

RESEARCH ARTICLE

S²RAM: Optimization of SRAM With Memory Access Patterns

WOONG CHOI , (Member, IEEE)

Department of Electrical Engineering, Sookmyung Women's University, Seoul 04310, South Korea

e-mail: woongchoi@sookmyung.ac.kr

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

ABSTRACT This paper presents a static sequential-random-access memory (S²RAM) designed to enable low-power and high-performance operations for workloads involving sequential memory accesses. Considering the address configuration and internal operation of column-interleaved SRAM, we propose an optimized memory structure and peripheral circuitry that can exploit access patterns as opportunities for low-cost operation. In the proposed S²RAM, word-line (WL) activation is restricted by using bit-line (BL) as temporary storage in case of memory access in which only column addresses are sequentially changed. In order to prevent the BL leakage-induced data corruption, a BL clamper is also proposed. The optimized design method of the controller and assist scheme for using reconfigurable access modes in the proposed S²RAM is also presented. The proposed 16Kb S²RAM macro has been implemented using a 28nm CMOS technology. Numerical results show that the proposed S²RAM saves up to 76% of operating energy compared to conventional SRAM at the cost of 8% increased area.


INDEX TERMS SRAM, low-power, clamper, leakage, assist, burst operation.

I. INTRODUCTION

As the speed difference between the processor and main memory continues to increase, more and more embedded memory is required for various memory subsystems. This problem, known as the memory wall [1], has been addressed by adding multiple levels of embedded memory to enable low latency in the on-chip processing units [2]. Static random-access memories (SRAMs) are commonly used for embedded memory due to their excellent speed and compatibility with logic process technologies [3]. This hierarchical SRAM-based on-chip storage is utilized in traditional processors as well as emerging devices for deep learning and vision applications. In recent deep learning hardware accelerators [4], [5], [6], SRAM is used at various storage levels such as local registers, scratchpads, and global buffers. In many of these devices, SRAM occupies a significant portion of

the total chip area and power consumption, requiring high operating speeds.

Previous studies have presented several SRAM design techniques to improve power, performance, and area (PPA). To optimize area, low-area bitcells [7], [8] with a reduced number of transistors have been proposed, but their use is limited due to low operational stability and insufficient area reduction. As a result, there is a limited exploration of the low-area design space, and high-density or high-performance cells provided by foundries are often selectively used depending on the target specification. To improve SRAM performance, various studies [9], [10], [11] have been conducted to reduce overestimated operating margins or process variations. These variation-aware designs are also used for power reduction and require additional examination for relevance to actual chip characteristics. In advanced technology nodes, the focus is placed on reducing the delay exacerbated by significant back-end wire RC load to improve SRAM performance [12], [13], [14]. To minimize

The associate editor coordinating the review of this manuscript and approving it for publication was Ye Zhou .

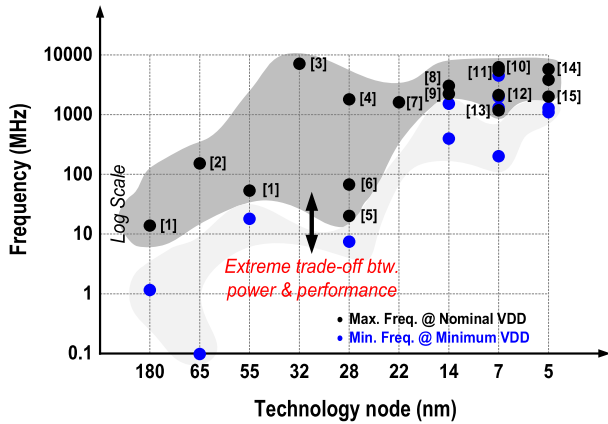


FIGURE 1. SRAM operating voltage and frequency reported in the recent test chips for various emerging applications.

SRAM power, research [15], [16], [17] is mainly focused on lowering the supply voltage while ensuring operational stability. This is commonly achieved through SRAM assist, which is essentially used below the 28nm technology node due to operational instability even near nominal supply voltages. The adoption of read-decoupled 8T [18] and 10T [19] SRAM facilitates supply voltage scaling, but is a reluctant option due to increased area.

In emerging devices, where all aspects of PPA are important, the design space for SRAM optimization is limited [14]. As shown in Fig. 1, scaling down the supply voltage to reduce SRAM power consumption significantly reduces operating speed. To address this trade-off, sequential SRAM [20] has been proposed, which leverages the sequential memory access pattern of the augmented reality (AR) applications to enable low-power and high-speed memory operation. Sequential SRAM can effectively reduce dynamic power without compromising performance, but its limitation of only supporting monotonic access patterns makes it unusable in computing systems that must support typical workloads. For this reason, sequential SRAM cannot be used as a cache or shared memory in multiprocessor computing systems [21], [22].

In this paper, we present static sequential-random-access memory (S²RAM), which supports both random and sequential accesses and reduces dynamic power consumption without performance degradation for workloads involving sequential memory access. The proposed S²RAM leverages the features of memory address configuration and column-interleaving structures to eliminate unnecessary bit-line (BL) swing. In cases of consecutive memory access where only the BL addresses change sequentially, we utilize BL as a temporary storage and restrict WL activation to reduce dynamic power consumption. To prevent operational failures caused by BL leakage, a BL clamber that collaborates with SRAM assist circuits is also proposed. For SRAM assist, which has significantly different constraints in random access

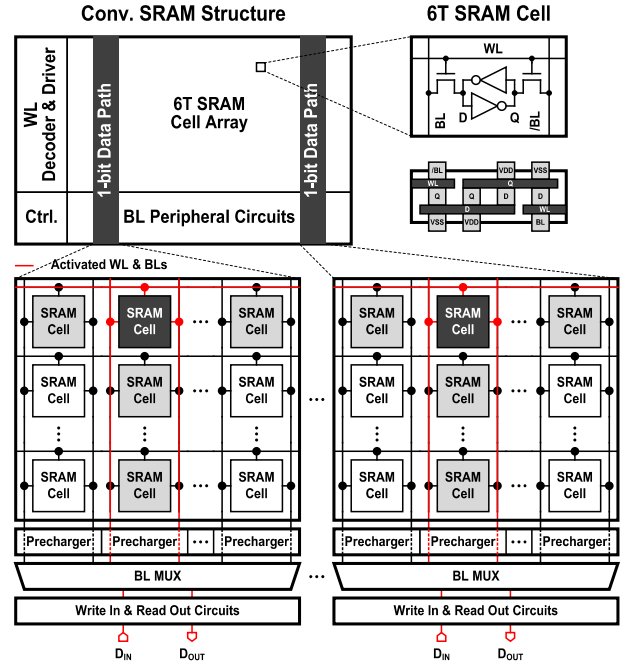


FIGURE 2. Conventional SRAM with column-interleaving structure.

and sequential access, the BL clamber enables implementation of the assist scheme optimized for the proposed S²RAM.

The rest of the paper is organized as follows. Section II provides an overview of the conventional column-interleaved SRAM. The proposed S²RAM is presented in Section III. In Section IV presents numerical results, including comparison with the state-of-the-art. Section V concludes this work.

II. PRELIMINARIES

This section provides an overview of standard SRAM structures and address configurations to understand how to leverage memory access patterns for SRAM optimization.

A. COLUMN-INTERLEAVING BASED SRAM

Fig. 2 illustrates the column-interleaving structure of a typical SRAM. SRAM consists of a cell array that stores data, a word-line (WL) decoder and driver that activates cell regions for address values, peripheral circuitry for BLs through which read and write data travel, and a controller that orchestrates the entire block. When focusing on the 1-bit data path of the cell array and BL peripheral circuits, it can be observed that a bundle of several BL pairs is connected to write-in and read-out circuits through one BL multiplexer (BL MUX). This structure, called column-interleaving or bit-interleaving, is commonly used in standard SRAMs for i) soft-error immunity and ii) efficient integration of BL peripherals. Fig. 3 compares SRAMs with and without the column-interleaving structure when four different data (A[3:0] to D[3:0]) are stored in the cell array. With the column-interleaving structure, one data can be stored as far apart as the number of inputs of the BL MUX, whereas

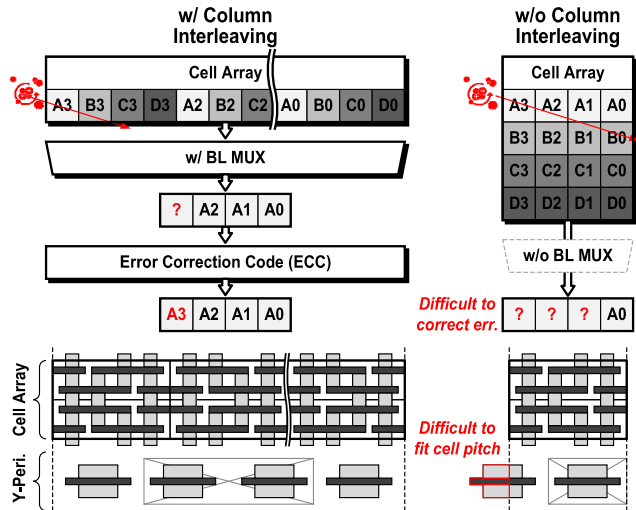


FIGURE 3. Advantages of column-interleaving in the conventional SRAM.

without it, each bit of one data is stored in a physically adjacent cell. These structural difference causes a difference in the number of error bits for one data when a single event upset (SEU) occurs in the cell array. In column-interleaved SRAM, error correction is possible using lightweight error correction codes (ECC), such as BCH codes. On the other hand, non-column-interleaved SRAM requires expensive ECC due to its relatively large number of error bits. Additionally, the column-interleaving structure enables efficient integration of the cell array and peripheral circuits by arranging the BL peripheral circuits for several columns rather than for each cell.

