A 4 + 2T SRAM for Searching and In-Memory Computing With 0.3-V V_{DDmin}

Qing Dong, *Student Member, IEEE*, Supreet Jeloka, *Student Member, IEEE,* Mehdi Saligane, Yejoong Kim, *Student Member, IEEE,* Masaru Kawaminami, Akihiko Harada, Satoru Miyoshi, Makoto Yasuda, David Blaauw, *Fellow, IEEE*, and Dennis Sylvester, *Fellow, IEEE*

***Abstract*—This paper presents a 4 + 2T SRAM for embedded searching and in-memory-computing applications. The proposed SRAM cell uses the n-well as the write wordline to perform write operations and eliminate the write access transistors, achieving 15% area saving compared with conventional 8T SRAM. The decoupled differential read paths significantly improve read noise margin, and therefore reliable multi-word activation can be enabled to perform in-memory Boolean logic functions. Reconfigurable differential sense amplifiers are employed to realize fast normal read or multi-functional logic operations. Moreover, the proposed 4 + 2T SRAM can be reconfigured as binary content-addressable memory (BCAM) or ternary content-addressable memory (TCAM) for searching operations, achieving 0.13 fJ/search/bit at 0.35 V. The chip is fabricated in 55-nm deeply depleted channel technology. The area efficiency is 65% for a 128 x 128 pushed-rule array including all peripherals such as column-wise sense amplifier for read/logic and row-wise sense amplifier for BCAM/TCAM operations. Forty dies across five wafers in different corners are measured, showing a worst-case read/write V_{DDmin} of 0.3 V.**

***Index Terms*—Binary content-addressable memory (BCAM), deeply depleted channel (DDC), in-memory computing (IMC), memory, SRAM, ternary content-addressable memory (TCAM).**

I. Introduction

Conventional von Neumann (CVN) architecture continuously transfers data between memory banks and compute elements, incurring substantial energy and latency costs that can dominate system power and performance. As shown in Fig. 1, to perform Boolean logic function on two words, a CVN architecture involves three major steps: 1) two memory read cycles; 2) data movement from memory hierarchy to computing element's registers; and 3) computation within the ALU. Both power consumption and latency are largely dominated by the two memory accesses and the data movement (Fig. 2) [1]–[3]. To minimize the power and latency, in-memory computing (IMC) has been recently proposed for data processing directly inside an on-chip memory macro [4], [5]. As shown in Fig. 1, IMC activates multiple rows simultaneously and computes the logical functions directly on the bitline (BL). Computation results are available immediately at the end of a memory access. Therefore, only one cycle is required instead of multiple cycles of latency [Fig. 2(a)]. As a result, IMC can reduce the energy costs associated with both data movement and memory access, as shown in Fig. 2(b). Since memory banks are typically very wide (many words per word line), it also potentially provides inherently parallel computation.

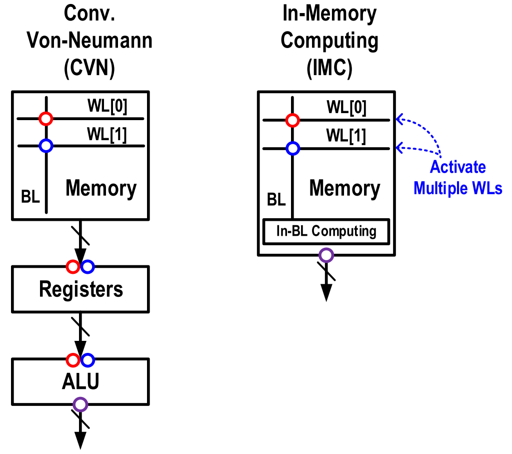


Fig. 1. Data flow required for reading two words to perform logic operation in CVN architecture and IMC system.

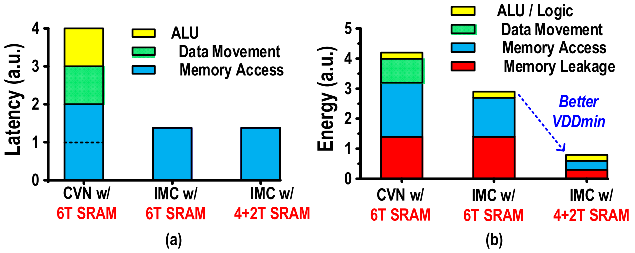


Fig. 2. Comparison between CVN and IMC in terms of (a) latency and (b) energy for a two-word Boolean operation.

Several silicon-validated works have recently been proposed that implement IMC, with most approaches using conventional 6T SRAM [6], [7]. With parallel operation of all rows to perform scalar product, a machine-learning classifier can be realized using standard 6T SRAM array [6]. By activating two wordlines (WLs) in a standard 6T SRAM array,

Manuscript received August 7, 2017; revised October 1, 2017; accepted November 13, 2017. Date of publication December 11, 2017; date of current version March 23, 2018. This paper was approved by Guest Editor Makoto Ikeda. *(Corresponding author: Qing Dong.)*

Q. Dong is with TSMC, San Jose, CA 95134 USA (e-mail: qingdong@umich.edu).

S. Jeloka is with ARM, Austin, TX 78735 USA.

M. Saligane, Y. Kim, D. Blaauw, and D. Sylvester are with the University of Michigan, Ann Arbor, MI 48105 USA.

M. Kawaminami and A. Harada are with Fujitsu Electronics America, Inc., Sunnyvale, CA 94085 USA, and also with Mie Fujitsu Semiconductor Ltd., Kuwana 2-100-45, Japan.

S. Miyoshi is with Fujitsu Electronics America, Inc., Sunnyvale, CA 94085 USA.

M. Yasuda is with Mie Fujitsu Semiconductor Ltd., Kuwana 2-100-45, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2776309

AND/NOR operation of the two words can be directly obtained on the BLs [7]. However, conventional 6T SRAM suffers from degraded read noise margins when multiple rows are activated. Therefore, WL under-drive is required to improve the noise margin [7], which lowers the read performance. Moreover, V_{DDmin} is limited to approximately 0.7 V because of the degraded read noise margin. As a result, overall power consumption is high. Due to the poor V_{DDmin} scalability of 6T SRAM, the memory access and leakage still dominate the system energy consumption of an IMC-based system [Fig. 2(b)]. It is well known that the 8T SRAM topology improves read noise margin by decoupling read and write paths [8]. The decoupled read/write paths enable independent optimization of read and write operations, significantly enhancing operating margins. Larger margins obtained from the isolated read/write paths lower the SRAM V_{DDmin}, which translates to reduced power consumption. However, the additional two transistors incur 30% area overhead or more [9]–[11]. In addition, because of the single read BL, it only provides the results of the AND operation while other logic functions cannot be supported in 8T SRAM. To address this problem, 10T SRAM has to be used for IMC [12], sacrificing area for functionality.

In this paper, we propose a 4 + 2T SRAM cell with better noise margins than 6T SRAM and less area overhead than 8T SRAM [13]. It also supports multiple computing functions with low energy consumption due to its voltage scalability. The proposed 4 + 2T SRAM cell uses the n-well as a write WL (WWL), eliminating the access transistors and resulting in a 4T-core memory cell. Two decoupled read paths (2T) significantly improve read noise margin, enabling reliable multi-word activation for IMC while limiting area overhead to 15% over 6T SRAM in the same technology. Using dual sense amplifiers per column, Boolean logic functions (AND, OR, and XOR) between the two activated words can be realized. Furthermore, with separated read BLs (RBL/RBLB) and read WLs (RWL/RWLB), the SRAM can be configured as a binary content-addressable memory (BCAM) or a ternary content-addressable memory (TCAM), enabling searching operations. The memory cell is designed using pushed rules in 55-nm deeply depleted channel (DDC) technology, which offers a high body coefficient and low process variation [14]–[17]. Measurement of 40 dies across five corner wafers shows a worst-case V_{DDmin} of 0.3 V for SRAM operations. The BCAM achieves 0.13 fJ/search/bit at 0.35-V power supply.

