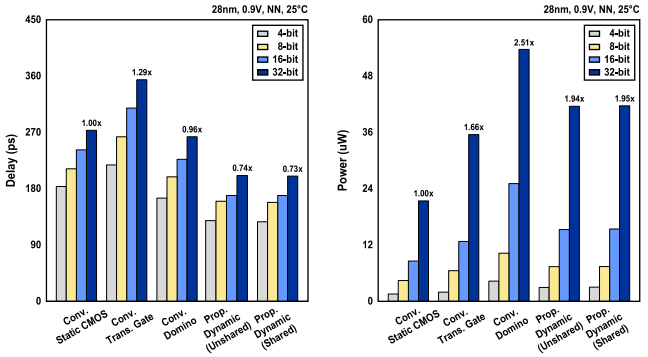
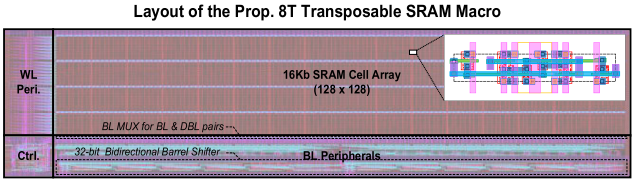


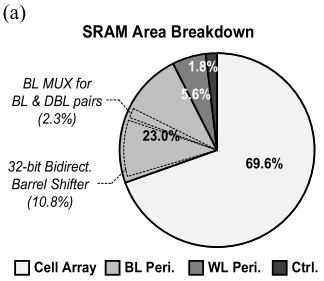
FIGURE 18. Delay and power comparison of the bidirectional barrel shifter.



(a)

Design Summary

Process	28nm CMOS
Voltage	0.9V
Capacity	16Kb (128 WLs, 128 BLs)
Structure	4:1 Column-Interleaving
Cell Type	Prop. 8T SRAM Cell
Cell Size	2.056 x 0.348 (0.715) μm^2
Macro Area	263.168 x 46.632 (12272.1) μm^2
Routing	5 Metal Layers



(b)

FIGURE 19. (a) Macro-level layout, (b) design summary, and (c) area breakdown of the proposed TP-SRAM.

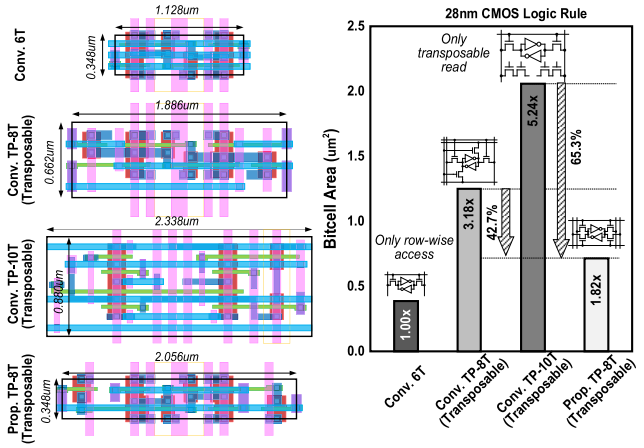


FIGURE 20. Comparisons of the TP-SRAM bitcell area with the conventional 6T SRAM.

TABLE 3. Comparison of bitcells supporting transposable function.

	Conv. 6T	Conv. TP-8T [1]	Conv. TP-10T [22]	Prop. TP-8T
1)Bitcell				
2)Size (PU:PG:PD)	1.0 : 1.3 : 1.6	1.0 : 1.3 : 1.6	1.0 : 1.3 : 1.3	1.0 : 1.3 : 1.6
Transposable	Not supported	Both read & write	Only read	Both read & write
Area (um^2)	0.39 (1.00x)	1.25 (3.18x)	2.06 (5.24x)	0.72 (1.82x)
3)RSNM (mV)	170 / 15.6	170 / 15.5	367 / 10.4	170 / 15.5
3)WSNM (mV)	601 / 48.4	599 / 49.5	614 / 44.3	599 / 49.5
4)Assist conflict	WL-based	WL&CVDD-based	WL-based	WL-based

1) The part responsible for the transposable function is marked with a red line.
2) The conventional TP-10T, which has two decoupled-read ports, reduced the size of the PD transistor to improve write-ability.
3) The mean and standard deviation of SNM are measured under 0.9V, 25°C, and Monte=2000. (left: mean / right: std.)
4) When using column-interleaving structure

increases as the number of bits increases because there is no need for spacing between the pull-up and pull-down paths and adjacent MUX selection parts.