However, the column-interleaving structure introduces half-selection issues that make SRAM design challenging. As shown in Fig. 4, an SRAM cell placed under the activated WL but not selected by the BL MUX is referred to as a row-wise half-selected (H/S) cell. When WL is activated, row-wise half-selected cells require read operations to retain the stored data. This inevitably leads to BL swings in the unselected columns, increasing dynamic power ((1) in Fig. 4). Additionally, using row-wise SRAM assist reduces the reliability of half-selected cells ((2) in Fig. 4).

B. DATA PATH OF SRAM ADDRESS AND SEQUENTIAL SRAM

In column-interleaved SRAM, the input address determines which WL and BL MUX select (BLMS) signals will be activated, as shown in Fig. 5. Here, the WL decoding address is indicated as WLADDR, and the BL decoding address is indicated as BLADDR. To enable WL, the signal X_{11:0} first decoded from the pre-decoder inside the SRAM controller is passed to the final decoder (WLDEC) adjacent to the cell array. The least significant bits (LSBs) within WLADDR determine which WL within the WLDEC is activated, and the most significant bits (MSBs) determine which WLDEC block is activated. Therefore, when the WLADDR value

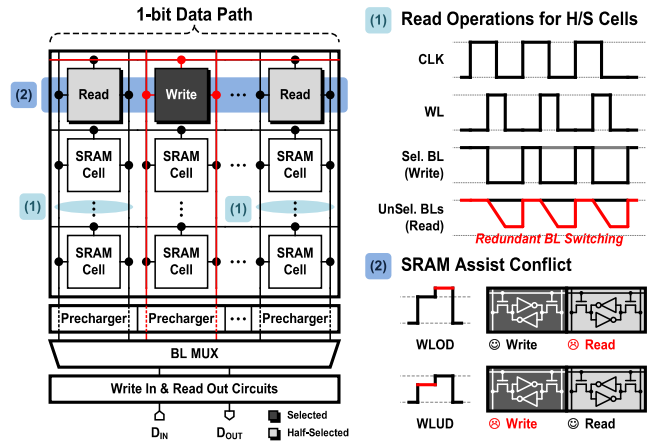


FIGURE 4. Unnecessary BL swing and half-select issues in the column-interleaving-based SRAM structure.

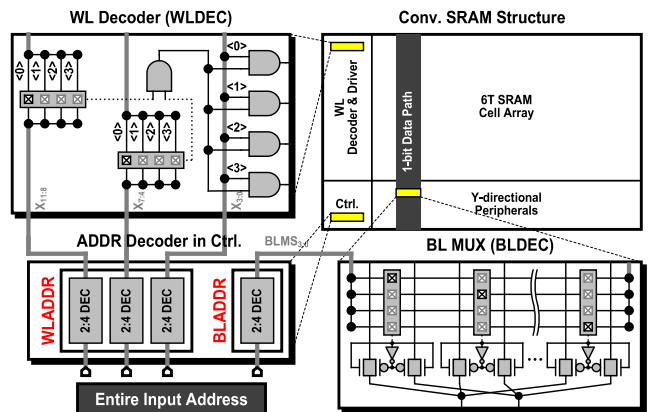


FIGURE 5. Simplified data path of memory address in the column-interleaved SRAM.

sequentially increases or decreases, the row index of the activated WL also increases or decreases sequentially. Similarly, BLADDR determines which BL pair is selected in the BL MUX. When BLADDR increases or decreases sequentially, the column index for the selected BL increases or decreases sequentially. The order of WLADDR and BLADDR in the entire input address can be set in two ways, as shown in Fig. 6. When the BLADDR is located on the MSB side (left case in Fig. 6), if the input address (ADDR) increases sequentially, only the WLADDR increases until there is a change in the BLADDR. On the other hand, if BLADDR is placed on the LSB side (right case in Fig. 6), when ADDR increases sequentially, WLADDR is maintained for a certain period and BLADDR is switched every cycle.

III. PROPOSED S²RAM

A. THE SEQUENTIAL OPERATION PRINCIPLE

In the proposed S²RAM, as illustrated in the right case of Fig. 6, memory addresses are configured with WLADDR on the MSB side and BLADDR on the LSB side. When memory addresses change sequentially, this address configuration results in a period in which the activated WL is

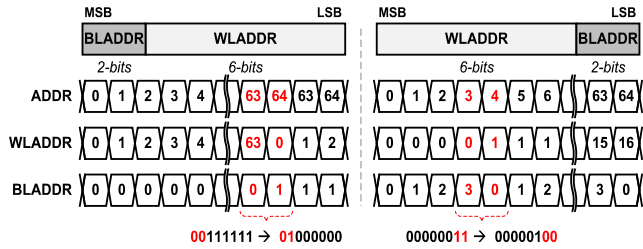


FIGURE 6. Available address configuration and sequential memory access.

constant and the selected BL pairs change sequentially. Like conventional sequential SRAM [20], the proposed approach uses these periods to reduce BL switching, which causes significant power consumption in SRAM. Fig. 7 presents the sequential read and write operation principle of the proposed S²RAM. In the figure, shaded circuit blocks represent parts that should be activated, and unshaded circuit blocks indicate parts that should be disabled. For the proposed sequential S²RAM read operation, in the first read cycle where the leftmost BL pair is selected, the read operation is performed in the same way as for the conventional SRAM. At this time, data from all bit cells connected to the activated WL is read into each BL regardless of whether it is selected by the BL MUX. This means that the data required for the subsequent cycles has already been read. Therefore, the BL precharge and WL activation are omitted in subsequent cycles to eliminate unnecessary BL switching. Similarly, in sequential S²RAM write operation, WL is activated only in the last cycle of the section in which the WL to be activated is fixed. In the preceding write cycles within the section, only the BL MUX and write circuit are activated to sequentially transmit write data to the BL. The proposed S²RAM writes multiple data at a time by activating WL only in the last cycle when write data for all BLs are ready. In summary, the proposed S²RAM uses BL as temporary storage to read or write data at a time, thereby eliminating unnecessary power consumption during sequential memory access.

B. OVERALL ARCHITECTURE AND KEY FEATURES

Fig. 8 shows the overall architecture of the proposed S²RAM. The three key features of the proposed S²RAM are i) reconfigurable access mode depending on workloads, ii) BL clamper turns BL into stable temporary storage, and iii) collaborative S²RAM assist scheme. The first key feature, reconfigurable access mode, is configured to support both sequential and non-sequential (random) memory access. As shown in Fig. 8, when the memory access address sequentially increases or decreases, the internal control signals are reconfigured using the MODE and SEQ signals. In random access mode (MODE: ‘1’), the WL enable (WLEN) and BL precharge (BLPC) signals are activated every cycle, just like conventional SRAM. On the other hand, in sequential access mode (MODE: ‘0’), WLEN and BLPC signals are activated only when needed. In order to support both cases

where the memory access address is sequentially increased or sequentially decreased, a SEQ signal indicating the sequential access direction is adopted in addition to the MODE signal. In sequential read mode (MODE: ‘0’, READ: ‘1’), when memory addresses increase sequentially, the BLMS signal (BLMS₀), which selects the first BL column, is activated first. On the other hand, when memory addresses are sequentially decreased, the BLMS signal (BLMS_{N-1}), which selects the last BL column, is activated first. As shown in Fig. 8, the SEQ signal controls the BLPC and WLEN signals for sequential direction by selecting which of the BLMS signals is activated first. Meanwhile, in sequential write mode (MODE: ‘0’, READ: ‘0’), the BLPC signal is continuously deactivated, and the WLEN signal is controlled to be activated only in the last cycle. In this way, the proposed S²RAM is reconfigured into an operation mode suitable for the system workload through the added MODE and SEQ signals.