The remainder of this paper is organized as follows. Section II generally introduces the structure, operation, and layout of the proposed 4 + 2T SRAM. Section III presents the n-well-based write method and analyzes the disturbances. Section IV compares the normal differential read operation with in-memory logic operations. Section V describes the configurations of BCAM and TCAM in detail. Section VI discusses measurement results of the proposed design, and finally, conclusions are presented in Section VII.

II. 4 + 2T SRAM Cell

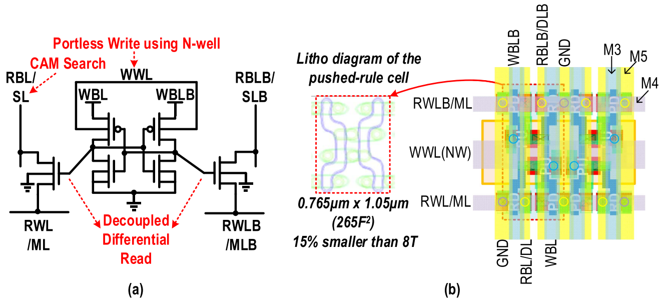


Fig. 3. 4 + 2T SRAM cell. (a) Schematic. (b) Layout.

Fig. 3(a) shows the schematic of the proposed 4 + 2T SRAM cell, which has six transistors. Four of them form the cross-coupled inverter pair storing data internally and the other two are access transistors for read. The design includes two gate-connected differential read ports, isolating storage nodes with RWL and RBL/RBLB. Unlike the conventional 6T SRAM in which BLs are shorted to storage nodes when WLs are activated during a read, the proposed 4 + 2T SRAM has much weaker impact on the storages nodes from RWL and RBL/RBLB. Therefore, the read noise margin can be significantly improved. As a result, reliable multi-word activation for IMC is achieved. Moreover, with the differential RBL and RBLB available, multiple in-memory logic functions (AND/OR/XOR) can all be realized.

In this proposed 4 + 2T SRAM cell, cross-coupled inverters have separated V_{DD} terminals, which serve as complementary write BLs (WBL/WBLB), respectively. To perform the actual write, we lower one of the two supply lines while at the same time lowering the voltage on n-well, which acts as the WWL. This reduction of the n-well voltage lowers the threshold voltage of the PMOSs, which allows for the data input on the WBL and WBLB to be written into the storage nodes. The write margin can be significantly improved with DDC technology, which offers the dual advantages of low process variation and high body coefficient [14]–[17]. Detailed write operation will be discussed in Section III.

Content-addressable memory (CAM) searches its entire memory for the input data. If the data is found, a CAM returns all address locations of matching words. A BCAM searches for an exact match, while a TCAM allows for partial matching with "don't care" bits in the memory. To realize BCAM or TCAM using the 4 + 2T cell, we separate the RWL of the two access transistors into matching lines (ML/MLB). In addition, the RBL and RBLB are reconfigured to supply search data inputs: SL and SLB. BCAM can be realized with a single 4 + 2T SRAM cell; while TCAM requires two cells. Section V further describes the detailed configurations.

Fig. 3(b) shows the layout and lithographic simulation of the proposed 4+2T SRAM cell. It is designed using pushed design rules in 55-nm DDC technology. The layout is lithographically symmetric and the sources of both read-access transistors and pull-down transistors can be shared with adjacent cells to save area. Thanks to significantly improved process variation of DDC technology, bent poly is tolerable. All six transistors are minimum-sized high-Vt (HVT) ultra-low-leakage device for leakage reduction. The cell size is 0.765 by 1.05 μm, which is 265F^2 when normalized to technology node feature size (F).

TABLE I
DETAILED OPERATION TABLE

		WWL	WBL	WBLB	RWL/ML	RWLB/MLB	RBL/SL	RBL/SLB

Memory Operations	WRITE	GND(Sel.)
		VDDH(Unsel.)
			GND(Write0)
			VDD(Write1)
				VDD(Write0)
				GND(Write1)
					VDD	VDD	Floating	Floating

	READ	VDD*	VDD	VDD	GND	GND	Precharge(VDD)	Precharge(VDD)
	HOLD	VDD*	VDD	VDD	VDD	VDD	Floating	Floating

CAM Operations		VDD*	VDD	VDD	Precharge(VDD)	Precharge(VDD)	VDD(Search 0)
							GND(Search1)
								GND(Search 0)
								VDD(Search1)

Logic Operations	AND	VDD*	VDD	VDD	GND	VDD	Precharge(VDD)	Floating
	OR	VDD*	VDD	VDD	VDD	GND	Floating	Precharge(VDD)
	XOR	VDD*	VDD	VDD	GND	GND	Precharge(VDD)	Precharge(VDD)

*Can also be kept at VDDH.

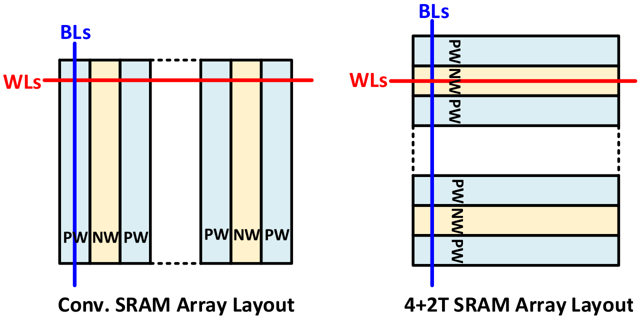


Fig. 4. Rotated well orientation for 4 + 2T SRAM array.

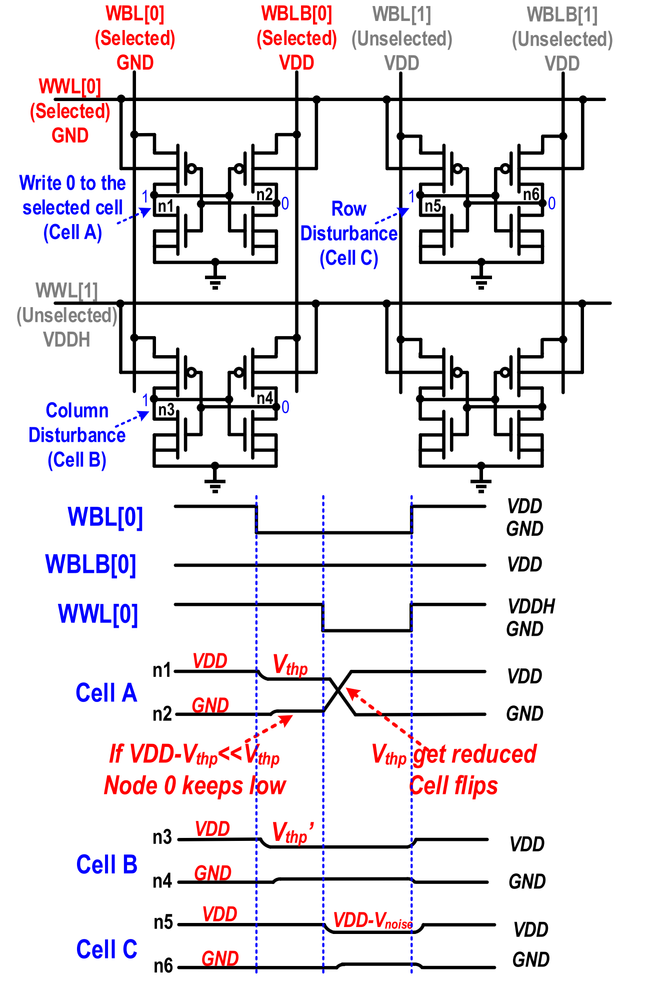


Fig. 5. n-well-based write method and analysis of disturbances.