IV. NUMERICAL RESULTS

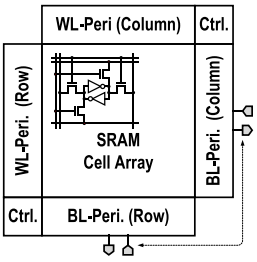
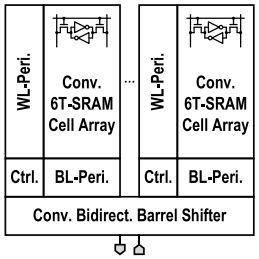
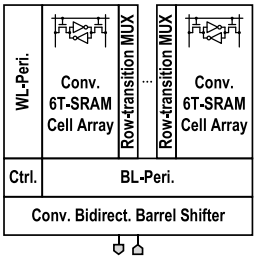
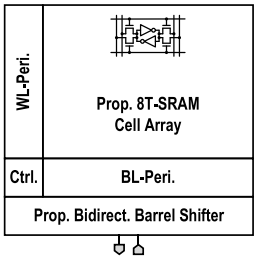
To enable hardware-efficient orthogonal data access, the proposed TP SRAM adopts i) a novel 8T SRAM bitcell with a diagonal WL bridge, and ii) a low-complex TP addressing scheme. We also proposed a dynamic gate-based MUX with duality to reduce the hardware cost of bi-directional barrel shifters. In the following subsections, a post-layout simulation based comprehensive analysis is carried out to verify the effectiveness of the proposed low-cost TP-SRAM scheme.

A. BIDIRECTIONAL BARREL SHIFTER

As an applicable option for a bidirectional barrel shifter, various logic gate-based designs in Table 1 are compared with the proposed dynamic gate-based design. The area comparison of unit MUX based on the layout shown in Fig. 15(b) is presented in Fig. 17. Although the distance between the supply rail and the transistor can be narrowed in the conventional gate layout in Fig. 15(b), the area of the unit MUX is compared based on Fig. 15(b), considering the number of metal routing tracks and the buffer layout to be followed. The proposed unit MUX (unshared structure) has 37% and 51% smaller areas than the conventional static CMOS-based MUX and domino logic-based MUX, respectively. Since it is not possible to share the pull-up and pull-down paths of the proposed dynamic gate for the unit MUX, that area of the shared structure is not counted in Fig. 17. Fig. 17 also shows an area comparison for a bidirectional barrel shifter. Each bidirectional barrel shifter includes the area of a two-stage inverter in common as a buffer of the output stage. In the case of the proposed bidirectional barrel shifter, when the number of inverting MUX stages is odd, the area of the 3-stage inverter is included. The proposed bidirectional barrel shifter with shared pull-up and pull-down paths in a 32-bit configuration has an area reduced by 35% compared to the conventional static CMOS-based designs. In the proposed 32-bit bidirectional barrel shifter, the area reduction before and after sharing the pull-up and pull-down paths is about 9%.

A comparison of delay and dynamic power for a bidirectional barrel shifter is shown in Fig. 18. Based on a 32-bit configuration at nominal supply voltage (0.9V), the pro-posed dynamic gate-based design achieves a 26% reduction in delay compared to the conventional static CMOS gate-based design. Compared to the static CMOS gate, the delay reduction of the domino logic gate is very sensitive to

TABLE 4. Comparison with the state-of-the-art.

Type	Conv. Non-integrated [1]	Conv. Multi-Bank [2]	Conv. Row-transition MUX [3]	Prop. Transposable 8T SRAM
Technology	28nm CMOS			
Cell type	Transpose 8T-SRAM Cell	Conv. 6T-SRAM Cell	Conv. 6T-SRAM Cell	Prop. 8T-SRAM Cell
[1)]Array structure				
Transposable	Supported by TP 8T-SRAM Cell	Supported by Diagonal WL & Transposable Addressing Scheme		
[1)]Cell Array	21388 um^2 (1.00x)	6732 um^2 (0.31x)	[2)]6732 um^2 (0.31x)	12272 um^2 (0.57x)
Array Efficiency	67%	14%	23%	68%
Barrel Shifter	-	Conv. Domino-Logic Based	Conv. Domino-Logic Based	Prop. Dynamic Gate Based
Addr. Scheme	Conv.	Conv. Transposable (Complex)	Conv. Transposable (Complex)	Prop. Transposable (Simple)
Drawback	Doubled Peri. & Separated I/Os	Repeated Peripherals	Significant WL Transition Delay	-
[3)]Power	383 uW (1.01x)	646 uW (1.71x)	379 uW (1.00x)	316 uW (0.83x)
[3)]Write Delay	701 ps (0.55x)	356 ps (0.28x)	1275 ps (1.00x)	506 ps (0.40x)
[3)]Read Delay	733 ps (0.48x)	545 ps (0.36x)	1528 ps (1.00x)	746 ps (0.49x)
[3)]Total Area	32085 um^2 (1.09x)	47692 um^2 (1.61x)	29570 um^2 (1.00x)	18086 um^2 (0.61x)