In the proposed sequential S²RAM operation, BLs are used as data storage while the WL to be activated is fixed. The second key feature, the BL clamper, allows the BL to act as stable storage during sequential access. Sequential data temporarily stored in BL may be altered by leakage current. The corresponding worst-case scenario is presented in Fig. 9. During sequential read operations, the data to be read in the last cycle is most affected by leakage because it has already been stored in the BL from the first cycle. The worst leak in the last column of a sequential read occurs when all the data in the other rows has the opposite value as the data read. Similarly, for sequential write, data stored in BL in the first cycle must be retained until WL is activated in the last cycle. In this case, the worst leakage occurs when the data stored in the BL and all cell data in the first column have opposite values. The BL voltage change according to the worst leakage current is shown in Fig. 10. As shown in Fig. 10, in a configuration of 128 cells per BL, there is only one more leakage path for the write operations, so the BL voltage changes during sequential read and write operations are nearly identical. Additionally, as shown in Fig. 10(a), the BL voltage decrease on the high-level side (red line) appeared faster than the BL voltage increase on the low-level side (blue line). This is because the gate-source voltage between WL and the cell storage node is different on the left BL and right BL. In sequential read operations (Fig. 10(a)), the differential BL voltage reaches the sensing margin over time, which can result in read failures. Similarly, for sequential write operations (Fig. 10(b)), write failures may occur because BL deviates from the full-rail voltage. Additionally, as shown in Fig. 10(c) and Fig. 10(d), the BL voltage retention degrades as the number of cells per BL increases and the supply voltage decreases.

To compensate for the instability caused by leakage during sequential S²RAM operation, a second key feature, a BL clamper, is designed as shown in Fig. 11. For sequential reads, the BL clamper for all columns is activated immediately after data is read into BL during the first cycle. Using a full-swing BL clamper will cause unnecessary swings in the differential

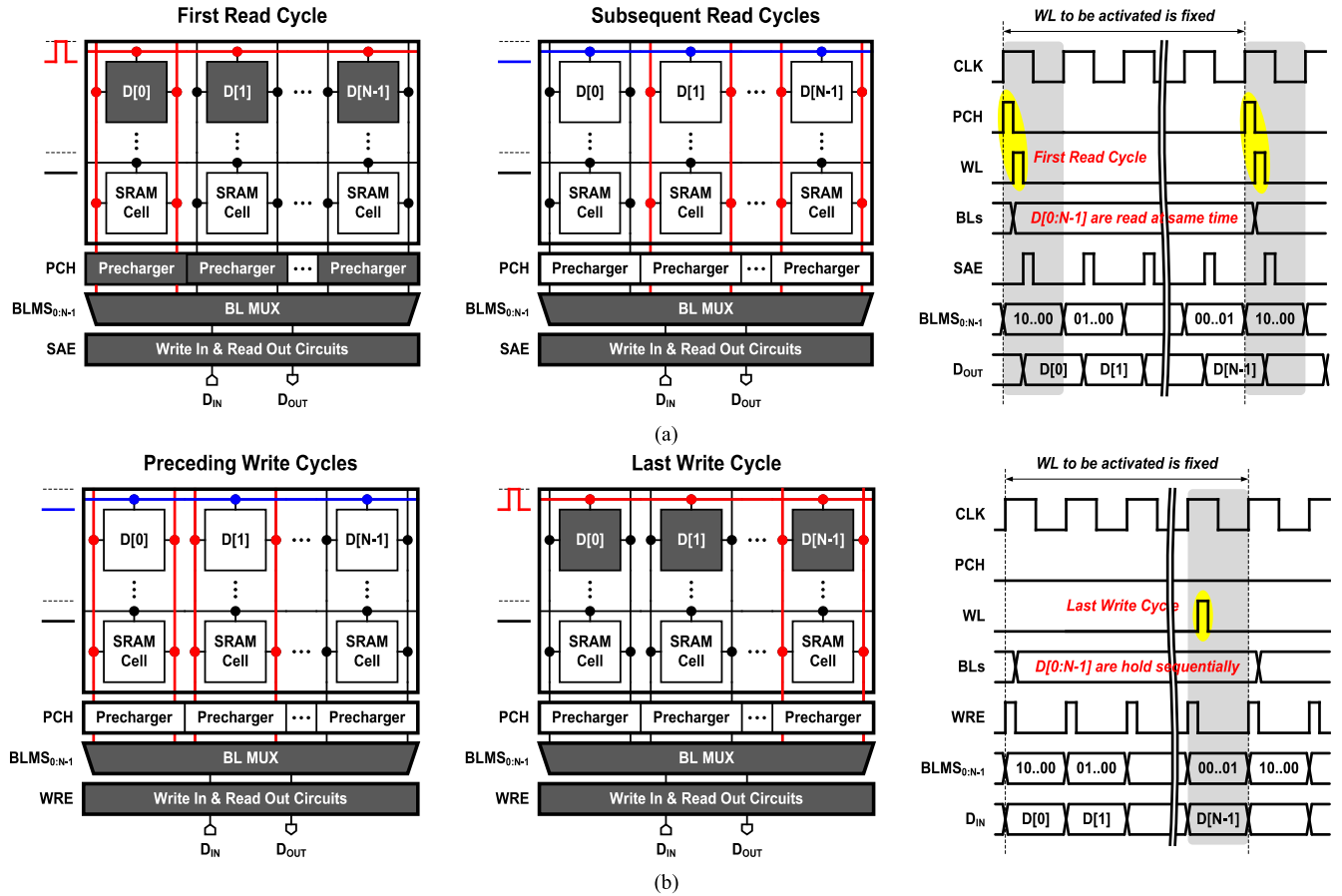


FIGURE 7. (a) Sequential read and (b) write operation principle of the proposed S²RAM.

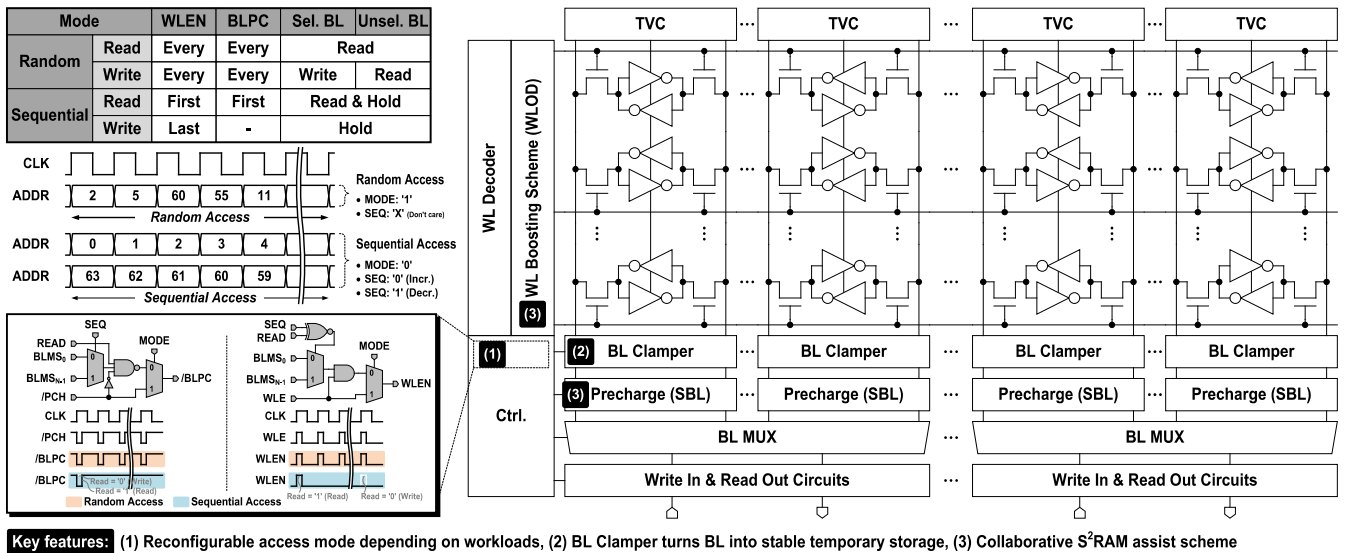


FIGURE 8. Overall architecture and key features of the proposed burst SRAM.