The proposed 4 + 2T SRAM achieves 15% smaller cell area than conventional 8T SRAM while maintaining comparable robustness. Since, we have separated all WLs and BLs in order to realize multiple functions, up to five metal layers are used in this cell design, which has two more metal layers than conventional 6T SRAM. Detailed metal usage is shown in Fig. 3(b). In conventional 6T and 8T SRAM array layouts, WLs run in horizontal while n-wells are vertical, as shown in Fig. 4. Since the proposed 4 + 2T SRAM uses n-wells as the WLs, the array orientation is rotated so that NW runs in the same direction as WLs.

The proposed 4+2T SRAM can therefore be reconfigured as a basic memory, BCAM/TCAM, and logic-in-memory. Table I summarizes the voltages applied on each terminal for basic memory, CAM, and logic operations. Write operation employs two supply voltages, whereas other operations use a single supply voltage.

III. WRITE METHOD

A. n-Well-Based Cell Write

Due to the strong body effect in DDC technology [14]–[17], the n-well can be used as WWL. Fig. 5 shows the write scheme applied to the 4T structure in a 2 × 2 array. Cell A is the selected cell, while cell B and cell C are the half-selected cell. In this proposed 4 + 2T SRAM, two supply voltages are necessary during write operation for write assist. In standby mode, both WBL[0] and WBLB[0] are set to V_{DD}, and WWL[0] stays at another voltage V_{DDH}, which is higher than V_{DD}. To write 0 into the node $n1$ of the selected cell A, the selected WBL[0] is lowered from V_{DD} to GND, while WBLB[0] remains at V_{DD}. Then, the voltage on the node $n1$ will be reduced to V_{thp} from V_{DD}. To determine whether the cell can be flipped, we compare the voltage on the left node $n1$, which is V_{thp}, with the turn-on voltage of the right PMOS, which is $V_{\mathrm{DD}} - V_{\mathrm{thp}}$. If $V_{\mathrm{DD}} - V_{\mathrm{thp}}$ is less than V_{thp}, then the right PMOS will not be turned on and the node $n2$ stays low. Since the ultra-low-leakage PMOS device used in this SRAM has high threshold voltage, cell would not be flipped during this stage. Once WWL[0] gets lowered from V_{DDH} to GND, then the selected PMOS device becomes much stronger due to its forward body bias. V_{thp} is therefore significantly reduced. As a result, the right PMOS turns on and flips the cell within hundreds of picoseconds. After write, both WWL and WBL are reset to their standby voltage. This completes the write sequence of the selected cell A.

B. Disturbance Analysis

Successful write needs to consider both selected write and half-select disturbance issues. There are two types of half-select disturbances that must be taken into account in this proposed 4 + 2T SRAM: column wise (cell B) and row wise (cell C).

For the column disturbance, cell B might be flipped because its WBL[0], which is shared with the selected cell A,

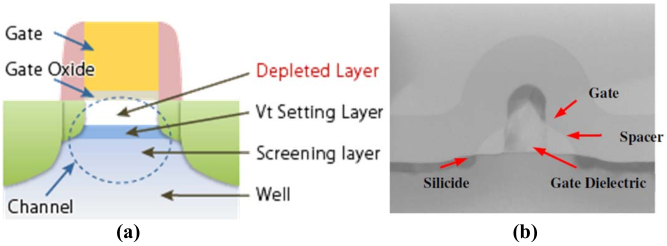


Fig. 6. (a) Structure and (b) TEM of DDC device.

is lowered to ground. This reduction of voltage on WBL[0] reduces the voltage of the left node $n3$ of cell B to V_{thp}. As mentioned above, V_{thp} has to be high enough such that the cell would not flip. Therefore, we use the higher voltage V_{DDH} for the unselected WWL to reverse body bias the PMOSs in unselected rows, such that their V_{thp} are increased to avoid this type of disturbance.

Conversely, all cells in the selected row have stronger PMOS devices during a write, increasing the chance of un-intended write into their internal nodes. Specifically, cell C might be flipped due to noise on its WBL and WBLB. Both of these signals remain at a common V_{DD}. However, local coupling noise can generate small voltage differences. Since both PMOSs are forward body biased, even a small voltage difference might affect the robustness. To avoid row disturbance, the V_{DD} on WBL/WBLB must be high enough to suppress the coupling noise. This noise limits the write V_{DDmin}. Moreover, forward body-biased PMOSs on the selected row can generate diode currents flowing through V_{DD} to GND. To avoid unexpected power consumption, the V_{DD} is limited to 0.8 V in this paper.

Since the proposed write method of the 4 + 2T SRAM applies forward body bias to the selected row and reverse body bias to the unselected rows, it relies heavily on body coefficient. Because DDC technology has two main advantages: strong body effect and low process variation [14]–[17], the proposed write method can benefit from the DDC technology employed in this paper.

C. DDC Technology

Fig. 6 shows the structure and TEM photograph of an ultra-low-leakage device in 55-nm DDC technology [14]–[17]. It has a very low-doped layer to reduce random dopant fluctuation. A Vt setting layer enables multiple threshold voltage devices. The screening layer terminates the depletion in the channel and also serves to smooth the depletion layer across the device, affording excellent σVt and short channel effects. Together, these three deep layers produce a strong body coefficient. Fig. 7(a) shows the body effect compared with BULK technology [15]. DDC technology provides much higher body factor than conventional BULK technology. Moreover, the body leakage is much smaller, enabling low-power operation when PMOSs are forward body biased during write operation. Fig. 7(b) shows the Vt variation comparison [15], [16]. The red ellipses show the Vt distribution of DDC technology. Compared with conventional BULK technologies in blue, DDC technology significantly improves the process variation with the low-doped channel and highly doped screen layer.

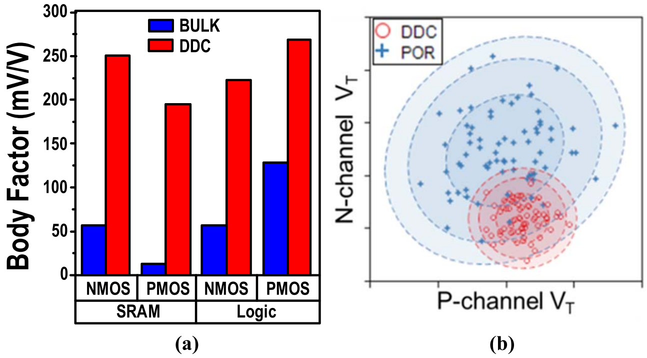


Fig. 7. (a) Strong body effect and (b) low process variation of DDC technology.

D. Simulation Results of Write Margin

We performed Monte Carlo simulation of the dynamic write margin for the proposed 4 + 2T SRAM in different technologies, as shown in Fig. 8. In this simulation, we apply noise source on the feedback loop. At each given noise value, we run Monte Carlo of transient write simulations and check the number of write failures. Then the noise source is increased for next round of transient Monte Carlo simulations. Finally, a cumulative distribution function curve of write failure rate across noise is obtained. By derivation, the probability density functions curve of write failure rate across noise can be used to estimate the mean/sigma of dynamic write margin. All variation sources are included in this simulation. The green square represents the voltage combination of V_{DD} and V_{DDH} with write margin of more than 5σ; while other colors means less than 5σ write margin due to different type of write failures. Fig. 8(a) shows the simulation result from a conventional 40-nm BULK technology with weak body factor; while Fig. 8(b) is simulation result from 55-nm DDC technology with significantly improved body effect. The proposed write method can still function in conventional 40-nm BULK technology with over 5σ write margin. However, it requires significantly higher V_{DDmin} and higher voltage difference between V_{DD} and V_{DDH}. The simulated write margin in 55-nm DDC technology is improved by over 2 compared with conventional 40-nm BULK technology. Other technologies with strong body effect such as FD-SOI may also achieve good write margin with this SRAM cell topology.