1) Based on 4:1 Column-Interleaving.
2) In the conventional row-transition MUX (RTM) based approach [3], RTMs are not included for cell array area
3) Reanalyzed based on 28nm CMOS technology (Post-Layout Sim.) and 16Kb SRAM (128 WLs & 128 BLs) due to the different technology node and lack of reported H/W cost

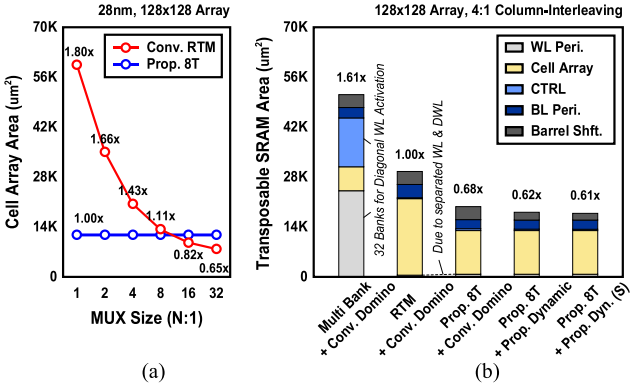


FIGURE 21. Comparison of the (a) TP-SRAM cell array area and (b) overall macro area.

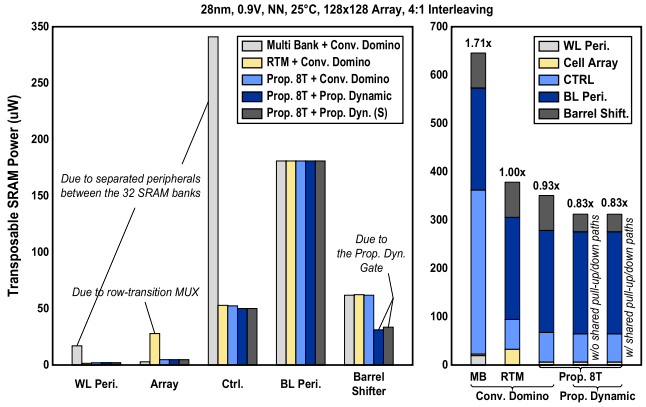


FIGURE 22. Power comparison of the TP-SRAM.

the strength of the keeper, and it was confirmed that doubling the length of the keeper PMOS reduced the delay by 4%. The power consumption of the bidirectional barrel shifter was measured as the average value of 4 cycles of left rotation and 4 cycles of right rotation by making the ratio of 0 and 1 of the input bits the same. Similar to the delay result, the proposed design shows little difference in power consumption with or without a shared path. At a 0.9V supply, power is increased by 94% over static CMOS-based designs, which is a 22.9% reduction over conventional domino logic-based design.

B. TRANSPOSABLE SRAM

For macro-level comparison, the proposed 16Kb (128 WLs and 128 BLs) TP-SRAM has been implemented using 28nm CMOS technology. The layout and design summary for this is shown in Fig. 19. Based on the proposed low-complexity TP addressing scheme (Fig. 9 and Fig. 13), fully functional macros can be implemented without additional circuitry for address decoding. As mentioned in Fig. 7, the separated BL and DBL pair for the proposed 8T TP-SRAM cell requires an additional BL MUX, which has an area overhead of 2.3%.