BL voltage, which is already close to the sensing margin after WL activation. On the other hand, during sequential reads, the proposed BL clamper, which includes a PUN NMOS pull-up

path, prevents excessive BL swing along with a read assist scheme that lowers the precharge voltage. In sequential write operations, activating the write driver while the full-swing

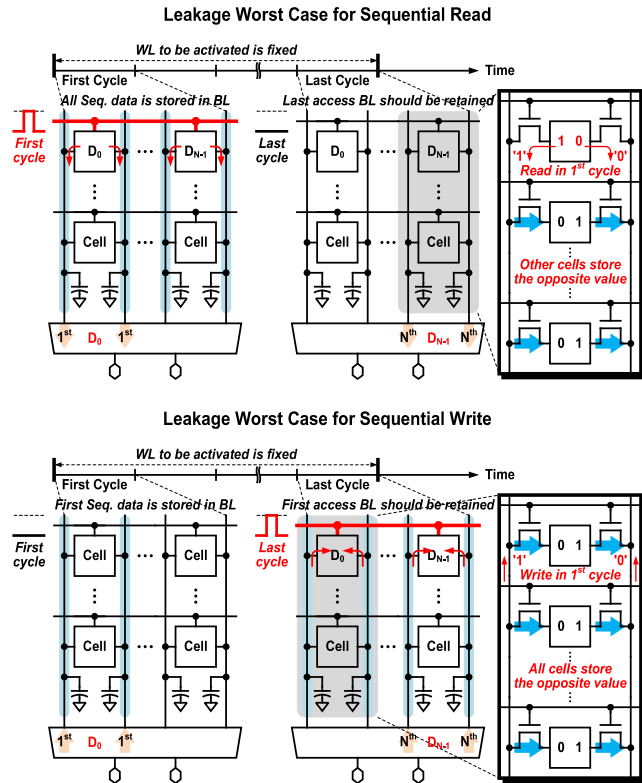


FIGURE 9. Leakage worst cases for sequential read and write in the proposed S²RAM.

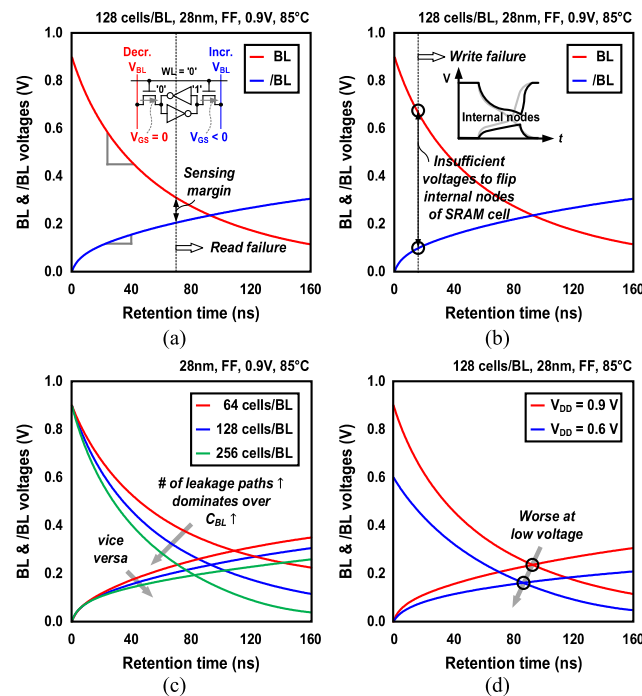


FIGURE 10. BL voltage characteristics in worst leakage scenarios; (a) sequential read and (b) write, (c) sequential writes at different number of cells per BL, and (d) sequential writes at different supply voltage.

clamper is enabled may cause the BL voltage not to change to the write value or cause significant delays. On the other

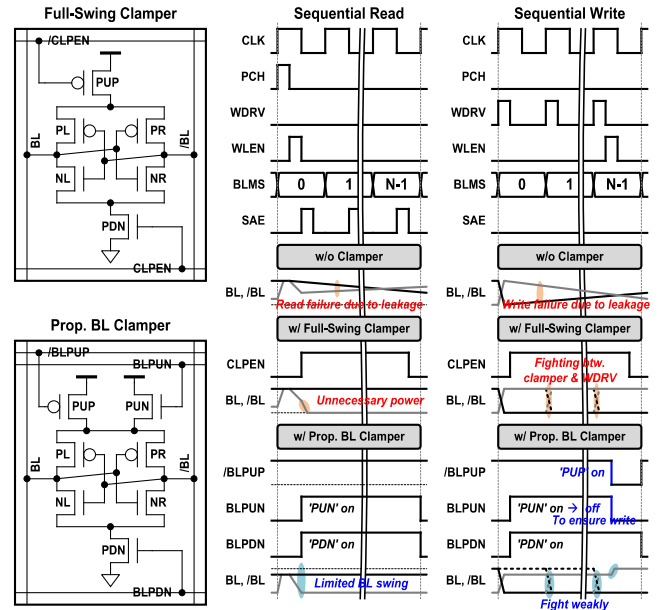


FIGURE 11. BL clamper for leakage compensation in the proposed S²RAM.

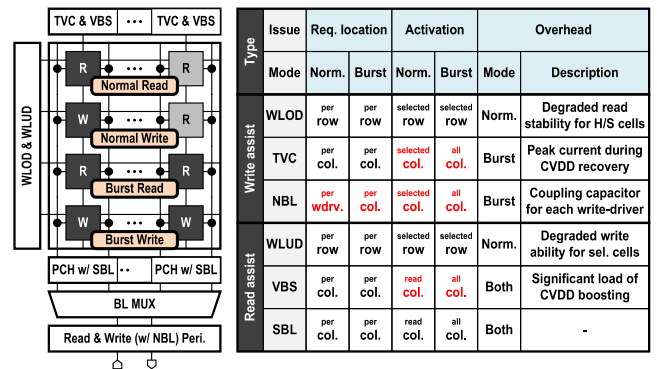


FIGURE 12. Considerations for the assist technique in the proposed S²RAM.

hand, the proposed BL clamper operates a pull-up network with an NMOS transistor (PUN), so it can alleviate conflicts with the write driver during sequential write operations. In our design, conflicts with the write driver are further mitigated through sequential activation of the BL clamper pull-down network, as shown in Fig. 13. Additionally, in the last cycle of sequential write operation, the high-level BL voltage ($V_{DD} - V_{TH,PUN}$) is raised to the supply voltage (V_{DD}) level to ensure write operation.

In general, SRAM assist is essential at advanced technology nodes due to unstable operation even near the nominal supply voltage [10], [12], [13], [14]. Considerations for determining the assist scheme of the proposed S²RAM are presented in Fig. 12. The operation of the proposed S²RAM is divided into normal read and write in random access mode and burst read and write in sequential access mode. Here, in sequential read and write, WL is activated only in the first and last cycle, respectively, hence the name ‘Burst’. When the WL-based assist (or WL Over-Drive (WLOD) [23]

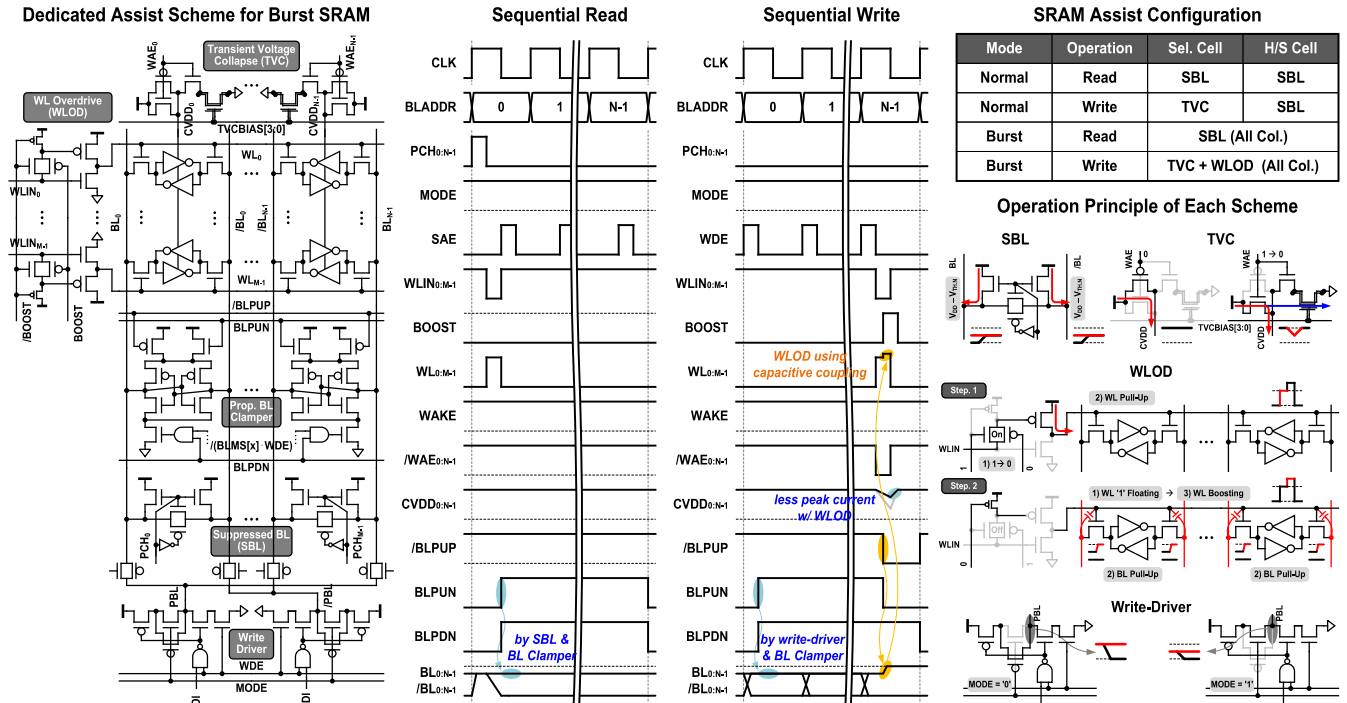


FIGURE 13. Assist configuration for the proposed S²RAM.