IV. Read Operation and In-Memory Computing

A. Differential Read

Basic read operation can be realized with a single decoupled read-port similar to a 5T or 7T SRAM [18], [19]. The proposed design uses the decoupled differential read ports to accelerate read speed and enable multiple logic operations. During a normal read, one pair of RWL/RWLB is activated (pulled low), and one of RBL/RBLB discharges while the other remains high [Fig. 9(a)]. The two small column-wise sense amplifiers

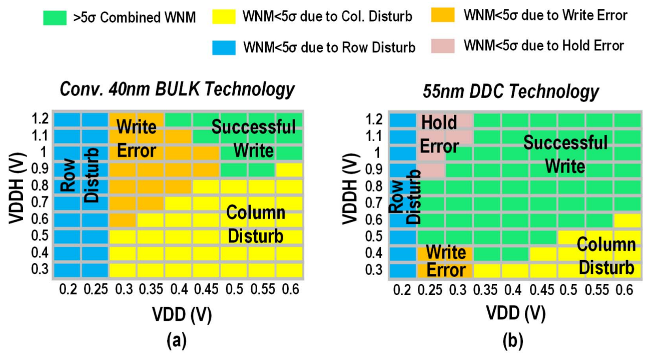


Fig. 8. Comparison of simulated write margin between (a) conventional 40-nm BULK technology and (b) 55-nm DDC technology.

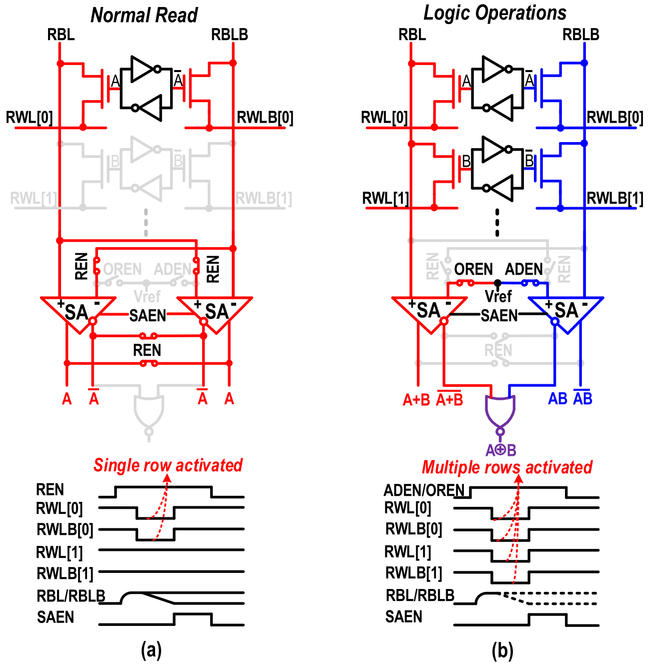


Fig. 9. Comparison between (a) normal read activating a single row and (b) Boolean logic operations (AND/OR/XOR) with multiple rows activated.

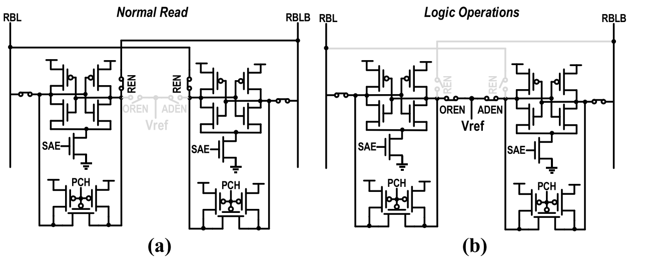


Fig. 10. (a) Normal differential read, where two small sense amplifiers are connected in parallel to form a larger one. (b) Logic operations, where each small sense amplifier evaluates the result for each RBL.

are connected in parallel to form a larger sense amplifier, accelerating the read operation as shown in Fig. 10(a). For the readout circuitry, a differential cross-coupled sense amplifier instead of single-ended inverter is employed to accelerate read speed and reduce access power consumption. Similar to 5T or 7T SRAM [18], [19], the sneaking currents are generated from unselected RWLs to the RBL if the unselected storage nodes stay at 1. Then the discharged RBL cannot be fully lowered to GND and saturates at $V_{\mathrm{DD}} - V_{\mathrm{thn}}$ instead, incurring short circuit current if single-ended inverter is used as sensing circuitry. Therefore, we use differential cross-coupled sense amplifier so that the readout circuitry can quickly distinguish small voltage difference between RBL and RBLB before the sneaking currents appear. Also the utilization of the HVT ULL device helps to reduce the saturation voltage of RBL ($V_{\mathrm{DD}} - V_{\mathrm{thn}}$) and hence minimizes the sneaking currents. Thanks to the significantly improved process variation in DDC technology, the sense amplifier does not require large transistor size to alleviate mismatch, which is necessary in conventional CMOS technology. We use transistors with only twice of minimum size in the core part of the small differential sense amplifier. According to Monte Carlo simulations in 55-nm DDC technology, the standard deviation of the input offset for the small sense amplifier is still less than 15 mV. Since each sense amplifier is compact, the area overhead of using two sense amplifiers is <5% compared to a normal SRAM using a large differential sense amplifier.

B. In-Memory-Logic Operations

For logic operations [Fig. 9(b)], first of all, both RBL and RBLB are precharged to V_{DD} which is the same as precharge phase in normal read operation. Then, two pairs of WLs are activated simultaneously. If one of A and B is 1, the RBL discharges, and therefore RBL represents the NOR operation of A and B. RBLB is connected to the complementary nodes. Similarly, RBLB remains high only if both cell nodes (A and B) store 1, and RBLB therefore provides the AND result of A and B. RBL and RBLB are separately evaluated with the two small differential sense amplifiers as shown in Fig. 10(b). Once the RBL/RBLB is lowered by a small voltage ($< V_{\mathrm{thn}}$) from V_{DD}, the differential sense amplifier will quickly compare the voltage on RBL/RBLB with an external voltage reference (V_{ref}) from PAD before the sneaking currents appear. Because of the differential outputs of sense amplifiers, NAND/AND and NOR/OR results are simultaneously evaluated. Furthermore, a NOR gate between the two sense amplifier outputs generates the XOR result of A and B. All Boolean logic functions are computed in a single read cycle.

V. CAM Configurations

A. BCAM Mode

Conventional BCAM uses specialized 10T bit cells [20], [21], including a storage part (6T SRAM) and a dynamic XNOR part (4T access transistors), which consumes over 2× larger area than a corresponding SRAM. A new BCAM is proposed in [7] that can operate with standard push-rule 6T SRAM cells, significantly reducing area consumption. However, this type of BCAM requires transposed data storage and two cycles per write, complicating word organization and lowering write performance. We propose a BCAM/TCAM configured from 4+2T SRAM which only requires one cycles per write as conventional CAM while achieving comparable area saving with [7].