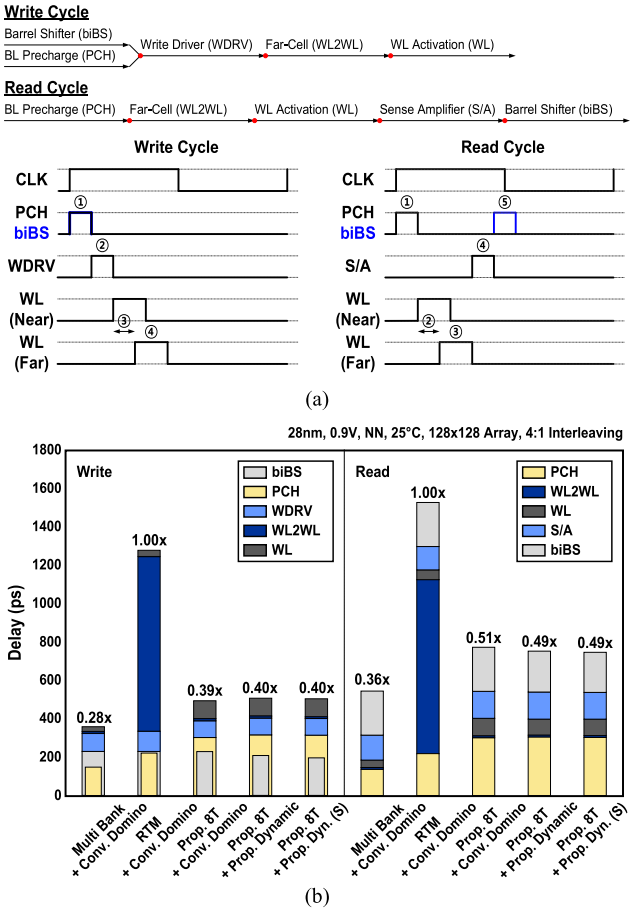


FIGURE 23. (a) Operation sequence and (b) delay comparison of the TP-SRAM.

The ratio of the proposed bidirectional barrel shifter in the total TP-SRAM macro is 10.8%, which increases further as the number of data I/O bits increases. Fig. 20 shows the area comparison of bitcells that enable the TP memory access function. Compared to the conventional 6T bitcell, the conventional TP-8T bitcell with additional orthogonal WL and BL pairs, conventional TP-10T bitcell with two decoupled read ports, and the proposed 8T bitcell have 5.24x, 3.18x and 1.82x larger areas, respectively. The proposed TP-8T bitcell with diagonal WL enables TP memory access function with a 42.7% and 65.3% reduced area compared to the conventional TP-8T bitcell and TP-10T bitcell, respectively. As shown in Table 3, the decoupled read port-based TP-10T SRAM cell [22] is advantageous in improving static noise margin (SNM) compared to other TP SRAM cells, but its advantages are overshadowed by the significant area increase and the lack of support for the write transposable function. When considering the use of an assist scheme, WL-based assist is difficult to use in all cell types due to half-select issues in the column-interleaved structure. Also, in the conventional TP-8T, the cell supply voltage (CVDD) is parallel to the bit line pair in only one direction, so the use of the CVDD-based assist scheme is limited.

For a fixed-size array, the conventional TP-SRAM [3] mentioned in Fig. 4 includes different numbers of row-transition MUXs (RTMs) according to the column-interleaving structure. Similar to the DWL bridge of the proposed TP-SRAM, since the RTMs are located on the column block boundary, more RTMs are required when the size of the BL MUX is small. On the other hand, the pro-posed 8T TP-SRAM maintains a constant area regardless of the column-interleaving structure. For a 16Kb cell array composed of 128 WLs and 128 BLs, Fig. 21(a) shows the area comparison between the conventional RTM-based cell array and the proposed cell array. When the BL MUX size is smaller than 16, the proposed 8T SRAM has a smaller cell array area, and the area of the conventional RTM-based cell array decreases exponentially as the BL MUX size increases. Although the conventional RTM-based cell array has a smaller area when the BL MUX is greater than 8, a large BL MUX requires several cycles for column data access, as shown in Case 3 in Fig. 11. The macro-level area comparison of the TP-SRAM in the 4:1 column-interleaving structure is shown in Fig. 21(b). Conventional multi-bank-based TP-SRAM [2] has a large area overhead due to redundancy for the controller and WL peripherals in 32 separate banks. Compared to RTM-based TP-SRAM, the proposed approach shows an area reduction of up to 39% due to optimization in the cell array and bidirectional barrel shifter.