WL UnderDrive (WLUD) [24] is used in the normal write operation, either the write operation of the selected cell or the read operation of the half-selected cell degrades the operation stability. On the other hand, in sequential mode, WL-based assist does not cause conflicts between read and write as all cells connected to an activated WL perform the same operation. SRAM assist applied in the column direction does not cause conflicts between reads and writes, but requires relatively more lines to be activated. In column-wise read assist, cell supply voltage booting (VBS) [27] must also be applied to the half-selected (H/S) cells performing the read operation, which imposes a significant burden due to the large load capacitance. On the other hand, suppressed BL (SBL) [16], another column-wise read assist, can be implemented without significant overhead because it lowers the precharge voltage below VDD. During ‘Burst Read’ operation in the proposed S²RAM, data for all columns is read in the same way as ‘Normal Read’. On the other hand, ‘Burst Write’ writes to all columns, unlike ‘Normal Write’, which writes only to selected columns. Therefore, for ‘Burst Write’, the negative BL (NBL) [28] circuit, mounted only on the write driver in ‘Normal Write’, must be mounted in all columns, which greatly increases the area due to the coupling capacitor. On the other hand, transient voltage collapse (TVC) [26] basically installs relatively small assist circuits in all columns, so there is no additional area increase for ‘Burst Write’. However, unlike ‘Normal Write’, ‘Burst Write’ requires TVC circuitry to be enabled for every column, which introduces significant overhead in driving the assist circuitry.

The assist circuits for the proposed S²RAM taking these constraints into account are presented in Fig. 13. For normal

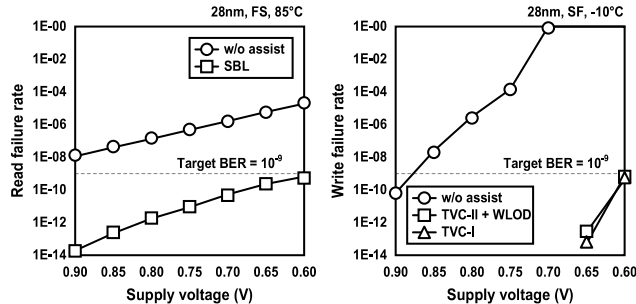
operations, NMOS pull-up precharge-based SBL is used as read assist, and ratioed pull-down based TVC is adopted as write assist. To reduce the overhead caused by TVC for all columns in ‘Burst Write’, capacitive-coupling-based WLOD [30] is used with TVC. The detailed operating principle of the WLOD is illustrated in Fig. 13. As a first step, the disabled BOOST signal turns on the transmission gate of the WL driver, and the inverted WL decode signal (WLIN) is sent to the last inverter stage to turn on the WL. After that, the activated BOOST signal blocks all pull-up passes of the last inverter stage, making WL floating. At this time, the proposed BL clamper raises the high-level (VDD-V_{TH,PUN}) BLs to full-VDD, and a capacitive coupling occurs due to parasitic capacitance between BL and WL. WLOD, used only in ‘Burst Write’, does not cause conflicts between read and write operations. The third key feature, the proposed assist scheme, is that the SBL, the write driver, and the BL clamper with NMOS-based pull-up network cooperate to prevent unnecessary BL swings. To avoid contention between the write drive and the BL clamper, column selection signal (/BLMS[x]-WDE)) is used in the BL clamper. In normal operations, the BL clamper is disabled by the BLPUP, BLPUN, and BLPDN signals, and the write driver operates with full rail voltage through the MODE signal. At this time, BL precharge and WL are controlled as in the conventional SRAM.

C. DETAIL ASSIST CONFIGURATION OF THE PROPOSED S²RAM

SRAM belonging to a high-replication circuit requires a very low failure rate of around 10⁻⁸ to 10⁻⁶ [29]. Since the intensity and duration of the assist circuit affect the SRAM

ADM/WRM conditions	SBL	TVC-I & TVC-II	WLOD
<ul style="list-style-type: none"> Tech.: 28nm CMOS Indep. ΔV_{TH} for each Tr. σV_{TH}: 30mV, ΔT_{AST}: 20FO4 BL cap: 10fF Target BER: 10^{-9} @ 0.6V WL pulse-width: 50FO4 			
	$\Delta V_{AST}: 0.63VDD$	$\Delta V_{AST}: 0.60/0.78VDD$	$\Delta V_{AST}: 1.05VDD$

(a)



(b)

(c)

FIGURE 14. Operation failure rate with SRAM assist; (a) configuration of ADM and WRM BER estimation, (b) read failure rate, and (c) write failure rate.

operation failure rate [15], the detailed configurations necessary for the proposed assist circuit are investigated. Fig. 14 presents the simulation conditions and results of access disturbance margin and write margin (ADM/WRM), which are mainly used in the industry to measure SRAM stability [29]. In this simulation, an assist configuration that meets the target BER of 10^{-9} at a supply voltage is 0.6V is examined under worst-case operating conditions. For read operations, SBL based on the precharge voltage lowered to 0.63 VDD satisfies the constraints. If only TVC is used in write operations, the constraints are met by setting the assist duration to 20 FO4 and lowering the VDD to 60%. The proposed combination of TVC and WLOD for ‘Burst Write’ operation requires assist intensity of voltage levels of 0.78 VDD and 1.05 VDD respectively, based on 20 FO4 assist duration.

The proposed assist circuits shown in Fig. 13 are configured to satisfy the above operating conditions. The assist intensities resulting from different PVT conditions are shown in Fig. 15. In the case of SBL, the required BL voltage is generated using an NMOS-based pull-up precharge circuit. For WLOD, the BL at the intermediate voltage held by the BL clumper is raised to the VDD level to create the over-driven WL voltage. On the other hand, for TVC, the required cell supply voltage (CVDD) is generated through pulse width modulation by a control signal (WAE in Fig. 13). For this reason, compared to SBL and WLOD, changes in TVC are relatively larger in PVT variations. Also, the proposed S²RAM requires different TVC voltages for ‘Normal Write’ and ‘Burst Write’. To counter this, the TVC includes a bias signal (TVCBIAS in Fig. 13) that allows the pull-down rate to be adjusted. Fig. 15 shows that all assist configurations except WLOD when VDD is reduced by 10% produce the target voltage. However, as VDD decreases, the write ability of the SRAM cell itself increases, so a WLOD close to the target voltage can be used for the target BER.

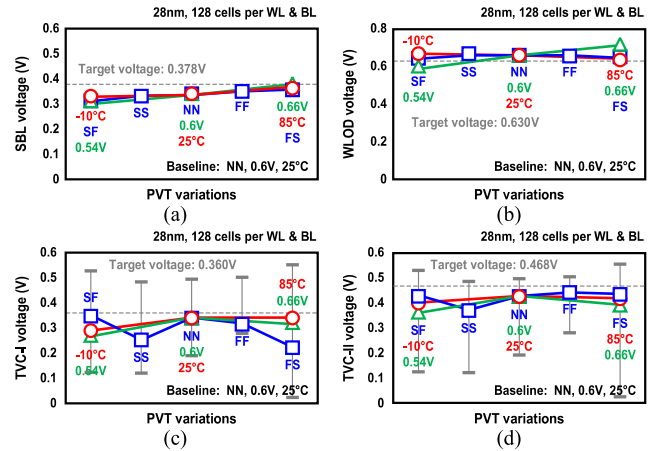


FIGURE 15. Assist intensity of the (a) SBL, (b) WLOD, (c) TVC-I (for normal mode), and (d) TVC-II (for sequential mode) in PVT variations.

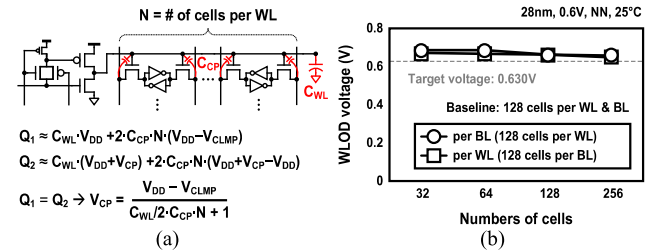


FIGURE 16. (a) Derivation of WLOD voltage, and (b) assist intensity of WLOD in various array size.