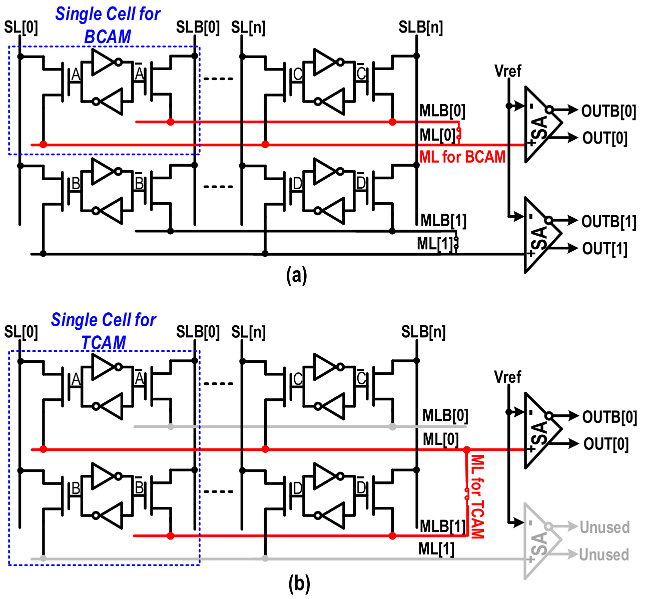


Fig. 11. CAM operations: BL/BLB are reconfigured as SL/SLB; RWL/RWLB act as ML. (a) BCAM uses only one 4 + 2T cell. (b) TCAM uses two cells.

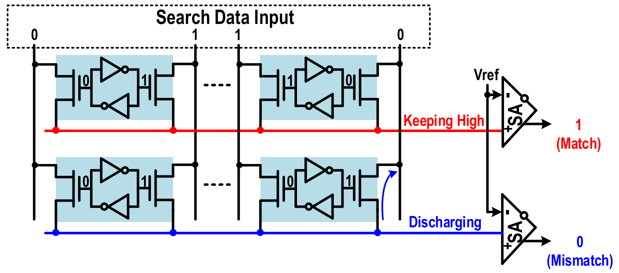


Fig. 12. BCAM search example.

Fig. 11 shows the BCAM/TCAM configurations using the 4 + 2T SRAM. In CAM mode, the RBL/RBLB supply the search data input SL/SLB, and the RWL/RWLB function as ML/MLB. For BCAM operation [Fig. 11(a)], ML and MLB in a row are shorted together as a single ML. If all the input data along the row match the stored data, ML remains high; otherwise ML discharges. Each ML has a row-wise sense amplifier with one side connected to the V_{ref} to evaluate the results, similar to a conventional BCAM [20], [21]. The row-wise sense amplifier has the same circuit structure as column-wise small differential sense amplifier used for logic operations.

Fig. 12 shows an example of the BCAM searching. As the top row matches with the input data, no discharging path generated from ML to SLs. And therefore, the top ML keeps high and the output of the row-wise sense amplifier is 1. The bottom ML discharges since the mismatched bottom right bit creates a discharging path through the right access transistor, resulting in a "0" on the sense amplifier output.

Unlike the previous 6T BCAM [7] that requires transposed data storage with two cycles per update, the proposed BCAM stores data in a normal row-wise fashion instead of column wise such that single cycle per write is realized. Moreover, read margin during searching is not degraded when multiple rows are activated, in contrast to [7].

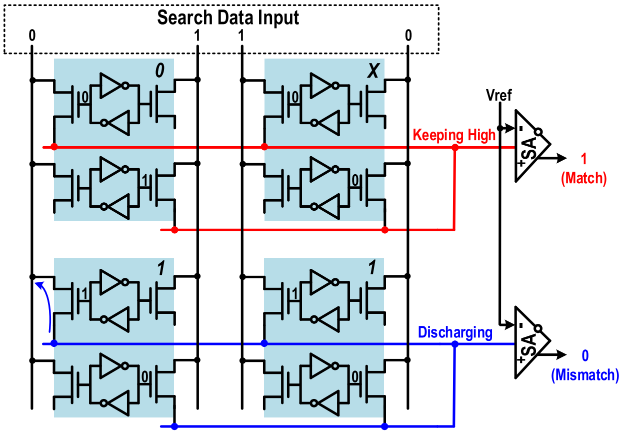


Fig. 13. TCAM search example.

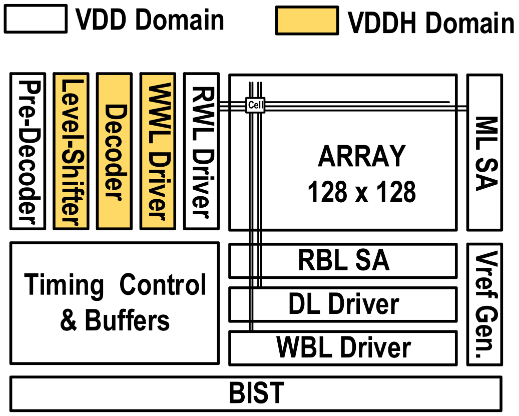


Fig. 14. Detailed block diagram of the proposed SRAM.

B. TCAM Mode

The TCAM cell represents 1/0/X [Fig. 11(b)]. Since it has three states, two SRAM cells are required, spanning two rows. By connecting ML[0] and MLB[1], cell A and cell B can be combined as a single TCAM cell representing 1/0/X when the AB cells store 00/11/01, respectively. The searching and sensing method of TCAM is the same as BCAM.

Fig. 13 shows a TCAM searching example. The bottom ML discharges due to the mismatched bit on the bottom left. The top row has a "don't care" bit, marked as X with both access transistors isolating ML with SLs. As a result, the top ML remains high, resulting 1 on the sense amplifier output.

VI. Measurement Results

Fig. 14 shows the detailed block diagram. We designed a 128 by 128 array that can be configured as SRAM, BCAM, TCAM, and IMC. The yellow shaded blocks are in V_{DDH} domain while all other blocks are in V_{DD} domain. Level converters are inserted between the pre-decoder and the on-pitch NAND3 gates in the main decoder. Therefore, we only need 16 level converters instead of 128 on-pitch level converters, reducing total area by 3%. The proposed SRAM was fabricated in 55-nm DDC technology. Die photograph is shown in Fig. 15. The area efficiency is 65% for a 128 × 128 pushed-rule array including all peripherals such as column-wise sense

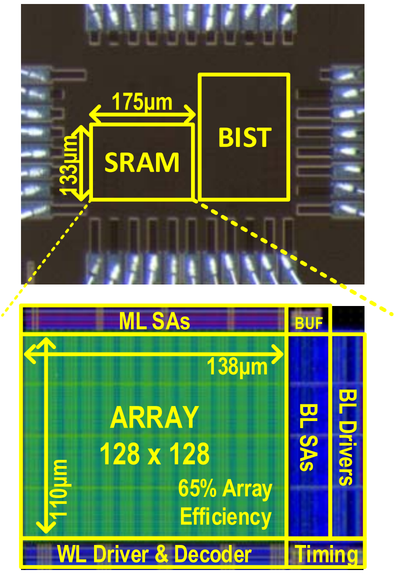


Fig. 15. Die photograph in 55-nm DDC technology. Array efficiency is 65% for the 128 × 128 multi-functional pushed-rule SRAM.

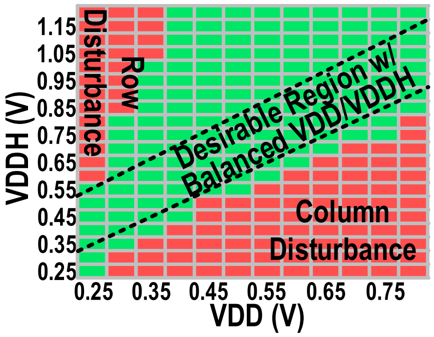


Fig. 16. Measured write shmoo plot of 16-kb array.