Fig. 22 shows the average power for row-wise read/write operations and column-wise read/write operations to compare TP-SRAM power consumption. In the conventional multi-bank-based design, the power overhead due to redundant WL peripherals and controllers in each separate bank is noticeable. On the other hand, in the cell array, the RTM-based design consumes about 5.65 times more power than the proposed 8T TP-SRAM due to the switching of the MUX included in the middle of the array. As a result, the proposed TP-SRAM consumes up to 17% less power than conventional RTM-based designs. To compare the operation delay of TP-SRAM, the operation sequence including data relocation (bidirectional barrel shifter) for read and write operations can be defined as in Fig. 23(a). In the write operation, the rotation operation of the bidirectional barrel shifter for input data and the precharge operation for the half-selected cells are simultaneously performed. Also, for RTM-based TP-SRAM, it takes 31 RTMs to activate the WL of the column block farthest from the WL driver. For this reason, the proposed TP-SRAM, as shown in Fig. 23(b), shows 61% and 52% faster operation delay than RTM-based design for write and read operations, respectively. The comparison results with the state-of-the-art works are summarized in Table 4. The proposed 16Kb TP-SRAM implemented in 28nm CMOS technology has 17% reduced power, 52% faster operation delay, and 39% smaller area compared to the state-of-the-art.

V. CONCLUSION

This paper presents a novel hardware-efficient TP-SRAM design method that uses an improved addressing scheme to simplify address decoding and minimize data reordering

costs. By utilizing 8T SRAM bitcells and diagonal WL bridges, the excessive area overhead of the TP-SRAM is reduced. Furthermore, the proposed bidirectional barrel shifter, which uses inverting multiplexers and complementary dynamic gates, eliminates the hardware inefficiencies of traditional domino logic gate-based designs. Numerical results indicate that the proposed 16Kb TP-SRAM has reduced power consumption by 17%, faster operation latency by 52%, and a smaller area by 39% compared to the state-of-the-art.

REFERENCES

[1] J.-s. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha, and D. J. Friedman, "A 45 nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, San Jose, CA, USA, Sep. 2011, pp. 1–4.

[2] J. Kim, J. Koo, T. Kim, and J.-J. Kim, "Efficient synapse memory structure for reconfigurable digital neuromorphic hardware," *Frontiers Neurosci.*, vol. 12, p. 829, Nov. 2018.

[3] J. Koo, J. Kim, S. Ryu, C. Kim, and J.-J. Kim, "Area-efficient transposable 6T SRAM for fast online learning in neuromorphic processors," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, Apr. 2019, pp. 1–4.

[4] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[5] X. Qiao, J. Song, X. Tang, H. Luo, N. Pan, X. Cui, R. Wang, and Y. Wang, "A 65 nm 73 kb SRAM-based computing-in-memory macro with dynamic-sparsity controlling," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 6, pp. 2977–2981, Jun. 2022.

[6] X. Si, "A local computing cell and 6T SRAM-based computing-in-memory macro with 8-b MAC operation for edge AI chips," *IEEE J. Solid-State Circuits*, vol. 56, no. 9, pp. 2817–2831, Sep. 2021.

[7] F. Tan, Y. Wang, Y. Yang, L. Li, T. Wang, F. Zhang, X. Wang, J. Gao, and Y. Liu, "A ReRAM-based computing-in-memory convolutional-macro with customized 2T2R bit-cell for AIoT chip IP applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 9, pp. 1534–1538, Sep. 2020.

[8] Y. Huang, Y. He, J. Wang, J. Yue, L. Zhang, K. Zou, H. Yang, and Y. Liu, "Bit-aware fault-tolerant hybrid retraining and remapping schemes for RRAM-based computing-in-memory systems," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 7, pp. 3144–3148, Jul. 2022.

[9] L. Han, P. Huang, Y. Wang, Z. Zhou, Y. Zhang, X. Liu, and J. Kang, "Efficient discrete temporal coding spike-driven in-memory computing macro for deep neural network based on nonvolatile memory," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 11, pp. 4487–4498, Nov. 2022.

[10] A. Jaiswal, R. Andrawis, A. Agrawal, and K. Roy, "Functional read enabling in-memory computations in 1Transistor—1Resistor memory arrays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 12, pp. 3347–3351, Dec. 2020.