Fig. 16 shows the effect of array size on capacitive-coupling based WLOD voltage. The total charge amount of WL can be approximated as the sum of the charge of the WL capacitance (C_{WL}) and the charge of all capacitances between WL and BL (C_{CP}). Since there is no direct charge inflow into WL before or after capacitive coupling, the total charge on WL is conserved. Therefore, the WLOD voltage can be derived as shown in Fig. 16(a). Here, Q_1 and Q_2 represent the total WL charge before and after capacitive coupling, respectively. V_{CP} and V_{CLMP} represent the capacitive-coupled WL voltage and the initial BL voltage held by the BL clumper, respectively. Also, N represents the number of cells per WL. As a result, the lower the initial voltage (V_{CLMP}) by the BL clumper, the larger ‘ $V_{DD}-V_{CLMP}$ ’, so it can be seen that as the BL voltage change increases, the over-driven WL voltage (V_{CP}) also increases. Additionally, as the number of cells per WL (N) increases, the WL capacitance (C_{WL}) increases, but the WLOD voltage remains almost constant due to the increase in the number of columns (N) where the capacitive coupling effect occurs (Fig. 16(b)).

As mentioned earlier, SRAM assist for ‘Burst Write’ must be applied to all columns, so it is more burdensome than ‘Normal Write’. This means that excessive peak currents generated during ‘Burst Write’ can disrupt the supply of stable voltage to adjacent circuits. Furthermore, there may be significant delays in CVDD recovery that must be completed

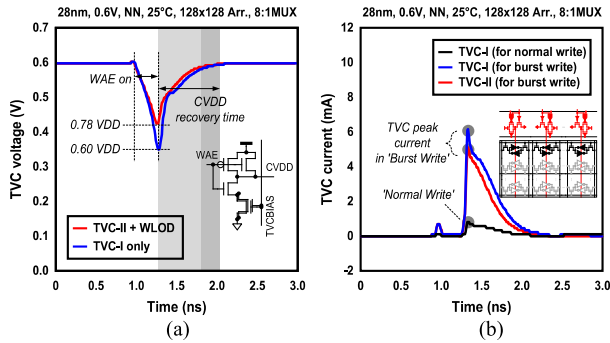


FIGURE 17. Waveform of (a) TVC voltage, and (b) current.

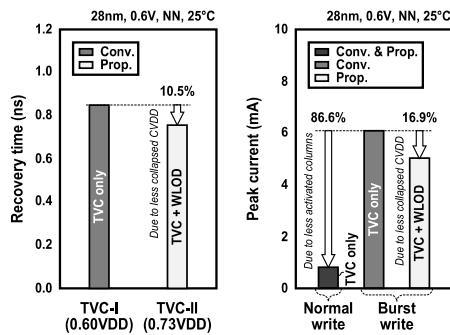


FIGURE 18. Comparison results of write assist in the proposed S²RAM.

before transitioning to the next operating cycle. The proposed TVC and WLOD combination for ‘Burst Write’ alleviates these peak current and CVDD recovery time issues. To verify the superiority of the proposed assist approach, the CVDD recovery time, TVC peak current, and the write assist power are compared with the monotonic write assist scheme (TVC only). Fig. 17 and Fig. 18 show comparison results for a 128 × 128 cell array (8:1 BL MUX) designed in a 28nm CMOS process. As shown in Fig. 17(a), the CVDD recovery time was measured from the end of the write assist enable (WAE) signal until the CVDD reached 99% of the original voltage level. Immediately after disabling the WAE signal, the TVC circuit generates a peak current to recover CVDD, as shown in Fig. 17(b). Numerical results show that the proposed combination of TVC and WLOD in ‘Burst Write’ improves the CVDD recovery time by 10.5% compared to using TVC alone. As shown in Fig. 18, a ‘Burst Write’ operation is approximately 8 times higher than a ‘Normal Write’ in terms of peak current due to the 8:1 MUX configuration.

IV. NUMERICAL RESULTS

The proposed S²RAM is implemented as a fully-functional 16kb macro using 28nm CMOS technology. The layout and design summary of the proposed S²RAM is presented in Fig. 19. As shown in Fig. 14, the target operating voltage range (BER < 10⁻⁹) is 0.6V to 0.9V, including NMOS pull-up based SBL for read assist and TVC and WLOD for write assist. In a 128 × 128 cell array configuration, the BL clamper added to use BL as temporary storage in sequential

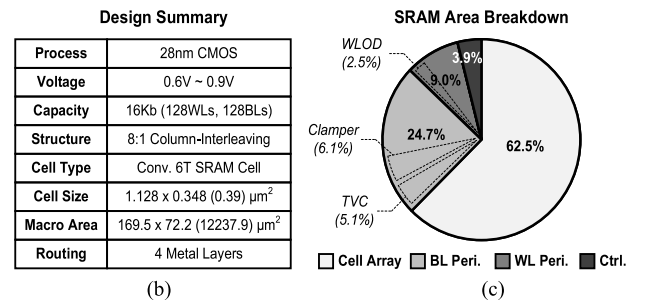
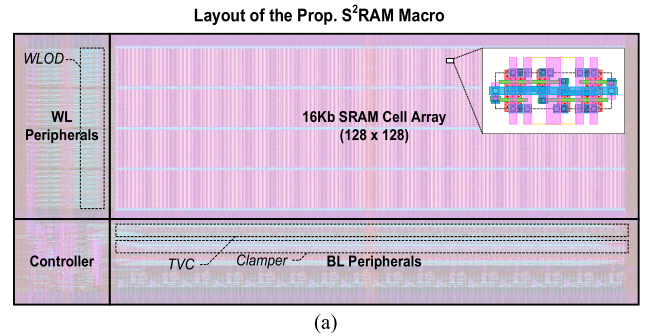
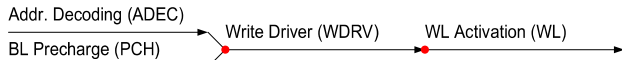


FIGURE 19. (a) Macro-level layout, (b) design summary, and (c) area breakdown of the proposed S²RAM.

memory access occupies 6.1% of the total area. As presented in Fig. 13, the adopted SBL circuit has no area overhead compared to the typical PMOS pull-up based precharge circuit. Meanwhile, WLOD based on capacitive-coupling and TVC based on ratioed-circuit occupy 2.5% and 5.1% of the total area, respectively. Based on post-layout simulation, minimum operation delay and dynamic power are compared with the conventional 16kb SRAM, which can only support random access. For comparison in the same operating voltage range, the conventional SRAM includes TVC and SBL as an assist scheme. To compare the minimum operation delay, the basic operation sequence of read and write operations is defined as shown in Fig. 20(a). For SRAM assist, based on the results in Fig. 14, only SBL is activated at a supply voltage of 0.9V, and both SBL and TVC are activated at 0.6V. As can be seen from the waveform in Fig. 13, the BL clamper of the proposed S²RAM is controlled to be activated only when necessary. As shown in Fig. 20(b), at a supply voltage of 0.9V, the operation delays of the conventional SRAM and the proposed S²RAM operating in random access mode are almost the same. When the proposed S²RAM operates in sequential access mode, WL is activated only in the last cycle of the write operation and only in the first cycle of the read operation. Therefore, in other cycles without WL activation, minimum operation delay is reduced by 12% for write and 25% for read. At 0.6V supply voltage (Fig. 20(c)), the BL precharge delay is larger than the address decoding delay due to the pull-up path driven by the NMOS. Similarly, the proposed S²RAM in sequential access mode operates the write-driver via NMOS pull-up path, unlike random access mode which uses PMOS pull-up path, resulting in 5.8% increase in write-driver delay.