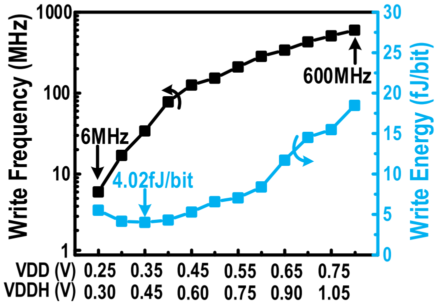


Fig. 17. Write frequency and energy across $V_{\mathrm{DD}}/V_{\mathrm{DDH}}$.

amplifier for read and row-wise sense amplifier for CAM operations.

Fig. 16 shows the measured cell write margin. The green region indicates voltage combination with over 5σ write margin. The balanced green region has at least 200-mV margin, which is sufficient for robust write. Fig. 17 shows the write frequency and energy across V_{DD} and V_{DDH}. At 0.8-V V_{DD}, the write frequency is 600 MHz. The minimum supply voltage is 0.25/0.30 V for $V_{\mathrm{DD}}/V_{\mathrm{DDH}}$, respectively. And the optimal write energy is 4.02 fJ/bit at V_{DD} of 0.35 V.

Fig. 18 shows the CAM frequency and energy across V_{DD}. The measured CAM frequency is the worst case with only 1-bit mismatch; while the energy is measured with half-mismatch data pattern. V_{DDmin} is ~0.35 V for CAM operation, at which the optimal energy/search is 0.13 fJ/bit for BCAM. TCAM achieves the same frequency as BCAM but consumes twice the energy/search/bit.

Fig. 19 shows the optimized frequency and energy across V_{DD} for normal read and logic operations. V_{DDmin} for read is ~0.25 V, whereas it is ~0.35 V for logic operations since they employ single-port sensing and half-strength sense amplifiers. The optimal read energy is 3.96 fJ/bit at 0.35 V; the energy at V_{DDmin} (0.25 V) is higher because the leakage energy overhead exceeds the reduction in dynamic energy. The logic frequency is 30% slower due to weaker sense amplifier and

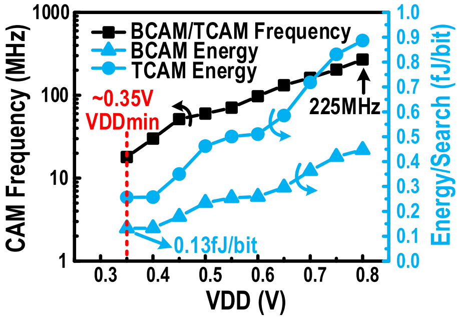


Fig. 18. BCAM/TCAM frequency and energy across V_{DD}.

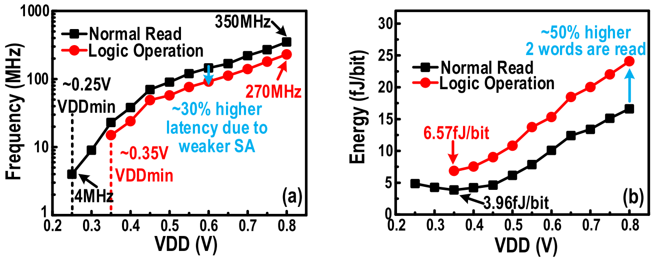


Fig. 19. (a) Frequency and (b) energy comparison between normal read and logic operation.

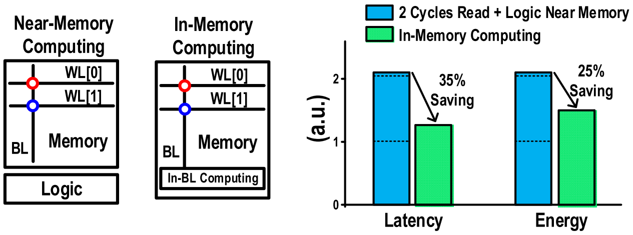


Fig. 20. Compared with near-memory-computing, IMC saves latency and energy by 35% and 25%, respectively.

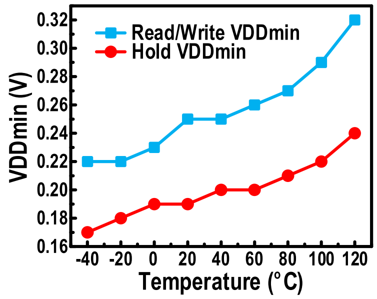


Fig. 21. V_{DDmin} across temperature.

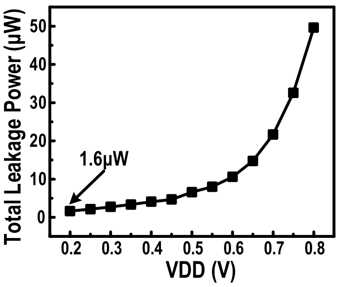


Fig. 22. Leakage power across V_{DD}.

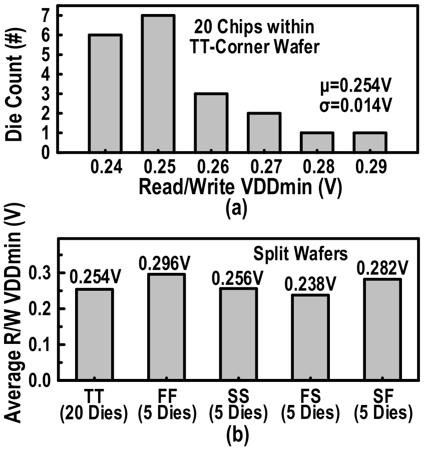


Fig. 23. (a) Within-wafer and (b) split-wafer V_{DDmin} distribution.

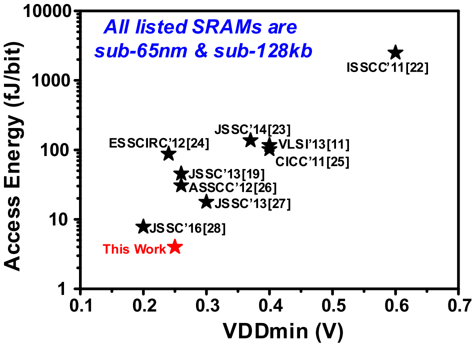


Fig. 24. Comparison with other small-scale sub-65-nm SRAMs in terms of V_{DDmin} and access energy consumption.

TABLE II

COMPARISONS WITH OTHER DECOUPLED SRAM WORKS

		This work		Decoupled SRAM Work		
				VLSI'15[18]	VLSI'12[19]	ISSCC'08[8]
Technology		55nm DDC		40nm	65nm	65nm
Cell Type		4+2T		5T	7T	8T
Cell Area Scaled to 6T		1.12x		0.93x	1.15x	1.3x
Array Size		128 x 128 (16kb)		4Mb	256 x 128 (32kb)	256 x 128 x 8 (256kb)
Array Efficiency		65%		55%	46%	NA
Read/Write VDDmin (V)		0.25		0.38	0.26	0.35
Write	Freq. (MHz)	600 (0.8V)	6 (0.25V)	NA	NA	0.025 (0.35V)
	Energy (fJ/bit)[1]	18.5 (0.8V)	5.5 (0.25V)	NA	NA	1240 (0.35V)
Read	Freq. (MHz)	350 (0.8V)	4 (0.25V)	100 (0.6V)	1.8 (0.26V)	0.025 (0.35V)
	Energy (fJ/bit)[1]	16.6 (0.8V)	4.9 (0.25V)	103 (0.6V)	44 (0.26V)	880 (0.35V)

[1] Divided by word length.