[11] K.-H. Li, C.-F. Hsu, Y.-S. Lin, S.-Y. Chien, and W.-C. Chen, "Configuration through optimization for in-memory computing hardware and simulators," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Rennes, France, Nov. 2022, pp. 1–6.

[12] S. Um, S. Kim, S. Kim, and H.-J. Yoo, "A 43.1TOPS/W energy-efficient absolute-difference-accumulation operation computing-in-memory with computation reuse," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 5, pp. 1605–1609, May 2021.

[13] S. Ha, S. Kim, D. Han, S. Um, and H.-J. Yoo, "A 36.2 dB high SNR and PVT/leakage-robust eDRAM computing-in-memory macro with segmented BL and reference cell array," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 5, pp. 2433–2437, May 2022.

[14] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing CNN accelerator employing charge-domain compute," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, Jun. 2019.

[15] Z. Xuan, C. Liu, Y. Zhang, Y. Li, and Y. Kang, "A brain-inspired ADC-free SRAM-based in-memory computing macro with high-precision MAC for AI application," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 4, pp. 1276–1280, Apr. 2023.

[16] R. Rafati, S. M. Fakhraie, and K. C. Smith, "A 16-bit barrel-shifter implemented in data-driven dynamic logic (D^3L)," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 10, pp. 2194–2202, Oct. 2006.

[17] R. Pereira, J. A. Michell, and J. M. Solana, "Fully pipelined TSPC barrel shifter for high-speed applications," *IEEE J. Solid-State Circuits*, vol. 30, no. 6, pp. 686–690, Jun. 1995.

[18] S, "Das and S. P. Khatri, "A timing-driven approach to synthesize fast barrel shifters," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 55, no. 1, pp. 31–35, Jan. 2008.

[19] S. M. Kang, "Domino-CMOS barrel switch for 32-bit VLSI processors," *IEEE Circuits Devices Mag.*, vol. CDM-3, no. 3, pp. 3–8, May 1987.

[20] P. Srivastava and E. Chung, "An asynchronous bundled-data barrel shifter design that incorporates a deterministic completion detection technique," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 3, pp. 1667–1671, Mar. 2022.

[21] N. Yadav, A. P. Shah, and S. K. Vishvakarma, "Stable, reliable, and bit-interleaving 12T SRAM for space applications: A device circuit co-design," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 3, pp. 276–284, Aug. 2017.

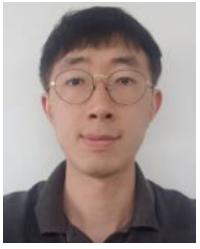
[22] Z. Lin, Z. Zhu, H. Zhan, C. Peng, X. Wu, Y. Yao, J. Niu, and J. Chen, "Two-direction in-memory computing based on 10T SRAM with horizontal and vertical decoupled read ports," *IEEE J. Solid-State Circuits*, vol. 56, no. 9, pp. 2832–2844, Sep. 2021.

[23] H. Jiang, X. Peng, S. Huang, and S. Yu, "CIMAT: A compute-in-memory architecture for on-chip training based on transpose SRAM arrays," *IEEE Trans. Comput.*, vol. 69, no. 7, pp. 944–954, Jul. 2020.

[24] J. Song, X. Tang, X. Qiao, Y. Wang, R. Wang, and R. Huang, "A 28 nm 16 kb bit-scalable charge-domain transpose 6T SRAM in-memory computing macro," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 5, pp. 1835–1845, May 2023.

[25] J.-W. Su, "Two-way transpose multibit 6T SRAM computing-in-memory macro for inference-training AI edge chips," *IEEE J. Solid-State Circuits*, vol. 57, no. 2, pp. 609–624, Feb. 2022.

DAIN CHON (Student Member, IEEE) received the B.S. degree in electronics engineering from Sookmyung Women's University, Seoul, in 2021, where she is currently pursuing the M.S. degree with the VLSI and System Laboratory. Her research interest includes low-power and high-performance embedded memory (SRAM) designs in advanced technology.



WOONG CHOI (Member, IEEE) received the B.S. and Ph.D. degrees in electronics engineering from Korea University, Seoul, South Korea, in 2011 and 2018, respectively. In 2018, he joined Samsung Electronics Ltd., Hwaseong-si, South Korea. Since 2019, he has been an Assistant Professor with the Department of Electronics Engineering, Sookmyung Women's University, Seoul. His current research interests include neural network accelerator and embedded memory designs in advanced technologies.