Write Cycle



Read Cycle

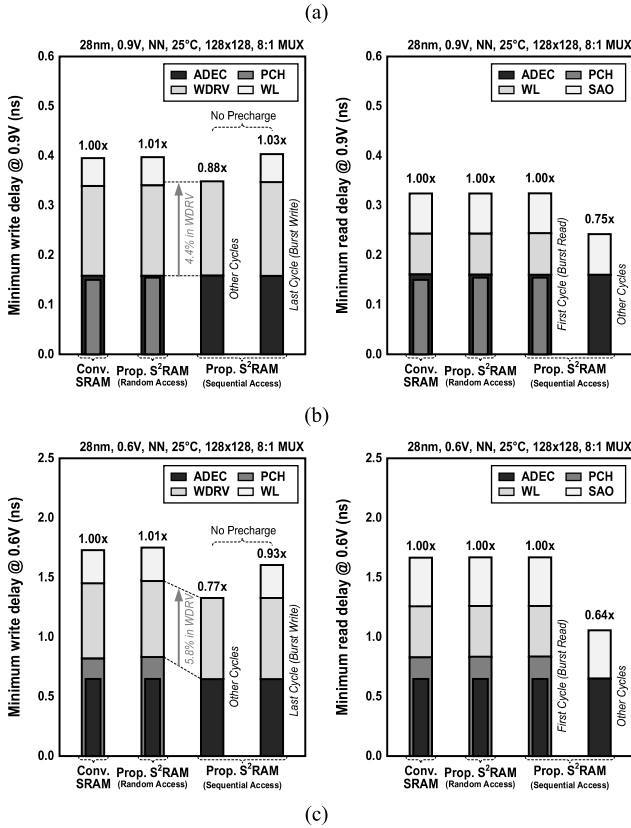
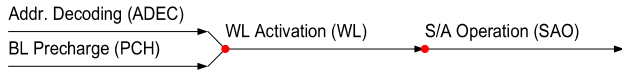


FIGURE 20. (a) Baseline of operation sequence, and comparison of minimum operation delay at (b) 0.9V and (c) 0.6V.

However, since the BL precharge operation is omitted in sequential write operations, the minimum write delay is reduced by 7% compared to the conventional SRAM.

To compare the dynamic power consumption, the average power for each block of the conventional SRAM and the proposed S²RAM is simulated over 16 operating cycles (half read and half write). As shown in Fig. 21, each block is divided into WL/BL peripherals, controller, cell array, BL switching, and assist circuit. Here, the average power of the cell array is measured separately to closely examine the overhead for recovering the cell supply voltage of the TVC. Similarly, to focus on the power associated with BL switching, the BL precharge and clamber (‘BL Switch’ in Fig. 21) are measured separately from the BL peripheral circuitry (‘BL-Peri.’ in Fig. 21). As shown in Fig. 21, the controller and BL switching are the main components of power consumption during SRAM operation. The proposed S²RAM shows increased controller power of 15.6% and 7.4%

TABLE 1. Comparison with the state-of-the-arts.

	Baseline	Sequential SRAM [20]	Prop. S ² RAM
Technology	28nm	7nm	28nm
Capacity	16Kb	8Mb	16Kb
Dedicated Scheme	BL-ADDR [0 1 2 3] PCH WL BL[0] BL[3]	BL-ADDR [p r 0 1 2 3 4] PCH WL BL[0] BL[3]	BL-ADDR [0 1 2 3] PCH WL BL[0] BL[3]
Purpose	-	Low Power w/o VDD ↓	Low Power w/o VDD ↓
Access	Random	Sequential Only	Random & Sequential
Assist	TVC + SBL	WLOD + ¹⁾ SBL	TVC + WLOD + SBL
Overhead	Excessive BL Power	Row Act. Cycles (RAC)	Reconfig. Controller
²⁾ Write Energy	2.65pJ/Byte @0.90V (1.0x) 2.20pJ/Byte @0.60V (1.0x)	³⁾ 0.91pJ/Byte @0.90V (0.34x) ³⁾ 0.78pJ/Byte @0.60V (0.35x) 0.560pJ/B @0.75V (reported)	0.86pJ/Byte @0.90V (0.32x) 0.89pJ/Byte @0.60V (0.40x)
²⁾ Read Energy	2.64pJ/Byte @0.90V (1.0x) 1.81pJ/Byte @0.60V (1.0x)	³⁾ 0.70pJ/Byte @0.90V (0.26x) ³⁾ 0.38pJ/Byte @0.60V (0.21x) 0.468pJ/B @0.75V (reported)	0.67pJ/Byte @0.90V (0.25x) 0.43pJ/Byte @0.60V (0.24x)
Area	11132 μm ² (1.00x)	⁴⁾ 11016 μm ² (0.97x)	12372 μm ² (1.08x)

1) In the original work, only WLOD is used.
2) 500MHz @0.9V, 333MHz @0.6V
3) Reanalyzed based on 28nm CMOS technology and 16Kb SRAM (128 W/Ls & 128 BLs) due to the different technology node
4) Due to different write assist scheme (sequential SRAM: MOS capacitor based WLOD)

at 0.9V and 0.6V, respectively, compared to the conventional SRAM in random access mode. However, this only represents about a 3~7% increase in total power consumption, and the proposed S²RAM provides significant power savings opportunities without performance degradation as it activates WL and control signals only when needed during sequential access operations. Additionally, the proposed S²RAM consumes less power during TVC recovery due to its collaborative assist scheme when the supply voltage is 0.6V, resulting in a 40.8% reduction in cell array power compared to conventional SRAM. As a result, the proposed S²RAM shows a total power reduction of 71% and 67% at 0.9V and 0.6V, respectively, compared to the conventional SRAM during sequential access operation. Fig. 22 shows the average power consumption change according to BL MUX configuration. The array configuration is fixed at 128 × 128 and the BL MUX based interleaved columns are incremented by an exponential of 2. That is, for an N:1 BL MUX, the number of I/O bits is 128 divided by N. As shown in Fig. 22, the larger the bundle unit of the BL MUX, the fewer BL peripherals are required, thus reducing the average power in 2N cycles. Also, in the proposed S²RAM, during sequential access operation, WL is activated only once every 4 cycles in 4:1 BL MUX configuration, but WL is activated only once every 32 cycles in 32:1 BL MUX configuration. For this reason, unlike conventional SRAM where power reduction occurs only in the BL peripherals, the proposed S²RAM exhibits reduced power consumption for all components when operating in sequential access mode. As a result, in the proposed S²RAM, as the number of interleaved columns increases, greater power savings can be achieved during sequential access (Fig. 22(c)).

Table 1 shows a comparison with the state-of-the-art. The main difference between the sequential SRAM in [20] and the proposed S²RAM is the WL activation method. In sequential

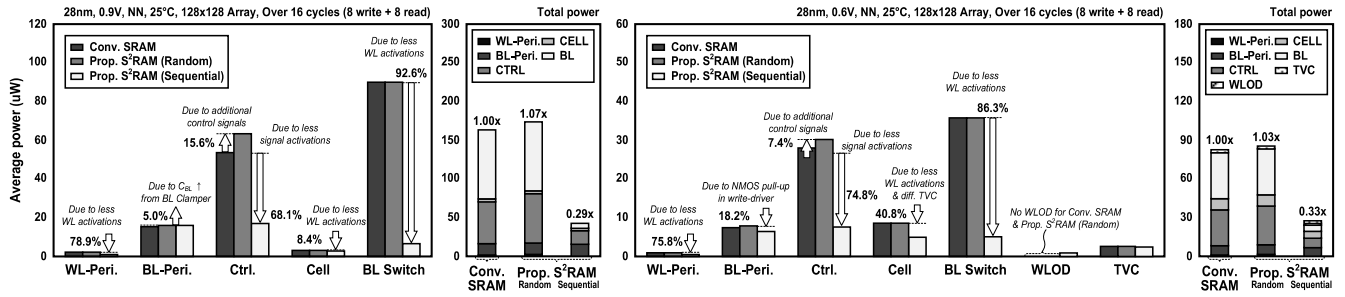


FIGURE 21. Comparison of average power consumption at 0.9V (left side) and 0.6V (right side).

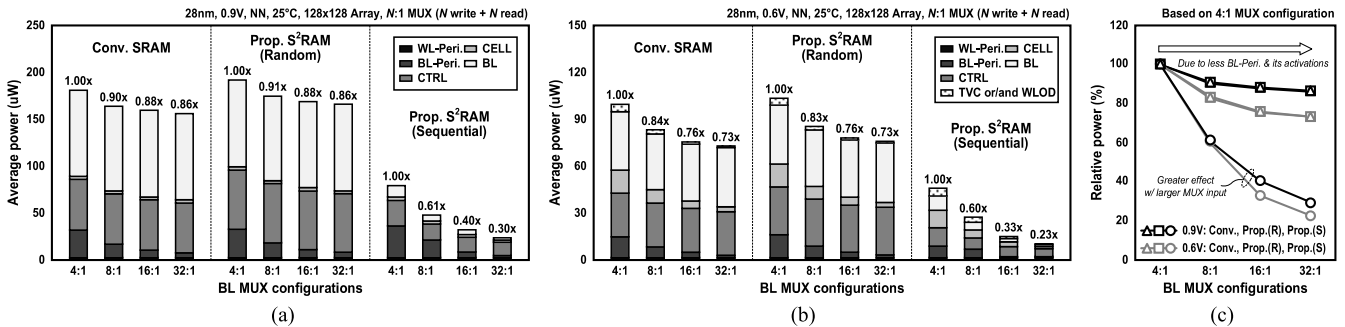


FIGURE 22. Average power depending on BL MUX configuration at (a) 0.9V and (b) 0.6V, and (c) relative power reduction rates.