TABLE III

COMPARISONS WITH OTHER CAM WORKS

		This work		CAM Work		
				VLSI'15[7]	ESSCIRC'11[20]	ESSCIRC'13[21]
Function		SRAM/CAM/Logic		SRAM/CAM/Logic	BCAM	BCAM
Technology		55nm DDC		28nm FDSOI	32nm	65nm
Cell Type		4+2T		6T	11T	10T
Cell Area Scaled to 6T		1.12x		1x	>2x	>2x
Array Size		128 x 128 (16kb)		64 x 64 (4kb)	64 x 64 x 4 (16kb)	128 x 128 (16kb)
Array Efficiency		65%		60%	NA	NA
CAM VDDmin (V)		0.35		0.75	0.5	0.8
BCAM	Freq. (MHz)	270 (0.8V)	18 (0.35V)	370 (1V)	NA	500 (1.2V)
	Energy (fJ/bit)[1]	0.45 (0.8V)	0.13 (0.35V)	0.6 (1V)	0.3 (0.5V)	0.77 (1.2V)
Logic	Freq. (MHz)	230 (0.8V)	15 (0.35V)	594 (1V)	NA	NA
	Energy (fJ/bit)[2]	24.1 (0.8V)	6.6 (0.35V)	NA	NA	NA

[1] Divided by array size.
[2] Divided by word length.

the energy of logic operation is 50% higher because of dual WL activation. However, the logic functions operate on two words simultaneously instead of a single word as in normal read and also performs computing in a single read cycle. Using the measured data in this paper, we compare IMC with near-memory-computing, which reads out two words in two cycles and computes in near-memory logic blocks as shown in Fig. 20 (left). The total latency of IMC is 35% smaller with at least 25% lower energy.

Fig. 21 shows the measured V_{DDmin} across temperature. Hold V_{DDmin} is 0.2 V at room temperature and 0.24 V at 120 °C. Fig. 22 shows the total leakage power across V_{DD}. At 0.2 V, which is the hold V_{DDmin}, the minimum leakage power is 1.6 μW at room temperature for a typical die.

We measured 40 dies in total. Twenty of them are from TT wafer and five each are from corner wafers. Fig. 23(a) shows the V_{DDmin} distribution of the 20 dies taken from a TT wafer. The average V_{DDmin} is 0.254 V with a standard deviation of 0.014 V. Fig. 23(b) shows the distribution of the average V_{DDmin} across split wafers. Even in the worst case, the V_{DDmin} is still below 0.3 V.

Fig. 24 compares this paper with other recently published works in terms of V_{DDmin} and access energy. All listed SRAMs here are fabricated in sub-65-nm CMOS technologies and their capacity is smaller than 128 kb for a fair comparison. With 55-nm DDC technology, we achieve the lowest access energy with good V_{DDmin} compared to those shown in the graph.

Table II summarizes the comparison of the proposed $4 + 2$T SRAM with other decoupled SRAM cells which has eight or less than eight transistors in the bitcell. We achieve comparable V_{DDmin} and lowest read energy with minimum area overhead among the listed works. Table III compares the proposed CAM with other CAM works. We also have better V_{DDmin} and energy with slightly improved area efficiency.

VII. CONCLUSION

A $4 + 2$T SRAM is proposed with 15% area saving compared to 8T SRAM. The differential decoupled read paths enable reliable multi-word simultaneous activation to perform Boolean logic functions. The proposed IMC saves latency and energy by 35% and 25%, respectively, compared with near-memory-computing. The BCAM configured from the 4 + 2T SRAM achieves 0.13 fJ/search/bit at 0.35 V. The chip is fabricated in 55-nm DDC technology and achieves 0.3-V read/write V_{DDmin} across corners.

ACKNOWLEDGMENT

The authors would like to thank Fujitsu for chip fabrication and support. They would also like to thank DDC development team members from MIFS for their great help, and V. Agrawal for his kind advice.

REFERENCES

[1] M. Gao and C. Kozyrakis, "HRL: Efficient and flexible reconfigurable logic for near-data processing," in *Proc. IEEE Symp. High Perform. Comput. Archit.*, Mar. 2016, pp. 126–137.

[2] M. Gao, G. Ayers, and C. Kozyrakis, "Practical near-data processing for in-memory analytics frameworks," in *Proc. IEEE Int. Conf. Parallel Archit. Compilation*, Sep. 2015, pp. 113–124.

[3] M. Horowitz, "Computing's energy problem," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

[4] M. Kang and N. R. Shanbhag, "In-memory computing architectures for sparse distributed memory," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 4, pp. 855–863, Aug. 2016.

[5] S. Aga, S. Jeloka, A. Subramaniyan, S. Narayanasamy, D. Blaauw, and R. Das, "Compute caches," in *Proc. IEEE Symp. High Perform. Comput. Archit.*, Feb. 2017, pp. 481–492.

[6] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[7] S. Jeloka, N. B. Akesh, D. Sylvester, and D. Blaauw, "A 28 nm configurable memory (TCAM/BCAM/SRAM) using push-rule 6T bit cell enabling logic-in-memory," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, Apr. 2016.

[8] N. Verma and A. P. Chandrakasan, "A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, Jan. 2008.

[9] J. Keane *et al.*, "5.6 Mb/mm^2 1R1W 8T SRAM arrays operating down to 560 mV utilizing small-signal sensing with charge-shared bitline and asymmetric sense amplifier in 14 nm FinFET CMOS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 308–309.

[10] M. Qazi, K. Stawiasz, L. Chang, and A. Chandrakasan, "A 512 kb 8T SRAM macro operating down to 0.57 V with an AC-coupled sense amplifier and embedded data-retention-voltage sensor in 45 nm SOI CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 350–351.

[11] J. Kulkarni *et al.*, "Dual-VCC 8T-bitcell SRAM array in 22 nm tri-gate CMOS for energy-efficient operation across wide dynamic voltage range," in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Jun. 2013, pp. 126–127.

[12] Y. Zhang *et al.*, "Recryptor: A reconfigurable in-memory cryptographic cortex-M0 processor for IoT," in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Jun. 2017, pp. 264–265.

[13] Q. Dong *et al.*, "A 0.3 V VDDmin 4+2T SRAM for searching and in-memory computing using 55 nm DDC technology," in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Jun. 2017, pp. 160–161.

[14] K. Fujita *et al.*, "Advanced channel engineering achieving aggressive reduction of VT variation for ultra-low-power applications," in *IEDM Tech. Dig.*, Dec. 2011, pp. 749–752.

[15] L. T. Clark *et al.*, "A highly integrated 65-nm SoC process with enhanced power/performance of digital and analog circuits," in *IEDM Tech. Dig.*, Dec. 2012, pp. 335–338.

[16] V. Agrawal *et al.*, "Low power ARM cortex-M0 CPU and SRAM using deeply depleted channel (DDC) transistors with VDD scaling and body bias," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2013, pp. 1–4.

[17] H. N. Patel *et al.*, "A 55 nm ultra low leakage deeply depleted channel technology optimized for energy minimization in subthreshold SRAM and logic," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2016, pp. 45–48.

[18] D. Jeon *et al.*, "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.

[19] M.-F. Chang *et al.*, "A sub-0.3 V area-efficient L-shaped 7T SRAM with read bitline swing expansion schemes based on boosted read-bitline, asymmetric-V$_{\mathrm{TH}}$ read-port, and offset cell VDD biasing techniques," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2558–2569, Oct. 2013.

[20] A. Agarwal *et al.*, "A 128×128b high-speed wide-and match-line content addressable memory in 32 nm CMOS," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2011, pp. 83–86.

[21] A. T. Do, C. Yin, K. S. Yeo, and T. T.-H. Kim, "Design of a power-efficient CAM using automated background checking scheme for small match line swing," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2013, pp. 209–212.

[22] M. E. Sinangil, H. Mair, and A. P. Chandrakasan, "A 28 nm high-density 6 T SRAM with optimized peripheral-assist circuits for operation down to 0.6 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 260–261.

[23] Y. Sinangil and A. P. Chandrakasan, "A 128 kbit SRAM with an embedded energy monitoring circuit and sense-amplifier offset compensation using body biasing," *IEEE J. Solid-State Circuits*, vol. 49, no. 11, pp. 2730–2739, Nov. 2014.