SRAM [20], sequential access operations are performed after row activation cycles (RACs), which consists of 2 to 3 cycles of BL precharge, WL activation, and differential BL voltage development steps. At this time, the activated WL continues until the WL address changes. For comparison, the 16kb sequential SRAM is implemented based on the same 28nm CMOS technology. Because the proposed S²RAM adopts a write assist scheme that considers both random and sequential access, it shows a 12% increased area and about 13% higher operating energy at 0.6V compared to sequential SRAM. However, the proposed S²RAM shows a similar level of power savings as sequential SRAM compared to the conventional SRAM (baseline in Table 1), and supports both random and sequential access.

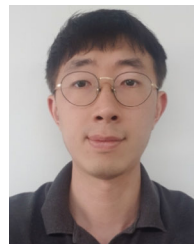
V. CONCLUSION

In this paper, we present S²RAM that supports both random and sequential access, optimized by considering memory access patterns and internal structure. A memory controller for workload-dependent reconfigurable access modes and a BL clamber for using BL as reliable temporary storage are proposed. For SRAM assist, which is essential for reliable SRAM operation in advanced technology nodes, an optimized assist scheme that allows BL clammers to operate cooperatively during sequential write operations is also proposed. The proposed 16kb S²RAM implemented in 28nm CMOS technology achieves a 60% to 76% reduction in operating energy and an 8% increase in area overhead compared to conventional SRAM.

REFERENCES

- [1] Y. Wang, H. Yu, L. Ni, G.-B. Huang, M. Yan, C. Weng, W. Yang, and J. Zhao, "An energy-efficient nonvolatile in-memory computing architecture for extreme learning machine by domain-wall nanowire devices," *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 998–1012, Nov. 2015.
- [2] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.
- [3] J. Hu, C. J. Xue, Q. Zhuge, W.-C. Tseng, and E. H.-M. Sha, "Data allocation optimization for hybrid scratch pad memory with SRAM and nonvolatile memory," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 21, no. 6, pp. 1094–1102, Jun. 2013.
- [4] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, and R. Boyle, "In-datacenter performance analysis of a tensor processing unit," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Toronto ON Canada, Jun. 2017, pp. 1–12.
- [5] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.
- [6] Y. Ju and J. Gu, "A systolic neural CPU processor combining deep learning and general-purpose computing with enhanced data locality and end-to-end performance," *IEEE J. Solid-State Circuits*, vol. 58, no. 1, pp. 216–226, Jan. 2023.
- [7] S. Nalam and B. H. Calhoun, "5T SRAM with asymmetric sizing for improved read stability," *IEEE J. Solid-State Circuits*, vol. 46, no. 10, pp. 2431–2442, Oct. 2011.
- [8] D. Jeon, Q. Dong, Y. Kim, X. Wang, S. Chen, H. Yu, D. Blaauw, and D. Sylvester, "A 23-mW face recognition processor with mostly-read 5T memory in 40-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 52, no. 6, pp. 1628–1642, Jun. 2017.
- [9] Y. Yang, J. Park, S. C. Song, J. Wang, G. Yeap, and S.-O. Jung, "Single-ended 9T SRAM cell for near-threshold voltage operation with enhanced read performance in 22-nm FinFET technology," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 11, pp. 2748–2752, Nov. 2015.

- [10] T. Song, W. Rim, S. Park, Y. Kim, G. Yang, H. Kim, S. Baek, J. Jung, B. Kwon, S. Cho, H. Jung, Y. Choo, and J. Choi, "A 10 nm FinFET 128 mb SRAM with assist adjustment system for power, performance, and area optimization," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 240–249, Jan. 2017.
- [11] Y. Yang, H. Jeong, S. C. Song, J. Wang, G. Yeap, and S.-O. Jung, "Single bit-line 7T SRAM cell for near-threshold voltage operation with enhanced performance and energy in 14 nm FinFET technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 7, pp. 1023–1032, Jul. 2016.
- [12] J. Chang, Y. H. Chen, W. M. Chan, S. P. Singh, H. Cheng, H. Fujiwara, J. Y. Lin, K. C. Lin, J. Hung, R. Lee, and H. J. Liao, "A 7 nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-V_{MIN} applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 206–207.
- [13] T. Song, J. Jung, W. Rim, H. Kim, Y. Kim, C. Park, J. Do, S. Park, S. Cho, H. Jung, B. Kwon, H.-S. Choi, J. Choi, and J. S. Yoon, "A 7 nm FinFET SRAM using EUV lithography with dual write-driver-assist circuitry for low-voltage applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 198–200.
- [14] T. Song, W. Rim, H. Kim, K. H. Cho, T. Kim, T. Lee, G. Bae, D.-W. Kim, S. Kwon, S. Baek, J. Jung, J. Kye, H. Jung, H. Kim, S.-M. Jung, and J. Park, "A 3 nm gate-all-around SRAM featuring an adaptive dual-BL and an adaptive cell-power assist circuit," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 338–340.
- [15] W. Choi and J. Park, "A charge-recycling assist technique for reliable and low power SRAM design," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 8, pp. 1164–1175, Aug. 2016.
- [16] K. Kim, T. W. Oh, and S.-O. Jung, "Bitline charge sharing suppressed bitline and cell supply collapse assists for energy-efficient 6T SRAM," *IEEE Access*, vol. 9, pp. 57393–57403, 2021.
- [17] K. Kozu, Y. Tanabe, M. Kitakami, and K. Namba, "Low power neural network by reducing SRAM operating voltage," *IEEE Access*, vol. 10, pp. 116982–116986, 2022.
- [18] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek, "An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 956–963, Apr. 2008.
- [19] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, Feb. 2009.
- [20] D. H. Morris, H. Liu, T. F. Wu, H. E. Sumbul, E. Ansari, A. Barachant, J. Reid, and E. Beigne, "Co-optimization of SRAM circuits with sequential access patterns in a 7 nm SoC achieving 58% memory energy reduction for AR applications," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Honolulu, HI, USA, Jun. 2022, pp. 216–217.
- [21] J. Burgess, "RTX on—The NVIDIA Turing GPU," *IEEE Micro*, vol. 40, no. 2, pp. 36–44, Mar. 2020.
- [22] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 tensor core GPU: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, Mar. 2021.
- [23] M. Nabavi and M. Sachdev, "A 290-mV, 3.34-MHz, 6T SRAM with pMOS access transistors and boosted wordline in 65-nm CMOS technology," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 656–667, Feb. 2018.
- [24] V. P. Hu, M.-L. Fan, P. Su, and C.-T. Chuang, "Analysis of GeOI FinFET 6T SRAM cells with variation-tolerant WLUD read-assist and TVC write-assist," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1710–1715, Jun. 2015.
- [25] P. Kolar, E. Karl, U. Bhattacharya, F. Hamzaoglu, H. Nho, Y.-G. Ng, Y. Wang, and K. Zhang, "A 32 nm high-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 76–84, Jan. 2011.
- [26] Z. Guo, D. Kim, S. Nalam, J. Wiedemer, X. Wang, and E. Karl, "A 23.6-Mb/mm² SRAM in 10-nm FinFET technology with pulsed-pMOS TVC and stepped-WL for low-voltage applications," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 210–216, Jan. 2019.
- [27] M. F. Chang, J. J. Wu, K. T. Chen, Y. C. Chen, Y. H. Chen, R. Lee, H. J. Liao, and H. Yamauchi, "A differential data-aware power-supplied (D²AP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin applications," *IEEE J. Solid-State Circuits*, vol. 45, no. 6, pp. 1234–1245, Jun. 2010.
- [28] H. Jeong, T. Kim, Y. Yang, T. Song, G. Kim, H.-S. Won, and S.-O. Jung, "Offset-compensated cross-coupled PFET bit-line conditioning and selective negative bit-line write assist for high-density low-power SRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 4, pp. 1062–1070, Apr. 2015.
- [29] W. Choi and J. Park, "Improved perturbation vector generation method for accurate SRAM yield estimation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 9, pp. 1511–1521, Sep. 2017.
- [30] J. Kulkarni, B. Geuskens, T. Karnik, M. Khellah, J. Tschanz, and V. De, "Capacitive-coupling wordline boosting with self-induced VCC collapse for write VMIN reduction in 22-nm 8T SRAM," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2012, pp. 234–236.



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, where he was involved with SRAM design in advanced technology nodes (14 nm FinFET ~ 3 nm GAA). Since 2019, he has been an Assistant Professor with the Department of Electronics Engineering, Sookmyung Women's University, Seoul. His current research interests include the neural network accelerator and embedded memory designs in advanced technologies.

• • •