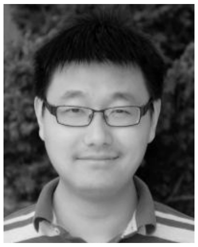
[24] Q. Li, B. Wang, and T. T. Kim, "A 5.61 pJ, 16 kb 9T SRAM with single-ended equalized bitlines and fast local write-back for cell stability improvement," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2012, pp. 201–204.

[25] S. Moriwaki, A. Kawasumi, T. Suzuki, T. Sakurai, and S. Miyano, "0.4 V SRAM with bit line swing suppression charge share hierarchical bit line scheme," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2011, pp. 1–4.

[26] B. Wang, T. Q. Nguyen, A. T. Do, J. Zhou, M. Je, and T. T. Kim, "A 0.2 V 16 Kb 9T SRAM with bitline leakage equalization and CAM-assisted write performance boosting for improving energy efficiency," in *Proc. IEEE Asian Solid-State Circuits Conf. (ASSCC)*, Nov. 2012, pp. 73–76.

[27] S. Lutkemeier, T. Jungeblut, H. K. O. Berge, S. Aunet, M. Porrmann, and U. Ruckert, "A 65 nm 32 b subthreshold processor with 9T multi-Vt SRAM and adaptive supply voltage control," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 8–19, Jan. 2013.

[28] A. T. Do, Z. C. Lee, B. Wang, I.-J. Chang, X. Liu, and T. T.-H. Kim, "0.2 V 8T SRAM with PVT-aware bitline sensing and column-based data randomization," *IEEE J. Solid-State Circuits*, vol. 51, no. 6, pp. 1487–1498, Jun. 2016.



Qing Dong (S'14) received the B.S. and M.S. degrees in microelectronis from Fudan University, Shanghai, China, in 2010 and 2013, respectively, and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA, in 2017.

He is currently with TSMC, San Jose, CA, USA. His current research interests include memory circuits design.

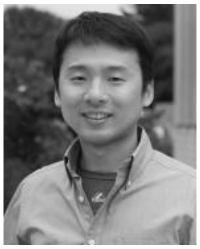
Dr. Dong was a recipient of the Best Paper Awards at the 2012 IEEE International Conference on Solid-State and Integrated Circuit Technology, the 2015 IEEE International Symposium on Circuits and Systems, and the 2016 IEEE Symposium on Security and Privacy.

Supreet Jeloka (S'15) received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology, Warangal, India, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2013 and 2017, respectively.

He is currently a Senior Research Engineer with ARM Research, Austin, TX, USA. His current research interests include memory design, in-memory computing, low-power circuit design, interconnect fabrics, and hardware security.

Mehdi Saligane received the B.S. and M.S. degrees in electrical engineering systems and control from the Ecole Polytechnique de Grenoble, Grenoble, France, in 2009 and 2011, respectively, and the Ph.D. degree in electrical engineering and computer science from the University of Aix-Marseille, Marseille, France, in 2016.

He was a Visiting Researcher at the Michigan Integrated Circuit Laboratory, University of Michigan, Ann Arbor, MI, USA, for two years. During 2010–2015, he was a Research Engineer with STMicroelectronics Central R and D, Crolles, France, where he focused on developing new adaptive solutions and ultra-low-power digital design. In 2015, he joined the Michigan Integrated Circuit Laboratory, University of Michigan, as a Research Investigator, where he has been a Research Fellow since 2017. His current research includes on-chip monitoring, adaptive techniques for variability tolerant designs, and near/sub-threshold energy efficient systems.

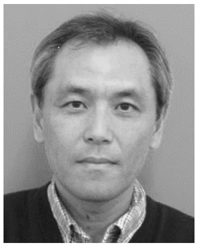


Yejoong Kim (S'08) received the bachelor's degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2008, and the master's and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2012 and 2015, respectively.

He is currently a Research Fellow with the University of Michigan and a Vice President of Research and Development, CubeWorks, Inc., Ann Arbor. His current research interests include subthreshold circuit designs, ultra-low-power SRAM, and the design of millimeter-scale computing systems and sensor platforms.

Masaru Kawaminami received the B.S. and M.S degrees in science and engineering (applied chemistry) from Waseda University, Tokyo, Japan, in 2002.

He is currently a Technical Marketing Specialist at the Business Development Division, Fujitsu Electronics America, Inc., Sunnyvale, CA, USA, where he was involved in expanding foundry businesses of Mie Fujitsu Semiconductor Ltd., to USA customers.



Akihiko Harada received the M.S. degree in science from Osaka University, Osaka, Japan, in 1995.

He was a Customer Support Engineer at Fujitsu Electronics America, Sunnyvale, CA, USA, from 2014 to 2017, where he was involved in the development of DDC technology. He is currently a Manager of Customer Engineering Division, Mie Fujitsu Semiconductor Ltd., Kuwana, Japan.

Satoru Miyoshi received the B.S. degree in material science from Yokohama National University, Yokohama, Japan, in 1988.

He joined Fujitsu Ltd., Kawasaki, Japan, as a Process Integration Engineer in 1988. He has been a Technology Marketing Director at Fujitsu Electronics America, Sunnyvale, CA, USA, since 2014, where he has been involved in the development of deeply depleted channel technology.

Makoto Yasuda received the B.S. degree in electronic physical engineering from Hiroshima University, Hiroshima, Japan, in 1992.

He joined Fujitsu Ltd., Kawasaki, Japan, as a Process Integration Engineer in 1992. He is currently with Mie Fujitsu Semiconductor Ltd., Kuwana, Japan, where he was involved in the development of deeply depleted channel technology.

David Blaauw (M'94–SM'07–F'12) received the B.S. degree in physics and computer science from Duke University, Durham, NC, USA, in 1986, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 1991.

He was the Manager of the High Performance Design Technology Group, Motorola, Inc., Austin, TX, USA. Since 2001, he has been the Faculty at the University of Michigan, Ann Arbor, MI, USA, where he is currently a Professor. He has authored or co-authored over 500 papers and holds 50 patents. His current research interests include VLSI design with including near-threshold and subthreshold design for ultra-low-power millimeter-scale sensor nodes.

Prof. Blaauw was a member of the ISSCC Technical Program Committee. He was the Technical Program Chair and the General Chair for the International Symposium on Low Power Electronic and Design. He was also the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference.

Dennis Sylvester (S'95–M'00–SM'04–F'11) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, CA, USA, in 1999.

He held research staff positions with the Advanced Technology Group, Synopsys, Mountain View, CA, USA; and the Hewlett-Packard Laboratories, Palo Alto, CA, USA; and Visiting Professorships at the National University of Singapore, Singapore, and Nanyang Technological University, Singapore. He is currently a Professor of electrical engineering and computer science with the University of Michigan, Ann Arbor, MI, USA, and the Director of the Michigan Integrated Circuits Laboratory, a group of ten faculties and more than 70 graduate students. He is the Co-Founder of Ambiq Micro, Austin, TX, USA. He has authored or co-authored over 375 articles along with one book and several book chapters. He has 20 U.S. patents. His current research interests include the design of millimeter-scale computing systems and energy-efficient near-threshold computing.

Prof. Sylvester was a recipient of the NSF CAREER Award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and eight Best Paper Awards and Nominations. He is a recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He serves on the Technical Program Committee of the IEEE International Solid-State Circuits Conference and previously served on the Executive Committee of the ACM/IEEE Design Automation Conference. He also serves as a Consultant and Technical Advisory Board Member for electronic design automation and semiconductor firms in his research areas. He has served as an Associate Editor for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, and the Guest Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II. His dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS Department.